INVESTIGATIONS IN MOLECULAR RECOGNITION: STATISTICAL TOOLS AND EXPERIMENTAL STUDIES

Thesis by

Richard E. Barrans Jr.

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1993

(Defended June 16, 1992)

Dedication

This dissertation, and the work described within, is respectfully dedicated to

Virgil Grissom

Edward White

Roger Chaffee

Vladimir Komarov

Valdislav Volkov

Georgi Dobrovolsky

Viktor Datsayev

Michael J. Smith

Francis R. Scobee

Ronald E. McNair

Ellison S. Onizuka

Sharon Christa Mcauliffe

Gregory Jarvis

Judith A. Resnik

and all others who give their lives in the exploration of space.

Acknowledgements

This thesis would never have come about without the examples, advice, encouragement, and assistance provided by many people. I will attempt to acknowledge those who have helped me the most, mindful that a brief mention here cannot repay the debts I have accumulated in the past seven years. I apologize for any inadvertant omissions.

The one person most directly responsible for this work is my advisor, Dennis Dougherty. He had the courage to let me explore statistical tangents of my own choosing and the patience to endure a stream of lukewarm laboratory results, always having faith that I would come up with something worthwhile. His efforts to instill in his group a desire to do things the right way and to be aware of developments outside of the direct sphere of our investigations have fostered an ideal environment for science. Furthermore, his steadfast refusal to let his career overwhelm all other aspects of his life provides an example that I hope to remember and follow.

My introduction to and tutelage in molecular recognition chemistry was provided by the founders of the project, Mike Petti and Tim Shepodd. They not only developed the synthesis of the family of ethenoanthracene hosts, but aslo carried out a number of careful studies that revealed the fascinating chemistry of these compounds. When the time came to publish their results, their gracious decision to include *me* as a co-author of the paper (because of my development of Multifit) was both flattering and humbling. Their regard for my efforts and their willingness to take me under their wing benefitted me enormously.

Similar thanks are also due to Dave Stauffer. Twice when he discovered something interesting, I managed to attach myself remora-like to his project and wind up

with my name on his paper. Those experiences emphasized to me in the strongest possible terms the value of collaboration.

None of the programs in the binding study analysis packege would have been developed without the motivation provided by the laboratory results of Petti, Shepodd, Stauffer, and all of the "pockets people" who followed. Pat Kearney, Alison McCurdy, Laura Mizoue, Leslie Jimenez, Liz Warner, Jon Forman, and Sandro Mecozzi have all provided me with intriguing results that stretch the limits of my data reduction and interpretive abilities. To them I sincerely say: you guys are brilliant. It has been an honor and a privelege to work with you.

I have also received considerable help from outside of the pockets side of the group. Mike Sponsler, Brian Masek, Gary Snyder, Rakesh Jain, and Frank Coms all helped me get oriented at Caltech; more recently, Dave Shultz and Piotr Kaszynski have been dependable sounding boards and sources of fresh ideas. Julian Pranata, Josh Jacobs, Bob Kumpf, and Dick Brown have been instrumental in making my association with computers a pleasant and fulfilling one.

On the statistical side of things, I have received considerable help from Gary Lorden and H. O. Adeyemi, who taught the statistics course at Caltech and who answered my frantic questions at inopportune times. I also received helpful criticisms and advice on a preliminary draft of Chapter 3 from Karen Roche.

My thanks also go to Helen Iams, Dennis Dougherty, Patrick Kearney, and Jon Forman for proofreading the initial drafts of this thesis. Any part that makes sense is surely the result of their input; all remaining errors and awkward sections are solely the responsibility of the author.

Any attempt to acknowledge those whose friendship helped me over the last seven years is necessarily incomplete, because my debts are to so many. My sanity,

such as it is, has been protected by the ever-changing gang at the Caltech weight room, the folks at Holliston Avenue UMC, and, of course, the members of the Dougherty group. Ed Stewart, my classmate, labmate, partner in bad music, and next-door-neighbor, was instrumental in making my graduate school experience the adventure that it was. Finally, I thank Helen Iams for her love, support, patience, time, and understanding.

Abstract

The free energies of formation of intermolecular complexes in solution are often estimated by fitting a nonlinear model to an NMR titration experiment. A regression procedure that assigns weights to each observation on the basis of expected measurement errors has been developed, and yields better parameter estimates than other methods in common use. Procedures for critically evaluating the fit of the model to the experimental data and for assigning confidence limits to the fitted parameters have also been developed. These employ Monte Carlo simulations of the NMR titration experiment to obtain probability distributions that are not available by theoretical means.

Aspects of the complexation behavior of a family of water-soluble macrocyclic cyclophanes are also described. Significant heat capacity effects, which are interpreted in terms of hydrophobic hydration, are seen in variable-temperature studies. The alkylation reactions of pyridine-based compounds are accelerated by complexation with these cyclophanes; an interpretation based on the dynamic properties of the alkylation reaction, the solvent, and the cyclophane is offered. In addition, accounts of efforts to make axially-substituted cyclophanes, to synthesize cyclophanes incorporating a diphosphine ligand, to append additional water-solubilizing groups to the cyclophanes, and to employ small cyclophanes as complexants for alkali metal cations are given.

TABLE OF CONTENTS

Chapter 1

Introduction

I. General

Bonding is the one concept central to all branches of chemistry. In its broadest sense, any attractive interaction between particles is a bond. Chemical bonds range in scope from well-defined covalent interactions between adjacent atoms in the same molecule to less oriented van der Waals and hydrophobic forces. The properties of systems ranging in size from the isolated diatomic hydrogen molecule to a bulk liquid arise directly from these interactions. In effect, all of chemistry is the study of bonding.[1]

The covalent bonds that determine the structure of individual molecules can be explained in considerable detail both qualitatively and from quantum mechanical principles. The same cannot be said for the intermolecular forces that govern the association of molecules in solution.[2] Such complexation processes defy simple theoretical explanation; their properties are predicted only by intensive computational procedures.[3] The forces involved are poorly understood and even poorly defined. For instance, the hydrophobic effect, which has long been held responsible for the

(1) *The Binding Force*; Banigan, Sharon, Ed.; Walker: New York, 1966.

(2) (a) Hobza, Pavel; Zahradník, Rudolf *Intermolecular Complexes*; Elsevier: New York, 1988. (b) Israelachvili, J. N. *Intermolecular and Surface Forces*; Academic: London, 1985. (c) Stone, A. J.; Price, S. L. "Some new ideas in the theory of intermolecular forces: anisotropic atom-atom potentials," *J. Phys. Chem.* **1988**, *92*, 3325–3335. (d) McLachlan, A. D. "Retarded dispersion forces between molecules," *Proc. Roy. Soc. London Ser. A.* **1963**, *271*, 387–401. McLachlan, A. D. "Retarded dispersion forces in dielectrics at finite temperatures," *Proc. Roy. Soc. London Ser. A.* **1963**, *274*, 80–90. McLachlan, A. D. "Three-body dispersion forces," *Mol. Phys.* **1963**, *6*, 423–427.

(3) (a) Kollman, Peter A.; Merz, Kenneth M. Jr. "Computer modeling of the interactions of complex molecules," *Acc. Chem. Res.* **1990**, *23*, 246–252. (b) Jorgensen, William L. "Computational insights on intermolecular interactions and binding in solution," *Chemtracts: Org. Chem.* **1991**, *4*, 91–119.

behavior of nonpolar solutes in water, is a subject of ongoing, and not always cordial, debate.[4–6]

Supramolecular chemistry, as defined by Lehn,[7] is the study of intermolecular complexes. Molecular recognition is the branch of this field concerned with exploring and exploiting intermolecular forces in order to direct the identities, affinities, and geometries of supramolecular complexes. The fundamental question of molecular recognition is why certain molecules prefer each other's company to that of the bulk solvent.

This question is of particular interest when the solvent is water. The most spectacular examples of molecular recognition are provided by biological systems. Molecular recognition in an aqueous environment is a critical component of many important biological processes, such as nerve signal transduction, enzyme action, immune response, and the sense of smell. The species responsible for these tasks are, by chemical standards, very large and poorly characterized. In order to understand the chemical processes of molecular recognition, it is necessary to scale these complex systems down to ones that can be studied in detail.

The first task for a physical-organic chemist wishing to study molecular recognition is to find a molecule small enough to study that still possesses some essential characteristic of an interesting but intractable biological system. The most essential molecular recognition characteristic, of course, is a tendency to associate with

---

(4) Muller, Norbert "Search for a realistic view of hydrophobic effects," *Acc. Chem. Res.* **1990**, *23*, 23–28.

(5) Privalov, Peter L.; Gill, Stanley J. "The hydrophobic effect: a reappraisal," *Pure Appl. Chem.* **1989**, *61*, 1097–1104.

(6) Shinoda, Kōzō "Iceberg formation and solubility," *J. Phys. Chem.* **1977**, *81*, 1300–1302. Shinoda, Kōzō; Kobayashi, Makoto; Yamaguchi, Nobuyoshi "Effect of 'Iceberg' formation of water on the enthalpy and entropy of solution of paraffin chain compounds: the effect of temperature on the critical micelle concentration of lithium perfluorooctanesulfonate," *J. Phys. Chem.* **1987**, *91*, 5292–5294.

(7) Lehn, Jean-Marie "Supramolecular chemistry: receptors, catalysts, and carriers," *Science* **1985**, *227*, 849–856.

another molecule. A potential "host" molecule thus must furnish an environment that some "guest" will prefer to the solvent.

Many different compounds have been put forth in answer to this challenge. The first class of molecules employed to serve as host molecules were the cyclodextrins. These cyclic oligomers of glucose are produced by enzymatic degradation of starch, and are commercially available. They come in three sizes: α, β, and γ, which are made of six, seven, and eight glucose units, respectively. Their hydrophilic alcohol groups principally point away from the macrocyclic cavity, the interior of which is composed of hydrocarbon and ether functionality. This creates a water-soluble molecule with a hydrophobic interior cavity, suitable for encapsulating small hydrophobic "guests." The cavity of α-cyclodextrin can accommodate a single benzene ring; γ-cyclodetrin can enclose two face-to-face benzene rings at once. The complexation behavior of the cyclodextrins has been studied extensively.[8]

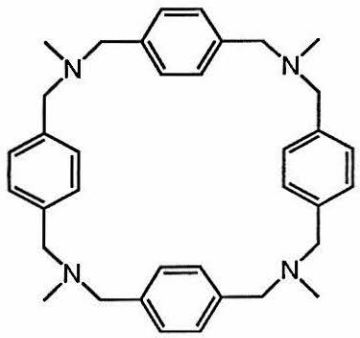
This family of only three members does not lend itself to extensive structure-activity studies. Furthermore, cyclodextrins' many identical hydroxide groups make them difficult to selectively functionalize. Nevertheless, studies of modified cyclodextrins sporting appended catalytic and other groups have been carried out.[9]

The need for more variable and diverse hosts has sparked the development of many classes of synthetic receptors. One of the first of these was Tabushi's xylylenediamine-based host **1**, a molecule of time-averaged D_{4h} symmetry. In moderately acidic media, in which the amines are protonated, this host binds the anionic guest 1-anilino-8-napthalenesulfonate (ANS).[10]

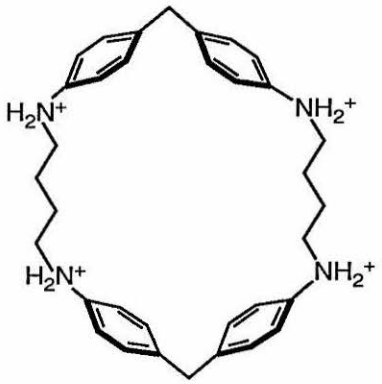
(8) Szejtli, J. *Cyclodextrin Technology*; Kluwer academic: Dordrecht, 1988. Szejtli, J. *Cyclodextrins and Their Inclusion Complexes*; Akadémiai Kiadó: Budapest, 1982. Saenger, Wolfram "Cyclodextrin inclusion compounds in research and industry," *Angew. Chem. Int. Ed. Engl.* **1980**, *19*, 344–362. Bender, M.; Komiyama, M. L. *Cyclodextrin Chemistry*; Springer-Verlag: New York, 1978.

(9) (a) Tabishi, Iwao; Yamamura, Kazuno; Nabeshima, Tatsuya "Characterization of regiospecific A,C- and A,D-disulfonate capping of β-cyclodextrin. Capping as an efficient production technique," *J. Am. Chem. Soc.* **1984**, *106*, 5267–5270. (b) Breslow, Ronald; Anslyn, Eric; Huang, Deeng-Lih "Ribonuclease mimics," *Tetrahedron* **1991**, *47*, 2365–2376.

(10) Tabushi, I.; Kuroda, Y.; Kimura, Y. "Strong hydrophobic binding by water soluble macrocyclic heterocyclophane," *Tetrahedron Lett.* **1976**, *37*, 3327–3330.



1



2

Shortly thereafter, Koga reported a similar but larger host, **2**, based on the 4,4'-diaminodiphenylmethane moiety. The molecule co-crystallized with 1,2,4,5-tetramethylbenzene (durene) from water, with the durene at the exact center of the host cavity.[11] This was in welcome contrast to structures of previously-obtained crystals containing purported host/guest pairs, in which host and guest molecules merely alternated.[12] This encouraging result provided the first unequivocal evidence that a synthetic host in aqueous solution could actually encapsulate a small hydrophobic guest. Koga's group has continued to investigate the properties of members of this series.[13]

Diederich employed the basic diphenylmethane skeleton of Koga's macrocycle in a refined series of hosts exemplified by **3** and **4**. The macrocyclization closure is provided by ethers instead of amines, and water-solubility is provided by piperdine

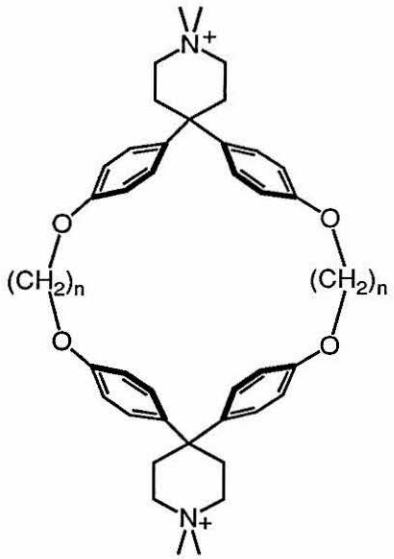
---

(11) Odashima, Kazunori; Itai, Akiko; Koga, Kenji "Biomimetic studies using artificial systems. 3. Design, syntheses, and inclusion complex forming ability of a novel water-soluble paracyclophane possessing diphenylmethane skeletons," *J. Org. Chem.* **1985**, *50*, 4478–4484.

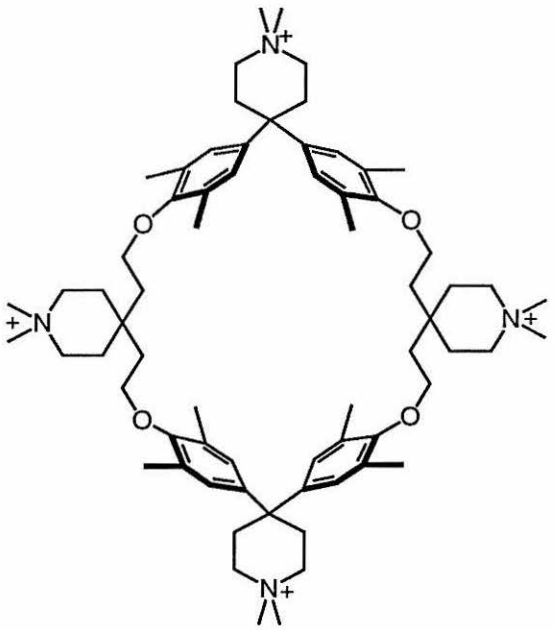
(12) Diederich, François "Complexation of neutral molecules by cyclophane hosts," *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 362–386.

(13) Miyake, Munehau; Kirisawa, Makoto; Koga, Kenji "Anionic cyclophanes as hosts for cationic Aromatic guests," *Tetrahedron Lett.* **1991**, *32*, 7295–7298.

rings linked in a *spiro* fashion to the macrocycle. The exclusion of the water-solubilizing groups from the binding site enforced by this *spiro* linkage requires the host to provide a more rigorously hydrophobic environment for its guests.[14]



3



4

Diederich first recognized the importance of ensuring that a host does not aggregate while its complexation behavior is studied.[14a] Host molecules, possessing both hydrophobic and hydrophilic regions, are similar to detergents. In water, they may be expected to associate into structures in which the hydrophilic regions are held outward into the solvent and the hydrophilic regions congregate in the interior. Such an aggregate, regardless of its morphology, may be called a *micelle*.[18] The

(14) (a) Diederich, François; Dick, Klaus "New water- soluble macrocycles of the paracyclophane type: aggregation behavior and host-guest interactions with hydrophobic substrates," *Tetrahedron Lett.* **1982**, *23*, 3167–3170. (b) Krieger, Klaus; Diederich, François "Structure of host-guest complexes of 1′,1″-dimethyldispiro[1,6,20,25-tetraoxa[6.1.6.1]paracyclophane-13,4′:32,4″-bispiperidine] with benzene and *p*-xylene," *Chem. Ber.* **1985**, *118*, 3620–3631.

(18) Tanford, Charles *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*; Wiley-Interscience: New York, 1980.

aggregation process may be approximated as the precipitation of a solid, with the micelle as the solid phase. The dissociation of a single molecule from a micelle,

$$H_n \rightleftharpoons H + H_{n-1},$$

is governed by the equilibrium constant

$$K = \frac{[H][H_{n-1}]}{[H_n]}.$$

If the micelle concentrations $[H_n]$ and $[H_{n-1}]$ are, by analogy to a condensed phase, taken as unity, this equilibrium expression reduces to $K = [H]$. According to this simple model, host will be monomeric in solution if its total concentration is below K. If its total concentration exceeds K, then the concentration of monomeric host will stay at its maximum value K and the remaining host molecules will be in micellar form. This concentration K is the *critical micelle concentration*, or cmc. Studies of molecular recognition with the host at a concentration above its cmc could be complicated by the formation of these additional micellar species. Consequently, studies should be carried out with a concentration of host below its cmc.

Diederich's group determined the cmc of their host **3** by finding the concentration above which its NMR spectrum began to change.[14a] It proved to be too low for convenient study, so they designed a new host, **4**, with more water-solubilizing groups. This host also included additional methyl substituents on the aromatic rings in an effort to both increase the hydrophobic surface area of the cavity and disrupt micellar packing. It indeed proved to be a highly soluble, efficient host.[15] This basic

(15) Diederich, François; Dick, Klaus "Inclusion complexes between a macrocyclic host molecule and aromatic hydrocarbons in aqueous solution," *Angew. Chem. Int. Ed. Engl.* **1983**, *22*, 715-716. Diederich, François; Dick, Klaus "A new water- soluble macrocyclic host of the cyclophane type: host-guest complexation with aromatic guests in aqueous solution and acceleration of the transport of arenes through an aqueous phase," *J. Am. Chem. Soc.* **1984**, *106*, 8024–8036. Diederich, François; Griebel, Dieter "^{1}H NMR investigations of host-guest complexation betwen macrocyclic hosts of the cyclophane type and aromatic guests in aqueous solution," *J. Am. Chem. Soc.* **1984**, *106*, 8037–8046. Diederich, François; Dick, Klaus; Griebel, Dieter "Water-soluble tetraoxa [*n*.1.*n*.1]-paracyclophanes: synthesis and host-guest interactions in aqueous solution," *Chem. Ber.* **1985**, *118*, 3588–3619. Diederich, François; Dick, Klaus "A water-soluble tetraoxa [7.1.7.1]paracyclophane: synthesis and host-guest interactions with alicyclic and cationic aromatic guest molecules in aqueous solution," *Chem. Ber.* **1985**, *118*, 3817–3829.

design has been employed extensively in a number of related molecules.[16]

Other groups have employed additional host designs to study molecular recognition in water.[12,17] Wilcox has developed a chiral, C_2-symmetric amine[19] that forms the basis of a family of cyclophane hosts.[20] Several other macrocyclic compounds have been studied as water-soluble hosts, including sulfonated calixarenes,[21] several designs from Vögtle *et al.*,[22] and cucurbitural, a condensate of urea, glyoxal, and formaldehyde.[23] In addition, Diederich and Wilcox have undertaken studies of their hosts in organic as well as aqueous solvents; studies of molecular recognition in or-

---

(16) see, for example: Ferguson, Stephen B.; Sanford, Elizabeth M.; Seward, Eileen M.; Diederich, François "Cyclophane-arene inclusion complexation in protic solvents. Solvent effects versus electron donor-acceptor interactions," *J. Am. Chem. Soc.* **1991**, *113*, 813–820. Smithrud, David B.; Wyman, Tara B.; Diederich, François N. "Enthalpically driven cyclophane-arene inclusion complexation: solvent-dependent calorimetric studies," *J. Am. Chem. Soc.* **1991**, *113*, 5420–5426. Diederich, François "Molecular recognition in aqueous solution: supramolecular complexation and catalysis," *J. Chem. Educ.* **1990**, *67*, 813–820.

(17) Diederich, François *Cyclophanes*; Monographs in Supramolecular Chemistry; Royal Society of Chemistry: Cambridge, 1991.

(19) Wilcox, Craig S; Cowart, Marlon D. "New approaches to synthetic receptors. Synthesis and host properties of a water soluble macrocyclic analog of Tröger's base," *Tetrahedron Lett.* **1986**, *27*, 5563–5566.

(20) See, for example: Adrian, James C. Jr.; Wilcox, Craig S. "General effects of binding site water exclusion on hydrogen bond based molecular recognition systems: a 'closed' binding site is less affected by environmental changes than an 'open' site," *J. Am. Chem. Soc.* **1992**, *114*, 1398–1403. Webb, Thomas H.; Suh, Hongsuk; Wilcox, Craig S. "Enantioselective and diastereoselective molecular recognition of alicyclic substrates in aqueous media by a chiral, resolved synthetic receptor," *J. Am. Chem. Soc.* **1991**, *113*, 8554–8555.

(21) Shinkai, Seji; Araki, Koji; Manabe, Osamu "Does the calixarene cavity recognise the size of guest molecules? On the 'hole-size selectivity' in water-soluble calixarenes," *J. Chem. Soc., Chem. Commun.* **1988**, 187–189. Shinkai, Seji "Calixarenes as new functionalized host molecules," *Pure Appl. Chem.* **1986**, *58*, 1523–1528.

(22) see, for example: Vögtle, Fritz; Müller, Walter M.; Werner, Ute; Losensky, Hans-Willi "Selective molecular recognition and separation of isomeric and partially hydrogenated arenes," *Angew. Chem. Int. Ed. Engl.* **1987**, *26*, 901–903. Merz, T.; Wirtz, H.; Vögtle, F. "Anionic host molecules with bicyclic carbon skeletons—synthesis and guest inclusion in aqueous solution," *Angew. Chem. Int. Ed. Engl.* **1986**, *25*, 567–569. Franke, F.; Vögtle, F. "'In-out' isomeric large cavities and their differing guest selectivity," *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 219–221.

(23) Mock, W. L.; Shih, N.-Y. "Structure and selectivity in host-guest complexes of cucurbituril," *J. Org. Chem.* **1986**, *51*, 4440–4446. Mock, W. L.; Shih, N.-Y. "Host-guest binding capacity of cucurbituril," *J. Org. Chem.* **1983**, *48*, 3618–3619. Mock, W. L.; Shih, N.-Y. "Organic ligand-receptor interactions between cucurbituril and alkylammonium ions," *J. Am. Chem. Soc.* **1988**, *110*, 4706–4710.

ganic media are also pursued by Rebek,[24] Hamilton,[25] Whitlock,[26] Zimmerman,[27] and Still.[28]

II. Dougherty Group Studies.

A. Design.

The Dougherty group has also been active in the field of molecular recognition for over ten years.[29] The host family we employ in our investigations is a

---

(24) see, for example: Rotello, Vincent; Hong, Jong In; Rebek, Julius Jr. "Sigmoidal growth in a self-replicating system," *J. Am. Chem. Soc.* **1991**, *113*, 9422–9423. Galán, Amalial; De Mendoza, Javier; Toiron, Catherine; Bruix, Marta; Deslongchamps, Ghislain; Rebek, Julius Jr. "A synthetic receptor for dinucleotides," *J. Am. Chem. Soc.* **1991**, *113*, 9424-9425. Nowick, James S.; Feng, Qing; Tjivikua, Tjama; Ballester, Pablo; Rebek, Julius Jr. "Kinetic studies and modeling of a self-replicating system," *J. Am. Chem. Soc.* **1991**, *113*, 8831-8839.

(25) see, for example: Garcia-Tellado, Fernando; Albert, Jeffrey; Hamilton, Andrew D. "Chiral recognition of tartaric acid derivatives by a synthetic receptor," *J. Chem. Soc., Chem. Commun.* **1991**, 1761–1763. Chang, Suk-Kyu; van Engen, Donna; Fan, Erkang; Hamilton, Andrew D. "Hydrogen bonding and molecular recognition: synthetic, complexation, and structural studies on barbiturate binding to an artificial receptor," *J. Am. Chem. Soc.* **1991**, *113*, 7640–7645. Hamilton, Andrew D. " Molecular recognition: design and synthesis of artificial receptors employing directed hydrogen bonding interactions," *J. Chem. Educ.* **1990**, *767*, 821–828.

(26) see, for example: Whitlock, B. J.; Whitlock, H W. "Concave functionality: design criteria for nonaqueous binding sites," *J. Am. Chem. Soc.* **1990**, *112*, 3910–3915. Haeg, M. E.; Whitlock, B. J.; Whitlock, H. W. "Anthraquinone-based cyclophane hosts: synthesis and complexation studies," *J. Am. Chem. Soc.* **1989**, *111*, 692–696.

(27) see, for example: Zimmerman, Stephen C.; Wu, Weiming; Zeng, Zijian "Complexation of nucleotide bases by molecular tweezers with active site carboxylic acids: effects of microenvironment," *J. Am. Chem. Soc.* **1991**, *113*, 196–201. Zimmerman, Stephen C.; Zeng, Zijian; Wu, Weiming; Reichert, David "Synthesis and structure of molecular tweezers containing active site functionality," *J. Am. Chem. Soc.* **1991**, *113*, 183–196.

(28) see, for example: Chapman, Kevin T.; Still, W. Clark "A remarkable effect of solvent size on the stability of a molecular complex," *J. Am. Chem. Soc.* **1989**, *111*, 3075–3077. Sanderson, Philip E. J.; Kilburn, Jeremy D.; Still, W. Clark "Enantioselective complexation of simple amides by a C_2 host molecule," *J. Am. Chem. Soc.* **1989**, *111*, 8314–8315.

(29) (a) Petti, Michael A.; Shepodd, Timothy J; Dougherty, Dennis A. "Design and synthesis of a new class of hydrophibic binding sites," *Tetrahedron Lett.* **1986**, *27*, 807–810. (b) Shepodd, Timothy J.; Petti, Michael A.; Dougherty, Dennis A. *J. Am. Chem. Soc.* **1986**, *108*, 6085–6087. (c) Shepodd, Timothy J; Petti, Michael A.; Dougherty, Dennis A. "Molecular recognition in aqueous media: donor-acceptor and ion-dipole interactions produce tight binding for highly soluble guests," *J. Am. Chem. Soc.* **1988**, *110*, 1983–1985. (d) Petti, Michael A.; Shepodd, Timothy J.; Barrans, Richard E. Jr.; Dougherty, Dennis A. "'Hydrophobic' binding of water-soluble guests by high-symmetry, chiral hosts. An electron-rich receptor site with a general affinity for quaternary ammonium compounds and electron-deficient π systems," *J. Am. Chem. Soc.* **1988**, *110*, 6825–6840. (e) Stauffer, David A.; Dougherty, Dennis A. "Ion-dipole effect as a force for molecular recognition in organic media," *Tetrahedron Lett.* **1988**, *47*, 6039–6042.

series of water-soluble macrocyclic cyclophanes based on the 9,10-ethenoanthracene (dibenzobarrelene) system. Figure 1 depicts this series of hosts.

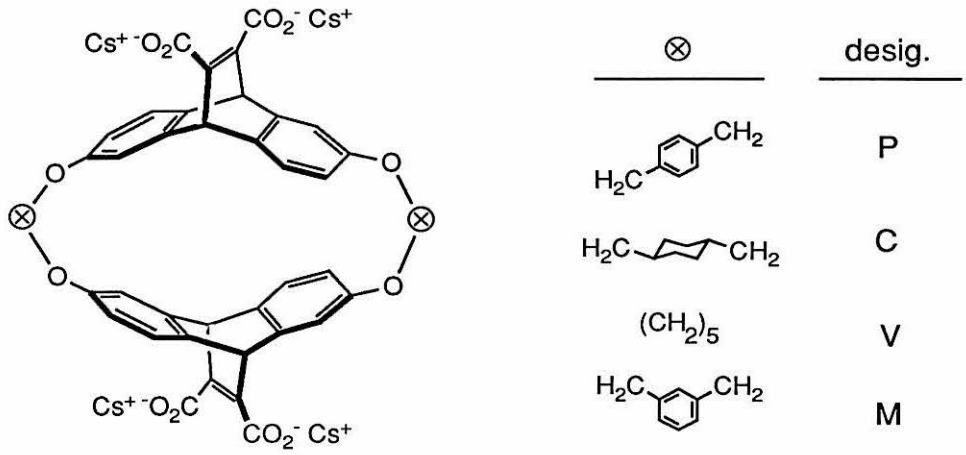


Figure 1. The family of ethenoanthracene-based hosts.

The members of this series are conventionally specified by a subscripted letter. The letter refers to the linker group: Figure 1 illustrates hosts P (*para*-xylyl linkers), C (cyclohexyldimethylene linkers), V (five-carbon chain linkers), and M (*ortho*-xylyl linkers). In addition, hosts IV, which has tetramethylene linkers, and O, which has *ortho*-xylyl linkers, have also been prepared. A subscript is used to specify the stereochemistry of the host, which will be discussed shortly. A host with C_{2h} symmetry is identified by the subscript "*meso*;" a host with D_2 symmetry by the subscript "*R*" or "*S*," depending on its absolute configuration. If it is racemic, it is identified by the subscript "*dl*." It is also our convention to refer to a macrocycle with methyl esters instead of free carboxylates by the subscript "E," assuming that the absolute stereochemistry is chiral. Thus, "P_E" is the tetramethyl ester of the D_2 diastereomer of host P.

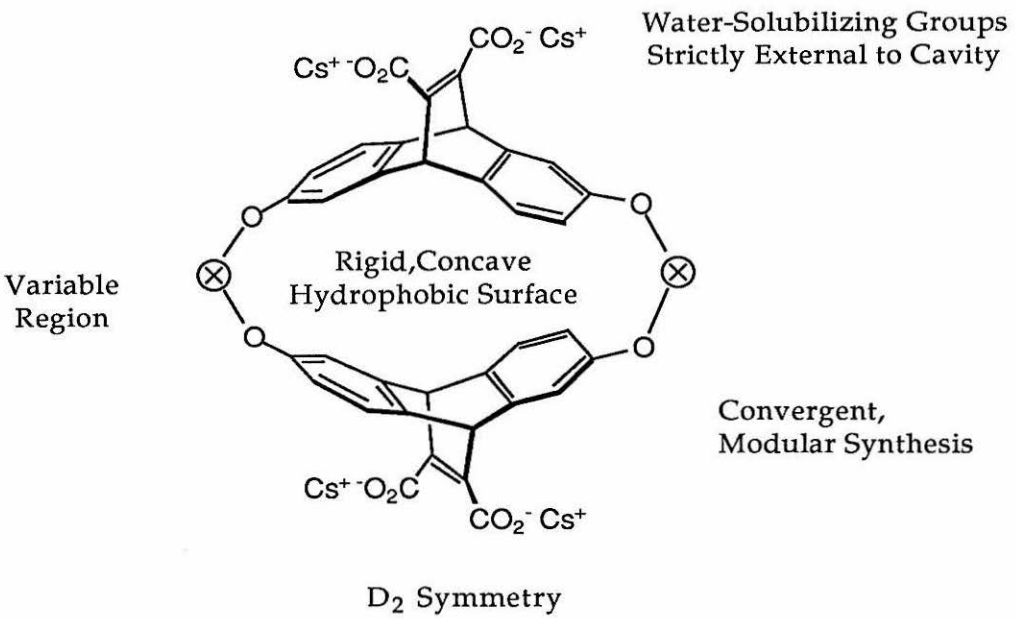


Figure 2. Positive attributes of ethenoanthracene-based hosts.

The ethenoanthracene unit provides a rigid, concave aromatic surface forming two walls and a corner of a binding cavity. This contrasts with the conformationally mobile diphenylmethane unit employed in other hosts. Each aromatic ring of diphenylmethane can freely rotate about the single bond connecting it to the central methylene carbon. A macrocyclic host composed of diphenylmethane units can thus adopt conformations in which the plane of an aromatic ring is perpendicular to the macrocyclic axis; in such a conformation, the macrocyclic cavity is poorly suited for binding. These aryl rotations must be inhibited in order to hold a guest within the cavity, exacting an entropic price.

With the ethenoanthracene unit, however, such aryl rotations are impossible, and are thus not missed when a guest is bound. Furthermore, the angle between the two benzene rings of ethenoanthracene is larger than the analogous angle of diphenylmethane. This makes the interior cavity of an ethenoanthracene-based macrocycle wider than that of a diphenylmethane-based macrocycle.

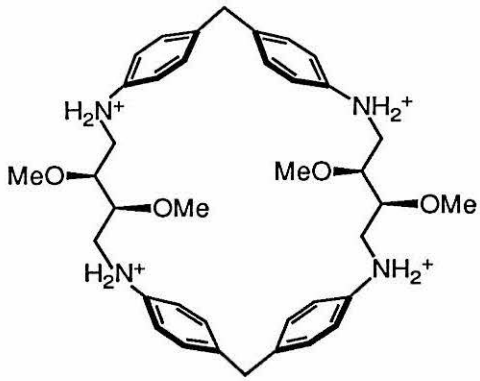
The bicyclic construction of the ethenoanthracene, in addition to providing a rigid framework for the aromatic rings, also furnishes a rigid scaffold for attaching water-solubilizing groups. Placement of these groups on the etheno bridge makes it impossible for them to be inside, or even near, the cavity. The entire ethenoanthracene moiety would need to rotate inward to place the carboxylates at such an interior position. Steric constraints render such conformations impossible even for hosts with moderately large linkers. The concave binding site and the water-solubilizing groups are always on opposite faces of the ethenoanthracene.

Furthermore, the ethenoanthracene unit is chiral when substituted at the 2- and 6-positions. If the two substituents are the same, one symmetry element, the C_2 axis, of the parent ethenoanthracene is preserved. A host made up of two such chiral units could be formed in two diastereomers: one with D_2 symmetry if both subunits have the same absolute configuration, and one with C_{2h} symmetry if they have opposite absolute configurations. Both of these diastereomers have high symmetry, which makes them detectable by NMR even at low concentrations. Each of its ethenoanthracene protons have three symmetry-identical counterparts that will resonate at the same frequency.

The D_2 diastereomer is especially interesting because of its chirality. This gives it the potential to display enantioselective complexation, possibly discriminating between enantiomers of a dissymmetric guest. The structure of these macrocycles makes this possibility even more plausible. The D_2 macrocycle is a grossly twisted structure, not simply an achiral framework perturbed by a chiral influence. An example of the latter has already been provided by Koga's group, who synthesized host **5**, made of diphenylmethane units connected by tartrate-derived linker groups.[30] If the chirally-placed methoxy groups of this host were to be replaced by hydrogens, it

(30) Takahashi, Ichiro; Odashima, Kazunori; Koga, Kenji "Diastereomeric host-guest complex formation by an optically active paracyclophane in water," *Tetrahedron Lett.* **1984**, *25*, 973–976.

would no longer be chiral. In contrast, no such simple substituent replacement can transform the D_2 ethenoanthracene into an achiral counterpart. The lack of mirror symmetry is an intrinsic property of the macrocyclic framework. These hosts have indeed shown modest enantioselectivities for some guests, but a directed, exhaustive enantioselectivity study has never been carried out.

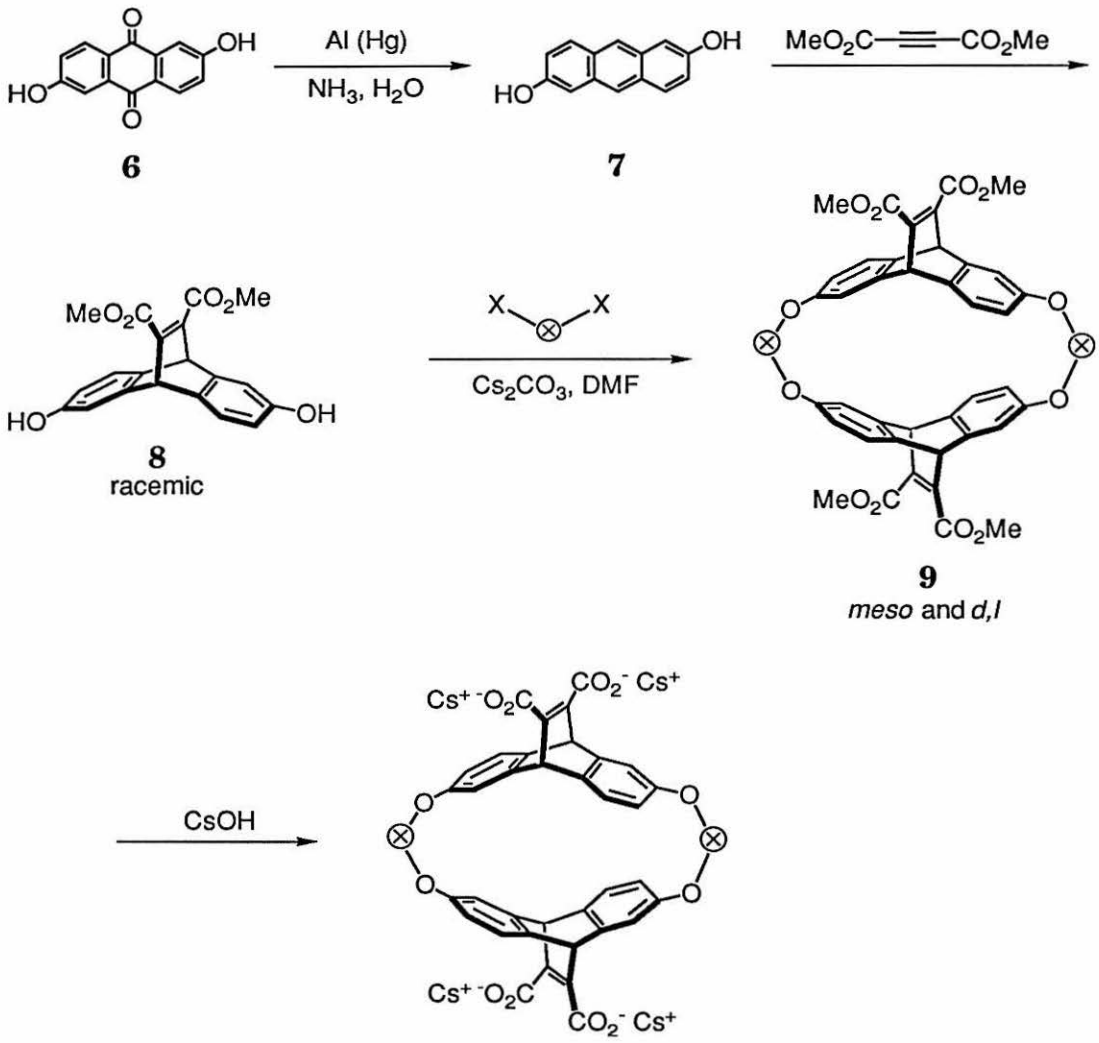


5

B. Synthesis.

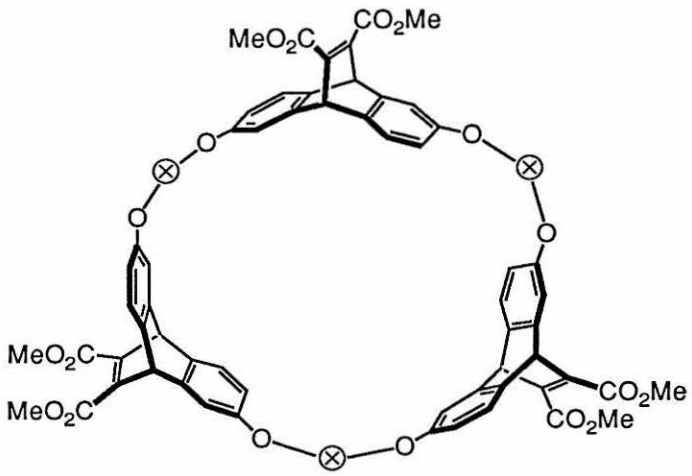
The syntheses of these hosts were developed by Timothy Shepodd and Michael Petti, the first graduate students to work on this project. The first and shortest route is shown in Scheme I. Commercially-available anthraflavic acid **6** (2,6-dihydroxyanthraquinone) is reduced by aluminum amalgam to the air-sensitive 2,6-dihydroxyanthracene **7**. This undergoes a Diels-Alder reaction with dimethylacetylenedicarboxylate (DMAD) in refluxing dioxane, to produce the racemic dihydroxyethenoanthracene **8**. A macrocyclization reaction is then performed at high dilution in dry DMF in the presence of excess cesium carbonate. The racemic diol reacts with α,ω-disubstituted electrophiles, such as *p*-xylylene dibromide or 1,5-dibromopentane, to form macrocycles such as **9** and higher oligomers such as **10**. Macrocycles of different sizes can be separated from each other and from other reaction products by flash

Scheme I



chromatography; the individual diastereomers of each size are only separable by high-pressure liquid chromatography (HPLC). This procedure offers no means to obtain the individual enantiomers of the chiral diastereomers.

One particularly distasteful step in this synthesis is the Diels-Alder reaction of 2,6-dihydroxyanthracene with DMAD. The dihydroxyanthracene is insoluble in most solvents, making it impossible to carry out the reaction at the high concentrations that would increase the reaction rate. It is only modestly soluble in re-

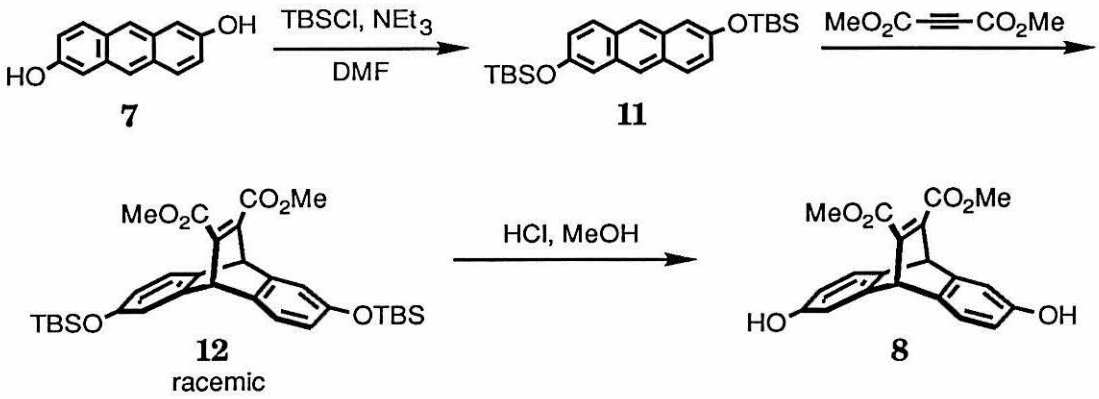


10

fluxing dioxane, the solvent eventually chosen for the reaction. Furthermore, the anthracene tends to oxidize, and DMAD is prone to polymerization and Michael addition. At the elevated temperatures of the reaction, the starting materials degrade fairly rapidly. At the end of the three-day course of the reaction, barely-soluble product must be isolated from a matrix of black tar.

This synthesis is improved considerably by the addition of two steps to the synthetic scheme. Protection of the dihydroxyanthracene as its bis-*t*-butyldimethylsilyl (TBS) ether **11** removes the problems of its air-sensitivity, nucleophilicity, and insolubility. This bis (silyl ether) is very soluble in nonpolar solvents, so the Diels-Alder reaction can conveniently be run at high concentration in refluxing toluene or xylenes over the course of one or two days. The masking of the hydroxyl groups also curtails their Michael addition to the DMAD, eliminating most of the troublesome side reactions. The product must still be isolated from a mass of black tar, but the isolation is easy and in higher yield. Removal of the silyl ether protecting groups unmasks the dihydroxyethenoanthracene, which can be used in a macrocyclization as before.

Scheme II

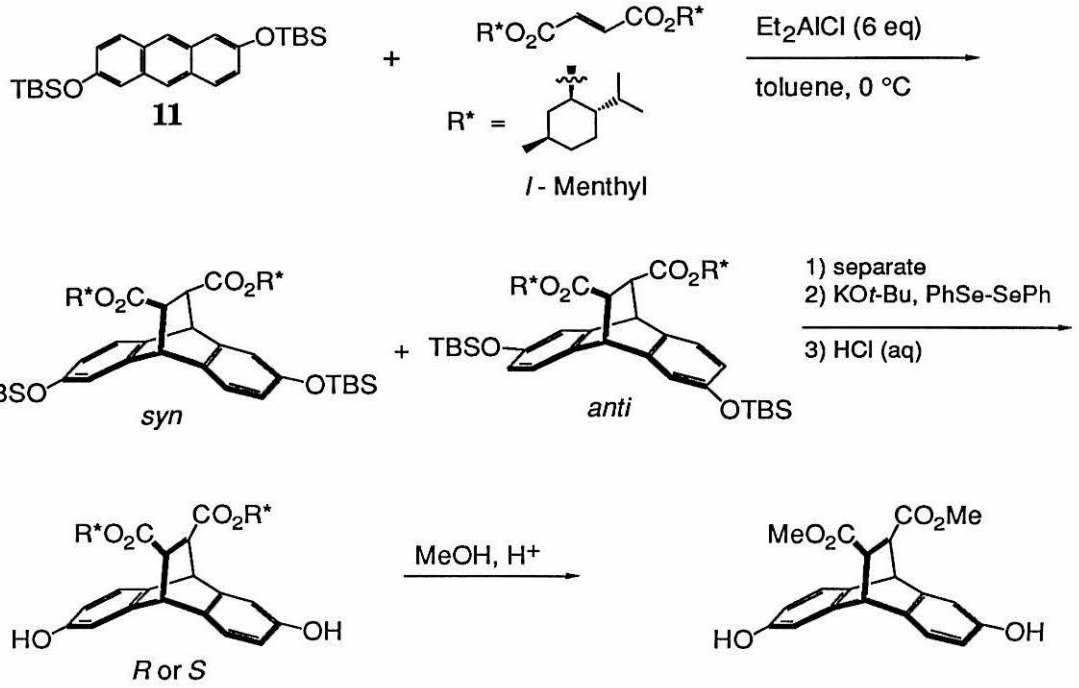


Tim Shepodd developed a means to obtain enantiomerically pure dihydroxyethenoanthracene, which is the key to the synthesis of enantiomerically pure host.[29d] This is accomplished by a diastereoselective Diels-Alder reaction using the chiral dienophile dimenthyl fumarate, aided by Lewis acid catalysis. This class of reaction was studied by Yamomoto[31] and found to proceed by the approach of only one of the two faces of the fumarate to the anthracene. If this pathway is followed in a reaction with 2,6-bis(*t*-butyldimethylsiloxy)anthracene, only two diastereomers would result from the addition of only one face of the dienophile to either face of the anthracene. These two products may be termed *syn* or *anti* according to the relative placement of the silyl and carboxylate groups.

The reality of this experiment is in exact accord with expectation: only two of the four possible diastereomers are formed. These diastereomers (*syn* and *anti*) are separated from each other by silica gel chromatography and low-temperature (−100 °C) crystallization from pentane. These products can then be elaborated separately, by identical sequences of reactions, to the opposite enantiomers of diol **8**. This diol, when reacted with an appropriate difunctional electrophile under

(31) Furuta, K.; Iwanaga, K.; Yamomoto, H. "Asymmetric Diels-Alder reaction. Cooperative blocking effect in organic synthesis," *Tetrahedron Lett.* **1986**, *27*, 4507–4510.

Scheme III



macrocyclization conditions, produces only the D_2 diastereomer of the resulting host (along with higher oligomers), enantiomerically pure.

C. Complexation Properties.

The aggregation behavior of these hosts is studied by finding the concentration above which their NMR spectra vary. All of them have cmc's below 1 mM, and some, such as M_R, have cmc's so low that they could not be determined. Such aggregation at low concentrations is detrimental to our investigation. In order to obtain reliable information about the host/guest complexation process, reactions of a host should be studied below its cmc. Studies carried out within the concentration range bounded below by the detection limit of the NMR and above by the cmc, however, have produced a wealth of data.

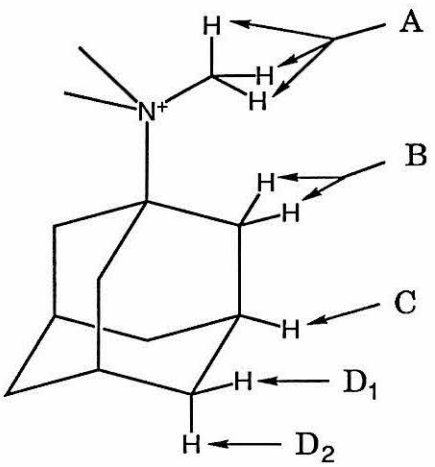
These tetraanionic ethenoanthracene-based macrocycles have indeed proven to be effective hosts for organic molecules in water. This section contains a brief,

interpretative overview of the binding behavior of hosts of this series and the forces we believe are responsible for these reactions.

Not surprisingly, sparingly-soluble aromatic compounds are brought into solution by these hosts. The first studies involved water-insoluble fluorescent guests such as anthracene and pyrene, whose solution spectra in the presence of host demonstrated that some of the dissolved material was in an environment less polar than water. Furthermore, a higher concentration was present than could be accounted for by their intrinsic solubilities.[29a] Any further study, such as a determination of the free energies of association, was stymied by the inconveniently low solubilities of such guests. More complete and informative studies had to await guests that were more soluble in water, yet still had an affinity for the hosts.

One of the most important such guests to be studied was 1-adamantyltrimethylammonium iodide (ATMA). This egg-shaped molecule has a large globular hydrophobic base and a cationic tip. Its D_3 symmetry makes it convenient for NMR study. It has five types of protons, denoted A, B, C, D_1, and D_2. Space-filling (CPK) models indicate that the adamantyl moiety has a good steric fit to hosts such as V, IV, P, and M. It was expected that the adamantyl portion would be bound in the host cavity by the hydrophobic effect, and that the trimethylammonium portion, which was present to lend water-solubility to the guest, would protrude from the cavity in order to associate with solvent.

Table I shows the free energies of association of several different hosts with ATMA, along with the D values of the ATMA protons in these complexes. D is the change that a proton's resonance undergoes upon binding: $D = \delta_{\text{free}} - \delta_{\text{bound}}$. This value contains information about the geometry of the host-guest complex. In an NMR experiment, the aromatic rings of the host perturb the magnetic field in their immediate vicinity, engendering a shielding region at their faces. Thus, guest protons in the interior of the host cavity, surrounded by the host aromatic rings, will be shifted upfield more than the protons protruding into the solvent.



ATMA

Table I. Characteristics of ATMA complexes with ethenoanthracene hosts in aqueous media[a].

host	D_A	D_B	D_C	D_{D_1}	D_{D_2}	$-\Delta G^o_{295}$
IV^b_{meso}	0.91	1.02	0.99	1.09	1.10	5.2
IV^b_{dl}	0.75	0.65	0.60	0.80	0.90	4.2
V^b_{meso}	2.51	2.84	0.96	1.10	0.97	4.6
V^b_{dl}	2.44	2.80	1.43	1.52	1.33	4.4
P^b_{dl}	1.85	2.90	1.19	1.30	0.76	6.9
C^c_l	1.25	2.64	0.92	1.02	0.41	5.4
V^c_{dl}	1.41	1.58	0.89	0.92	0.89	5.3
P^c_{dl}	1.95	3.09	1.22	1.23	0.75	6.7
M^c_{dl}	1.15	0.92	0.48	0.52	0.47	5.5

[a]From reference 29d. Values of D in ppm. [b]In phosphate buffer, pD $\simeq$ 7.5. [c]In borate-*d* buffer.

What actually occurred was not at all in accord with expectation. The shift patterns of ATMA with these guests suggest three different modes of binding. The first, exhibited by hosts IV, is characterized by similar shifting of all the ATMA protons. This indicates that there is no single preferred binding geometry; most likely, the guest is loosely associated with the host, which is collapsed into a "bowl" conformation. The second, exemplified by hosts V and M, shows large shifting of the protons *near the charged end* of the guest and smaller shifts of those at the base. We interpret this as a specific attraction between the cationic quaternary

ammonium group and the aromatic rings of the host. The similar shifting of the C, D_1, and D_2 protons in the guest's aliphatic portion show that the aliphatic part of the guest has no specific preferred orientation in the host cavity. The third and most oriented binding mode is displayed by hosts P and C. As in the second mode, the charged end of the guest is the deepest inside the host cavity. The aliphatic protons C, D_1, and D_2 show different shifting, however, suggesting an orientational preference. Specifically, the D_2 protons, which point roughly parallel to the C_3 axis of the guest, are shifted only slightly compared to the C and D_1 protons, which are more equatorial. This is consistent with a complex geometry in which ATMA is bound tip-first in the host cavity, with its C_3 axis coincident with the cavity C_2 axis of the host. In this geometry, the A and B protons, which are well within the host cavity, experience significant shielding, and the C and D_1 protons, at the edge of the cavity, are less shielded. The D_2 protons, which point away from the cavity and into the solvent, are the least shielded of all. These shifting patterns are illustrated pictorially in Figure 3, in which the protons shifted the most are colored white, and the others are shades of gray, the lightness of which is proportional to the D value.

This unexpected binding of ATMA's trimethylammonium group led to an investigation of other quaternary ammonium compounds. Ethenoanthracene hosts, especially host P, proved to be general receptors for such compounds. This indicated that some force other than the hydrophobic effect was operating. In fact, Michael Petti studied the guest (4-*t*-butylphenyl)trimethylammonium **13**, and found that, when this guest was complexed with host V_{dl} or P_m, its *N*-methyl protons were shifted more than its *t*-butyl protons. The hosts, given a choice between encapsulating a trimethylammonium group or a *tert*-butyl group, choose the trimethylammonium, leaving the nonpolar *tert*-butyl group in contact with the solvent![29d]

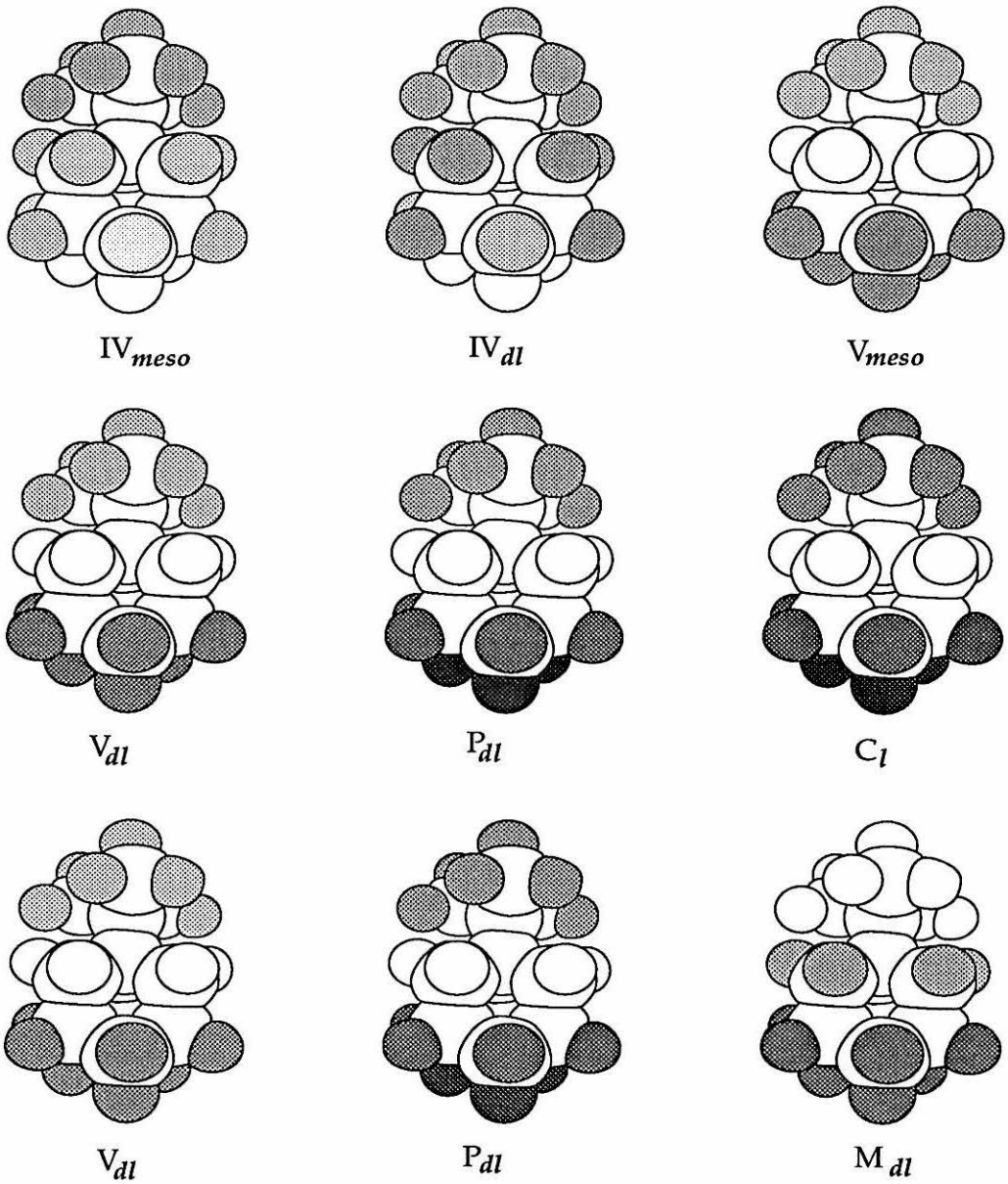


Figure 3. *D* values of ATMA protons in complexes with ethenoanthracene hosts. For each system, the shading of the protons indicates their relative *D* values. The most shifted proton is white, and the others are shades of gray corresponding to the ratios of their *D* values to that of the most shifted proton. The top six complexes were observed in pD $\simeq$ 7.5 phosphate buffer, and the bottom three in borate-*d* buffer.

One possible explanation for this general host affinity for quaternary ammonium groups is a Coulombic attraction between the positive charge of the guest and the

13

four negatively-charged carboxylate groups of the hosts. In an aqueous solvent, the attractive force between two oppositely-charged particles is attenuated by the large dielectric constant of water ($\varepsilon_{H_2O} = 78.54$), but when the cationic guest is in the host cavity, the interposing medium is not water, but a hydrocarbon. This has a lower dielectric constant than water ($\varepsilon_{phenol} = 9.78$, $\varepsilon_{toluene} = 2.38$),[32] making the force commensurately stronger.

While this dielectric effect is probably partly responsible for these hosts' affinities for cationic guests, it does not explain all of the observations. The most convincing evidence that the negatively-charged carboxylates are not the only cause of this attraction was found by David Stauffer in studies of the tetraester P_E in $CDCl_3$. This host is uncharged, yet it binds the same positively-charged guests in $CDCl_3$ that host P does in D_2O. Furthermore, no binding of neutral guests by P_E in chloroform was ever observed.[29e] This shows that there truly is an attraction between cationic guests and the macrocyclic host framework.

Support for this interpretation is lent by gas-phase studies of the interaction between NH_4^+ and benzene,[33] and by a statistical analysis of protein crystal structures that found that ammonium groups inside globular proteins are often closely associated with aromatic residues.[34] We hypothesize that the electron-rich π clouds

---

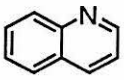
(32) *CRC Handbook of Chemistry and Physics*, 65th ed.; Weast, Robert C., Ed.; CRC: Boca Raton, FL, 1984; pp E-50–E-52.

(33) Meot-Ner (Mautner), M.; Deakyne, "Unconventional ionic hydrogen bonds. 2. $NH^+ \cdots \pi$. Complexes of onium ions with olefins and benzene derivatives," C. A. *J. Am. Chem. Soc.* **1985**, *107*, 469–474.

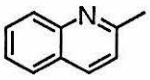
(34) Burley, S. K.; Petsko, G. A. "Amino-Aromatic interactions in proteins," *FEBS Lett.* **1986**, *203*, 139–143.

of the aromatic hosts polarize in response to a cationic guest, stabilizing its approach. Initially, we named this interaction the "ion-dipole" effect, but we realized upon further reflection that that this term had an unnecessarily specific and possibly misleading connotation. Hence, we now refer to it as the "cation-π" effect.

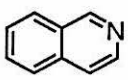
A related effect is seen in the complexation of neutral aromatic guests. Electron-deficient guests are always bound more strongly than comparable electron-rich guests. Table II shows some examples of this phenomenon. The first four guests in this table, **14**–**17**, incorporate the electron-poor pyridine system. All are bound more strongly than **18**, which is of similar size and shape, but contains the electron-rich pyrrole system. This can not be a consequence of less soluble guests most preferring the host cavity to water: of these five guests, **18** is the *least* soluble in borate-*d*. The last two guests show the effect of an electron-withdrawing nitro group. This effect is consistent with donor-acceptor π-stacking interactions between the aromatic systems of the electron-rich host and electron-deficient guest. Such effects had been previously observed by Diederich in organic, but not aqueous, media.[35]



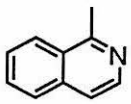
14



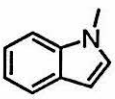
15



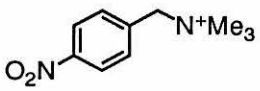
16



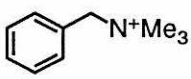
17



18



19



20

(35) Diederich, François; Dick, Klaus; Griebel, Dieter "Complexation of arenes by macrocycloc hosts in aqueous and organic solutions," *J. Am. Chem. Soc.* **1986**, *108*, 2273–2286.

Table II. Free energies of complexation of aromatic guests with ethenoanthracene hosts.[a]

	guests						
host	**14**	**15**	**16**	**17**	**18**	**19**	**20**
P	5.4	5.5	6.3	6.4	4.5	5.6*m*	5.1*m*
C	5.9	5.8	6.3	6.7	4.8	5.2	4.9
M	4.6	4.5	4.6	4.5	4.4	—	—
V	3.9	3.5	4.3	4.2	0.0[b]	5.3	4.6

[a]From reference 29d. Energies are $-\Delta G^{\circ}_{295}$ in kcal mol^{-1}. All hosts were the *dl* diastereomer except where designated (*m*), which was *meso*. [b]No change was observed in the guest NMR spectrum upon addition of host.

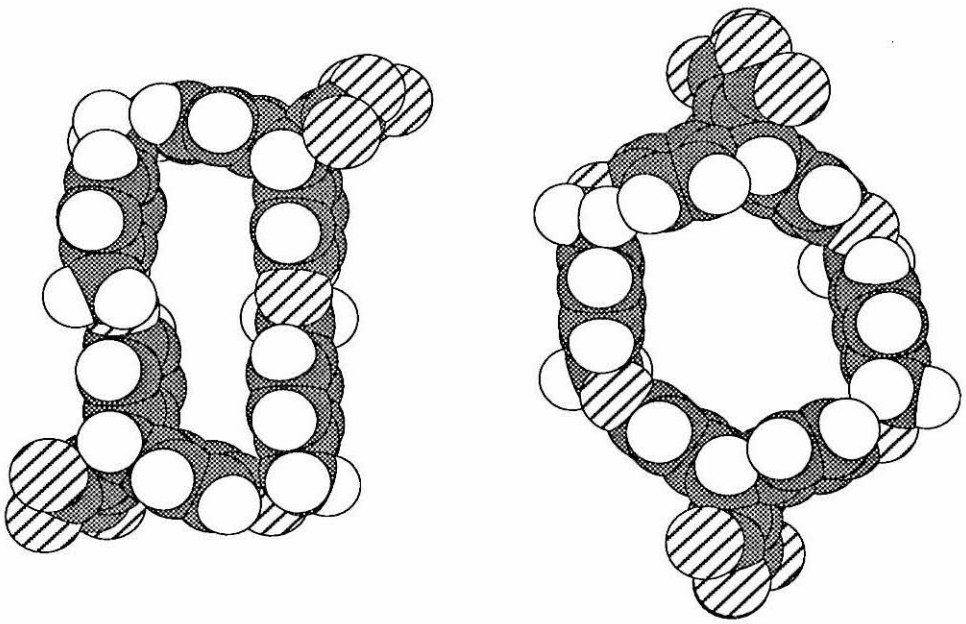


Figure 4. Host conformations. Left: rhomboid; Right: toroid.

The most intensely studied hosts of this family are the D_2 diastereomers of P and C. Two complexation geometries, known as the *toroid* and the *rhomboid* (Figure 4), are important for these compounds. The toroid conformation, which has D_2 symmetry, is an open structure with a nearly cylindrical cavity. This is the conformation adopted when binding globular guests such as ATMA and other trimethylammonium compounds. The rhomboid conformation is more flattened,

and has only C_2 symmetry. The cavity in this case is more a parallepiped, and is complementary in size and shape to a naphthalene molecule. Aromatic guests such as quinoline, indole, and 1-methyl-4-(dimethylamino)pyridinium are bound by hosts in this conformation.

These ethenoanthracene-based macrocycles have proven to be selective hosts, able to discriminate between guests on the basis of electronic as well as steric characteristics. This discrimination cannot be explained solely by guest/solvent interactions: certainly, specific forces between host and guest are operating. These early studies have demonstrated that ethenoanthracene-based hosts exhibit true molecular recognition, and have uncovered some of the reasons for this behavior.

Chapter 2

An Improved Method for Determining Bimolecular Association Constants from NMR Titration Experiments

Abstract: A nonlinear regression procedure for fitting estimates of an association constant and saturation shifts to NMR titration experiments under fast-exchange conditions is described. The method assigns weights to each observation by propagating measurement errors through the fitted model. A series of Monte Carlo studies simulating a variety of possible experimental conditions has shown this method to be significantly superior to other methods in common use.

I. Introduction

A. Quantitative Measurements.

The purpose of research in molecular recognition is to understand how and why intermolecular association processes occur. Comparison of different systems enables the clever investigator to recognize underlying themes and to piece together isolated observations, creating a consistent understanding of a process.

Useful comparisons require quantitative measurements. It is not enough to merely say that compounds H and G associate in solution: in order to compare the complexation of H and G with that of H and J, one needs objective information about both systems. What is the stoichiometry of each system? What is the strength of each interaction? What are the geometries of the intermolecular complexes? Consideration of a generic molecular recognition process reveals the physical measurements needed to answer these questions.

The simplest possible association equlibrium is one in which two separate species come together to form a binary complex.

$$H + G \rightleftharpoons H{\cdot}G \tag{1}$$

This interaction is governed by the usual laws of chemical equilibrium. In particular, the concentrations of free host [H], free guest [G], and host/guest complex [H·G] are interrelated by an equilibrium constant K.

$$K = \frac{[\mathrm{H{\cdot}G}]}{[\mathrm{H}][\mathrm{G}]} \tag{2}$$

The magnitude of the equilibrium constant is a measure of the strength of the attraction. This association constant contains the same information as two other common measures of intermolecular affinity. These are the dissociation constant K_d, and the free energy of complexation $\Delta G°$. The dissociation constant is simply the equilibrium constant of the reverse of reaction 1, and is the reciprocal of K. The free energy of complexation is the difference between the energy of a mole of complex H·G and the energy of a mole of each of the dissociated species H and G. This is related to K by the Boltzmann distribution.

$$\Delta G° = -RT \ln K \tag{3}$$

In this equation, R is the molar Boltzmann constant, 1.98719 cal mol^{-1} K^{-1}, and T is the absolute temperature.

B. Determinimg K.

Since the strength of the interaction is the most important characteristic of a complexation reaction, it is imperative that there be a good way to measure it. Because intermolecular complexation in general is such a studied topic, there are a multitude of ways to carry out such a measurement. I will briefly describe some techniques as they apply to our systems; for a comprehensive overview, turn to Connors.[1]

---

(1) Connors, Kenneth A. *Binding Constants*; Wiley-Interscience: New York, 1987.

1. Direct methods. Any of the quantities K, K_d, or ΔG° can be obtained by determining [H], [G], and [H·G]. Conversely, these concentrations can be predicted from a knowledge of K.

$$\text{total host concentration} = [\text{H}]_0 = [\text{H}] + [\text{H·G}] \tag{4}$$

$$\text{total guest concentration} = [\text{G}]_0 = [\text{G}] + [\text{H·G}] \tag{5}$$

Substituting equations 4 and 5 into 2 and solving for [H·G] yields

$$[\text{H·G}] = \frac{1}{2}\left\{[\text{H}]_0 + [\text{G}]_0 + 1/K - \sqrt{\left([\text{H}]_0 + [\text{G}]_0 + 1/K\right)^2 - 4[\text{H}]_0[\text{G}]_0}\right\} \tag{6}$$

as the physically meaningful root. Conceptually, the most direct way to measure K would be to combine host and guest together in a reaction vessel and count the number of molecules in the free and bound states. Technically, this task is not simple. To directly measure the concentrations of compounds existing together in solution, there must be not only a way to relate some observable quantity to a concentration, but also a means to discern one species from another. Optical absorption methods in principle provide a means for measuring concentrations, because absorbance is directly proportional to concentration. The absorbance due to uncomplexed host or guest alone provides enough information to calculate all of the concentrations in solution, even if the extinction coefficient of the host/guest complex is unknown. In practice, however, electronic absorption bands are so broad that the spectra of different species overlap. If both the shape and extinction coefficient of the absorbance of the complex is unknown, a single spectrum does not contain enough information to specify any of the solution concentrations.[2]

---

(2) However, estimation of K by nonlinear fitting of a set of spectra is possible.

NMR spectroscopy has very desirable qualities for concentration determination. First of all, spectrometer settings can be adjusted so that the integral of a compound's NMR signal is directly proportional to the concentration of the compound. Furthermore, NMR signals are narrow and well-resolved, especially on high-field spectrometers. Thus, overlapping signals are less of a problem than in electronic excitation spectroscopy.

In an NMR spectrum of a solution containing host and guest, there should be two resonances for each interacting species. One, at δ_{free}, will be from the species in its uncomplexed state, which can easily be identified by comparison to a spectrum of the species alone in solution. The other signal, at δ_{bound}, is from the complexed species. The relative populations of the free and bound states can be determined from the relative integrals of these two signals.

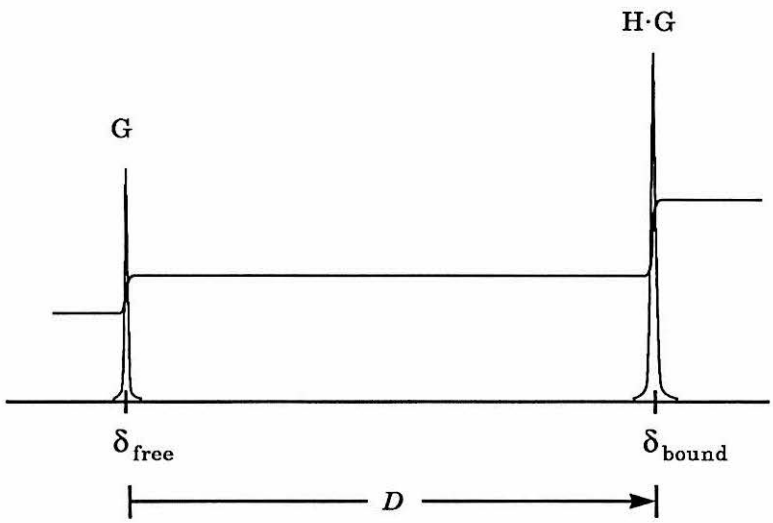


Figure 1. Determining the fraction of a species in the complexed state by comparing the areas under the NMR signals from the two environments.

In fact, each proton of both host and guest will provide this information, furnishing several independent measurements of [H·G]. Additionally, the magnitude of the change in the peak position, D, contains information about the geometry of the bound complex. Thus, a single spectrum can be very rich in information, revealing both the strength and the geometry of the interaction.

If, however, the complexation/decomplexation rate is faster than the difference between the resonant frequencies of the nuclei in the two environments, separate signals will not be observed. Instead, the NMR signal of a proton of a rapidly interconverting species will appear as a single resonance. Its position will be between the free and bound resonance positions δ_{free} and δ_{bound}, weighted by the relative population of each state.

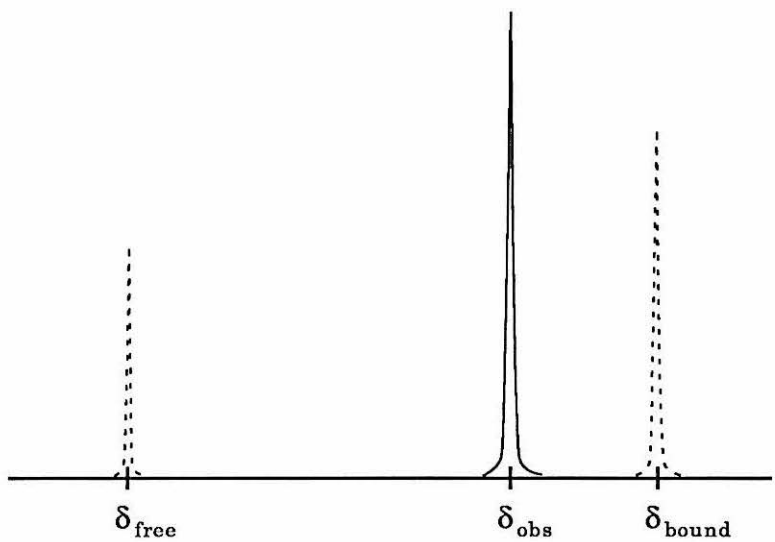


Figure 2. Appearance of the NMR spectrum of an associating species under fast-exchange conditions.

If the observed proton belongs to the guest species, it peak position is described by equation 7.

$$\delta_{\text{obs}} = \delta_{\text{free}}\frac{[\text{G}]}{[\text{G}]_0} + \delta_{\text{bound}}\frac{[\text{H}\cdot\text{G}]}{[\text{G}]_0} \tag{7}$$

If it is part of the host species, the position is similarly given by equation 8.

$$\delta_{\text{obs}} = \delta_{\text{free}}\frac{[\text{H}]}{[\text{H}]_0} + \delta_{\text{bound}}\frac{[\text{H}\cdot\text{G}]}{[\text{H}]_0} \tag{8}$$

For generality, we shall call the observed species "S." Using this notation, both equations 7 and 8 are specific cases of equation 9.

$$\delta_{\text{obs}} = \delta_{\text{free}}\frac{[\text{S}]}{[\text{S}]_0} + \delta_{\text{bound}}\frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0} \tag{9}$$

According to equations 4 and 5, $[\text{S}]_0 = [\text{H}\cdot\text{G}] + [\text{S}]$. Making this substitution into equation 9 yields

$$\begin{aligned}
\delta_{\text{obs}} &= \delta_{\text{free}}\frac{[\text{S}]_0 - [\text{H}\cdot\text{G}]}{[\text{S}]_0} + \delta_{\text{bound}}\frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0} \\
&= \delta_{\text{free}}\left(1 - \frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0}\right) + \delta_{\text{bound}}\frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0} \\
&= \delta_{\text{free}} - (\delta_{\text{free}} - \delta_{\text{bound}})\frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0}.
\end{aligned}$$

It is convenient to express δ_{obs} in terms of the maximum upfield shift D and fraction of species bound F.

$$D \equiv \delta_{\text{free}} - \delta_{\text{bound}} \tag{10}$$

$$F \equiv \frac{[\text{H}\cdot\text{G}]}{[\text{S}]_0} \tag{11}$$

The expression for the observed signal then is

$$\delta_{\text{obs}} = \delta_{\text{free}} - DF. \tag{12}$$

As a greater fraction of the species becomes bound, δ_{obs} moves steadily from δ_{free} to δ_{bound}. If δ_{free} and D are both known independently, δ_{obs} very simply leads to F, [H·G], and ultimately K.

$$F = \frac{\delta_{free} - \delta_{obs}}{D} \tag{13}$$

If either the free or bound chemical shifts are unknown, however, a single NMR spectrum does not impart enough information to determine the concentrations. Ordinarily, δ_{bound} is not independently known. In such cases, a set of samples is prepared in which $[H]_0$ and $[G]_0$ vary, so that δ_{obs} appears at different frequencies in different samples. These observed frequencies follow equation 12, which, when fully expanded, is equation 14.

$$\delta_{obs} = \delta_{free} - D\,\frac{[H]_0 + [G]_0 + 1/K - \sqrt{\left([H]_0 + [G]_0 + 1/K\right)^2 - 4[H]_0[G]_0}}{2[S]_0} \tag{14}$$

D and K are the only quantities in this equation that cannot be independently measured. However, estimates of D and K may be evaluated by substituting them, along with measured values of the explanatory variables $[H]_0$, $[G]_0$, and δ_{free}, into this equation. This generates δ_{calc}, the predicted value of the response variable δ_{obs}. If the model is correct, if the measurements are performed without error, and if the estimates of K and D are equal to the true values of these parameters, then the predicted δ_{calc} and measured δ_{obs} will be identical. Measurement errors, however, make the predictions unlikely to exactly equal the observations, even if the model and parameters are correct.

2. Least-squares estimation. Nonetheless, a good model should duplicate the observed data as closely as possible. This means that the residuals, the differences between the actual observations and the predictions of the model, should be

small. This is reflected in the criteria used for finding parameter estimates. Most often, parameters of physical models are determined by M-estimates, which are parameter values that minimize some loss score. In *least-squares* estimation, this score is often the unweighted sum of squared residuals. For an experiment involving observations of P protons in each of N samples, this score is defined as

$$\mathrm{SSR} = \sum_{p=1}^{P} \sum_{i=1}^{N} (\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2 \,. \tag{15}$$

Minimization of SSR is not the only useful criterion for evaluating parameter estimates. The real objective of parameter estimation is not merely to find a model to fit some data set, but to find the true parameter values. It is thus desirable for estimation procedures to reliably determine values near the true ones. The performance of any estimator may thus be evaluated by its *bias* and *variance*. Bias is the difference between the expectation, or mean, of the estimator and the true value of the parameter. An ideal estimator is unbiased, that is, its expectation is exactly the true parameter value. The variance of a random variable is the second moment about its mean, that is, the average square of the difference between a random occurrence of the variable and the variable's true mean. This is a measure of the spread of the variable: a small variance means that the variable's distribution is very compact. A good estimator thus has a small bias and a small variance.

Least-squares estimators are used primarily because of their convenience and desirable properties in special cases.[3] If the predictive model $y = g(\mathbf{x}, \theta)$ for the relation between the explanatory variables $\mathbf{x}$ and the response variable y is a linear

(3) Seber, G. A. F.; Wild, C. J. *Nonlinear Regression*; Wiley: New York, 1989; Chapter 12.

function of its parameters θ, and if measurement uncertainties arise only in the response variables,

$$y_{\text{obs}} = g(\mathbf{x}, \theta) + \varepsilon_i \,,$$

then the least-squares estimator of θ is unbiased. Even if the predictive model $g(\mathbf{x}, \theta)$ is a nonlinear function of θ, the least-squares estimate of θ is still consistent, that is, it converges to the true value of θ in the limit of infinite sample size. Furthermore, if the measurement errors are normally distributed, then a least-squares parameter estimator is a maximum likelihood estimator as well; that is, it returns the parameter value that most likely gave rise to the observed data. Finally, the least-squares parameter estimates for linear models are mathematically easy to find.

Such regularity conditions are not fulfilled by the model of equation 14. This model is not a linear function of the parameter K. In addition, there are measurement errors in all of the measured variables, not only in the response variables.[4] As a result, there is no guarantee that least-squares parameter estimates are good by any standards.

This chapter reports my efforts to develop good estimates for D and K from NMR titration experiments. Statistical theory is unable to identify estimators that have optimal properties for this model. Consequently, the search has been qualitative and empirical rather than rigorously theoretical.

---

(4) In fact, the determination of the response variables δ_{obs} is among the most precise of all measurments performed in an NMR titration binding study.

II. Fitting Methods.

A. NMRfit.

When I first joined the molecular recognition project, the parameters K and D were assigned by minimizing the unweighted SSR of equation 16.

$$\mathrm{SSR} = \sum_{i=1}^{N} (\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2 \tag{16}$$

This is simpler than the SSR of equation 15; it counts only the observations of a single proton. If a binding study was performed in which the resonances of more than one proton were followed, K and D were determined separately for each proton. According to the NMR titration model, the observations of all protons should give the same estimate of K. In fact, measurement errors ensure that this will never occur exactly. When a binding study of a system with P observed protons is carried out, P different estimates of the association constant are returned. A "best" estimate of the association constant was typically devised by averaging the estimates from the individual protons, or by disregarding all protons but the one the model was best able to fit.

B. Multifit.

As a first step up from this simple analysis, I developed a Pascal program (Multifit) to fit a single binding constant to all of the observations in a binding study. If P protons are observed in a study, there are $P + 1$ adjustable parameters in the predictive model: K and the P D's. The association constant K returned by this procedure is the true least squares estimate. It is more reliable than an estimate from any single proton, because it is based on more information. Furthermore, the D values returned are consistent with each other, which is not necessarily the case if the protons are fitted separately. This allows confident comparison of the D values

of different protons. Such an ability is essential for describing the geometry of the complex.

This procedure, nonetheless, had some undesirable properties. The most severe is its assignment of equal weights to all observations. For instance, it is common for different protons to have values of D that are very different in magnitude. Protons with large absolute values of D will change their peak positions in an NMR binding study much more significantly than will protons with smaller D's. An error in sample host concentration, for instance, will cause the predicted resonance of a proton with a large D to be further from the truth than the predicted resonance of a proton with a smaller D. Since the resonance of a proton with a large D is more likely to be predicted inaccurately, it should be possible to penalize residuals from its observations less severely than residuals from the observations of other protons. The unweighted least-squares estimation procedure places the greatest relative importance on fitting the observations of the proton whose signals move the most. This is not the best use of all of the information available in the experiment.

C. Emul.

1. Design. To combat this drawback, I developed a *weighted* least-squares fitting program (Emul). This program minimizes SSR*, a different loss score than the SSR of Multifit.

$$\mathrm{SSR}^\star = \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2}{\sigma_{pi}^2} \tag{17}$$

In this equation, σ_{pi} is the estimated inaccuracy in predicting the resonance of proton p in sample i. This value is determined by a first-order approximation of the influence of each measurement error on the eventual magnitude of the residual $(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})$. If this residual is affected by L measured variables x_j, each of

which can be thought of as a random variable with variance $\sigma^2_{x_j}$, then the estimate of σ^2_{pi} is[5]

$$\sigma^2_{pi} = \sum_{j=1}^{L} \sigma^2_{x_j} \left(\frac{\partial(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})}{\partial x_j} \right)^2 . \tag{18}$$

This is expected to provide a good weighting factor because the derivative $\partial(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})/\partial x_j$ tells how much a change in the variable x_j changes the residual $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})$. The value σ_{x_j} is the standard deviation of the measurement of x_j. Thus, multiplying the expected deviation of x_j by the effect of x_j on $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})$ gives the expected magnitude of $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})$ as a result of the error in x_j. The expected total magnitude of $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})$ is the sum of all the deviations arising from each of the variables x_j. When random variables are added together, the variance of the resulting sum is equal to the sum of the variances of the original variables. Thus, the variance of $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})$ is given by the sum in equation 18. This weighting factor σ^2_{pi} is the expected value of $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})^2$.

2. Execution.

The first step toward creating a procedure to minimize SSR* was to develop the means to evaluate SSR* itself. This required identifying the fundamental random variables x_j, their uncertainties $\sigma^2_{x_j}$, and the derivatives $\partial(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})/\partial x_j$.

The identities of the fundamental random variables of an experiment depend on the design of the experiment itself. Strictly, every measurement performed is a random variable. Binding studies are typically performed in the Dougherty group by combining stock solutions of host and guest together with additional buffer in an NMR sample tube, and recording the spectrum. The values of $[H]_0$ and $[G]_0$

(5) Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill; New York, 1969; p 60.

are then altered by adding more host solution, guest solution, or buffer, and the NMR spectrum is again recorded. The steps of adding solution and recording the spectrum are repeated several times, and the spectra of uncomplexed host and guest are measured independently.

The fundamental random variables contributing to a single observation are:

- The host and guest concentrations, $[H]_s$ and $[G]_s$, of every stock solution used to make up the sample,
- the volume V_a of each solution aliquot added to the sample tube,
- The calibration I of the delivery devices (pipets or syringes) employed to add the aliquots, and
- the NMR peak position measurements $\delta_{\mathrm{obs}\,pi}$ and $\delta_{\mathrm{free}\,p}$.

The stock solution concentrations are perhaps the most significantly mis-measured quantities in a binding study. Ordinarily, these concentrations are determined by NMR integration against a known standard, such as 3,3-dimethylglutaric acid or potassium hydrogen phthalate. NMR integrations are notoriously imprecise, and are not considered valid to within less than about five percent. Aliquot volumes are determined by two related but independent random variables: delivery device precision and delivery device accuracy. Precision is the reproducibility of volumes added by a device. The precision errors of aliquots added by the same device are independent and identically distributed with a mean of zero. Accuracy is a measure of the likely calibration error of the delivery device. The volumes of all aliquots delivered by a single device will be mis-measured by the same proportional amount; for instance, they all may be two percent too low. Thus, the difference between an aliquot's true and measured volumes is (measured value) $\times$ (calibration error) $+$ (reproducibility error). The final fundamental random variables considered are the

NMR measurements. Because NMR signals are well-resolved and reproducible, the errors in these variables are very small. The principal source of such error is the digitization of the spectrum: the peak position cannot be known more specifically than the distance between two points. Another possible but not always present contributor to peak position measurement error is peak width. If a peak is very broad, it is difficult to tell exactly where its center lies.

Once the fundamental random variables have been identified, it is necessary to determine their impacts upon the observations according to equation 18. This task is tedious but straightforward: it requires only differentiation of $(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})$ with respect to each of the fundamental variables. Substitution of equations 12 and 6 into equation 18 put it in terms of the the fundamental random variables $\delta_{\mathrm{obs}\,pi}$ and $\delta_{\mathrm{free}\,p}$, and the not-so-fundamental random variables $[\mathrm{H}]_0$ and $[\mathrm{G}]_0$. These concentrations can in turn be expressed in terms of fundamental random variables. In any sample solution created by adding aliquots of other solutions together, the total host and guest concentrations are given by equations 20 and 21.

$$V = V_t + IV_a \tag{19}$$

$$[\mathrm{H}]_0 = \left([\mathrm{H}_t]_0 V_t + [\mathrm{H}_a]_0 IV_a\right)/V \tag{20}$$

$$[\mathrm{G}]_0 = \left([\mathrm{G}_t]_0 V_t + [\mathrm{G}_a]_0 IV_a\right)/V \tag{21}$$

V is the total sample volume; it is the sum of V_t, the volume of solution in the sample tube before addition of the most recent aliquot, and IV_a, the volume of the most recent aliquot. The calibration of the delivery device adding the most recent aliquot is I; its "measured" value is unity. The nominal volume of the aliquot is V_a. $[\mathrm{H}_t]_0$ and $[\mathrm{G}_t]_0$ are the total host and guest concentrations of the sample before the addition of the most recent aliquot, and $[\mathrm{H}_a]_0$ and $[\mathrm{G}_a]_0$ are the total

host and guest concentrations of the added solution. The only random variables in equations 19–21 that are necessarily fundamental are V_a and I. All of the other variables, however, can eventually be decomposed into fundamental variables if they are not fundamental themselves. V_t, $[H_t]_0$, and $[G_t]_0$ are the V, $[H]_0$, and $[G]_0$ of the sample previously in the tube; consequently, these values are all zero for the first solution in a tube. The added solutions may be but are not required to be stock solutions. If they are stock solutions, then $[H_a]_0$ and $[G_a]_0$ are the fundamental random variables $[H]_s$ and $[G]_s$; if they are not, then they are ultimately composed of stock solutions added together. The details of determining these derivatives are given in the Appendix to this chapter.

D. Method of Creswell and Allred.

Another popular method for finding K from NMR titration experiments was developed independently by Creswell and Allred[6] and by Horman and Dreux,[7] and is currently championed by Wilcox.[8] This method involves using δ_{free} as an adjustable parameter instead of an independently-measured variable. It is claimed to be superior to methods in which δ_{free} is directly measured, because the parameter estimates are unaffected by errors in the determination of δ_{free}. In other methods (such as Emul), if δ_{free} is measured erroneously, the model is systematically compromised. The method of Creswell and Allred determines δ_{free} from the entire data set, instead of relying on a single measurement.

---

(6) Creswell, Clifford J.; Allred, A. L. "Thermodynamic constants for hydrogen bond formation in the chloroform-benzene-cyclohexane system," *J. Phys. Chem.* **1962**, *66*, 1469–1472.

(7) Horman, Ian; Dreux, Bernard "Estimation of Association constants of bimolecular organic complexes," *Anal. Chem.* **1983**, *55*, 1219–1221.

(8) Wilcox, Craig S. "Design, synthesis, and evaluation of an efficacious functional group dyad. Methods and limitations in the use of NMR for measuring host-guest interactions," In *Frontiers in Supramolecular Organic Chemistry and Photochemistry*, Schneider, H.-J.; Dürr, H., Ed.; VCH: Weinheim, 1990.

III. Comparison of Fitting Methods.

A. Design.

Although weighted least-squares fitting carried out by propagation of errors intuitively appears worthwhile, there is no theoretical proof that its parameter estimates are better than any others. In order to compare different fitting schemes to each other, I have tested them on data sets generated by Monte Carlo simulation experiments.[9] Each data set is fitted by the regression procedures being compared, creating parameter estimates from each procedure. The behavior of the estimates over a large number of data sets provides an empirical basis for the comparison of the different procedures.

Such comparisons were carried out for a variety of experimental designs, covering the range of binding constant values that can reasonably be determined from NMR titration experiments. Five basic types of experimental design were modeled: (1) adding aliquots of host stock solution to a sample tube containing guest; (2) adding aliquots of guest stock solution to a sample tube containing host; (3) adding aliquots of diluent to a sample tube containing both host and guest; (4) a Job or continuous variation study, in which $[H]_0 + [G]_0$ is the same in all samples, and the mole fraction of each species is varied in equal steps from 0 to 1; and (5) making $[H]_0$ the same in all samples, changing only the concentration of guest. Each experiment involved fifteen observed samples; in each of these samples $[H]_0$ was between 10 and 200 μM, and $[G]_0$ was between 10 and 500 μM. Two protons were followed, one from the host and the other from the guest. D of the host proton was -100 Hz, and D of the guest proton was $+500$ Hz. For each experimental design, four

(9) The generation of these Monte Carlo data sets is described in the Experimental Section of this chapter.

association constants K were considered: 10^3, 10^4, 10^5, and 10^6 M^{-1}. Each of these twenty experiments was designed to provide a good measure of the association constant by keeping the fraction of the minor component bound between 0.2 and 0.8.[8] These sets are summarized in Table I. When the association constant was 10^3, the method of continuous variation (design 4) proved to be an extremely poor experimental design. Small simulated measurement errors led to a preponderance of terrible parameter estimates. As a result, this set was not included in the large study; only the remaining nineteen sets were used.

Table I. Names of experimental designs simulated in Monte Carlo studies.

	$K^a =$			
design	10^3	10^4	10^5	10^6
adding host	H3	H4	H5	H6
adding guest	G3	G4	G5	G6
adding diluent	D3	D4	D5	D6
continuous variation	J3*	J4	J5	J6
constant $[H]_0$	V3	V4	V5	V6

[a]In M^{-1}. *Not included in simulations.

Except as specified otherwise, measurement errors were as follows. The standard deviation of stock solution concentration measurements was 5%, and the standard deviation of NMR peak position measurements was 0.5 Hz. Aliquot volume errors and delivery device calibration errors depended on the delivery device used. Delivery device accuracy and precision error distibutions were adapted from the specifications for Eppendorf Varipette 4810 piston stroke pipettes.

B. Testing Error Propagation.

The first comparisons performed were to assess the importance of propagating the measurement errors in each of the fundamental explanatory variables. In each

of these comparisons, 1500 Monte Carlo replications of each of the nineteen experiments under consideration were performed. The data set from each replication was subjected to three types of least-squares fit. The first method minimized the sum of squares of the unweighted residuals, SSR (equation 15). The third minimized the sum of squares of the weighted residuals, SSR* (equation 17), with the weights calculated by propagation of all measurement errors according to equation 18. The second also minimized a weighted sum of squares, but with the errors in one type of measurement not propagated. This was to determine if propagating each of the different types of measurement error was beneficial or detrimental. If it is disadvantageous to propagate a certain type of measurement error, then this abbreviated procedure should perform better than the one with full propagation. The value of propagating each type of measurement error was tested in this way.

C. Evaluating Performance.

Six measures of performance of the fitting methods were calculated under each experimental condition. These measures were the medians and standard deviations of the three fitted parameters K, D_1, and D_2. These provide a way to evaluate the bias and variance of the parameter estimates from the fitting procedures. Medians were evaluated in preference to means because the median is a more robust measure of central tendency. The performances of the three fitting procedures with respect to each of these six measures were compared and ranked. The method with the best performance in a measure received the rank of 1, the second best received the rank of 2, and the worst the rank of 3. If some procedures performed indistinguishably well (if they were tied), each received the same rank, which was the average of the ranks they would have received if they had been slightly different.

Let us take as an example the replications of experiment H3, in which the association constant is 10^3 M^{-1} and the protocol follows experimental design 1. The distribution of the estimates of K from the unweighted minimization had a standard deviation of 138; from both the fully- and partially-weighted procedures, the standard deviation of this same estimate was 149. Thus, the unweighted procedure received a rank of 1, and the others both received ranks of 2.5. The *medians* of all three of these distributions were 1.00×10^3 M^{-1}, however, so each procedure received a rank of 2 for this measure.

Each study thus produced $19 \times 6 = 114$ sets of rankings of these three fitting procedures. In order to determine if one fitting procedure performs significantly better overall than any of the others, these rankings have been evaluated by a Friedman-Cochran-McNemar test.[9] This nonparametric statistical test is designed to determine if there is a significant difference between s subjects that have been ranked by N independent judges. This is evaluated by a statistic related to the variance in the sums of the N ranks received by each subject. Let us define the total rank R_i if the ith subject:

$$R_i = \sum_{j=1}^{N} \text{rank}_{ij} ,$$

that is, the sum of the N ranks received by subject i. If there is no difference between subjects, these ranks will all have been assigned randomly and uniformly, so that all sums R_i are about the same. The test statistic

$$Q = \left(\frac{12N}{s(s+1)} \sum_{i=1}^{s} R_i^2 \right) - 3N(s+1)$$

(9) Lehmann, Erich Leo *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day: San Francisco, 1975; p 265.

will be distributed as a χ^2 variable with $s-1$ degrees of freedom. A very large Q rejects the null hypothesis that the subjects are indistinguishable.

If there are ties in the rankings by a judge, the value of Q will be artificially lowered, making this statistic not exactly follow the χ^2_{s-1} distribution. To correct for this effect, the modified statistic $Q^\star$, based on rank sums $R_i^\star$ that may contain ties, is used.

$$Q^\star = \frac{\frac{12}{Ns(s+1)}\sum_{i=1}^{s} R_i^{\star 2} - 3N(s+1)}{1 - \frac{\sum_{j=1}^{N}\sum_{i=1}^{e_j}(d_{ij}^3 - d_{ij})}{Ns(s+1)}}.$$

The index d_{ij} here is the number of subjects assigned rank i by judge j. Each judge j assigns e_j distinct ranks. If there are no ties, then $e_j = s$; if some subjects are tied, then $1 \leq e_j < s$. If each rank assigned by judge j is different, then $\sum_{i=1}^{s}(d_{ij}^3 - d_{ij})$ reduces to $\sum_{i=1}^{s}(1-1) = 0$, making $Q^\star$ identical to Q. If there are any ties, however, the denominator of $Q^\star$ becomes less than 1. In this way, $Q^\star$ corrects for the lowering of the sum of squares by ties.

In this evaluation of fitting methods, the subjects are the fitting methods and the judges are the sets of experimental conditions. It is most informative to make paired comparisons of fitting methods, that is, to compare one method to one other. With three fitting methods, there are $\binom{3}{2} = 3$ such comparisons to be made. These comparisons can be carried out by the Friedman-Cochran-McNemar test, with $s = 2$. Since s is 2 instead of 3, slightly different ranks from those assigned from the full set of three fitting methods must be used. These new ranks are easily derived from the old ranks: the subject with the lowest rank is assigned a new rank of 1, and the subject with the highest rank is assigned a new rank of 2. If the subjects are tied, both receive new ranks of 1.5. These ranks are summed and squared to obtain $R_i^{\star 2}$

and $Q^\star$. If the two subjects are indistinguishable, $Q^\star$ will follow the χ^2_1 distribution. The null hypothesis of indistinguishability is rejected if $Q^\star$ falls above some cutoff for this distribution. The 95% cutoff for this distribution, for instance, is 3.84.

The performance of these three fitting methods according to the six different measures can be conveniently summarized in the following manner. If one method performs significantly better than another, that is, if $Q^\star$ from the head-to-head comparison is greater than 3.84, then the winning method receives a score of $+1$ and the losing method receives a score of -1. If no significant difference is found between the two methods, each receives a score of 0. The scores a method receives in its comparisons to the other two methods are added together to give a total score for that measure.

For example, let us examine the standard deviations of K estimates in the test of propagating aliquot volume reproducibility errors. In this comparison, full propagation of errors proved to be significantly superior both to no error propagation at all and to propagation of all errors except for aliquot volume reproducibility errors. Furthermore, the partial propagation method was significantly better than the method of no propagation at all. The scores assigned are thus $1 + 1 = 2$ to the full propagation method, $1 - 1 = 0$ to the partial propagation method, and $-1 - 1 = -2$ to the no propagation method. This is illustrated in Figure 3.

	none	partial	full	total
none	—	-1	-1	-2
partial	+1	—	-1	0
full	+1	+1	—	+2

Figure 3. Chart showing the scores assigned to the three fitting methods in their direct comparisons according to some measure. The cell in row r and column c contains the score given to method r when compared to method c. The "total" column contains the sum of scores for each row.

Table II shows the total scores given to the three fitting methods for each of the six measures of fitting method performance. The final column gives the sum of scores assigned by these six measures for each of the fitting methods. Comparison of the total scores for the competing methods reveals which method performs best overall. Six separate studies are summarized in this table. In all of these studies, the fitting method that does not propagate errors performs worse overall than the method employing full error propagation. In no case does propagation of a subset of the measurement errors perform better overall than does full propagation. Consequently, I believe that this full error-propagation method is justified. There is no indication that propagating fewer measurement errors would produce a better estimation procedure.

D. Other Fitting Procedures.

A similar series of Monte Carlo studies was also performed to compare a larger class of fitting procedures. In these studies, five fitting methods were compared. These methods were: (1) Fitting the entire data set at once by adjusting a single association constant, and the free chemical shift and saturation shift ($\delta_{\text{free}\,p}$ and D_p)

Table II. Relative performances of fitting methods in a test of error propagation

	K		D_1		D_2		
propagation	med	sdev	med	sdev	med	sdev	total
			aliquot volume errors				
none	0	−2	0	−2	0	−2	−6
partial	0	0	0	1	0	1	2
full	0	2	0	1	0	1	4
			device calibration errors				
none	0	−2	−2	−2	0	−2	−8
partial	0	1	1	1	0	1	4
full	0	1	1	1	0	1	4
			stock solution concentration errors				
none	0	−1	−1	−1	0	−1	−4
partial	0	0	0	0	0	−1	−1
full	0	1	1	1	0	2	5
			all NMR spectrometer errors				
none	0	−1	0	0	0	−1	−2
partial	−1	0	0	0	−1	0	−2
full	1	1	0	0	1	1	4
			δ_{free} errors only				
none	0	−2	0	0	0	−2	−4
partial	0	1	0	0	0	1	2
full	0	1	0	0	0	1	2
			δ_{obs} errors only				
none	1	−2	1	0	0	0	0
partial	−2	1	−2	0	0	0	−3
full	1	1	1	0	0	0	3

for each proton. All observations are weighted equally, and no use is made of an independent measurement of a proton's free chemical shift. When P protons are observed, this method has $2P + 1$ adjustable parameters. This is the method of Creswell and Allred, generalized to accomodate more than one proton. (2) Fitting the entire data set by adjusting the same parameters as in method 1, but treating an independent measurement of a proton's free chemical shift δ_{free} as an additional observation. This adds one squared residual term $(\delta_{\text{free calc }p} - \delta_{\text{free obs }p})^2$ to the fit score SSR for every proton observed. This method is intermediate between methods 1 and 4. (3) Fitting the observations for each proton separately. The association

constant K and saturation shift D_p are optimized for the observations on a single proton, and this process is carried out for each proton. After all observations have been modeled in this manner, the estimates of K from each proton are averaged to give the overall "best" estimate of K. This is the method of NMRfit. (4) Fitting the entire data set by adjusting a single association constant K and the saturation shifts D_p of all observed protons. All observations are weighted equally. This procedure has $P + 1$ adjustable parameters. This is the method employed by Multifit. (5) Fitting the entire data set by adjusting the association constant K and the saturation shifts D_p of each observed proton. Observations are assigned weights by propagating all measurement errors according to equation 18. This is the method used by Emul.

These studies were carried out in a manner similar to that used for testing the propagation of the different classes of measurement errors. Every Monte Carlo data set produced was fitted by each of the five fitting methods. The performances of the fitting methods according to the fitted parameter means and standard deviations were evaluated and ranked for each set of experimental conditions. Head-to-head comparisons of pairs of fitting methods were evaluated by using the Friedman-Cochran-McNemar $Q^\star$ statistic, based on the relative ranks of the two compared methods. Since five subjects were evaluated, there were $\binom{5}{2} = 10$ pairwise comparisons for each performance measure. Each method received a score of 1, 0, or -1 from each of its pairwise comparisons, which were added together to give a total score for the method. These total scores, and the sums of these scores over the six performance measures, are reported in Table III.

Table III. Relative performances of fitting methods.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	0	−2	−1	−2	0	−2	−7
2	0	1	0	1	0	0	2
3	0	0	0	−3	0	0	−3
4	0	−1	0	0	0	−1	−2
5	0	2	1	4	0	3	10

Clearly, the superior method for fitting NMR data under the experimental conditions considered is 5, the method that assigns weights by propagating measurement errors. It is equally clear that 1, which eschews experimental measurement of uncomplexed chemical shifts, is the worst method.

The performance of method 1 would probably improve if the experiments were designed to sample the entire range of chemical shift values for all the protons observed. This would require the fraction bound of each species to range from near zero to near unity in each study. Experimental conditions often prohibit such observations if one is unwilling to measure the spectra of host and guest individually. For example, if the asociation constant of a given host/guest pair is 10^6 M^{-1}, both species are 80% bound if the total concentration of each is 20 μM. It is not practical to reduce the fractions bound by making the sample more dilute, because NMR is not sensitive enough to detect lower concentrations. Raising the concentration of one species so that it swamps the other would allow observation of the major component in the almost entirely free state, and of the minor component in the almost entirely bound state. Such an observation is informative for determining δ_{free} of the major component and δ_{bound} of the minor component, but it contains practically no information about the association constant.[8] Ironically, the very experimental

design favored by Wilcox, in which the ratio of host concentration to guest concentration is the same in all samples, is the least likely to cover the entire range of binding if the association constant is large.

The performance of these fitting methods when one experimental error is anomalously large has also been evaluated in Monte Carlo studies. One such set of experimental conditions modeled was that in which two independent stock solutions of the same intended concentration were used for one of the species. In the experimental protocol in which host stock solution aliquots are added to the sample, for instance, every other such aliquot was taken from the second stock solution. This design was contrived to test the performance of the fitting methods when the stock solution concentration behaves more like a random error and less like a systematic error. All the experimental designs were perturbed in this fashion, except for the design in which aliquots of diluent are added to the sample. The number of judges in this study, N, was therefore 15 instead of 19. The outcomes of the head-to-head comparisons between fitting methods are presented in Table IV.

Table IV. Relative performances of fitting methods when a duplicate stock solution is used.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	0	−1	0	−2	0	−1	−4
2	0	−1	0	1	0	−1	−1
3	0	−1	0	−3	0	−1	−5
4	0	−1	0	0	0	−1	−2
5	0	4	0	4	0	4	12

Method 5 is again the best performer under these experimental conditions. Method 1 is no longer the worst performer; it has been eclipsed by method 3, in which separate estimates of K from the individual protons are averaged to give the

overall estimate. Apparently, this method is more vulnerable than the others to vagaries in the stock solution concentrations. This difference may also be a random fluctuation: visual inspection of the five methods under this set of experimental conditions and the set summarized in Table III does not reveal any qualitative differences between these two sets.

The effect of imprecise NMR measurements was investigated in a series of studies. Table V summarizes the results from a study considering a single bad spectrum. In every experiment in this study, the observations of both the host and guest protons have a standard deviation of 20 Hz in the observations of the second sample. As may be expected, method 5 performs exceedingly well under these conditions.

Table V. Relative performances of fitting methods when the standard deviation of the second observation for each proton is 20 Hz.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	−1	−2	−4	−3	−4	−3	−17
2	1	1	1	1	1	0	5
3	−3	−2	1	−3	1	0	−6
4	1	−1	1	2	1	−1	3
5	2	4	1	3	1	4	15

Another study investigated the effect of extremely imprecise measurements of δ_{free} of both protons. In this study, the standard deviation of these measurements was 20 Hz. Since method 1 does not use these measurements, it could be expected to perform well under such conditions. The outcome of this study is summarized in Table VI.

Table VI. Relative performances of fitting methods when the standard deviation of every δ_{free} is 20 Hz.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	1	−1	1	−3	1	−1	−2
2	1	0	1	2	2	0	6
3	1	−1	1	−2	1	1	1
4	1	−1	1	2	0	0	3
5	−4	3	−4	1	−4	0	−8

In this instance, the performance of method 5 is the worst. Although the large measurement errors in δ_{free} were propagated to assign weights to the observations, this method was unable to obtain good parameter estimates. Like methods 3 and 4, it has only one opportunity to estimate δ_{free}, and that is in the measurement itself. If the measurement is bad, so is the estimate of δ_{free}, and no subsequent observations can improve it. Still, the inferior performance of this method in comparison to methods 2 and 3 indicates that the propagation of errors is in fact detrimental to the parameter estimation when δ_{free} is poorly known. On the other hand, the method of Creswell and Allred still does not perform better than methods 3 or 4. Even these experimental conditions, which adversely affect to all methods but method 1, do not allow this method to triumph. Instead, the best performer is 2, which considers the measurement of δ_{free} to be just another observation.

This effect was further investigated by making the measurement of δ_{free} of only the *host* proton imprecise. This study is summarized in Table VII. In this case, propagation of errors appears superior to the method of Creswell and Allred. Method 2 is still superior overall, but the margin between all methods has narrowed considerably.

Table VII. Relative performances of fitting methods when the standard deviation of δ_{free} of the host proton only is 20 Hz.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	0	0	0	−2	0	−1	−3
2	2	0	0	2	0	0	4
3	−1	0	0	−2	0	1	−2
4	0	0	0	1	0	0	1
5	−1	0	0	1	0	0	0

Two more variations in NMR observation uncertainties were studied. In these, *all* of the sample resonances δ_{obs} were assigned an uncertainty of 5 Hz.[10] In the first case, the free chemical shifts of both protons were assigned uncertainties of only 0.5 Hz; in the second case, the free chemical shifts were assigned uncertainties of 5 Hz as well. The relative performances of the fitting methods under these two cases are summarized in Tables VIII and IX, respectively.

Table VIII. Relative performances of fitting methods when the standard deviations of all NMR sample observations δ_{obs} are 5 Hz, but the standard deviations of δ_{free} measurements of both protons are 0.5 Hz.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	−2	−1	0	−4	−2	−1	−10
2	3	2	0	1	1	1	8
3	−3	−3	0	−1	−1	−3	−11
4	2	1	0	2	3	2	10
5	0	1	0	2	−1	1	3

In the first of these cases, in which measurement could have given good estimates of δ_{free}, the method of Creswell and Allred performs comparatively poorly. The best performances are by methods 2 and 4, which consider all observations

(10) Assigning an uncertainty of 20 Hz to all of these measurements led to data sets that could not be adequately fit by any method.

Table IX. Relative performances of fitting methods when the standard deviations of all NMR measurements are 5 Hz.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	−2	−1	−2	−4	0	−1	−10
2	3	2	1	1	3	2	12
3	−2	−3	1	0	−2	−3	−9
4	3	1	3	3	2	1	13
5	−2	1	−3	0	−3	1	−6

but do not propagate errors. Method 5, which propagates measurement errors, has somewhat intermediate performance. In the second case, in which all NMR observations are equally poor, the best performers are again methods 2 and 4. Methods 1 and 3 are again the worst, but the performance of method 5 has descended almost to their level. It should be noted that *all* methods performed poorly in these cases.

These results indicate that propagation of errors is unable to compensate for large uncertainties in δ_{free} or for large and similar measurement errors in all of the values of δ_{obs}. It is unarguably inappropriate to take a measured value of δ_{free} as the final word if that measurement is imprecise; clearly, a method such as 2 is then a better choice. However, NMR peak position measurements are typically *not* imprecise. Peak positions referenced to an internal standard are very reproducible. Even if a peak is broad, an assignment of its center is seldom uncertain to more than a small fraction of the peak width. Therefore, the most relevant cases to consider when evaluating fitting methods are those in which the NMR errors are negligible. In such cases, method 5 appears to be superior.

The effect of non-normal measurement errors was also investigated. In this study, the "experimental" errors were drawn from a Cauchy distribution instead of from a normal distribution. The probability density of a Cauchy distribution is

described by a Lorentzian function; a normal probability density is described by a Gaussian function. These functions are qualitatively similar: they are both bell-shaped and centered on zero. The Cauchy generating procedure was scaled to the normal generating procedure used in the other studies. This was arranged so that the probability of obtaining a value within one standard deviation of zero was the same in both distributions. Figure 4 graphically compares these two distributions to each other. Overall, the Cauchy distribution is much more diffuse than the normal distribution: extremely large values are more likely to arise from it. In fact, the Cauchy distribution is so diffuse that its mean value is undefined.[11] The relative performances of the fitting methods with these non-normal errors are summarized in Table X. The variances of the parameter estimates from *all* fitting methods under *all* of these experimental conditions were too large for meaningful comparisons, so only the medians have been analyzed. Although all methods performed exceedingly poorly, method 5 is clearly the worst under these conditions. The best is method 4, which differs from method 5 only in that it performs no propagation of errors.

Table X. Relative performances of fitting methods when the measurement errors follow a Cauchy distribution.

method	K med	D_1 med	D_2 med	total
1	0	1	−3	−2
2	0	1	1	2
3	1	1	2	4
4	3	1	3	7
5	−4	−4	−3	−11

(11) Hoel, Paul G.; Port, Sidney C.; Stone, Charles J. *Introduction to Probability Theory*; Houghton Mifflin: Boston, 1971; p 174.

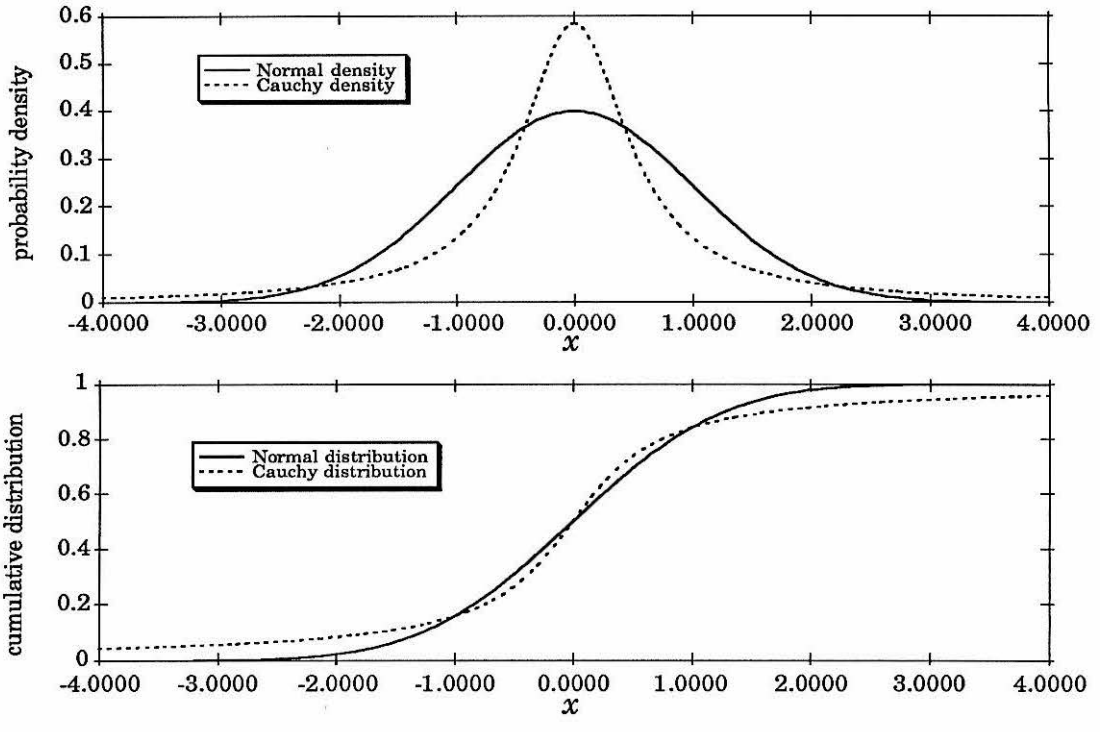


Figure 4. The Cauchy and normal distributions. The Cauchy distribution here is scaled to give the same probability as the standard normal distribution of obtaining a value between -1 and 1. **Top:** Probability densities. **Bottom:** Cumulative distribution functions. These are the integrals of the densities, and represent the probability of obtaining a value between $-\infty$ and x. Off-center values are clearly more likely to arise from the Cauchy distribution.

A set of less severe non-normal experimental errors was also tested. This used the same scaled Cauchy generator, but variates falling more than five "standard deviations" from the center were rejected. This distribution is still much more diffuse than the normal distribution, but it is considerably less pathological than the full Cauchy distribution. The results of this comparison are summarized in Table XI. The best performers in this case are methods 2, 4, and 5.

These two tests of non-normal errors indicate that method 5 is more adversely affected by non-normal measurement errors than are the other methods. However, when the error distributions are not hideously pathological, method 5 is still able to

Table XI. Relative performances of fitting methods when the measurement errors follow a truncated Cauchy distribution.

	K		D_1		D_2		
method	med	sdev	med	sdev	med	sdev	total
1	−2	−1	0	0	0	−3	−6
2	2	−1	0	0	0	1	2
3	0	−1	0	−1	0	1	−1
4	2	−1	0	1	0	1	3
5	−2	4	0	0	0	0	2

perform well. The actual measurement errors in NMR titration studies are probably not truly normal, but they are probably closer to normality than are the Cauchy or truncated Cauchy distributions considered here. If there were reason to suspect that the true measurement errors are substantially non-normal, a fitting procedure based on a more robust criterion than least-squares should be used; in such a case, none of the methods tested here would be good choices. Normal measurement errors apparently favor method 5.

E. Conclusions

We have developed a method for determining the association constant K and saturation shifts D of a host/guest pair from variable-concentration NMR titration experiments. This method minimizes a sum of squared *weighted* residuals; weights are calculated by propagating measurement errors according to equations 17 and 18. Monte Carlo studies simulating realistic measurement errors and a variety of experimental designs demonstrate that this method performs well in comparison to other methods.

IV. Experimental Section.

A. Monte Carlo Comparisons.

A Monte Carlo comparison test is carried out in the following manner:

Repeat R times:

- A random error is drawn for each delivery device used in the experiment. This error is drawn from a normal distribution with a mean of zero and a variance determined by the distribution of calibration errors for that type of device. The Monte Carlo value of I, the device calibration, is obtained by adding 1 to this error. In the unlikely event that e_I is so large and negative that I is less than zero, another calibration error is drawn. The magnitudes of these calibration errors are obtained from the device manufacturer. Typically, small devices have greater calibration errors than do large devices.

$$e_I \sim N(0, \sigma_I^2)$$

$$I = 1 + e_I$$

- Errors are drawn for the host and guest concentrations of each stock solution. These errors are assumed to be from a normal distribution with a mean of zero. The variances are determined by the specified uncertainties in the concentration measurements, and are expressed as a fraction of the total concentration. Each drawn error is added to 1 and the sum is multiplied by the measured concentration ($[\tilde{H}]_s$ or $[\tilde{G}]_s$) to give the Monte Carlo stock solution concentrations. If such a sampled concentration is less than zero, another error is drawn.

$$e_H \sim N(0, \sigma_H^2)$$

$$e_G \sim N(0, \sigma_G^2)$$

$$[\mathrm{H}]_S = [\tilde{\mathrm{H}}]_S(1 + e_H)$$

$$[\mathrm{G}]_S = [\tilde{\mathrm{G}}]_S(1 + e_G)$$

- Samples are created by adding solution aliquots to the sample tube. The volume of each aliquot is determined by multiplying the appropriate delivery device's Monte Carlo calibration value I by the measured aliquot volume $\tilde{V}_a$, and then adding to this value a reproducibility error. The reproducibility error will be an absolute volume, say 0.03 μl, and is drawn from the distribution associated with the appropriate delivery device. This distribution has a mean of zero and a variance specified by the device manufacturer. Typically, small devices have smaller reproducibility errors than do large devices.

$$e_V \sim N(0, \sigma_V^2)$$

$$V_a = I\tilde{V}_a + e_V$$

This gives the Monte Carlo aliquot volume. The Monte Carlo sample volume is obtained by adding this aliquot volume to the volume V_t of the sample previously in the tube. The Monte Carlo host and guest concentrations are determined from these volumes and from the concentrations of the combined solutions.

$$V = V_t + V_a$$

$$[\mathrm{H}]_0 = ([\mathrm{H}_t]_0 V_t + [\mathrm{H}_a]_0 V_a)/V$$

$$[\mathrm{G}]_0 = ([\mathrm{G}_t]_0 V_t + [\mathrm{G}_a]_0 V_a)/V$$

- An error is drawn for the measurement of δ_{free} for each proton. This error is from a normal distribution with a mean of zero and a variance determined by the peak width of the observed signal and the separation between data points

in the frequency-domain spectrum. The Monte Carlo value of δ_{free} is obtained by adding this error to the measured value.

$$e_{\delta_{\text{free}\,p}} \sim N(0, \sigma^2_{\delta_{\text{free}\,p}})$$

$$\delta_{\text{free}\,p} = \tilde{\delta}_{\text{free}\,p} + e_{\delta_{\text{free}\,p}}$$

- An error is drawn for the observed chemical shift of each proton recorded in a sample. Each Monte Carlo "error-free" observation $\hat{\delta}_{\text{obs}\,pi}$ is generated by applying equation 14 to the assumed parameter values and the Monte Carlo values of $[\text{H}]_0$, $[\text{G}]_0$, and δ_{free}. To this result is added the random observation error.

$$e_{\delta_{\text{obs}\,pi}} \sim N(0, \sigma^2_{\delta_{\text{obs}\,pi}})$$

$$\delta_{\text{obs}\,pi} = \hat{\delta}_{\text{obs}\,pi} + e_{\delta_{\text{obs}\,pi}}$$

- Once the data set is generated, it is subjected to analysis by each regression procedure under examination. Each procedure generates estimates for every adjustable parameter.

After this set of R replications of the experiment is complete, the medians and standard deviations of the parameter estimates from each regression procedure are calculated, and the performances of the different procedures are compared with each other. Parameter estimate medians are evaluated by their distances from the true parameter value, and parameter estimate standard deviations are evaluated by size.

B. Random Numbers.

Uniform deviates were generated by the supplied linear congruence generator drand48 (standard C library in the IRIS) and shuffled.[12] The generator was initialized at the start of each run with the value of the current system time. Normal variates were constructed from the uniform variates by the Box-Muller method,[13] and Cauchy variates were constructed from the uniform variates by a tangent transformation.[14] The generated Cauchy and normal variates were verified by the Kolmogorov-Smirnov test to follow their intended theoretical distributions.

C. Experimental Conditions.

The nineteen sets of experimental conditions simulated in the Monte Carlo comparison studies are summarized in Tables XII–XXX. The accuracy and precision of each device is reported in the "Delivery Devices" table; host and guest concentrations of each stock solution, and their associated uncertainties, are in the "Stock Solutions" table. The "Samples" table describes the composition of each sample. Each row reports an aliquot. The "device" column reports the delivery device used for the addition, "V_a" is the aliquot volume, and "Solution" is the identity of the solution added. The "#" column contains the name of the new solution defined by the addition of the aliquot. Aliquots are combined together in the sequence listed; fresh sample tubes are indicated by intrusive "tube x" entries between rows. Sample names consisting of a number only denote samples at which NMR spectra were "recorded;" sample names consisting of a letter alone or a letter and a number were for set-up only.

---

(12) Press, W. H.; Flannery, D. P.; Teukolsky, S. A.; Vetterling, W.T *Numerical Recipes: the Art of Scientific Computing*; Cambridge University: New York, 1986; pp 192–195.

(13) Reference 12, p 202.

(14) Reference 11, p 122.

Table XII. Data set H3.

Association Constant: 10^3 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
1000λ	1%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	1000λ	375	buffer
b	1000λ	20	guest
1	10λ	5	host
2	10λ	5	host
3	10λ	5	host
4	10λ	5	host
5	10λ	5	host
6	10λ	5	host
7	10λ	5	host
8	10λ	5	host
9	10λ	5	host
10	10λ	5	host
11	10λ	5	host
12	10λ	5	host
13	10λ	5	host
14	10λ	5	host
15	10λ	5	host

[a]In μl.

Table XIII. Data set H4.

Association Constant: 10^4 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	235	buffer
b	250λ	160	guest
1	10λ	5	host
2	10λ	5	host
3	10λ	5	host
4	10λ	6	host
5	10λ	6	host
6	10λ	6	host
7	10λ	6	host
8	10λ	7	host
9	10λ	7	host
10	10λ	7	host
11	10λ	7	host
12	10λ	7	host
13	10λ	8	host
14	10λ	8	host
15	10λ	8	host

[a]In μl.

Table XIV. Data set H5.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
250λ	1%	1
1000λ	0.6%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	400	5%	0	0
guest	0	0	400	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	1000λ	358	buffer
b	250λ	32	guest
1	10λ	10	host
2	10λ	5	host
3	10λ	5	host
4	10λ	5	host
5	10λ	6	host
6	10λ	6	host
7	10λ	6	host
8	10λ	6	host
9	10λ	6	host
10	10λ	7	host
11	10λ	7	host
12	10λ	7	host
13	10λ	7	host
14	10λ	7	host
15	10λ	7	host

[a]In μl.

Table XV. Data set H6.

Association Constant: 10^6 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
20λ	6%	0.04
1000λ	1%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	500	5%	0	0
guest	0	0	500	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	1000λ	380	buffer
b	20λ	12	guest
1	10λ	8	host
c	10λ	8	buffer[b]
2	10λ	1	host
3	10λ	2	host
4	10λ	3	host
5	10λ	4	host
6	10λ	5	host
7	10λ	6	host
8	10λ	7	host
9	10λ	8	host
10	10λ	8	host
11	10λ	8	host
12	10λ	8	host
13	10λ	8	host
14	10λ	8	host
15	10λ	8	host

[a]In μl. [b]To correct for a miscalculation.

Table XVI. Data set G3.

Association Constant: 10^3 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
20λ	6%	0.04
100λ	1%	0.2
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	200	buffer
b	100λ	80	host
1	250λ	120	guest
2	10λ	3	guest
3	10λ	5	guest
4	10λ	8	guest
5	20λ	11	guest
6	20λ	13	guest
7	20λ	16	guest
8	20λ	19	guest
9	100λ	21	guest
10	100λ	24	guest
11	100λ	27	guest
12	100λ	29	guest
13	100λ	32	guest
14	100λ	35	guest
15	100λ	37	guest

[a]In μl.

Table XVII. Data set G4.

Association Constant: 10^4 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
20λ	6%	0.04
100λ	1%	0.2
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	200	buffer
b	100λ	80	host
1	250λ	120	guest
2	10λ	3	guest
3	10λ	5	guest
4	10λ	8	guest
5	20λ	11	guest
6	20λ	13	guest
7	20λ	16	guest
8	20λ	19	guest
9	100λ	21	guest
10	100λ	24	guest
11	100λ	27	guest
12	100λ	29	guest
13	100λ	32	guest
14	100λ	35	guest
15	100λ	37	guest

[a]In μl.

Table XVIII. Data set G5.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
$100a\lambda$	5%	0.04
100λ	1%	0.2
1000λ	0.6%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	500	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	1000λ	360	buffer
b	100λ	30	host
1	100λ	10	guest
2	$100a\lambda$	5	guest
3	$100a\lambda$	5	guest
4	$100a\lambda$	5	guest
5	$100a\lambda$	5	guest
6	$100a\lambda$	10	guest
7	$100a\lambda$	10	guest
8	$100a\lambda$	10	guest
9	$100a\lambda$	15	guest
10	$100a\lambda$	15	guest
11	$100a\lambda$	15	guest
12	$100a\lambda$	20	guest
13	$100a\lambda$	30	guest
14	$100a\lambda$	40	guest
15	$100a\lambda$	55	guest

[a]In μl.

Table XIX. Data set G6.

Association Constant: 10^6 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
10λ	5%	0.04
20λ	6%	0.04
100λ	1%	0.2
1000λ	0.6%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	100	5%	0	0
guest	0	0	250	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	1000λ	324	buffer
b	100λ	60	host
1	20λ	16	guest
2	10λ	1	guest
3	10λ	2	guest
4	10λ	3	guest
5	10λ	4	guest
6	10λ	5	guest
7	10λ	6	guest
8	10λ	7	guest
9	10λ	8	guest
10	10λ	9	guest
11	10λ	10	guest
12	20λ	11	guest
13	20λ	12	guest
14	20λ	13	guest
15	20λ	14	guest

[a]In μl.

Table XX. Data set D3.

Association Constant: 10^3 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
100λ	1%	0.2
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	120	guest
b	100λ	80	host
1	250λ	120	buffer
2	100λ	11	buffer
3	100λ	23	buffer
4	100λ	34	buffer
5	100λ	46	buffer
6	250λ	57	buffer
7	250λ	69	buffer
8	250λ	80	buffer
9	250λ	90	buffer
10	250λ	104	buffer
11	250λ	114	buffer
12	250λ	126	buffer
13	250λ	137	buffer
14	250λ	149	buffer
15	250λ	160	buffer

[a]In μl.

Table XXI. Data set D4.

Association Constant: 10^4 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
100λ	1%	0.2
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	120	guest
b	100λ	80	host
1	250λ	120	buffer
2	100λ	11	buffer
3	100λ	23	buffer
4	100λ	34	buffer
5	100λ	46	buffer
6	250λ	57	buffer
7	250λ	69	buffer
8	250λ	80	buffer
9	250λ	90	buffer
10	250λ	104	buffer
11	250λ	114	buffer
12	250λ	126	buffer
13	250λ	137	buffer
14	250λ	149	buffer
15	250λ	160	buffer

[a]In μl.

Table XXII. Data set D5.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
100λ	1%	0.2
250λ	1%	1

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	240	buffer
b	100λ	80	guest
1	100λ	80	host
2	100λ	11	buffer
3	100λ	23	buffer
4	100λ	34	buffer
5	100λ	46	buffer
6	100λ	57	buffer
7	100λ	69	buffer
8	100λ	80	buffer
9	100λ	91	buffer
10	250λ	103	buffer
11	250λ	114	buffer
12	250λ	126	buffer
13	250λ	137	buffer
14	250λ	149	buffer
15	250λ	160	buffer

[a]In μl.

Table XXIII. Data set D6.

Association Constant: 10^6 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	6%	0.04
100λ	1%	0.2
250λ	1%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	100	5%	0	0
guest	0	0	100	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution
a	250λ	240	buffer
b	100λ	80	guest
1	100λ	80	host
2	20λ	4	buffer
3	20λ	8	buffer
4	20λ	11	buffer
5	20λ	15	buffer
6	20λ	19	buffer
7	100λ	23	buffer
8	100λ	27	buffer
9	100λ	30	buffer
10	100λ	34	buffer
11	100λ	38	buffer
12	100λ	42	buffer
13	100λ	46	buffer
14	100λ	50	buffer
15	100λ	53	buffer

[a]In μl.

Table XXIV. Data set J4.

Association Constant: 10^4 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	6%	0.04
100λ	1%	0.2
1000λ	0.6%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	320	buffer	a9	1000λ	320	buffer
b1	100λ	75	host	b9	100λ	35	host
1	20λ	5	guest	9	100λ	45	guest
		tube 2				tube 10	
a2	1000λ	320	buffer	a10	1000λ	320	buffer
b2	100λ	70	host	b10	100λ	30	host
2	20λ	10	guest	10	100λ	50	guest
		tube 3				tube 11	
a3	1000λ	320	buffer	a11	1000λ	320	buffer
b3	100λ	65	host	b11	100λ	25	host
3	20λ	15	guest	11	100λ	55	guest
		tube 4				tube 12	
a4	1000λ	320	buffer	a12	1000λ	320	buffer
b4	100λ	60	host	b12	100λ	20	host
4	20λ	20	guest	12	100λ	60	guest
		tube 5				tube 13	
a5	1000λ	320	buffer	a13	1000λ	320	buffer
b5	100λ	55	host	b13	20λ	15	host
5	100λ	25	guest	13	100λ	65	guest
		tube 6				tube 14	
a6	1000λ	320	buffer	a14	1000λ	320	buffer
b6	100λ	50	host	b14	20λ	10	host
6	100λ	30	guest	14	100λ	70	guest
		tube 7				tube 15	
a7	1000λ	320	buffer	a15	1000λ	320	buffer
b7	100λ	45	host	b15	20λ	5	host
7	100λ	35	guest	15	100λ	75	guest
		tube 8					
a8	1000λ	320	buffer				
b8	100λ	40	host				
8	100λ	40	guest				

[a]In μl.

Table XXV. Data set J5.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	6%	0.04
100λ	1%	0.2
1000λ	0.6%	2

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	500	5%	0	0
guest	0	0	500	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	320	buffer	a9	1000λ	320	buffer
b1	20λ	8	host	b9	100λ	45	host
1	100λ	72	guest	9	100λ	35	guest
		tube 2				tube 10	
a2	1000λ	320	buffer	a10	1000λ	320	buffer
b2	20λ	13	host	b10	100λ	49	host
2	100λ	67	guest	10	100λ	31	guest
		tube 3				tube 11	
a3	1000λ	320	buffer	a11	1000λ	320	buffer
b3	20λ	17	host	b11	100λ	54	host
3	100λ	63	guest	11	100λ	26	guest
		tube 4				tube 12	
a4	1000λ	320	buffer	a12	1000λ	320	buffer
b4	100λ	22	host	b12	100λ	58	host
4	100λ	58	guest	12	100λ	22	guest
		tube 5				tube 13	
a5	1000λ	320	buffer	a13	1000λ	320	buffer
b5	100λ	26	host	b13	20λ	63	host
5	100λ	54	guest	13	20λ	17	guest
		tube 6				tube 14	
a6	1000λ	320	buffer	a14	1000λ	320	buffer
b6	100λ	31	host	b14	10λ	67	host
6	100λ	49	guest	14	20λ	13	guest
		tube 7				tube 15	
a7	1000λ	320	buffer	a15	1000λ	320	buffer
b7	100λ	35	host	b15	10λ	72	host
7	100λ	45	guest	15	20λ	8	guest
		tube 8					
a8	1000λ	320	buffer				
b8	100λ	40	host				
8	100λ	40	guest				

[a]In μl.

Table XXVI. Data set J6.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
100λ	1%	0.2
1000λ	0.6%	5

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	100	5%	0	0
guest	0	0	100	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	280	buffer	a9	1000λ	280	buffer
b1	100λ	40	host	b9	100λ	63	host
1	100λ	80	guest	9	100λ	57	guest
		tube 2				tube 10	
a2	1000λ	280	buffer	a10	1000λ	280	buffer
b2	100λ	43	host	b10	100λ	66	host
2	100λ	77	guest	10	100λ	54	guest
		tube 3				tube 11	
a3	1000λ	280	buffer	a11	1000λ	280	buffer
b3	100λ	46	host	b11	100λ	69	host
3	100λ	74	guest	11	100λ	51	guest
		tube 4				tube 12	
a4	1000λ	280	buffer	a12	1000λ	280	buffer
b4	100λ	49	host	b12	100λ	71	host
4	100λ	71	guest	12	100λ	49	guest
		tube 5				tube 13	
a5	1000λ	280	buffer	a13	1000λ	280	buffer
b5	100λ	51	host	b13	100λ	74	host
5	100λ	69	guest	13	100λ	46	guest
		tube 6				tube 14	
a6	1000λ	280	buffer	a14	1000λ	280	buffer
b6	100λ	54	host	b14	100λ	77	host
6	100λ	66	guest	14	100λ	43	guest
		tube 7				tube 15	
a7	1000λ	280	buffer	a15	1000λ	280	buffer
b7	100λ	57	host	b15	100λ	80	host
7	100λ	63	guest	15	100λ	40	guest
		tube 8					
a8	1000λ	280	buffer				
b8	100λ	60	host				
8	100λ	60	guest				

[a]In μl.

Table XXVII. Data set V3.

Association Constant: 10^3 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	5%	0.04
100λ	1%	0.2
250λ	1%	1
1000λ	0.6%	5

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	356	buffer	a9	1000λ	291	buffer
b1	100λ	40	host	b9	100λ	40	host
1	20λ	4	guest	9	100λ	69	guest
		tube 2				tube 10	
a2	1000λ	351	buffer	a10	1000λ	260	buffer
b2	100λ	40	host	b10	100λ	40	host
2	20λ	9	guest	10	100λ	100	guest
		tube 3				tube 11	
a3	1000λ	344	buffer	a11	250λ	243	buffer
b3	100λ	40	host	b11	100λ	40	host
3	20λ	16	guest	11	250λ	117	guest
		tube 4				tube 12	
a4	1000λ	336	buffer	a12	250λ	224	buffer
b4	100λ	40	host	b12	100λ	40	host
4	100λ	24	guest	12	250λ	136	guest
		tube 5				tube 13	
a5	1000λ	327	buffer	a13	250λ	204	buffer
b5	100λ	40	host	b13	100λ	40	host
5	100λ	33	guest	13	250λ	156	guest
		tube 6				tube 14	
a6	1000λ	316	buffer	a14	250λ	193	buffer
b6	100λ	40	host	b14	100λ	40	host
6	100λ	44	guest	14	250λ	177	guest
		tube 7				tube 15	
a7	1000λ	304	buffer	a15	250λ	160	buffer
b7	100λ	40	host	b15	100λ	40	host
7	100λ	56	guest	15	250λ	200	guest
		tube 8					
a8	1000λ	291	buffer				
b8	100λ	40	host				
8	100λ	69	guest				

[a]In μl.

Table XXVIII. Data set V4.

Association Constant: 10^4 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	5%	0.04
100λ	1%	0.2
250λ	1%	1
1000λ	0.6%	5

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	356	buffer	a9	1000λ	291	buffer
b1	100λ	40	host	b9	100λ	40	host
1	20λ	4	guest	9	100λ	69	guest
		tube 2				tube 10	
a2	1000λ	351	buffer	a10	1000λ	260	buffer
b2	100λ	40	host	b10	100λ	40	host
2	20λ	9	guest	10	100λ	100	guest
		tube 3				tube 11	
a3	1000λ	344	buffer	a11	250λ	243	buffer
b3	100λ	40	host	b11	100λ	40	host
3	20λ	16	guest	11	250λ	117	guest
		tube 4				tube 12	
a4	1000λ	336	buffer	a12	250λ	224	buffer
b4	100λ	40	host	b12	100λ	40	host
4	100λ	24	guest	12	250λ	136	guest
		tube 5				tube 13	
a5	1000λ	327	buffer	a13	250λ	204	buffer
b5	100λ	40	host	b13	100λ	40	host
5	100λ	33	guest	13	250λ	156	guest
		tube 6				tube 14	
a6	1000λ	316	buffer	a14	250λ	193	buffer
b6	100λ	40	host	b14	100λ	40	host
6	100λ	44	guest	14	250λ	177	guest
		tube 7				tube 15	
a7	1000λ	304	buffer	a15	250λ	160	buffer
b7	100λ	40	host	b15	100λ	40	host
7	100λ	56	guest	15	250λ	200	guest
		tube 8					
a8	1000λ	291	buffer				
b8	100λ	40	host				
8	100λ	69	guest				

[a]In μl.

Table XXIX. Data set V5.

Association Constant: 10^5 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	5%	0.04
100λ	1%	0.2
1000λ	0.6%	5

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	1000	5%	0	0
guest	0	0	1000	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	376	buffer	a9	1000λ	337	buffer
b1	20λ	20	host	b9	20λ	20	host
1	20λ	4	guest	9	100λ	43	guest
		tube 2				tube 10	
a2	1000λ	372	buffer	a10	1000λ	331	buffer
b2	20λ	20	host	b10	20λ	20	host
2	20λ	8	guest	10	100λ	49	guest
		tube 3				tube 11	
a3	1000λ	363	buffer	a11	250λ	325	buffer
b3	20λ	20	host	b11	20λ	20	host
3	20λ	12	guest	11	100λ	55	guest
		tube 4				tube 12	
a4	1000λ	363	buffer	a12	1000λ	319	buffer
b4	20λ	20	host	b12	20λ	20	host
4	20λ	17	guest	12	100λ	61	guest
		tube 5				tube 13	
a5	1000λ	358	buffer	a13	250λ	313	buffer
b5	20λ	20	host	b13	20λ	20	host
5	100λ	22	guest	13	100λ	67	guest
		tube 6				tube 14	
a6	1000λ	353	buffer	a14	1000λ	307	buffer
b6	20λ	20	host	b14	20λ	20	host
6	100λ	27	guest	14	100λ	73	guest
		tube 7				tube 15	
a7	1000λ	348	buffer	a15	1000λ	300	buffer
b7	20λ	20	host	b15	20λ	20	host
7	100λ	32	guest	15	100λ	80	guest
		tube 8					
a8	1000λ	343	buffer				
b8	20λ	20	host				
8	100λ	37	guest				

[a]In μl.

Table XXX. Data set V6.

Association Constant: 10^6 M^{-1}

Delivery Devices:

name	accuracy	precision[a]
20λ	5%	0.04
100λ	1%	0.2
1000λ	0.6%	5

[a]In μl.

Stock Solutions:

name	$[H]_s^a$	stdev	$[G]_s^a$	stdev
host	400	5%	0	0
guest	0	0	200	5%
buffer	0	0	0	0

[a]In μM.

Samples:

#	device	V_a^a	Solution	#	device	V_a^a	Solution
		tube 1				tube 9	
a1	1000λ	370	buffer	a9	1000λ	362	buffer
b1	20λ	10	host	b9	20λ	10	host
1	100λ	20	guest	9	100λ	28	guest
		tube 2				tube 10	
a2	1000λ	369	buffer	a10	1000λ	361	buffer
b2	20λ	10	host	b10	20λ	10	host
2	100λ	21	guest	10	100λ	29	guest
		tube 3				tube 11	
a3	1000λ	368	buffer	a11	250λ	360	buffer
b3	20λ	10	host	b11	20λ	10	host
3	100λ	22	guest	11	100λ	30	guest
		tube 4				tube 12	
a4	1000λ	367	buffer	a12	1000λ	359	buffer
b4	20λ	10	host	b12	20λ	10	host
4	100λ	23	guest	12	100λ	31	guest
		tube 5				tube 13	
a5	1000λ	366	buffer	a13	1000λ	358	buffer
b5	20λ	10	host	b13	20λ	10	host
5	100λ	24	guest	13	100λ	32	guest
		tube 6				tube 14	
a6	1000λ	365	buffer	a14	1000λ	357	buffer
b6	20λ	10	host	b14	20λ	10	host
6	100λ	25	guest	14	100λ	33	guest
		tube 7				tube 15	
a7	1000λ	364	buffer	a15	1000λ	356	buffer
b7	20λ	10	host	b15	20λ	10	host
7	100λ	26	guest	15	100λ	34	guest
		tube 8					
a8	1000λ	363	buffer				
b8	20λ	10	host				
8	100λ	27	guest				

[a]In μl.

Appendix. Details of the Fitting Procedures.

I. Background

A. The Model.

All of the fitting methods examined in this chapter seek to minimize some squared error score. It is not always obvious from the score how this minimization is carried out, so this Appendix describes how each of the five fitting methods arrives at its optimal parameter set.

The model for an observed chemical shift under fast-exchange conditions is

$$\delta_{\text{obs}} = \delta_{\text{free}} - DF\,. \tag{12}$$

F, the fraction bound, is the ratio of the total concentration of complex to the concentration of the species of interest; $F = [\text{H}{\cdot}\text{G}]/[\text{S}]_0$. The concentration of complex is a function of the association constant K and the total concentrations of host and guest.

$$[\text{H}{\cdot}\text{G}] = \frac{1}{2}\left\{[\text{H}]_0 + [\text{G}]_0 + 1/K - \sqrt{\left([\text{H}]_0 + [\text{G}]_0 + 1/K\right)^2 - 4[\text{H}]_0[\text{G}]_0}\right\} \tag{6}$$

This model is common to all of the fitting methods under consideration.

B. Least-squares estimation methods.

It is appropriate to review several general techniques for least squares estimation. All of the methods considered here adjust some model parameters θ to minimize a fit score of the form

$$\text{SSR} = \sum_{p=1}^{P}\sum_{i=1}^{N}\frac{\left(y_p(\mathbf{x}_i,\theta) - y_{pi}\right)^2}{\sigma_{pi}^2}\,.$$

Here, y_{pi} is the observation of proton p in sample i; $y_p(\mathbf{x}_i)$ is the predicted value of this variable based on the model, the ith set of explanatory variables $\mathbf{x}_i$, and the adjustable parameters K and D_p. The methods that are able to minimize this score depend on the particular functional forms of $y_p(\mathbf{x}_i, \theta)$ and σ_{pi}^2.

By calculus, the way to find the value of some variable x that minimizes a function $f(x)$ is to differentiate $f(x)$ with respect to x, and solve the resulting equation for x when the derivative $\partial f(x)/\partial x$ equals zero. In least-squares estimation, the function to be minimized is the fit score SSR, and the variables to be optimized are the adjustable parameters in θ. It is thus necessary to differentiate SSR with respect to the individual components of θ.

$$\frac{\partial \mathrm{SSR}}{\partial \theta_l} = \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{\big(y_p(\mathbf{x}_i, \theta) - y_{pi}\big)}{\sigma_{pi}^2} \left[2 \frac{\partial y_p(\mathbf{x}_i, \theta)}{\partial \theta_l} - \frac{\big(y_p(\mathbf{x}_i, \theta) - y_{pi}\big)}{\sigma_{pi}^2} \frac{\partial \sigma_{pi}^2}{\partial \theta_l} \right] \tag{22}$$

If the weighting factors σ_{pi}^2 are independent of the parameters θ, then the second term in this summation is zero. The derivative is then the simpler

$$\frac{\partial \mathrm{SSR}}{\partial \theta_l} = 2 \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{\big(y_p(\mathbf{x}_i, \theta) - y_{pi}\big)}{\sigma_{pi}^2} \frac{\partial y_p(\mathbf{x}_i, \theta)}{\partial \theta_l} . \tag{23}$$

If the weighting factors are equal to unity, the derivative is the still simpler

$$\frac{\partial \mathrm{SSR}}{\partial \theta_l} = 2 \sum_{p=1}^{P} \sum_{i=1}^{N} \big(y_p(\mathbf{x}_i, \theta) - y_{pi}\big) \frac{\partial y_p(\mathbf{x}_i, \theta)}{\partial \theta_l} . \tag{24}$$

The ability to solve equations 22–24 for θ_l hinges on the ability to differentiate $y_p(\mathbf{x}_i, \theta)$ and σ_{pi}^2 with respect to θ_l, and to solve the resulting equations for θ_l. If $\partial(\sigma_{pi}^2)/\partial\theta_l = 0$ and $y(\mathbf{x}_i, \theta)$ is a linear combination of the components of θ, this problem reduces to a system of simultaneous linear equations, which can be solved directly.[15] If the model function is a *nonlinear* function of these parameters, however, this system of equations must be solved numerically.

(15) Dunteman, George H.; *Introduction to Linear Models*; Sage: Beverly Hills, 1984; pp 157–175.

Many methods exist for such nonlinear regression.[16] If the predictive function $y(\mathbf{x}_i, \theta)$ cannot be differentiated with respect to θ_l, then no information about the shape of the (θ, SSR) surface is available at a single parameter value θ. In such a case, the optimum parameter set must be found by a robust procedure, such as the simplex method or Brent's method. Brent's method is convenient for finding the minimum of SSR as a function of a single parameter. It operates by modeling the (θ, SSR) surface as a parabola drawn through three values of SSR evaluated at three different values of θ, and choosing the value of θ at the minimum of the parabola as its next guess. It also employs a number of safeguards against (θ, SSR) surfaces with pathological properties.[17] If a model that cannot be differentiated contains two or more nonlinear parameters, the simplex method is the optimization procedure of choice. This method is extremely stable, but also very slow. It should thus be employed only as a last resort.

When information about local derivatives of the SSR surface is available, the Levenberg-Marquardt method may be employed. This algorithm is a stabler version of the Gauss-Newton method. These methods operate by evaluating the first and second derivatives of SSR with respect to each of the parameters θ_l, and modeling the SSR surface as a paraboloid with these derivatives. The next guess for the parameters is the value of θ at the minimum of this paraboloid. This model is very efficient near minima on the SSR surface, but it is unstable farther away. The Levenberg-Marquardt procedure accommodates such instability by increasing the relative importance of the first derivative over the second derivative when the parabolic approximation is performing badly, so that the parameters are improved

(16) Reference 3, Chapter 14.

(17) Reference 12, pp 283–284.

by a steepest-descent procedure. An excellent description of this technique can be found in Press *et al.*[18]

If the predictive model $y(\mathbf{x}_i, \theta)$ is a nonlinear function of some of the parameters and a linear function of the rest, the optimum values of the linear parameters at any arbitrary values of the nonlinear parameters can be found by simple linear regresssion techniques. The linear parameters are then effectively functions of the nonlinear parameters.[19] This greatly facilitates the estimation of the nonlinear parameters, and should be employed whenever possible.

II. The Methods under Consideration.

A. Method 1. The method of Creswell and Allred.

In this method, there are two adjustable parameters for each proton observed in addition to the global association constant. These parameters, δ_{free} and D, are linear parameters. This is obvious by inspection of the model relation, equation 12. At any arbitrary value of K, the values of F_i can be calculated without recourse to any other parameters. Consequently, equation 12 can be thought of as a linear equation of δ_{obs} as a function of the independent variable F. The best values of the parameters $\delta_{\text{free}\,p}$ and D_p are those that optimize the fit of this equation to the data points $(F_i, \delta_{\text{obs}\,pi})$. This is merely a matter of fitting a straight line to the $(F_i, \delta_{\text{obs}\,pi})$ data. The y-intercept of this fitted line is $\delta_{\text{free}\,p}$, and the slope is $-D_p$.

Finding the best estimate of K is not as easy. F is a decidedly nonlinear function of K, so a Levenberg-Marquardt procedure must be used for the least-squares fitting. This requires knowledge of the first and second derivates of the fit

(18) Reference 12, pp 521–525.

(19) Fujita, Iwao " Automatic analytical differentiation for nonlinear least-squares calculations," *Comput. Chem.* **1988**, *12*, 209–211.

score SSR with respect to K. The first derivative can be obtained in the folllowing manner:

$$\frac{\partial \mathrm{SSR}}{\partial K} = \frac{\partial}{\partial K} \sum_{p=1}^{P} \sum_{i=1}^{N} (\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2$$
$$= \sum_{p=1}^{P} \sum_{i=1}^{N} 2(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}) \frac{\partial}{\partial K} \delta_{\mathrm{calc}\,pi} \,.$$

It is now necessary to derive the expression for $\partial \delta_{\mathrm{calc}\,pi}/\partial K$.

$$\frac{\partial}{\partial K} \delta_{\mathrm{calc}\,pi} = \frac{\partial}{\partial K} (\delta_{\mathrm{free}\,p} - D_p F_i)$$
$$= \frac{\partial}{\partial K} \delta_{\mathrm{free}\,p} - F_i \frac{\partial D_p}{\partial K} - D_p \frac{\partial F_i}{\partial K} \tag{25}$$

This requires the calculation of three new derivatives. It is most convenient to begin with $\partial F_i/\partial K$.

$$\frac{\partial F_i}{\partial K} = \frac{\partial}{\partial K} \left(\frac{[\mathrm{H{\cdot}G}]}{[\mathrm{S}]_0} \right) = \frac{1}{[\mathrm{S}]_{0i}} \frac{\partial}{\partial K} [\mathrm{H{\cdot}G}]_i \tag{26}$$

$$\frac{\partial}{\partial K} [\mathrm{H{\cdot}G}]_i = \frac{1}{2} \frac{\partial}{\partial K} \left([\mathrm{H}]_{0i} + [\mathrm{G}]_{0i} + 1/K - \sqrt{([\mathrm{H}]_{0i} + [\mathrm{G}]_{0i} + 1/K)^2 - 4[\mathrm{H}]_{0i}[\mathrm{G}]_{0i}} \right)$$
$$\vdots$$
$$= \frac{1}{2K} \left(\frac{[\mathrm{H}]_{0i} + [\mathrm{G}]_{0i} + 1/K}{\sqrt{([\mathrm{H}]_{0i} + [\mathrm{G}]_{0i} + 1/K)^2 - 4[\mathrm{H}]_{0i}[\mathrm{G}]_{0i}}} - 1 \right) \tag{27}$$

To evaluate $\partial \delta_{\mathrm{free}\,p}/\partial K$ and $\partial D_p/\partial K$, it is necessary to know the functional forms of the best least-squares estimates of $\delta_{\mathrm{free}\,p}$ and D_p. These are too cumbersome to fit on a line without employing some simplifying notation. Let us define the shorthand terms

$$\alpha = \sum_{i=1}^{N} F_i^2 \,; \qquad \beta = \sum_{i=1}^{N} F_i \,; \qquad \gamma = \sum_{i=1}^{N} F_i \, \delta_{\mathrm{obs}\,pi} \,; \qquad \zeta = \sum_{i=1}^{N} \delta_{\mathrm{obs}\,pi} \,.$$

Using this notation, the least-squares estimates of D_p and $\delta_{\mathrm{free}\,p}$ are

$$D_p = -\frac{\zeta\alpha - \gamma\beta}{N\alpha - \beta^2} \qquad \text{and} \qquad \delta_{\mathrm{free}\,p} = \frac{1}{\alpha}(\gamma + D_p\beta) \,.$$

The derivatives of these quantities with respect to K are

$$\frac{\partial D_p}{\partial K} = -\frac{\partial}{\partial K}\left(\frac{\zeta\alpha - \gamma\beta}{N\alpha - \beta^2}\right)$$
$$= \frac{(\zeta\alpha - \gamma\beta)\left(N\frac{\partial\alpha}{\partial K} - 2\beta\frac{\partial\beta}{\partial K}\right) - (N\alpha - \beta)\left(\zeta\frac{\partial\alpha}{\partial K} - \gamma\frac{\partial\beta}{\partial K} - \beta\frac{\partial\gamma}{\partial K}\right)}{(N\alpha - \beta)^2} \quad (28)$$

and

$$\frac{\partial\delta_{\text{free}\,p}}{\partial K} = \frac{\partial}{\partial K}\left(\frac{\gamma + D_p\beta}{\alpha}\right)$$
$$= \frac{1}{\alpha}\left[\left(\frac{\partial\gamma}{\partial K} + \beta\frac{\partial D_p}{\partial K} + D_p\frac{\partial\beta}{\partial K}\right) - \frac{1}{\alpha}(\gamma + D_p\beta)\frac{\partial\alpha}{\partial K}\right]. \quad (29)$$

Obtaining the derivatives of α, β, and γ with respect to K is trivial.

The next step is to find the second derivative of SSR with respect to K. When the weighting factors σ_{pi}^2 are independent of the adjustable parameter values, the first derivative $\partial\text{SSR}/\partial K$ can be expressed as in equation 23. Differentiating this expression with respect to parameter θ_k gives the formula for the second derivative $\partial^2\text{SSR}/\partial\theta_l\theta_k$ in equation 30.

$$\frac{\partial^2}{\partial\theta_l\theta_k} = 2\sum_{p=1}^{P}\sum_{i=1}^{N}\frac{1}{\sigma_{pi}^2}\left[\frac{\partial y_p(\mathbf{x}_i,\theta)}{\partial\theta_l}\frac{\partial y_p(\mathbf{x}_i,\theta)}{\partial\theta_k} - \big(y_p(\mathbf{x}_i,\theta) - y_{pi}\big)\frac{\partial^2 y_p(\mathbf{x}_i,\theta)}{\partial\theta_l\theta_k}\right] \quad (30)$$

It is not actually required to evaluate the full expression for this second derivative. It is usually advantageous to omit the term in this expression that contains the second differential element $\partial^2 y_p(\mathbf{x}_i,\theta)$.[18] This is because, in the region of the SSR surface near the minimum, where the approximation by a paraboloid is most valid, the individual residual terms $\big(y_p(\mathbf{x}_i,\theta) - y_{pi}\big)$ will be evenly distributed about zero, cancelling each other out. There is thus no justification for including this term in the expression. The approximation for the second derivative in the Creswell and Allred model then is

$$\frac{\partial^2\text{SSR}}{\partial K} \simeq 2\sum_{p=1}^{P}\sum_{i=1}^{N}\left(\frac{\partial\delta_{\text{calc}\,pi}}{\partial K}\right)^2.$$

The actual value of this expression is easily obtained by substituting the formulas from equations 25-29.

B. Method 2. Modified method of Creswell and Allred.

This method is identical to the method of Creswell and Allred, with one small but significant modification: the set of (F, δ_{obs}) points to be fitted with a straight line includes the point $(0, \delta_{\text{free obs}})$, the observed chemical shift of the free species. The fraction bound for this observation is necessarily zero.

Implementation of this method as a Levenberg-Marquardt procedure is nearly identical to the implementation of the unmodified method. The only differences requiring special accomodation are that the sum $\zeta = \sum_{i=1}^{N} \delta_{\text{obs}\,pi}$ now includes an additional contribution from $\delta_{\text{free obs}}$, and that the value N in equation 28 is larger by 1 than it is in the unmodified procedure.

C. Method 3. Fitting each proton separately.

In this procedure, the observations for each proton are treated as an independent experiment. K and D_p are adjusted to give a best fit to each proton alone. The actual protocol for minimizing SSR is identical to that employed in Method 4.

D. Method 4. K and D's as adjustable parameters; no weighting.

This method is qualitatively similar in principle and in execution to the method of Creswell and Allred. The best linear parameters D to fit a data set given any arbitrary value of K can be obtained by linear fitting in a direct manner.

As with the method of Creswell and Allred, the predictive relation (equation 12) may be considered as a linear equation relating the dependent variable δ_{obs} to the

independent variable F. Since δ_{free} is a measured quantity instead of an adjustable parameter, it is convenient to rearrange this equation to

$$\delta_{\text{free}\,p} - \delta_{\text{obs}\,p} = D_p F\,.$$

This is a one-parameter linear equation; the y-intercept is zero, and the slope is the parameter D_p. The best D_p to fit this equation to a set of $(F, \delta_{\text{free}} - \delta_{\text{obs}})$ data is

$$D_p = \frac{\sum_{i=1}^{N} F_i(\delta_{\text{free}\,p} - \delta_{\text{obs}\,pi})}{\sum_{i=1}^{N} F_i^2}\,. \tag{31}$$

In order to find K by a Levenberg-Marquardt minimization procedure, it is necessary to differentiate SSR with respect to K.

$$\begin{aligned}
\frac{\partial \text{SSR}}{\partial K} &= \frac{\partial}{\partial K} \sum_{p=1}^{P} \sum_{i=1}^{N} (\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})^2 \\
&= 2 \sum_{p=1}^{P} \sum_{i=1}^{N} (\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi}) \frac{\partial \delta_{\text{calc}\,pi}}{\partial K} \\
\frac{\partial \delta_{\text{calc}\,pi}}{\partial K} &= \frac{\partial \delta_{\text{free}\,p}}{\partial K} - F_i \frac{\partial D_p}{\partial K} - D_p \frac{\partial F_i}{\partial K} \\
&= -F_i \frac{\partial D_p}{\partial K} - D_p \frac{\partial F_i}{\partial K}
\end{aligned}$$

The expression for $\partial F_i / \partial K$ has already been derived (equation 26). The expression for $\partial D_p / \partial K$ is obtained by differentiating equation 31.

$$\begin{aligned}
\frac{\partial D_p}{\partial K} &= \frac{\partial}{\partial K} \left(\frac{\sum_{i=1}^{N} F_i(\delta_{\text{free}\,p} - \delta_{\text{obs}\,pi})}{\sum_{i=1}^{N} F_i^2} \right) \\
&= \frac{1}{\sum_{i=1}^{N} F_i^2} \left\{ \sum_{i=1}^{N} (\delta_{\text{free}\,p} - \delta_{\text{obs}\,pi}) \frac{\partial F_i}{\partial K} \right. \\
&\qquad \left. - \frac{2}{\sum_{i=1}^{N} F_i^2} \left(\sum_{i=1}^{N} F_i (\delta_{\text{free}\,p} - \delta_{\text{obs}\,pi}) \right) \sum_{i=1}^{N} F_i \frac{\partial F_i}{\partial K} \right\}
\end{aligned}$$

The second derivative $\partial^2 \text{SSR} / \partial K^2$ can be approximated by

$$\frac{\partial^2 \text{SSR}}{\partial K^2} = 2 \sum_{p=1}^{P} \sum_{i=1}^{N} \left(\frac{\partial \delta_{\text{calc}\,pi}}{\partial K} \right)^2$$

for the same reasons as with method 1.

E. Method 5. Weighting of observations by propagating measurement errors.

1. The optimization method. In principle, it is possible to use a Levenberg-Marquardt procedure to find the parameter values that minimize the weighted fit score SSR*, which is defined in equations 17 and 18. To do this, however, it is necessary to expand equation 22 by substituting into it the expression for σ_{pi}^2. Because each weighting factor σ_{pi}^2 is a sum of many terms, the expression for the derivative is quite lengthy. Furthermore, the second derivative cannot be approximated as conveniently as it can when the weighting factors are constant.

In practice, a Levenberg-Marquardt algorithm that determines the derivatives of SSR* with respect to the adjustable parameters works quite poorly. When a trial parameter set is near the SSR* minimum, the derivatives calculated by this procedure often specify a succeeding trial parameter set in the wrong direction from the minimum. This means that the calculated derivative is of the wrong sign. A failure of this type should be impossible. I have thoroughly checked both my derivation of the formulas for the derivatives and my implementation of these derivatives in the computer code, and have located no errors. I therefore conclude that so many operations are involved in the calculation of these derivatives that roundoff errors are significant. These errors are of greatest consequence near the SSR* minimum, because the first derivatives $\partial \mathrm{SSR}^\star / \partial \theta_l$ are near zero in that region. Thus, roundoff errors in the thousands of operations involved in evaluating these derivatives easily overwhelm these intrinsically small values. Consequently, this

fitting procedure now uses Brent's method instead of the accident-prone Levenberg-Marquardt algorithm. Discussion of this procedure will follow the discussion of the propagation of errors.

2. Propagation of errors. The fundamental random variables can be divided into two categories: those that affect $[\mathrm{H}]_0$ and $[\mathrm{G}]_0$, and those that do not. This categorization is important in the computation of σ_{pi}^2. The latter category includes only the NMR observations δ_{obs} and δ_{free}; all other variables belong to the former group. The formula for σ_{pi}^2 in equation 18 requires a knowledge of the derivatives $\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})/\partial x_j$, where the x_j's are fundamental random variables. Such a derivative can be reduced to

$$\frac{\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial x_j} = (1-F_i)\frac{\partial \delta_{\mathrm{free}\,p}}{\partial x_j} - \frac{\partial \delta_{\mathrm{obs}\,pi}}{\partial x_j} - \frac{D_p}{[\mathrm{S}]_{0i}}\left(\frac{\partial [\mathrm{H{\cdot}G}]_i}{\partial x_j} - F_i\frac{\partial [\mathrm{S}]_{0i}}{\partial x_j}\right). \quad (32)$$

In this equation, $\delta_{\mathrm{free}\,p}$ and $\delta_{\mathrm{obs}\,pi}$ are themselves fundamental random variables. Thus, the derivatives of these quantities with respect to any fundamental variables other than themselves are zero. Furthermore, there is no dependence of F on these variables, so, when x_j is $\delta_{\mathrm{free}\,p}$ or $\delta_{\mathrm{obs}\,pi}$, equation 32 reduces to

$$\frac{\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial \delta_{\mathrm{free}\,p}} = 1 - F_i \qquad \text{or} \qquad \frac{\partial(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial \delta_{\mathrm{free}\,p}} = 1\,.$$

Consequently, the expression for σ_{pi}^2 (equation 18) can be rewritten

$$\sigma_{pi}^2 = (1-F_i)\sigma_{\delta_{\mathrm{free}\,p}}^2 + \sigma_{\delta_{\mathrm{obs}\,pi}}^2 + D_p^2 \sum_{j=3}^{L} \frac{\sigma_{x_j}^2}{[\mathrm{S}]_{0i}^2}\left(\frac{\partial [\mathrm{H{\cdot}G}]_i}{\partial x_j} - F_i\frac{\partial [\mathrm{S}]_{0i}}{\partial x_j}\right)^2.$$

It is computationally convenient to define the two terms

$$W_{pi} = (1-F_i)\sigma_{\delta_{\mathrm{free}\,p}}^2 + \sigma_{\delta_{\mathrm{obs}\,pi}}^2 \quad (33)$$

and

$$Q_i = \sum_{j=3}^{L} \frac{\sigma_{x_j}^2}{[\mathrm{S}]_{0i}^2}\left(\frac{\partial [\mathrm{H{\cdot}G}]_i}{\partial x_j} - F_i\frac{\partial [\mathrm{S}]_{0i}}{\partial x_j}\right)^2 \quad (34)$$

so that

$$\sigma_{pi}^2 = W_{pi} + D_p^2 Q_i \,. \tag{35}$$

There is a W associated with every observation, but a Q only for every sample.

The only task remaining in the derivation of an expresssion for σ_{pi}^2 is to evaluate the derivatives contained in the expression for Q_i (equation 34). Let us turn our attention to $\partial[\mathrm{H}\cdot\mathrm{G}]_i/\partial x_j$. For notational convenience, I shall define

$$A = [\mathrm{H}]_0 + [\mathrm{G}]_0 + 1/K - 4[\mathrm{H}]_0[\mathrm{G}]_0 \,. \tag{36}$$

Now $\partial[\mathrm{H}\cdot\mathrm{G}]_i/\partial x_j$ is

$$\begin{aligned}\frac{\partial[\mathrm{H}\cdot\mathrm{G}]_i}{\partial x_j} &= \frac{1}{2}\Big([\mathrm{H}]_0 + [\mathrm{G}]_0 + 1/K - \sqrt{A}\Big) \\ &= \frac{1}{2}\left\{\frac{\partial[\mathrm{H}]_{0i}}{\partial x_j} + \frac{\partial[\mathrm{G}]_{0i}}{\partial x_j} - \frac{1}{\sqrt{A}}\left[\Big([\mathrm{H}]_{0i} + [\mathrm{G}]_{0i} + 1/K\Big)\left(\frac{\partial[\mathrm{H}]_{0i}}{\partial x_j} + \frac{\partial[\mathrm{G}]_{0i}}{\partial x_j}\right)\right.\right. \\ &\qquad \left.\left. - 2\Big([\mathrm{H}]_{0i}\frac{\partial[\mathrm{G}]_{0i}}{\partial x_j} + [\mathrm{G}]_{0i}\frac{\partial[\mathrm{H}]_{0i}}{\partial x_j}\Big)\right]\right\}.\end{aligned}$$

The only derivatives that remain to be evaluated are $\partial[\mathrm{H}]_{0i}/\partial x_j$ and $\partial[\mathrm{G}]_{0i}/\partial x_j$. The equations for $[\mathrm{H}]_0$ and $[\mathrm{G}]_0$ have already been described in equations 19–21.

$$V = V_t + IV_a \tag{19}$$

$$[\mathrm{H}]_0 = \Big([\mathrm{H}_t]_0 V_t + [\mathrm{H}_a]_0 IV_a\Big)/V \tag{20}$$

$$[\mathrm{G}]_0 = \Big([\mathrm{G}_t]_0 V_t + [\mathrm{G}_a]_0 IV_a\Big)/V \tag{21}$$

The derivatives of the quantities V, $[\mathrm{H}]_0$, and $[\mathrm{G}]_0$ are then as follows:

$$\begin{aligned}\frac{\partial V}{\partial x_j} &= \frac{\partial V_t}{\partial x_j} + I\frac{\partial V_a}{\partial x_j} + V_a\frac{\partial I}{\partial x_j} \\ &= \frac{\partial V_t}{\partial x_j} + \frac{\partial V_a}{\partial x_j} + V_a\frac{\partial I}{\partial x_j}\end{aligned}$$

$$\frac{\partial[\mathrm{H}]_0}{\partial x_j} = \frac{1}{V}\left\{V_t\frac{\partial[\mathrm{H}_t]_0}{\partial x_j} + [\mathrm{H}_t]_0\frac{\partial V_t}{\partial x_j} + V_a I\frac{\partial[\mathrm{H}_a]_0}{\partial x_j} + [\mathrm{H}_t]_0 I\frac{\partial V_a}{\partial x_j} + [\mathrm{H}_a]_0 V_a\frac{\partial I}{\partial x_j}\right\}$$
$$- \frac{1}{V^2}\left([\mathrm{H}_t]_0 V_t + [\mathrm{H}_a]_0 I V_a\right)\frac{\partial V}{\partial x_j}.$$

We can ignore factors of I, the delivery device calibration, because its "measured" value is always unity. We can also recognize that $\left([\mathrm{H}_t]_0 V_t + [\mathrm{H}_a]_0 I V_a\right)/V = [\mathrm{H}]_0$, and make that substitution.

$$\frac{\partial[\mathrm{H}]_0}{\partial x_j} = \frac{1}{V}\left\{V_t\frac{\partial[\mathrm{H}_t]_0}{\partial x_j} + [\mathrm{H}_t]_0\frac{\partial V_t}{\partial x_j} + V_a\frac{\partial[\mathrm{H}_a]_0}{\partial x_j} + [\mathrm{H}_a]_0\frac{\partial V_a}{\partial x_j} + [\mathrm{H}_a]_0 V_a\frac{\partial I}{\partial x_j}\right.$$
$$\left. - [\mathrm{H}]_0\frac{\partial V}{\partial x_j}\right\}$$

Similar derivation reveals the expression for $[\mathrm{G}]_0$.

$$\frac{\partial[\mathrm{G}]_0}{\partial x_j} = \frac{1}{V}\left\{V_t\frac{\partial[\mathrm{G}_t]_0}{\partial x_j} + [\mathrm{G}_t]_0\frac{\partial V_t}{\partial x_j} + V_a\frac{\partial[\mathrm{G}_a]_0}{\partial x_j} + [\mathrm{G}_a]_0\frac{\partial V_a}{\partial x_j} + [\mathrm{G}_a]_0 V_a\frac{\partial I}{\partial x_j}\right.$$
$$\left. - [\mathrm{G}]_0\frac{\partial V}{\partial x_j}\right\}$$

This is almost all of the information necessary to propagate the errors. All of these remaining derivatives are in terms of $\partial V_t/\partial x_j$, $\partial V_a/\partial x_j$, $\partial I/\partial x_j$, $\partial[\mathrm{H}_a]_0/\partial x_j$, $\partial[\mathrm{H}_t]_0/\partial x_j$, $\partial[\mathrm{G}_a]_0/\partial x_j$, and $\partial[\mathrm{G}_t]_0/\partial x_j$. Three of these derivatives, $\partial V_t/\partial x_j$, $\partial[\mathrm{G}_t]_0/\partial x_j$, and $\partial[\mathrm{H}_t]_0/\partial x_j$, are merely the solution derivatives $\partial V/\partial x_j$, $\partial[\mathrm{G}]_0/\partial x_j$, and $\partial[\mathrm{H}]_0/\partial x_j$ of the solution already in the tube before the most recent addition. Thus, they can just be carried forward from the previous solution. This leaves only $\partial I/\partial x_j$, $\partial V_a/\partial x_j$, $\partial[\mathrm{H}_a]_0/\partial x_j$, and $\partial[\mathrm{G}_a]_0/\partial x_j$, the derivatives with respect to x_j of the calibration of the delivery device used to add the aliquot, the volume of the aliquot, the total host concentration of the added solution, and the total guest concentration of the added solution.

The derivative $\partial I/\partial x_j$ has very simple behavior. The only fundamental random variable upon which I has any dependence is I itself. Thus, the derivative is zero if x_j is anything other than I, and one if it is I.

$$\frac{\partial I}{\partial x_j} = \begin{cases} 0, & \text{if } x_j \neq I; \\ 1, & \text{if } x_j = I. \end{cases}$$

The derivative $\partial V_a/\partial x_j$ has similarly simple behavior. Just as with I, V_A is a fundamental random variable. Its distribution is determined by the delivery device used, but V_a is independent of all other random variables.

$$\frac{\partial V_a}{\partial x_j} = \begin{cases} 0, & \text{if } x_j \neq V_a; \\ 1, & \text{if } x_j = V_a. \end{cases}$$

The derivatives $\partial[\mathrm{H}_a]_0/\partial x_j$ and $\partial[\mathrm{G}_a]_0/\partial x_j$ are more interesting. If the added solution is a sample solution, these derivatives are already known: they are just $\partial[\mathrm{H}]_0/\partial x_j$ and $\partial[\mathrm{G}]_0/\partial x_j$ of that solution. If the added solution is a stock solution, however, it is no longer possible to defer calculation of the derivative. As may be expected, the behavior in this case is very simple. Stock solution concentrations are fundamental random variables in their own right, leading to the familiar-looking expressions

$$\text{if } [\mathrm{H}_a]_0 = [\mathrm{H}]_\mathrm{s}, \text{ then } \frac{\partial[\mathrm{H}_a]_0}{\partial x_j} = \frac{\partial[\mathrm{H}]_\mathrm{s}}{\partial x_j} = \begin{cases} 0, & \text{if } x_j \neq [\mathrm{H}]_\mathrm{s}\,; \\ 1, & \text{if } x_j = [\mathrm{H}]_\mathrm{s}\,; \end{cases}$$

$$\text{if } [\mathrm{G}_a]_0 = [\mathrm{G}]_\mathrm{s}, \text{ then } \frac{\partial[\mathrm{G}_a]_0}{\partial x_j} = \frac{\partial[\mathrm{G}]_\mathrm{s}}{\partial x_j} = \begin{cases} 0, & \text{if } x_j \neq [\mathrm{G}]_\mathrm{s}\,; \\ 1, & \text{if } x_j = [\mathrm{G}]_\mathrm{s}\,. \end{cases}$$

3. Minimizing SSR*. As previously stated, Levenberg-Marquardt proved to be a poor method for determining the parameters that minimize SSR*. Nevertheless, the protocol I use to minimize SSR* is conceptually similar to the protocol for minimization of SSR without propagation of measurement errors. At every trial value of K, the optimum values of all D's are determined. K is varied, with re-determination

of all D's at each new value, until SSR* is minimized. The only difference from the unweighted protocol is the actual computational method employed to perform the optimization.

Because of the complicated nature of the weighting factors, the D's cannot properly be treated as linear parameters. To determine the best D's for a given K, then, some nonlinear regression procedure must be used. A Levenberg-Marquardt procedure is a good first method to try. Although L.-M. proved to be inappropriate for optimizing K, the derivatives of SSR* with respect to the D's are not nearly as computationally intensive as those with respect to K. These derivatives can be calculated by applying the model to equation 22 to yield equation 37.

$$\frac{\partial \mathrm{SSR}^\star}{\partial D_p} = \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \left[2 \frac{\partial (\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\partial D_p} - \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \frac{\partial \sigma_{pi}^2}{\partial D_p} \right] \tag{37}$$

$$= \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \left[2 \frac{\partial}{\partial D_p} (\delta_{\mathrm{free}\,p} - D_p F_i - \delta_{\mathrm{obs}\,pi}) - \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \frac{\partial}{\partial D_p} (W_{pi} + D_p^2 Q_i) \right]$$

$$= \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \left[-2F_i - \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} (2 D_p Q_i) \right]$$

$$= -2 \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \left[F_i + D_p Q_i \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})}{\sigma_{pi}^2} \right]$$

The second derivative $\partial^2 \mathrm{SSR}^\star / \partial D_p^2$ must also be derived.

$$\frac{\partial^2 \mathrm{SSR}^\star}{\partial D_p^2} = -2 \sum_{i=1}^{N} \left[\left(F_i + D_p Q_i \frac{\partial \delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\partial \sigma_{pi}^2} \right) \frac{\partial}{\partial D_p} \left(\frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) + \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \frac{\partial}{\partial D_p} \left(F_i - 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) \right]$$

$$= -2 \sum_{i=1}^{N} \left[\left(F_i + D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) \left(\frac{-F_i}{\sigma_{pi}^2} - 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^4} \right) + \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \left\{ 0 + D_p Q_i \left(\frac{-F_i}{\sigma_{pi}^2} - 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^4} \right) + Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right\} \right]$$

$$= -2 \sum_{i=1}^{N} \left[\frac{-1}{\sigma_{pi}^2} \left(F_i + D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) \left(F_i + 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) - \frac{1}{\sigma_{pi}^2} \left(D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) \left(F_i + 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) + Q_i \left(\frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right)^2 \right]$$

$$= -2 \sum_{i=1}^{N} \left[\frac{-1}{\sigma_{pi}^2} \left(F_i + 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) \left(F_i + 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right) + Q_i \left(\frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right)^2 \right]$$

$$\frac{\partial^2 \mathrm{SSR}^\star}{\partial D_p^2} = \sum_{i=1}^{N} \left[- \frac{1}{\sigma_{pi}^2} \left(F_i + 2 D_p Q_i \frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right)^2 + Q_i \left(\frac{\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi}}{\sigma_{pi}^2} \right)^2 \right]$$

This Levenberg-Marquardt scheme for optimizing D_p works fairly well. Occasionally, it will fail to move the trial value of D_p in the proper direction when it is close to its optimum value. This is probably, once again, a result of roundoff error. If this happens, the calculated value of $\partial \mathrm{SSR}^\star / \partial D_p$ will decrease, but the resulting value of $\mathrm{SSR}^\star$ will *increase*. This signals that the method is failing, and the fitting procedure

will then abandon the L.-M. fitting and optimize D_p by Brent's method. The value of K that leads to the lowest possible SSR* value is also found by using Brent's method.

Chapter 3

Hypothesis Testing and Parameter Confidence Intervals for Nonlinear Models with Measurement Error. Application to Molecular Recognition Studies.

Abstract: The association free energies of intermolecular complexes are estimated in molecular recognition studies by fitting a nonlinear model to NMR titration data. Because the model is nonlinear, standard methods for evaluating the fit of the model to the data and assigning confidence limits to the fitted parameters cannot be applied. The presence of significant measurement errors further compromise the utility of theoretical approaches. A set of computer programs that employ Monte Carlo simulation of the nonlinear model and measurement errors has been developed to accomplish both goals. Although intended for molecular recognition studies, the methods are valid for nonlinear models in general.

I. Introduction

A. The Model.

In our studies of molecular recognition, we probe the complexation of molecules in solution. In the simplest case, this involves the general reaction $H + G \rightleftharpoons H{\cdot}G$, where H is a macrocyclic receptor ("host"), G is a smaller organic species ("guest"), and H·G is the host/guest complex. The felicity of this interaction is quantified by the equilibrium constant K, $K = [H{\cdot}G]/([H][G])$. In this expression [H] is the concentration of free host, [G] the concentration of free guest, and [H·G] the concentration of the host/guest complex. The total concentration of host, $[H]_0$, is $[H] + [H{\cdot}G]$, and the total concentration of guest, $[G]_0$, is $[G] + [H{\cdot}G]$.

Like other investigators in the field, we study this complexation with NMR spectroscopy. Because the complexation/dissociation reaction is fast on the NMR timescale, an interacting species does not exhibit separate resonances for its free and bound states. Instead, a proton's signal appears between where its free and bound resonances would have been if the reaction were slower, weighted by the fraction of

time it spends in each state. If the fraction of a species bound is F, the fraction free is $(1 - F)$. If the observed proton is part of the host, $F = [\mathrm{H{\cdot}G}]/[\mathrm{H}]_0$; if it is part of the guest, $F = [\mathrm{H{\cdot}G}]/[\mathrm{G}]_0$. The equilibrium constant and mass balance expressions combine to give equation 1, the expression for the concentration of the complex.

$$[\mathrm{H{\cdot}G}] = 1/2\left\{[\mathrm{H}]_0 + [\mathrm{G}]_0 + 1/K - \sqrt{([\mathrm{H}]_0 + [\mathrm{G}]_0 + 1/K)^2 + 4[\mathrm{H}]_0[\mathrm{G}]_0}\right\} \quad (1)$$

F thus determined predicts the observed proton resonance.

$$\delta_{\mathrm{obs}} = \delta_{\mathrm{free}}(1 - F) + \delta_{\mathrm{bound}}(F) \quad (2)$$

If D is defined as $\delta_{\mathrm{free}} - \delta_{\mathrm{bound}}$, equation 2 rearranges to equation 3.

$$\delta_{\mathrm{obs}} = \delta_{\mathrm{free}} - DF \quad (3)$$

F is a function of $[\mathrm{H}]_0$, $[\mathrm{G}]_0$, and the association constant K. Thus, the observed signal δ_{obs} is a function of δ_{free}, $[\mathrm{H}]_0$, $[\mathrm{G}]_0$, D, and K. This is the *response variable.* The quantities δ_{free}, $[\mathrm{H}]_0$, and $[\mathrm{G}]_0$, which can be measured independently, are the *explanatory variables*, and are represented as a vector $\mathbf{x}$.[1] The complete set of such vectors is $\mathbf{X}$. D and K, which are known only to Nature, are parameters, and can be represented as a vector θ. For generality, this relation between the parameters, explanatory variables, and response variables shall be represented by the function g.

$$\delta_{\mathrm{obs}} = g(\delta_{\mathrm{free}}, [\mathrm{H}]_0, [\mathrm{G}]_0, D, K) = g(\mathbf{x}, \theta) \quad (4)$$

These parameters θ are experimentally determined by taking observations at a variety of values of $\mathbf{x}$, and finding the parameter vector that best satisfies equation 4. The most satisfactory parameter vector for a data set $(\mathbf{X}, \mathbf{y})$ is denoted $\theta(\mathbf{X}, \mathbf{y})$. In a least squares sense, this is the parameter set that minimizes the fit score SSR.

$$\mathrm{SSR} = \sum_{i=1}^{N} \left(g(\mathbf{x}_i, \theta) - \delta_{\mathrm{obs}\,i}\right)^2 \quad (5)$$

(1) This is a simplification; see Appendix A.

SSR is the sum of squared residuals between the model's prediction and each of the N observations.[2]

If the predictive model g is a linear function of the parameters, the parameter vector corresponding to the minimum SSR can be found in a single analytical step. If the predictive model is a *nonlinear* function of the parameters, however, this is not possible. Because F is a nonlinear function of K, the model of equation 3 is a nonlinear model. Consequently, finding the parameters $\theta(\mathbf{X}, \mathbf{y})$ must be accomplished by nonlinear regression.

This chapter describes our efforts to overcome some of the theoretical drawbacks associated with nonlinear regression. We have focused on the molecular recognition model, but the discussion is largely applicable to nonlinear models in general. To maintain the generality of the discussion, we shall refer to the predictive model as $g(\mathbf{x}, \theta)$, the set of explanatory variables as $\mathbf{X}$, and the set of response variables as $\mathbf{y}$. Specific aspects of the molecular recognition model will be mentioned only if their uniqueness requires special consideration, or to provide clarifying examples.

B. Nonlinear Regression and Measurement Errors.

Parameter estimation in nonlinear models suffers from several drawbacks in comparison to standard linear fitting. The most familiar is that the best-fit parameters, $\tilde{\theta}$, cannot be found in a single step, as they can in the linear case. Instead, an initial guess for the parameters must be sequentially refined until some convergence criterion is satisfied. This is computationally intensive, and is not even certain to find the best parameter set.

---

(2) The fit score we actually use in our studies is more complicated. Each residual is inversely weighted by the estimated uncertainty in its corresponding prediction. This uncertainty is estimated to first order by propagating the measurement errors in the explanatory variables through the function $g(\mathbf{x}, \theta)$. See Bevington, P. R. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill; New York, 1969; p 60, and Chapter 2 of this thesis. The generality of this discussion, however, is not affected by the exact form of SSR, so we shall continue to refer to the simple definition in equation 5.

In most cases, however, nonlinear fitting is comfortably within reach of modest personal computers. Problems with multiple or nonexistent local minima occur most often in very poorly fitting data sets, and in regions of the SSR surface far from the global minimum. With proper care in programming, these problems are minimized.[3,4]

Other problems in nonlinear models stem from the peculiar behavior of parameter estimates. These are compounded in our system by the presence of measurement errors. The underlying assumption of most linear and nonlinear regression procedures is that there is uncertainty in only the response variable. In such a case, the outcome of an experiment with N data points is described by equation 6.

$$y_i = g(\mathbf{x}_i, \theta) + \varepsilon_i, \ i = 1, \ldots, N \tag{6}$$

The errors ε_i are independent and identically distributed (iid) random variables from a normal distribution with a mean of zero and variance of σ^2.

Measurement errors in the explanatory variables $\mathbf{X}$ complicate the situation. If such measurement errors exist, the prediction of y_i will be based, not on the actual values of the explanatory variables $x_{i1}, x_{i2}, \ldots, x_{iL}$, but on our mistaken perceptions of them, $x_{i1} + e_{i1}, x_{i2} + e_{i2}, \ldots, x_{iL} + e_{iL}$. These perceived values shall be called $\tilde{\mathbf{x}}_i$; $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{e}_i$. The proper model for the value of $\mathbf{y}$ in response to the perceived values $\widetilde{\mathbf{X}}$ then is

$$y_i = g(\tilde{\mathbf{x}}_i - \mathbf{e}_i, \theta) + \varepsilon_i, \ i = 1, \ldots, N\,. \tag{7}$$

---

(3) Press, W. H.; Flannery, D. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: the Art of Scientific Computing*; Cambridge University: New York, 1986; pp 521–525.

(4) Seber, George Arthur Frederick; Wild, Christopher John; *Nonlinear Regression*; Wiley: New York, 1989; Chapter 14.

As when there are no measurement errors, the best estimates of the parameters θ are those that minimize SSR.[5]

$$\mathrm{SSR} = \sum_{i=1}^{N} \big(g(\tilde{\mathbf{x}}_i, \theta) - y_i\big)^2 \tag{8}$$

These best-fit parameters $\theta(\widetilde{\mathbf{X}}, \mathbf{y})$ are designated $\tilde{\theta}$, and their corresponding minimum SSR is $\widetilde{\mathrm{SSR}}$.

II. Interpreting Parametric Models

Two questions immediately arise in fitting any parametric model to a data set. The first concerns the fitness of the model: does the model adequately explain the observed data? The second question concerns the fitted parameters themselves: how good are they as estimates of the true parameter values? This latter question is meaningless if the model is a poor explanation of the data. Nothing is gained by studying the parameters of an inappropriate model.

A. Fitness of the Model.

In the linear case, the fitness of the model may be evaluated by examination of the residuals (difference between the data and the model). If there are measurement errors only in $\mathbf{y}$, and the variances of these measurement errors are known, then the distribution of a fit score such as $\widetilde{\mathrm{SSR}}$ can be predicted theoretically. Specifically, when a model with k adjustable parameters is fitted to an experiment with N observations,

$$\widetilde{\mathrm{SSR}} = \min\left(\sum_{i=1}^{N} \frac{\big(y(\mathbf{x}_i) - y_i\big)^2}{\sigma_i^2}\right) \sim \chi^2_{N-k} \ , \tag{9}$$

the sum of squares of the residuals, each divided by its expected measurement variance σ_i^2, is distributed as a χ^2 variable with $N - k$ degrees of freedom. If the

(5) This is true only approximately. In principle, it is possible to improve on a naïve estimator by incorporating some corrections for the measurement errors; see Stefanski, Leonard A. "The effects of measurement error on parameter estimation," *Biometrika* **1985**, *72*, 583–592.

best parameters do not enable the model to fit the data adequately, this failure will be betrayed by residuals larger than warranted by the measurement uncertainties. This may be quantified by a χ^2 test, in which $\widetilde{\mathrm{SSR}}$ is compared to a theoretical χ^2_{N-k} distribution. If a variable truly following this χ^2 distribution has a probability less than α of attaining a value as high or higher than the observed $\widetilde{\mathrm{SSR}}$, then one can state with $(1-\alpha)\times 100\%$ confidence that $\widetilde{\mathrm{SSR}}$ does *not* follow this distribution. With the same confidence, the model is rejected as an adequate explanation of the data. The critical regions of the χ^2 distributions can be read from a table. For example, if $N-k=14$, the probability under the model that $\widetilde{\mathrm{SSR}} > 23.68$ is exactly 5%. An observed $\widetilde{\mathrm{SSR}}$ of 24, then, rejects the model with 95% confidence.

A qualitiative but potentially more sensitive way to detect an inadequate model is to examine the pattern of the residuals when plotted against one of the explanatory variables. If the residuals show only random chatter evenly distributed about zero, the model is probably good. If the residuals have a noticeable pattern, such as curvature or oscillations, however, the response variables probably depend on the explanatory variables in a way not accounted for by the model. This is sufficient reason to suspect that the model is poor, even if SSR is well within its rejection cutoff. This procedure is also useful if a rejection cutoff cannot be estimated because the magnitudes of the errors in $\mathbf{y}$ are unknown.

In the case of a nonlinear model and errors in the explanatory variables, neither of these tests for the fitness of the model is reliable. Errors in the explanatory variables can affect the predictions of a nonlinear model in complex ways. Even if the measurement errors are normally distributed about zero, such behavior is not guaranteed for the residuals. In the general nonlinear case, the distribution of a fit score such as $\widetilde{\mathrm{SSR}}$ is not theoretically known. A simple test of the fit against a catalogued distribution is thus not possible.

The unpredictable behavior of the residuals in a nonlinear model also destroys the utility of residual plots. A pattern to the residuals no longer necessarily indicates that the model must be overthrown. An example is shown in Figure 1. This figure shows the residuals from a fit to a hypothetical NMR titration study. The "data" follow equation 3 exactly; the only measurement error is that the concentration of the host stock solution is actually about 5% lower than assumed. Although the model is valid, a noticeable pattern to the residuals is caused by a single error in measuring a quantity that affects all of the observations. This pattern to the residuals is *not* cause for rejection of the model.

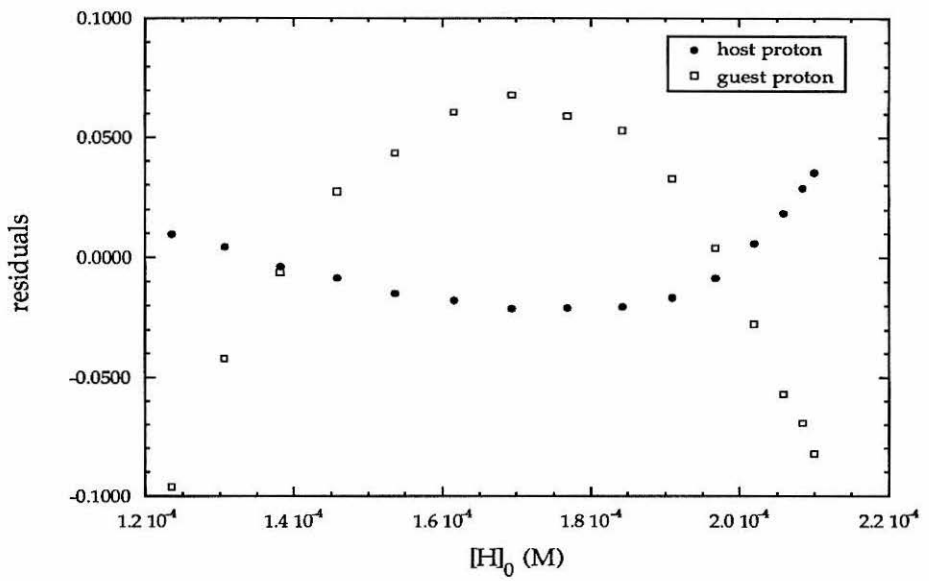


Figure 1. Plot of residuals against $[H]_0$ for a simulated data set containing error only in the measurement of the host stock solution concentration. The residuals here are dimensionless; see note 2.

B. Quality of the Fitted Parameters.

The most informative way to report the accuracy of fitted parameters is in the form of confidence intervals. The "confidence" associated with an interval is the

certainty that the specified region contains the true parameter value. The width of a given confidence interval is an inverse measure of the reliability of the parameter.

1. Distribution methods. The possibility of assigning confidence intervals to parameters implies that there exists some monotonically increasing confidence function $C(t)$ such that

$$C(t) = \text{Prob}(t > \theta_l) \ , \tag{10}$$

that is, $C(t)$ is the probability that t exceeds the true parameter value θ_l. If the exact functional form of $C(t)$ is known, confidence limits for θ_l can be assigned as in Figure 2. A 90% confidence region for θ_l, for example, would be $[C^{-1}(0.05), C^{-1}(0.95)]$. $C(t)$, as defined, is the cumulative distribution function of θ_l.

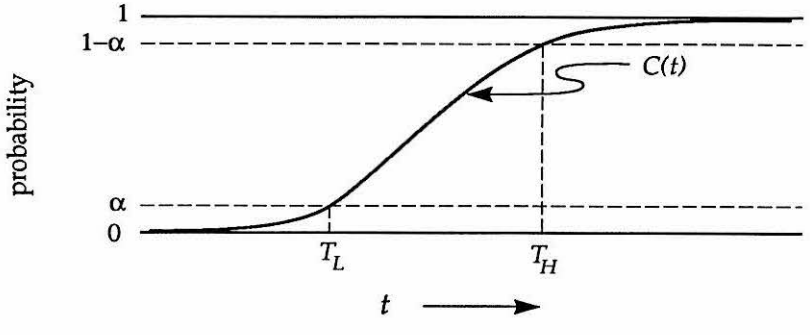


Figure 2. Constructing a confidence interval for the parameter θ_l. There is a probability of $(1 - 2\alpha)$ that the true parameter value lies between T_L and T_H.

Such a picture is fundamentally flawed because θ_l is not a random variable. There exists a single true value of θ_l, which is known to Nature but unknown to the experimenter. The true form of the function $C(t)$ is a step at the actual value of θ_l.

$$C(t) = \begin{cases} 0, & \text{if } t \leq \theta_l; \\ 1, & \text{if } t > \theta_l. \end{cases}$$

According to this function, any interval containing θ_l is a 100% confidence interval, and any interval not containing θ_l is a 0% confidence interval. Of course, if θ_l were known, there would be no need for confidence intervals.

Uncertainty in the estimated value of a parameter arises not because the actual parameter may take on a variety of values, but because estimates are imprecise. In order to assign confidence limits to a fitted parameter, it is necessary to find some other function containing "the same information" as the unknowable $C(t)$.

One candidate function is $C'(t)$, the probability of obtaining the observed experimental data set $(\widetilde{\mathbf{X}}, \tilde{\mathbf{y}})$ if t exceeds the true value of θ_l. To become a proper distribution function, this formula must be normalized to the total probability of obtaining the observed experimental data set at all values of t.

$$C'(t) = \frac{\int_{-\infty}^{t} \text{Prob}(\mathbf{y} = \tilde{\mathbf{y}} \mid \theta_l = v)dv}{\int_{-\infty}^{\infty} \text{Prob}(\mathbf{y} = \tilde{\mathbf{y}} \mid \theta_l = v)dv} \tag{11}$$

Practically, it is impossible to know the probability of finding a data set exactly like the one observed, so a more tractable substitute must be used instead of C'. A convenient such function which we shall call $C^{\star}(t)$ is the probability, if t is the true value of θ_l, that the best parameter value $\hat{\theta}_l$ returned from a fit to a resultant data set is greater than $\tilde{\theta}_l$.

$$C^{\star}(t) = \text{Prob}(\hat{\theta}_l > \tilde{\theta}_l \mid \theta_l = t) \tag{12}$$

The unrealizable goal of determining the probability in equation 11 has been replaced by the more attainable goal of determining the probability of observing a data set that gives a fitted parameter value $\hat{\theta}_l$ greater than $\tilde{\theta}_l$. This condenses all of the information about the actual data set into the summary statistic $\tilde{\theta}_l$, and all of the information about any hypothetical data set arising under the premise $\theta_l = t$ into the summary statistic $\hat{\theta}_l$. The ability to generate a data set similar to the experimental set is gauged by the tendency of fitted parameters to approximate the experimental ones. If a value of θ_l at t causes the best-fit parameters to be far removed from $\tilde{\theta}_l$, then the experimental data set probably would not have arisen if θ_l were t. This is the implicit meaning of most reported confidence intervals.

Another conceivable substitute for $C(t)$ is $C^{\dagger}(t)$.

$$C^{\dagger}(t) = \text{Prob}\Big(t > \theta_l(\mathbf{X},\mathbf{y}) \;\Big|\; \mathbf{x}_i \sim (\tilde{\mathbf{x}}_i + \mathbf{e}_i),\ \text{and}\ y_i \sim (\tilde{y}_i + \varepsilon_i)\Big) \qquad (13)$$

This is the probability that a data set $(\mathbf{X},\mathbf{y})$ will give rise to a best-fit parameter $\theta_l(\mathbf{X},\mathbf{y})$ less than t if the data sets are distributed as perturbations of the actual measured data set $(\widetilde{\mathbf{X}},\tilde{\mathbf{y}})$. This distribution is the best available estimate of the distribution that produced the measured data. If, according to this distribution, a data set giving a best-fit parameter value less than t is extremely unlikely to arise ($C^{\dagger}(t) \ll 1$), then it is also extremely unlikely that the true value of θ_l is less than t. Likewise, if a data set giving a best-fit parameter value *greater* than t is extremely unlikely to arise, then the true value of θ_l is probably not above t. Thus, confidence limits can be determined from $C^{\dagger}(t)$ according to Figure 2.

To find confidence intervals using function $C^{\star}(t)$ or $C^{\dagger}(t)$, the values of these functions at relevant values of t must be known. More precisely, their inverses $C^{\star -1}(p)$ and $C^{\dagger -1}(p)$ must be known for values of p near zero and one. In the linear case with no measurement errors in the explanatory variables, these two functions are not only known, but also interchangeable.

This is because the distribution of a fitted parameter $\hat{\theta}_l$ is known directly from the distribution of the measurement errors in $\mathbf{y}$. If these measurement errors are iid $\sim N(0,\sigma^2)$, and the true value of θ_l is T, then $\hat{\theta}_l$ is normally distributed $\sim N(T,\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll})$.[6] The variance of this distribution does *not* depend on the true value of θ_l.

The random variable in the function $C^{\star}$ is $(\hat{\theta}_l \mid \theta_l = t)$. This variable is distributed $\hat{\theta}_l \sim N(t,\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll})$. $\hat{\theta}_l$ can be transformed to a standard normal random variable $z^{\star}$.

$$z^{\star} \equiv \frac{\hat{\theta}_l - t}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}}$$

(6) Dunteman, George H.; *Introduction to Linear Models*; Sage: Beverly Hills, 1984; pp 157–175.

The function $C^\star$ can then be defined in terms of the known standard normal distribution function Φ. If z is a standard normal random variable, then $\Phi(t) = \text{Prob}(t > z)$.

$$C^\star(t) = \text{Prob}(\hat{\theta}_l > \tilde{\theta}_l | \theta_l = t) = \text{Prob}\left(\frac{t - \tilde{\theta}_l}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}} > z^\star\right) = \Phi\left(\frac{t - \tilde{\theta}_l}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}}\right)$$

Likewise, the random variable in the function $C^\dagger$ is $\Big(\theta_l(\mathbf{X}, \mathbf{y}) \;\Big|\; \mathbf{x}_i \sim (\tilde{\mathbf{x}}_i + \mathbf{e}_i)$, and $y_i \sim (\tilde{y}_i + \varepsilon_i)\Big)$. This variable is centered on $\tilde{\theta}_l$, so is distributed $\hat{\theta}_l \sim N(\tilde{\theta}_l, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll})$. Transforming $\hat{\theta}_l$ to a standard normal $z^\dagger$ enables the restatement of $C^\dagger$ in terms of Φ.

$$z^\dagger \equiv \frac{\hat{\theta}_l - \tilde{\theta}_l}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}}$$

$$C^\dagger(t) = \text{Prob}(t > \tilde{\theta}_l) = \text{Prob}\left(\frac{t - \tilde{\theta}_l}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}} > z^\dagger\right) = \Phi\left(\frac{t - \tilde{\theta}_l}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{ll}}}\right)$$

Thus, it is clear that in the linear case without errors in the explanatory variables, confidence limits are easy to obtain. Furthermore, the criteria of $C^\star$ and $C^\dagger$ lead to identical intervals.

The situation is naturally more complicated when the model is nonlinear and there are measurement errors in the explanatory variables. In such a case, no simple formula exists to determine the distributions of fitted parameters. There is thus no way to readily obtain confidence intervals from a catalogued function such as Φ, and no guarantee that the functions $C^\star(t)$ and $C^\dagger(t)$ are identical. Furthermore, in the nonlinear multiparameter case, the value of $C^\star(t)$ for parameter θ_l may depend on parameters $\theta_{j\neq l}$ other than θ_l.

3. Constant χ^2 surfaces. Another popular method for constructing confidence intervals is by inverting the χ^2 test for fitness of the model.[7] The χ^2 test tells the probability that the model is valid given the value of $\widetilde{\text{SSR}}$. By the same token, if a slightly different model is used, the magnitude of the SSR between it and the

(7) Reference 4, Section 5.4 (pp 202–203), Section 5.10 (pp 236–245).

data gives a score of the fitness of the new model. If the new model differs from the optimum model only in its parameter values, SSR tells how poorly the new parameters fit the data. A confidence region for the parameter θ_l can be drawn between low and high parameter values that give the same elevated SSR value. The level of this interval is determined by how far the elevated SSR exceeds the minimum $\widetilde{\text{SSR}}$. Typically, the limiting SSR score is scaled to $\widetilde{\text{SSR}}/(N-k)$, and the confidence level is chosen from the χ^2_{N-k} distribution.

For example, if a single-parameter model applied to an experiment with 12 data points leaves behind an $\widetilde{\text{SSR}}$ of 10.0, then the boundaries of the 95% confidence interval are those parameter values giving $\text{SSR} = \chi^{2\,(0.95)}_{12-1} \times 10.0/(12-1) = 196.8/11 = 17.89$.

This procedure is especially useful for delimiting joint parameter confidence regions in multiparameter models. It is usually even applicable to nonlinear models. There also exist methods based on this general technique to correct for artifacts of model nonlinearity.[8] Unfortunately, these have been theoretically verified only in the general case of no errors in the explanatory variables.

III. Monte Carlo Methods

Although the task is not as easy as in the linear case, protocols for both evaluating the model and assigning parameter confidence intervals can be developed even for cases as complex as molecular recognition studies. These procedures are in principle the same as procedures used in linear models. However, the distributions of fitted parameters cannot be identified as simple instances of known distributions. Instead, they must be found by Monte Carlo sampling before they can be applied to confidence intervals or a test of the model.

(8) Hamilton, David "Confidence regions for parameter subsets in nonlinear regression," *Biometrika* **1986**, *73*, 57–64.

A. Fitness of the Model.

The test of the model can be carried out in a manner analogous to a χ^2 test. Since the expected distribution of the fit score SSR is not known theoretically, the first step in this test is to find the distribution of SSR empirically. The model, including the measurement errors and the parameter vector $\tilde{\theta}$, defines a distribution for SSR. If this model is correct, then $\widetilde{\mathrm{SSR}}$ will comply with this distribution.

The real-world $\tilde{\theta}$ and $\widetilde{\mathrm{SSR}}$ distributions could in principle be mapped out by performing many independent experimental trials. This would involve, in the case of a molecular recognition study, making up new stock solutions and using a new set of syringes each time, in order to prevent any measurement errors from influencing more than one trial. In each trial, the true values of the explanatory variables will deviate from their measured values. (The true values deviate from the measured values rather than vice versa because the explanatory variables in these studies are *design points*, intended by the experimenter to have certain values. Measurement errors cause a failure to attain these values, but the experimenter's best estimates are still the intended values.) The response variables then will take on values dictated by the true model and the actual values of its explanatory variables. Finally, the imprecisely measured response and explanatory variables are fitted by the assumed model to produce a new best-fit $\tilde{\theta}$ and $\widetilde{\mathrm{SSR}}$. In time, this would eventually map out the real-world distributions of these quantities.

Performing such a multitude of experimental trials would be time-consuming and expensive. Fortunately, the SSR distribution of interest in a test of the model is not the real-world distribution, but the hypothetical distribution that *would* be the real-world distribution if the model and best-fit parameters were correct. Just as the real-world distribution can be found by performing many independent experimental trials, this hypothetical distribution can be explored by performing many independent Monte Carlo replications of the experiment. It is only necessary for the

replications to closely mimic the details of the actual experiment, including likely measurement errors. This is an example of a "parametric bootstrap" as defined by Efron.[9]

A computer can perform these replications automatically by randomly drawing appropriate values of the measurement errors in the explanatory variables and adding them to the experimental values, $\hat{x}_{ij} = \tilde{x}_{ij} + e_{ij}$. Applying the model to these sampled explanatory values gives the response values, which are in turn perturbed by random measurement errors, giving "measured" response values $\hat{y}_i = g(\hat{\mathbf{x}}_i, \tilde{\theta}) + \varepsilon_i$. This Monte Carlo data set $(\widehat{\mathbf{X}}, \widehat{\mathbf{y}})$ is subjected to a least squares fit, giving a best-fit parameter vector $\hat{\theta} = \theta(\widehat{\mathbf{X}}, \widehat{\mathbf{y}})$ and minimum fit score $\widehat{\mathrm{SSR}}$.

After many such replications have been performed, the experimental best-fit score $\widetilde{\mathrm{SSR}}$ is compared to the Monte Carlo distribution of $\widehat{\mathrm{SSR}}$. $\widehat{Q}$, the fraction of $\widehat{\mathrm{SSR}}$ values less than $\widetilde{\mathrm{SSR}}$, is the Monte Carlo estimate of Q, the actual probability that a minimum SSR at least as large as $\widetilde{\mathrm{SSR}}$ would arise. In other words, the probability that such a large $\widetilde{\mathrm{SSR}}$ came from this model distribution is only Q. Extremely low values of $\widehat{Q}$ (we use 0.001 as our cutoff) reject the model. We have written a Pascal program (Lucius) to perform this simulation and test on NMR titration experiments.

This Monte Carlo protocol for evaluating the model is summarized below:

repeat R times:

- for every explanatory variable x_{ij}, select a measurement error: $e_{ij} \sim N(0, \sigma_{ij}^2)$
- generate "actual" explanatory variables: $\hat{x}_{ij} = \tilde{x}_{ij} + e_{ij}$
- apply the model to these "actual" variables: $y_i = g(\hat{\mathbf{x}}_i, \tilde{\theta})$
- for every response variable y_i, select a measurement error: $\varepsilon_i \sim N(0, \sigma_i^2)$

(9) Efron, Bradley; *The Jackknife, the Bootstrap and Other Resampling Plans*; Society for Applied and Industrial Mathematics: Philadelphia, 1982; pp 29–30.

- generate "measured" response variables: $\hat{y}_i = y_i + \varepsilon_i$
- perform a least squares fit to the "measured" data set, obtaining $\hat{\theta} = \theta(\widetilde{\mathbf{X}}, \hat{\mathbf{y}})$ and $\widehat{\mathrm{SSR}}$

$\hat{Q}$ = (number of times $\widehat{\mathrm{SSR}} > \widetilde{\mathrm{SSR}}$)$/R$

reject the model if $\hat{Q} < 0.001$.

Failure to reject the model should not be construed as confirmation of the model. It is possible for a fundamentally incorrect model to give a fortuitously good fit to the data. If this happens, the test will not reject. In addition, the power of this test (its tendency to reject incorrect models) will increase roughly as $\sqrt{N}$. If several points are removed from a data set that rejects the model, the new truncated set may no longer reject. This does not necessarily mean that the model is now acceptable; it only means that there is not enough information to reject.

B. Parameter Confidence Intervals.

Confidence intervals for the fitted parameters can be assigned by either function $C^\star(t)$ (equation 12) or $C^\dagger(t)$ (equation 13). The distributions of the appropriate random variables in each case are determined by Monte Carlo sampling.

1. The confidence function $C^\star$. Determining the function $C^\star$ is the most complicated of the two, so it will be described first. Efron[9,10−12] has described a Monte Carlo technique, called the *bootstrap*, for assigning measures of error to statistical parameters. The method is general, but it is most useful when such measures cannot be obtained theoretically. When, as in the present case, the data are expected to follow a parametric model, a *parametric bootstrap* may be used.

(10) Efron, Bradley; Tibshirani, Robert "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science* **1986**, *1*, 54–77.

(11) Efron, Bradley "Better bootstrap confidence intervals" *J. Am. Statist. Assoc.* **1987**, *82*, 171–185.

(12) Efron, Bradley; Tibshirani, Robert "Statistical data analysis in the computer age," *Science* **1991**, *253*, 390–395.

This involves performing many Monte Carlo replications of the experiment, exactly as was described earlier for the test of the model. The distribution of data sets arising under the model, and of the parameters that best fit these data sets, can be mapped out in this way.

Let us define the bootstrap distribution function $A(t, T)$:

$$A(t, T) = \text{Prob}(t > \hat{\theta}_l | \theta_l = T)\,. \tag{14}$$

T in this case is a hypothetical true value of θ_l. $A(t, T)$ is the probability, if T *were* the true value, that a random best-fit $\hat{\theta}_l$ would be less than the index t. $\hat{\theta}_l$ is the least squares estimator of T; naturally, its distribution depends on T.

If the distribution of a given parameter estimator $\hat{\theta}_l$ is symmetrical and centered on the (assumed) "true" value $\tilde{\theta}_l$, then confidence intervals can be determined directly from it. In fact, in such a case the bootstrap distribution function $A(t, \tilde{\theta}_l)$ *is* the confidence function $C^\star(t)$. This can be intuitively proved as follows. Let us assume that the shape of the $\hat{\theta}_l$ distribution is independent of the true value of the parameter θ_l. The only effect of a change in θ_l is a translation of the distribution; every point on the distribution function is merely shifted the same distance along the horizontal axis. This assumption is reasonable given the stipulated nature of $A(t, \tilde{\theta}_l)$.

In the function $C^\star(t)$, the distribution of the random variable $\hat{\theta}_l$ is centered about t. The distribution of the random variable $\hat{\theta}_l - \tilde{\theta}_l$, then, is centered about $t - \tilde{\theta}_l$. By the same token, the random variable $\hat{\theta}_l$ in $A(t, \tilde{\theta}_l)$ is centered about $\tilde{\theta}_l$, so $t - \hat{\theta}_l$ is also centered about $t - \tilde{\theta}_l$. Since, under our assumptions, the only unique feature of a $\hat{\theta}$ distribution is its mean, these two random variables, $\hat{\theta}_l - \tilde{\theta}_l$ and $t - \hat{\theta}_l$, have identical distributions and are thus indistinguishable. Consequently, the functions $C^\star(t)$ and $A(t, \tilde{\theta}_l)$ are identical.

This scheme of using $A(t, \tilde{\theta}_l)$ as a substitute for $C^\star(t)$ is called the *percentile method*.[9,10] Unfortunately, this method is not appropriate if $A(t, \tilde{\theta}_l)$ is centered about a value other than $\tilde{\theta}_l$, or if it is not symmetrical. In the case of complexation studies, bootstrap parameter distributions are often slightly biased and always profoundly unsymmetrical.

Efron has refined the percentile method to correct for biased and unsymmetrical bootstrap distributions. The correction for bias is quite simple, requiring a knowledge of only the standard normal distribution function and the inverse of the bootstrap distribution $A^{-1}(\alpha, \tilde{\theta}_l)$. The correction for asymmetry, however, is considerably more involved, requiring determination of an *acceleration constant* for the distribution. It is not clear how to obtain this constant from a parametric bootstrap. Since the bootstrap distributions of fitted parameters from molecular recognition studies are profoundly lopsided, this inability to apply the acceleration correction is a crushing setback.

The confidence function $C^\star(t)$ can be mapped out, however, by extending the bootstrap philosophy. Bootstrap replications are a good way to uncover probability distributions that defy theoretical analysis. The function $C^\star(t)$ requires information about a large number of such distributions, and portions of it can be understood by performing several complete bootstrap studies. Such a study multiplies the computation-intensive nature of bootstrap methods several times over. It has none of the elegance of the theoretical methods applicable to linear models, or even of the bias and acceleration corrections to bootstrap methods. It substitutes brute force in place of such niceties. It is, however, completely general. This attribute outweighs its drawbacks when poorly-behaved models are considered.

The procedure involved in such a study is illustrated in Figure 3. A Monte Carlo study, comprising many replications, is performed under the assumption that $\theta_l =$

T_1. This generates the empirical distribution function $\widehat{A}(t, T_1)$, which is the bootstrap estimate of the true distribution function $A(t, T_1)$. Similar Monte Carlo studies assuming that $\theta_l = T_2$ and $\theta_l = T_3$ estimate the distribution functions $A(t, T_2)$ and $A(t, T_3)$, respectively. Each of the empirical distribution functions is evaluated where it crosses $\tilde{\theta}_l$, estimating $A(\tilde{\theta}_l, T_j)$. The points $\left(T_j, 1 - A(\tilde{\theta}_l, T_j)\right)$ are pieces to the puzzle of the $C^\star(t)$ curve. Through trial and error, the computer samples points in the important regions of the $C^\star(t)$ curve, eventually finding an estimate of the $(1 - 2\alpha) \times 100\%$ confidence interval for θ_l, which is $[\widehat{C}^{\star -1}(\alpha), \widehat{C}^{\star -1}(1 - \alpha)]$.

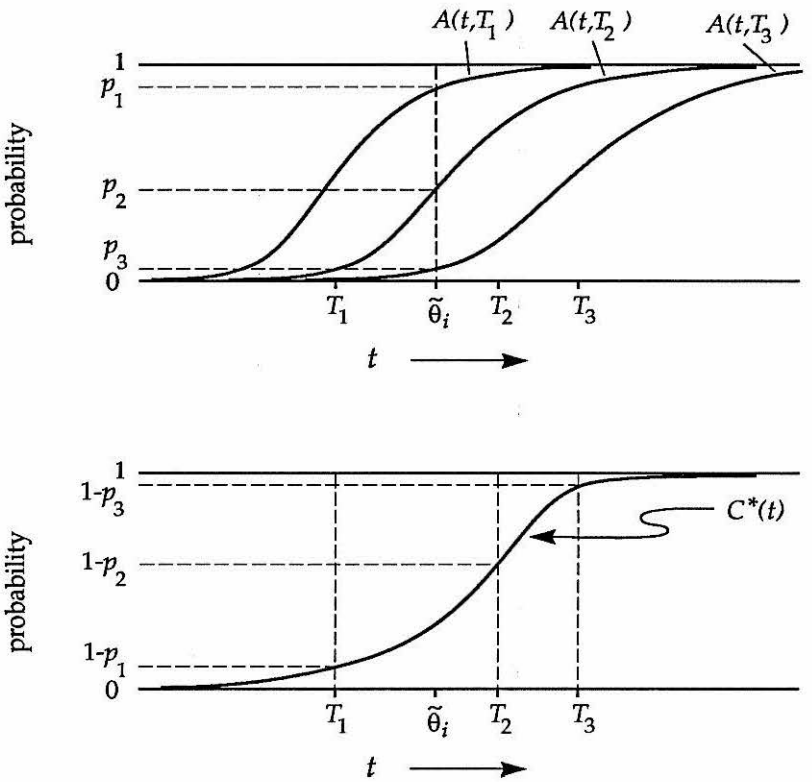


Figure 3. Determining the function $C^\star(t)$ from a series of Monte Carlo distributions $A(t, T)$. Top: p_1, p_2, and p_3 are the values at $t = \tilde{\theta}_l$ of $A(t, T_1)$, $A(t, T_2)$, and $A(t, T_3)$, respectively. Bottom: $1 - p_1$, $1 - p_2$, and $1 - p_3$ are the values at $t = T_1$, $t = T_2$, and $t = T_3$, respectively, of $C^\star(t)$.

There is one principal difficulty in obtaining the Monte Carlo distributions $\widehat{A}(t, T_j)$. In order to prepare data sets that would arise if $\theta_l = T_j$, it is necessary

to apply the model relation $g(\mathbf{x}, \theta)$ to the simulated "true" explanatory variables $\hat{\mathbf{x}}$. However, using this relation requires knowledge of all the parameters in θ, not just θ_l. This is not a problem in the test of the model, because all of the parameters in $\tilde{\theta}$ are known. If $\theta_l = T_j \neq \tilde{\theta}_l$, however, what values should be used in g for $\theta_{j \neq l}$? In order to explore $A(t, T_j)$, there must be some way to assign values to these "nuisance" parameters.

Consider the case when the parameter of interest is K and the nuisance parameters are the D's. How does the distribution function $A(t, \{K = T, D = ?\})$ respond to changes in the D values? If T is significantly smaller than $\widetilde{K}$ but the D's are held at their experimental values $\widetilde{D}$, the best-fit parameters to the resultant Monte Carlo data sets will tend toward $\widehat{K} = T$ and $\widehat{D} = \widetilde{D}$. There will be substantial variability in both $\widehat{K}$ and $\widehat{D}$, but their values will tend to be correlated. The fitted parameter values will be confined to a fairly small region of parameter space, as demonstrated by the scatter plots in Figure 4. If the D values are held in place at $\widetilde{D}$, even a slight change in K will render the region of parameter space surrounding $\tilde{\theta}$ inaccessible. As an example, the middle scatter plot in Figure 4 shows replications of the same experiment as in the upper scatter plot, but with a slightly higher K. The best-fit parameters cluster about the generating parameters $K = 12000\ \mathrm{M}^{-1}$ and $\widetilde{D}$. Although these best-fit $\widehat{K}$ and $\widehat{D}$ values in this plot fall on both sides of $\widetilde{K}$ and $\widetilde{D}$, the probability of obtaining a best-fit parameter vector near $(\widetilde{K}, \widetilde{D})$ is negligible. Clearly, such a parameter set $\{K = T, D = \widetilde{D}\}$ could not have led to the data at hand.

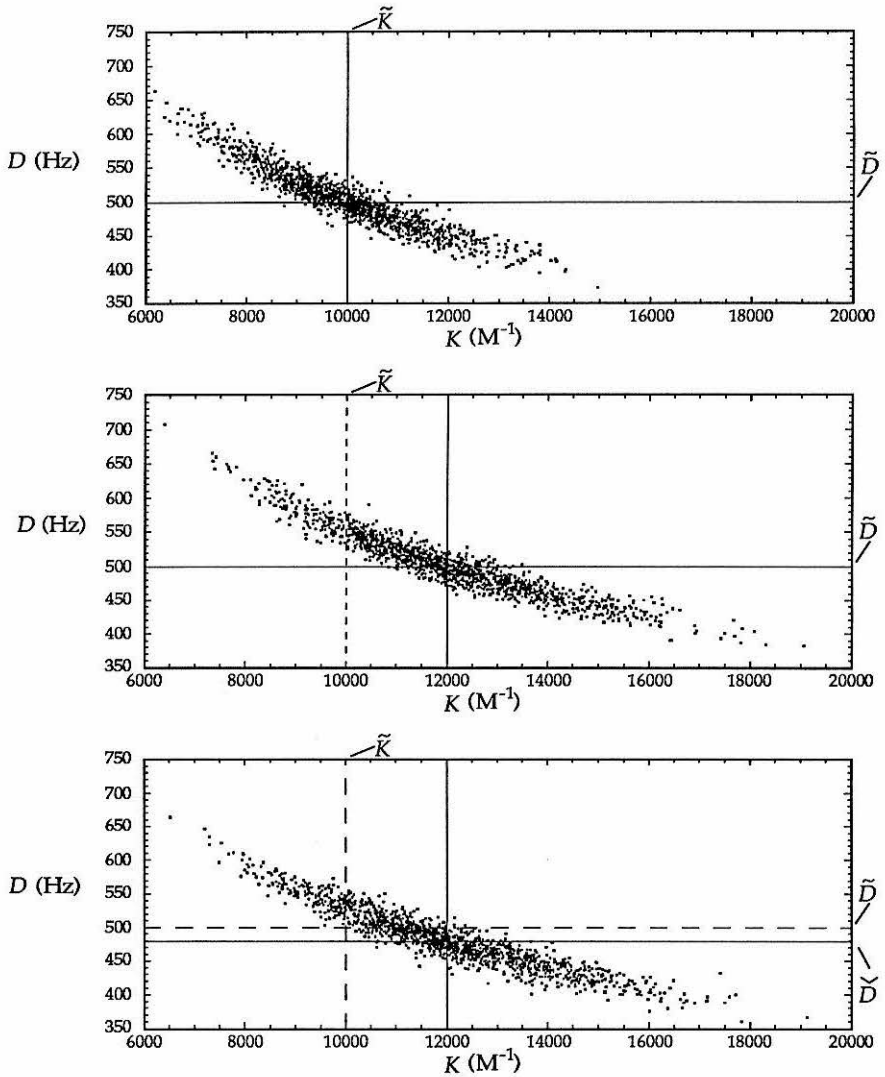


Figure 4. Scatter plots of best-fit parameter values to Monte Carlo data sets generated under different assumed parameters. Each simulation performed 1500 replications and all replications simulate the same experiment. Top: The assumed parameters are $K = \tilde{K}$ and $D = \tilde{D}$, which are as indicated. Middle: The assumed parameters are $K > \tilde{K}$ and $D = \tilde{D}$. Bottom: The assumed parameters are $K > \tilde{K}$ and $D = \check{D}$.

For any given parameter $\theta_l \neq \tilde{\theta}_l$, there must be a set of nuisance parameters $\theta_{j \neq l} \neq \tilde{\theta}_{j \neq l}$ that maximizes the chance of generating data sets that are best fit by parameter sets near $\tilde{\theta}$. In the specific case of the molecular recognition model, one would expect that a larger K could be largely compensated by a smaller D. This

would place $(\widetilde{K}, \widetilde{D})$ more into the narrow distribution of fitted parameters $(\widehat{K}, \widehat{D})$.

In order to find such parameters, the parameter of interest θ_l is held fixed at the desired value T. The nuisance parameters $\theta_{j\neq l}$ are then adjusted to give the best possible fit to the *experimental* data set. The set of best such values of $\theta_{j\neq l}$ is $\check{\theta}$. These are the values used in the Monte Carlo study at $\theta_l = T$. Because the generating function g and the consequent distribution function A depend on $\check{\theta}$, they should strictly be called $g(\mathbf{x}, \{T, \check{\theta}\})$ and $A(t, \{T, \check{\theta}\})$. However, we shall continue to use the simpler notation $g(\mathbf{x}, T)$ and $A(t, T)$, because $\check{\theta}$ is uniquely determined by T.

The lower scatter plot in Figure 4 shows best-fit parameter values obtained in such a Monte Carlo study. As in the middle plot, the generating function used 12000 M^{-1} for the value of K. The D values used were those that gave the best fit to the experiment when K was held fixed at that value. Thus, the data sets were generated from the distribution that would exist if $K = 12000$ and D is the best-fit value given this K.

There are a number of practical difficulties associated with executing this procedure, arising primarily from the stochastic nature of the individual points $\widehat{A}(\tilde{\theta}_l, t)$ making up $\widehat{C}^{\star}(t)$. The program we have developed to perform this type of study (Brutus) incorporates many safeguards to overcome these difficulties. Even in the most pathological cases, it is able to find the desired confidence limits. The computer time required for such a study is substantial. It is comparable to the time necessary for a conformational search or an ab initio quantum chemical calculation.

2. The confidence function $C^{\dagger}$. Determining the confidence function $C^{\dagger}(t)$ is considerably simpler. The idea behind this function is that all uncertainty in the fitted parameters is due to the measurement errors. Thus, thorough sampling of the conceivable values of the actual response and explanatory variables will generate a

distribution of data sets that may have actually occurred. The best-fit parameters to these data sets, then, cover the range of parameter values consistent with the experiment. This procedure is similar to, but quite distinct from, the parametric bootstrap. A parametric bootstrap generates wildly varying response variables and fits them to the model under the assumption that the explanatory variables were measured correctly. In contrast, the present method generates response values only slightly different from the ones measured in the actual experiment, and fits them to the model under the assumption that the explanatory variables are in fact somewhat different from the experimental measurements. The distributions of the fitted parameters $\hat{\theta}_l$ from a large number of Monte Carlo replications of the experiment are the confidence function estimates $\widehat{C}^{\dagger}(t)$. We have developed a Pascal program (Portia) that explores $C^{\dagger}$ for complexation studies in this manner.

The Monte Carlo protocol for exploring $C^{\dagger}$ is summarized below:

repeat R times:

- for every explanatory variable x_{ij}, select a measurement error: $e_{ij} \sim N(0, \sigma_{ij}^2)$
- generate possible explanatory variables: $\hat{x}_{ij} = \tilde{x}_{ij} + e_{ij}$
- for every response variable y_i, select a measurement error: $\varepsilon_i \sim N(0, \sigma_i^2)$
- generate possible response variables: $\hat{y}_i = \tilde{y}_i + \varepsilon_i$
- perform a least squares fit to the possible data set, obtaining $\hat{\theta} = \theta(\widetilde{\mathbf{X}}, \hat{\mathbf{y}})$ and $\widehat{\mathrm{SSR}}$.

Sort all R values of each parameter $\hat{\theta}_l$, so that $\{\hat{\theta}_{l1} \leq \hat{\theta}_{l2} \leq \cdots \leq \hat{\theta}_{lR}\}$.

The empirical confidence function is

$$\widehat{C}^{\dagger}(t) = \begin{cases} 0, & \text{if } t < \hat{\theta}_{l1}; \\ j/R, & \text{if } \hat{\theta}_{lj} \leq t < \hat{\theta}_{l\,j+1}; \\ 1, & \text{if } t \geq \hat{\theta}_{lR}. \end{cases}$$

The estimated $(1 - 2\alpha) \times 100\%$ confidence interval is $\left[\widehat{C}^{\dagger -1}(\alpha), \widehat{C}^{\dagger -1}(1-\alpha)\right]$.

This procedure is more akin to Efron's original nonparametric bootstrap than is the parametric bootstrap. Monte Carlo data sets, instead of being forced to conform to a parametric model, are built by adding errors to actual measurements. The primary difference between this procedure and the nonparametric bootstrap is that the experimental errors in the present case are drawn from independently-known distributions rather than estimated from the residuals.

3. Comparing $C^\star$ and $C^\dagger$. In our studies, the criteria of $C^\star$ and $C^\dagger$ do not produce identical confidence intervals. Typically, the confidence intervals produced by $C^\dagger$ are narrower than those from $C^\star$. We believe that this is because the replications used in finding $C^\dagger$ use more of the information about the actual experiment than do the replications used in finding $C^\star$. In $C^\dagger$, every single experimental measurement is reflected in the Monte Carlo data set. In $C^\star$, on the other hand, the only influence of the actual data is on the explanatory variables and the parameters fitted to the original experiment. With less information at its disposal, $C^\star$ is bound to be more conservative. Narrower confidence intervals, coupled with more modest requirements of computer time, make the procedure based on $C^\dagger$ the method of choice.

4. Constant χ^2 surfaces. It is not clear how parameter confidence intervals based on constant-χ^2 or SSR contours would be constructed in the present case. Constant-SSR contours are easy to locate, but assigning confidence levels to to the regions enclosed by such contours is more difficult. Even if the distribution of possible SSR scores is determined from a bootstrap $\widehat{\mathrm{SSR}}$ distribution instead of from the χ^2 distribution, the appropriate limiting values and normalizing scores are not obvious. In a binding study, a single response variable $\delta_{\mathrm{obs}\,i}$ (the NMR peak position of a proton) depends on only two parameters: K and D for that proton. The values of D for other protons do not influence it at all.

What limiting values of SSR, then, should be used to construct a confidence interval for a D? Should it be scaled to $\widetilde{\text{SSR}}$, or just to that part of $\widetilde{\text{SSR}}$ corresponding to measurements on the proton corresponding to that D? If the latter, what correction should be made for the fact that these measurements could be fit better if other protons were not included in the data set? What correction for degrees of freedom is appropriate when only two of the (number of observed protons) + 1 parameters actually apply to a single observation? Finally, what is the physical interpretation of confidence regions based on such contours? Lacking convincing answers to these questions, we have chosen not to pursue this type of parameter confidence interval.

IV. Conclusions

Although this discussion is based on one particular model for a specific chemical process, the general concept described is applicable to any nonlinear regression problem with measurement errors in the independent variables. Monte Carlo sampling allows empirical mapping of probability distributions that cannot be determined theoretically. Although these Monte Carlo methods suffer from the dual drawbacks of being computationally intensive and conceptually inelegant, they furnish an otherwise unavailable tool for critically evaluating nonlinear models and the parameters obtained from them.

Appendix A. Explanatory Variables

For convenience, we previously identified $[H]_0$, $[G]_0$, and δ_{obs} as the explanatory variables. In fact, complexation studies are conducted so that $[H]_0$ and $[G]_0$ of a given sample are actually functions of more fundamental random variables, and are often correlated with each other and with the concentrations of other samples. Typically, stock solutions of host and guest are combined in an NMR sample tube, and the NMR spectrum of the resulting sample is recorded. More host or guest stock solution, or more diluent, is then added to the tube to make another sample, whose spectrum is also recorded. $[H]_0$ and $[G]_0$ depend on the stock solution concentrations and on the volumes of all of the aliquots used in making the sample.

Thus, the true explanatory variables in the experiment, and the explanatory variables modeled by our Monte Carlo procedures, are the total host and guest concentrations $[H]_S$ and $[G]_S$ of each of the stock solutions, the calibration errors of the devices (pipets or syringes) used to add solution aliquots to the sample tubes, the volumes of the added aliquots, and NMR peak position measurements for δ_{obs} and δ_{free}. Not all the observations will depend on the same number of independent variables. Samples may comprise different numbers of aliquots, added by different devices and involving different stock solutions. For any two samples i and j, $\mathbf{x}_i$ and $\mathbf{x}_j$ may have different numbers of components.

Appendix B. How Many Monte Carlo Replications Must be Performed?

The number of Monte Carlo replications required depends on the accuracy desired for Q and the parameter confidence limits. The quantities $\widehat{Q}$, $\widehat{C}^{\star}(t)$, and $\widehat{C}^{\dagger}(t)$ are all random variables following a scaled binomial distribution $B(R,\rho)/R$. A binomial distribution $B(R,\rho)$ is the distribution followed by the number of successes occurring in R independent trials of an experiment with a probability ρ of success. It is possible to construct exact confidence intervals for these quantities. This is most easily done by approximating the scaled binomial variates as normal variates of the same mean and variance: $B(R,\rho)/R \simeq N(\rho, \rho(1-\rho)/R)$. An observed value p from such a distribution is an unbiased estimate of ρ. The distribution of p is approximately

$$p \sim \rho \pm z\sqrt{\rho(1-\rho)/R} \tag{15}$$

where z is a standard normal variate.

If we observe the value $\tilde{p}$, we would like to know what actual values of ρ could have engendered this observation. The ρ values so small and so large that the probability of obtaining a p as large (or small) as $\tilde{p}$ would be only a form the boundaries of the $(1-2a) \times 100\%$ confidence interval for ρ. In other words, if $\text{Prob}(p > \tilde{p}|\rho = \rho_L) = a$ and $\text{Prob}(p < \tilde{p}|\rho = \rho_H) = a$, the $(1-2a) \times 100\%$ confidence interval for ρ is $[\rho_L, \rho_H]$.

These boundaries ρ_L and ρ_H can be found with the help of the standard normal distribution function Φ. Let us define limiting values at significance a for a standard normal variable z. These values, $-z_a$ and z_a, have the properties $\Phi(-z_a) = a$ and $\Phi(z_a) = 1-a$. By equation 15,

$$\tilde{p} = \rho_L + z_a\sqrt{\rho_L(1-\rho_L)/R}$$

and

$$\tilde{p} = \rho_H - z_a\sqrt{\rho_H(1-\rho_H)/R}\,,$$

so the limits, ρ_L and ρ_H, of the $(1-2a)\times 100\%$ confidence interval for ρ will be found by solving equation 16 for ρ.

$$\tilde{p} = \rho \pm z_a\sqrt{\rho(1-\rho)/R} \tag{16}$$

$$\vdots$$

$$\rho = \frac{2R\tilde{p} + z_a^2 \pm z_a\sqrt{z_a^2 + 4R\tilde{p}(1-\tilde{p})}}{2(R+z_a^2)} \tag{17}$$

The number of replications R can be chosen to give acceptably narrow limits.

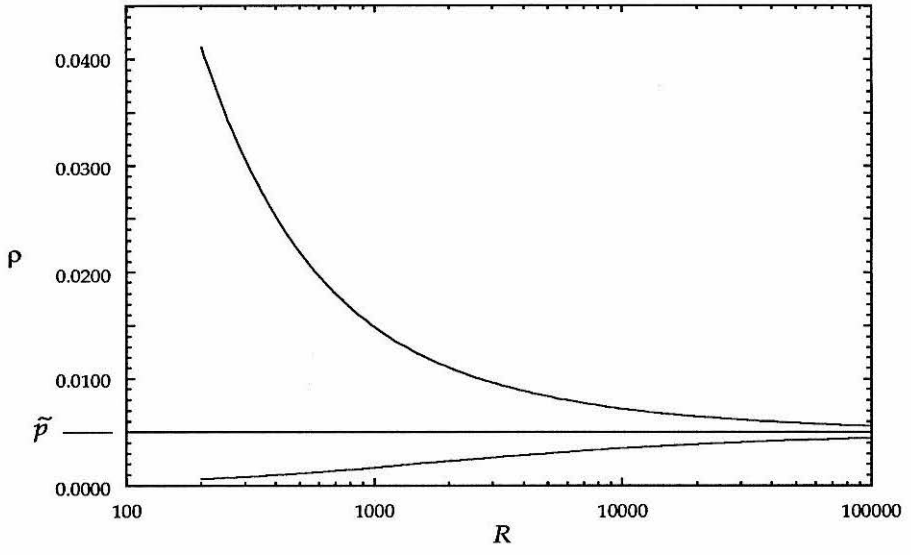


Figure 5. 99% confidence limits for ρ given that p, a scaled binomial variable $p \sim B(R,\rho)/R$, has a measured value $\tilde{p}$ of 0.005.

Figure 5 shows the 99% confidence limits for ρ, given that $\tilde{p} = 0.005$, as a function of R. Using equation 17, similar curves could easily be generated for other values of $\tilde{p}$ or a. When R is small, it is possible for ρ to be much larger than $\tilde{p}$. The profound lopsidedness of the confidence interval in this figure is a consequence of the variance of a binomial distribution growing rapidly as ρ moves away from

zero. This behavior has different consequences for the quantities $\widehat{Q}$, $\widehat{C}^{\star -1}(\alpha)$, and $\widehat{C}^{\dagger -1}(\alpha)$.

It would appear that a very large R would be required to establish that Q is larger or smaller than the rejection cutoff 0.001. In practice, however, this is not necessary. If the model is good, $\widehat{Q}$ comfortably exceeds 0.001; if it is bad, $\widehat{\mathrm{SSR}}$ is usually so far above the highest sampled value of $\widehat{\mathrm{SSR}}$ that it clearly exceeds any significant cutoff. For example, if $R = 500$, as few as three hits ($\widehat{Q} = 0.006$) would indicate with $> 99\%$ surety that $Q > 0.001$. Only in cases in which $\widehat{Q}$ is very nearly 0.001 are more replications actually required.

Confidence limits found using $C^{\dagger}$ become less certain as the confidence increases or R decreases. If $R = 5000$, for instance, what is nominally from this function a 99% parameter confidence limit is 99% likely to be between the actual 98.33% and 99.39% confidence limits. If $R = 1000$, it could range from 97.03% to 99.67%.

Confidence limits found using $\widehat{C}^{\star -1}$ have the same uncertainty properties as those found using $\widehat{C}^{\dagger -1}$. There is, however, an additional complication introduced by the method for finding $C^{\star -1}(\alpha)$. In order for $C^{\star -1}(\alpha)$ to be considered adequately determined, bracketing parameter values T_1 and T_2 must have been sampled such that $\widehat{C}^{\star}(T_1) < \alpha < \widehat{C}^{\star}(T_2)$. Furthermore, these values $\widehat{C}^{\star}(T_1)$ and $\widehat{C}^{\star}(T_2)$ must be within the interval $[\rho_L, \rho_H]$ for $\tilde{p} = \alpha$. In other words, $\rho_L \leq \widehat{C}^{\star}(T_1) < \alpha < \widehat{C}^{\star}(T_2) \leq \rho_H$. $C^{\star -1}(\alpha)$ is estimated by interpolating between T_1 and T_2.

In order to guarantee good sampling statistics for $\widehat{C}^{\star}(T_1)$, we arbitrarily require that, if $C^{\star}(T_1) = \alpha$, the estimated value $\widehat{C}^{\star}(T_1)$ must be unlikely to fall below $\alpha/2$. This is stated in relation 18. A $(1-2\alpha) \times 100\%$ confidence interval cannot be found unless there will be enough replications performed to satisfy this condition.

$$\alpha - z_a \sqrt{\alpha(1-\alpha)/R} \geq \alpha/2 \tag{18}$$

Solving this inequality for R gives inequality 19, the expression for the number of replications necessary to meet this condition.

$$R \geq 4z_a^2 \frac{1-\alpha}{\alpha} \tag{19}$$

Thus, to obtain 95% parameter confidence intervals at significance $a = 0.005$, at least 1036 replications must be performed. For 99% intervals, the number is 5283, and for 99.9% intervals, it is a whopping 53060.

Before attempting to find any confidence limits, the program checks that enough replications will be performed to adequately sample the requested confidence limits. If any requested limits are too ambitious, the program substitutes the most ambitious allowed limits in their place.

Chapter 4

The Binding Study Analysis Package

Chapters 2 and 3 of this thesis describe the philosophical and theoretical basis of the programs for analyzing NMR titration binding studies. This chapter provides a qualitative and practical guide to using these programs and interpreting their results. It is organized into three sections. The first provides an overview of what the programs do and how they operate. The second explains how to run the programs, giving thorough descriptions of input files and the user interfaces. The third describes how to use the information provided by these programs to obtain the best parameter estimates possible.

I. Overview

This package consists of four Pascal programs on the Silicon Graphics IRIS workstation. The first is Emul, which fits a simple parametric model to the binding data and generates an input file for the succeeding programs. The second is Lucius, which determines if experimental error alone can account for the discrepancies between the data and the fitted model. The final programs, Portia and Brutus, provide confidence limits for the fitted parameters.

A. Emul.

Emul estimates the association constant of a bimolecular complexation reaction. It does this by adjusting a set of parameters to fit NMR observations of solutions containing different concentrations of the complexing species. These parameters are K, the association constant, and D, the NMR shift change experienced

by one of the interacting species upon binding. There is only one K for a given system, but there are as many D's as there are protons observed. The model to which these parameters are applied is that of bimolecular association under fast-exchange conditions. To wit, the interaction of two species H and G (host and guest) is described by the reaction $\mathrm{H} + \mathrm{G} \rightleftharpoons \mathrm{H{\cdot}G}$, which obeys the equilibrium relation

$$K = \frac{[\mathrm{H{\cdot}G}]}{[\mathrm{H}][\mathrm{G}]}.$$

The NMR spectrum of one of these interacting species is a weighted average of its spectra in the complexed and free states; the weighting factor is F_i, the fraction of the species complexed in sample i. Thus, the appearance of proton p in sample i is described by

$$\delta_{\mathrm{obs}\,pi} = \delta_{\mathrm{free}\,p}(1 - F_i) + \delta_{\mathrm{bound}\,p}\,F_i$$

in which $\delta_{\mathrm{free}\,p}$ is the chemical shift in the complexed state, and $\delta_{\mathrm{bound}\,p}$ is the chemical shift of the same proton in the uncomplexed state. It is convenient to consider the NMR behavior in terms of the quantity D_p, the change in the resonance of proton p upon binding ($D_p = \delta_{\mathrm{free}\,p} - \delta_{\mathrm{bound}\,p}$). The descriptive equation is then

$$\delta_{\mathrm{obs}\,pi} = \delta_{\mathrm{free}\,p} - D_p F_i.$$

An NMR titration study in which P host and guest protons are followed in N different samples is fitted by a model with $P + 1$ adjustable parameters (K and the P D's) to minimize the squared error sum SSR.

$$\mathrm{SSR} = \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{(\delta_{\mathrm{calc}\,pi} - \delta_{\mathrm{obs}\,pi})^2}{\sigma_{pi}^2}$$

Emul actually performs *two* such minimizations; in one, the weighting factors σ_{pi}^2 are all equal to 1, and in the other, each is set equal to the expected magnitude of its

corresponding squared error term $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})^2$. This magnitude is estimated from the errors in all the measurements that make up the binding study. The actual formulas for estimating these values are thoroughly expounded in Chapter 2; the general idea is expressed pictorially in Figure 1. It is much easier computationally to minimize SSR when all of the weights are the same, so the unweighted procedure is performed first in order to obtain an initial guess for the minimization of the weighted SSR.

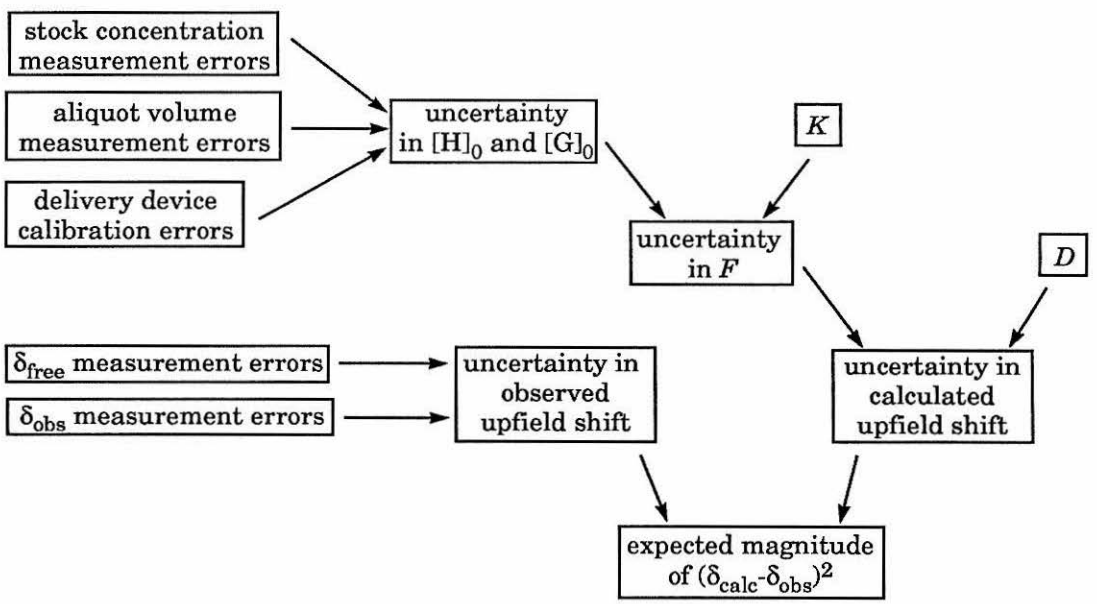


Figure 1. How the magnitudes of the squared residuals $(\delta_{\text{calc}\,pi} - \delta_{\text{obs}\,pi})^2$ are estimated from the different experimental errors in a binding study.

The formula for the NMR behavior is an intrinsically nonlinear function of the parameter K; consequently, it is not possible to directly find the parameters that minimize SSR from the data alone. Instead, the procedure needs to be primed with an initial guess for K, which can then be iteratively improved.

In order to carry out this minimization, Emul needs to know the design of the experiment and the NMR observations resulting from it. These are provided in an input file. It also needs an initial guess for K, which is provided by the user at run-time.

A number of output files are generated by this program. The *text output file* reports the parameters returned by the unweighted and weighted minimization procedures, and also reports the estimated degree of complexation in each of the samples. The optional *data summary file* reports the details of the internal workings of the experiment, and of the error-propagation calculations. A *tabular output file*, in a format readable by graphing programs such as Kaleidagraph and Cricketgraph, reports quantities such as the concentrations of each sample, the observed and calculated upfield shifts of the protons in each sample, and the differences (weighted and unweighted) between the fitted model and the actual experiment. This file is useful for making residuals plots, which are a good qualitative way to examine the model. Another output file, the *simulator input file*, acts as the input file for the remaining programs in the package. This file is in binary format, so it is not intelligible to the user. It contains all of the information from the input file, the values of the final fitted parameters, and a number of intermediate calculations for setting up the minimization procedures. This saves the later programs from having to repeat the same calculations. All the information in this file can be found in text form in the data summary file and the text output file.

B. Lucius.

This program provides a test of the fitted model. It is always possible for some physical processes to occur in the binding study that are not described by the model. For example, one of the species may form micelles, or higher-order intermolecular

complexes may exist in addition to the expected bimolecular complex. While it is never possible to prove that a given model is correct, it *is* possible to show that a model does *not* adequately describe the observed data. The purpose of this program is to find if the model fits the data as well as can be expected, given the assumed magnitudes of the experimental errors. It reports the probability that a hypothetical set of experimental data arising if the fitted model were true would be fit by the model even worse than the actual observed data set. If this probability is high, it means that the fit of the model to the observed data set is unexpectedly good; if it is low, it means that the model explains the observed data unexpectedly poorly.

A poor fit means any of three things: (1) Nothing is wrong; this is just one of those times that random errors have conspired to make the fit especially poor; (2) the assumed error bars are too small, because experimental inaccuracies are in fact far more severe than supposed; or (3) some other process is occurring. There is, unfortunately, no way to distinguish between these three possibilities without carrying out further studies. If it is just a case of uncharacteristically large measurement errors, the problem should not recur in a successive binding study. If the measurement errors tend to be larger than was optimistically supposed, this *may* be reflected in residuals plots that show large random chatter, with no discernable pattern. However, the design of NMR titration binding studies typically is such that several measurements affect many observations; that is, errors in these measurements are effectively *systematic* errors. Such errors may cause noticeable patterns to the residuals, *even if that measurement error is the only deficiency of the model.* See, for example, Figure 1 of Chapter 3. If some other process is occurring, it may leave a characteristic signature; for instance, aggregation processes are most significant at high concentrations.

Lucius uses Monte Carlo (random) sampling to find the probability that a data set arising under the fitted model will give a fit score SSR larger than that from the observed data set. It generates a multitude of simulated data sets by performing Monte Carlo replications of the binding study. The data sets it generates are like those Nature would generate if the fitted model were true; subjecting each of those data sets to Emul's least-squares regression procedure maps out the distribution of SSR scores and fitted parameter values that would arise in Nature if the binding study were repeated many times. The value of SSR from the actual data set ($\widetilde{\text{SSR}}$ in Chapter 3) is compared to the distribution of SSR scores from the Monte Carlo data sets. If a sizeable fraction of these Monte Carlo scores are larger than the experimental score, then the data do not provide cause for rejection of the model. If the experimental score is larger than the preponderance of the Monte Carlo scores, however, then the model and the observed data are incompatible. This is reported by Lucius as the statistic Q, which is the fraction of Monte Carlo SSR scores *exceeding* the experimental score. As a general rule, a value of Q less then 0.001 indicates that the model cannot be accepted.[1]

The Monte Carlo data sets are generated by following, as closely as possible, what Nature does in a real binding study. In a real experiment, the values of quantities such as stock solution concentrations and aliquot volumes are never perfectly known, and measuring devices are subject to both systematic bias and random fluctuations. The true values of the total host and guest concentrations of the samples are therefore different from the measured values. Consequently, the observed spectra will be different than the predicted spectra obtained by applying the true K

(1) This conservatively low value is chosen to allow for the possibility of non-normal error distributions. Lucius treats the experimental errors as normally-distributed random variables; if the errors actually follow a distribution with larger tails, Lucius will underestimate the likely values of SSR.

and D values to the measured concentrations. Adjusting K and D to minimize these differences will remove some, but not all, of this discrepancy. Thus, measurement errors result in a positive SSR. Monte Carlo replications of the binding study, in which the measurement errors are simulated by random numbers drawn from realistic distributions, will reveal the values of SSR that could result from random measurement error alone. If the experimental value of SSR is too large to have come from the simulated SSR distribution, then the assumed measurement errors are not the only factors contributing to its magnitude. The procedure for generating these Monte Carlo data sets is described in Chapter 3.

Lucius also determines the distributions of the parameter estimates that best fit the Monte Carlo data sets. These parameter sets are obtained in passing in order to arrive at the SSR values. These parameter estimate distributions have no direct relation to parameter confidence limits, but they *are* of value in evaluating the experimental design. These distributions are the range of parameter estimates that could be obtained in a binding study like the experimental one if the fitted model were true. If the range is very broad, then the experimental design is not a good one for precisely determining the parameter values. Conversely, if the distribution of parameter estimates is very narrow, then the experimental design allows good parameter estimation regardless of the particular measurement fluctuations encountered.

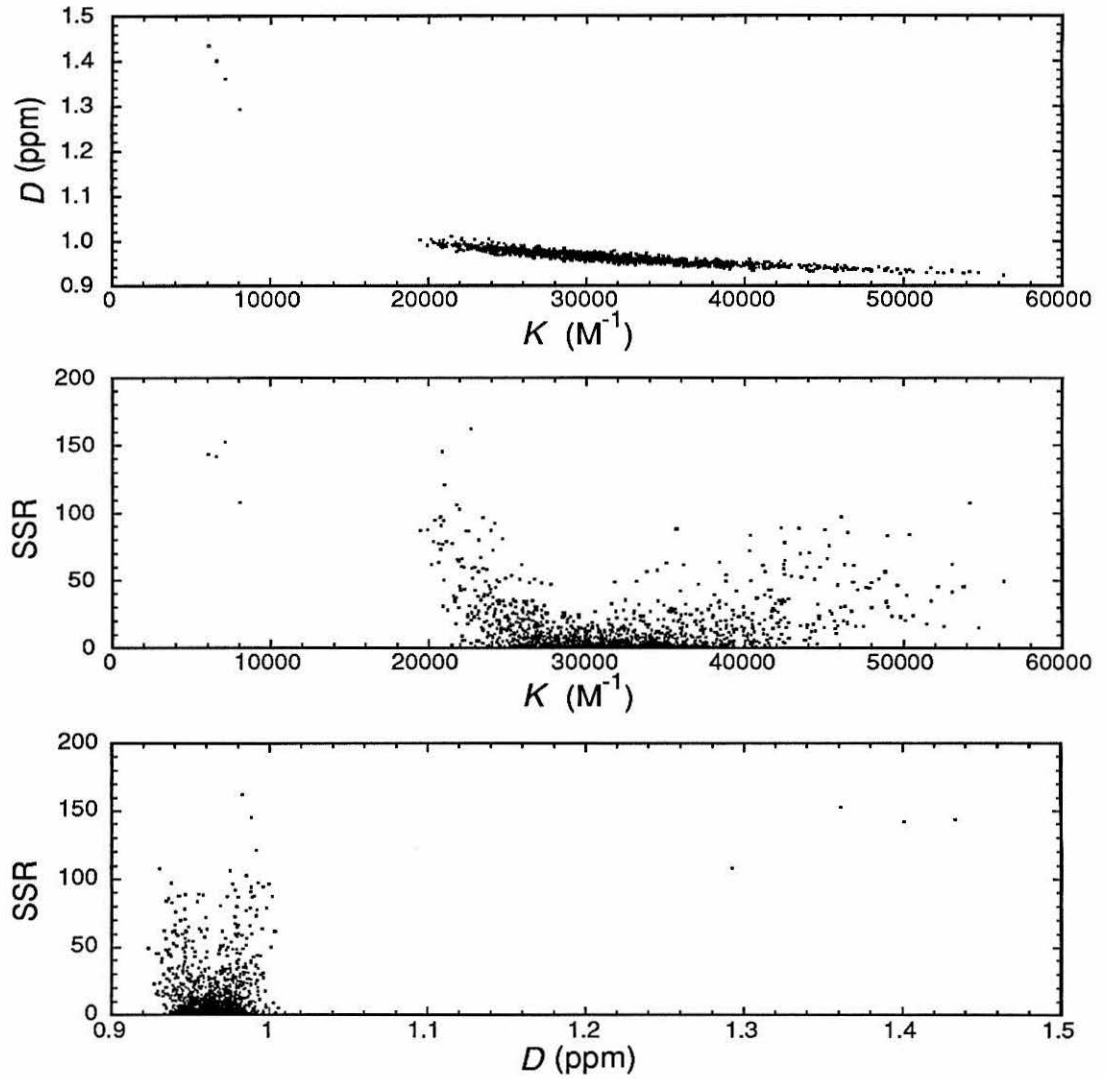


Figure 2. Several plots of the data from a Lucius scatter plot file.

Lucius generates three output files. One is a text file that reports the statistic Q, and summarizes the sampled distributions of SSR and of the parameter estimates. The other two are tabular files in a Kaleidagraph- or Cricketgraph-readable format. These files report the parameter and SSR distributions in two ways. In one file, the *scatter graph file*, the parameter estimates and SSR from fitting the ith Monte Carlo data set are in the ith line of the file. Plotting data from this file allows one

to see how (or if) the values of these quantities are correlated. For example, Figure 2 shows plots of several of these quantities against each other. It is obvious that estimated parameter values are strongly correlated; large estimates of K correspond to small absolute values of D, and vice versa. On the other hand, there is no obvious correlation between parameter estimates and SSR.

The other tabular output file, the *distribution file*, has the values of the individual parameters sorted in columns in ascending order. Another column of this file contains the index values i/R, where i is the row number and R is the total number of rows. Since the parameter estimates are in ascending order, the index i/R is the total fraction of estimates exceeded by the elements in row i. Plotting this index against the parameter values reveals the empirical cumulative distribution function of the parameter estimates. The cumulative distribution function F(x) for any random variable y is the probability that a random outcome of y will be less than x; all cumulative distribution functions are monotonically increasing functions of x that range from 0 to 1. The *empirical* cumulative distribution function is an estimate of this function compiled from a set of random observations of y. If the R observations of y are sorted in ascending order so that $y_i \leq y_{i+1}$ for $i = 1, 2, \ldots, R$, the empirical cumulative distribution function (c.d.f.) is defined as

$$F(x) = \begin{cases} 0, & \text{if } x < y_1; \\ i/R, & \text{if } y_i \leq x < y_{i+1}; \\ 1, & \text{if } x \geq y_R. \end{cases}$$

This exactly equals the true c.d.f. in the limit of infinite sample size. The distribution file also contains an estimate of the probability *density* function of each of the parameter estimates. The probability density is the derivative of the c.d.f., and is useful for visualizing the variable's distribution. A large value of the density indicates a high likelihood of observing the random variable in that vicinity. The

density functions estimated by Lucius are obtained by crude differentiation of the empirical c.d.f.'s, and tend to be quite noisy. Nonetheless, they do help give a feeling of the behavior of the parameter estimates. Figure 3 shows the distribution and density plots from one simulation study.

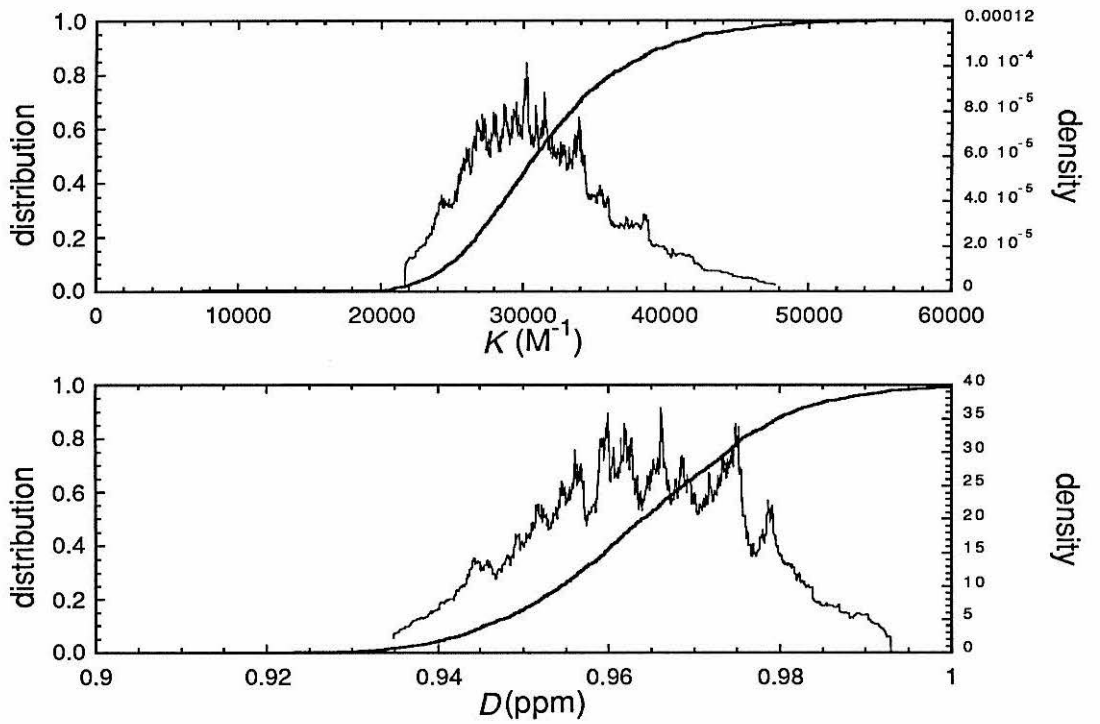


Figure 3. Parameter distribution and density plots from a Lucius parameter distribution file.

C. Portia.

The operation of Portia is very similar to the operation of Lucius. Instead of providing a test of the model, however, Portia provides confidence limits for the parameters of a model that has already been accepted. It does this by answering the question: given the distribution of measurement errors that affect the experiment, what range of parameter values is consistent with the experimental observations? This is investigated, as by Lucius, by Monte Carlo simulation of the experimental

measurement errors. Unlike Lucius, however, Portia does not create sets of hypothetical observations by applying the model to the randomly-generated concentrations. Instead, it finds the best parameter set for the actual observed data consistent with the model and the generated concentrations. The different approaches of Lucius and Portia are compared to each other and to the course of an actual binding study in Figures 4, 5, and 6.

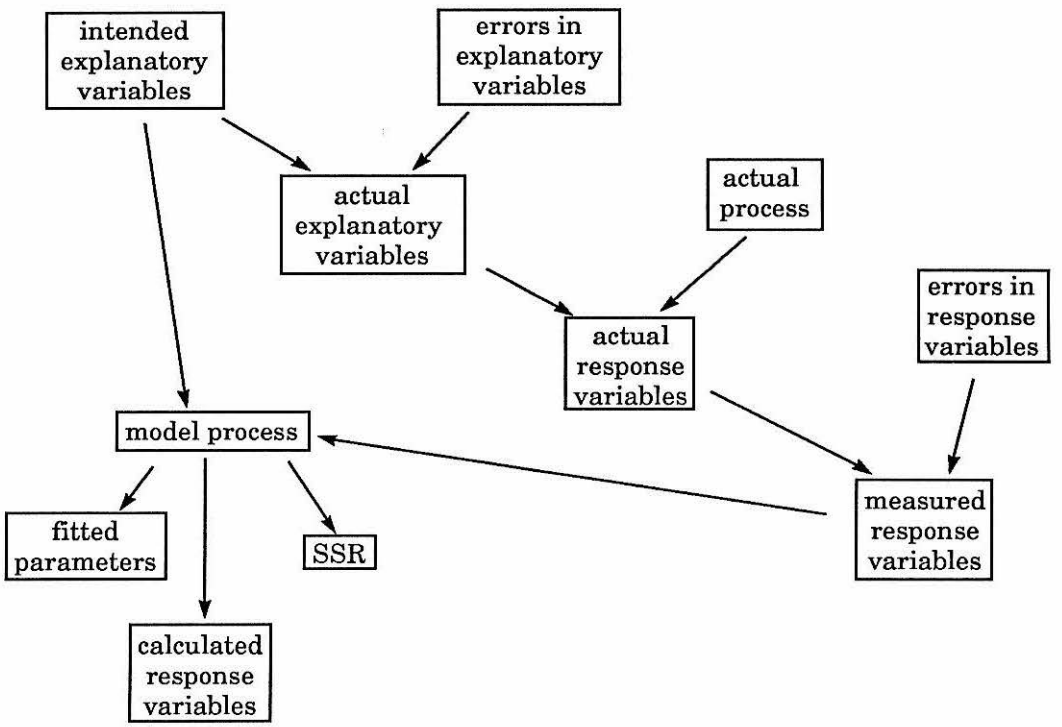


Figure 4. What occurs in an actual binding experiment. Experimental errors prevent the intended values of explanatory variables from being attained, and the actual values are unknown to the experimenter. The actual process uses the actual explanatory values to give the actual values of the response variables. These response variables are in their turn measured inaccurately. The only quantities available to the experimenter are the intended explanatory variables and the measured response variables. The model process, which is not necessarily the same as the actual process, is parametrically adjusted to reconcile these explanatory and response variables.

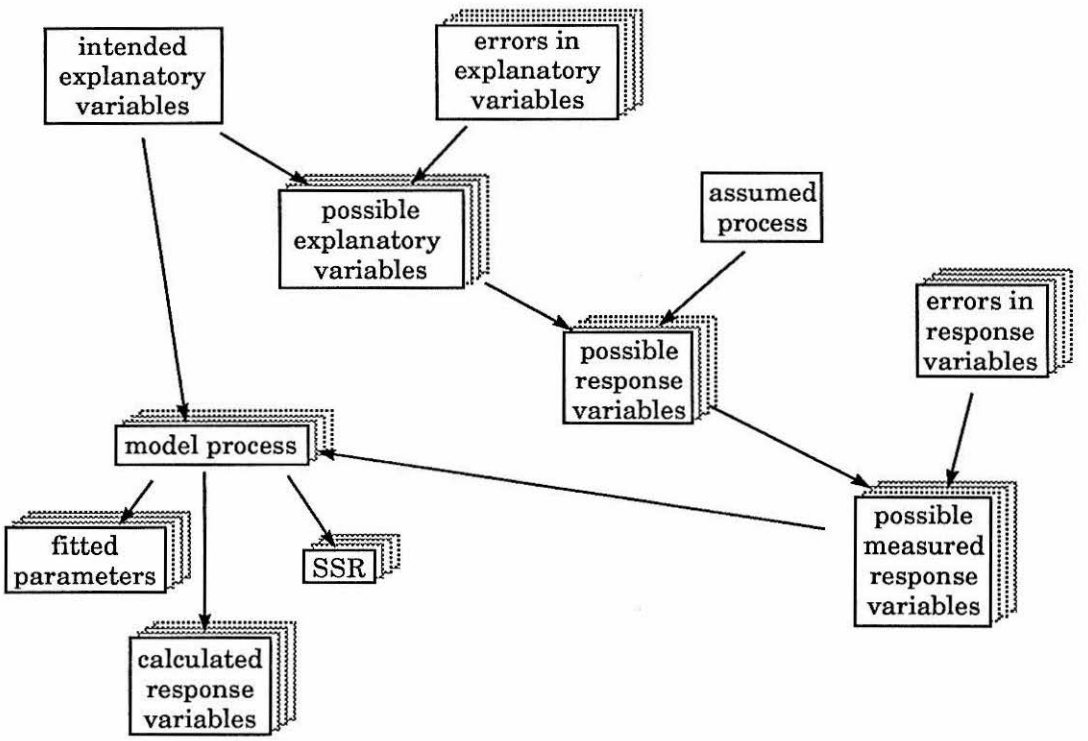


Figure 5. Lucius simulation of a binding study. The assumed process and intended explanatory variables are the same from replication to replication, but different sets of simulated measurement errors lead to different "measured" response variables, and ultimately to different SSR scores and fitted parameter sets. These are the SSR scores and fitted parameter sets that would arise in a series of repetitions of the binding study if the assumed model were true.

The output files of Portia correspond exactly to the output files of Lucius. A text output file summarizes the distributions of SSR and the fitted parameters, and also reports the statistic Q. The fitted parameter distributions may be thought of as parameter confidence limits: since the fitted parameter values are those consistent with the actual experiment, a region of parameter space into which these fitted values tend not to fall is unlikely to hold the true parameter value. I am unable to identify a meaningful interpretation of the distribution of SSR values and the Q statistic from Portia. These quantities are nonetheless reported, awaiting the day that such an interpretation is developed.

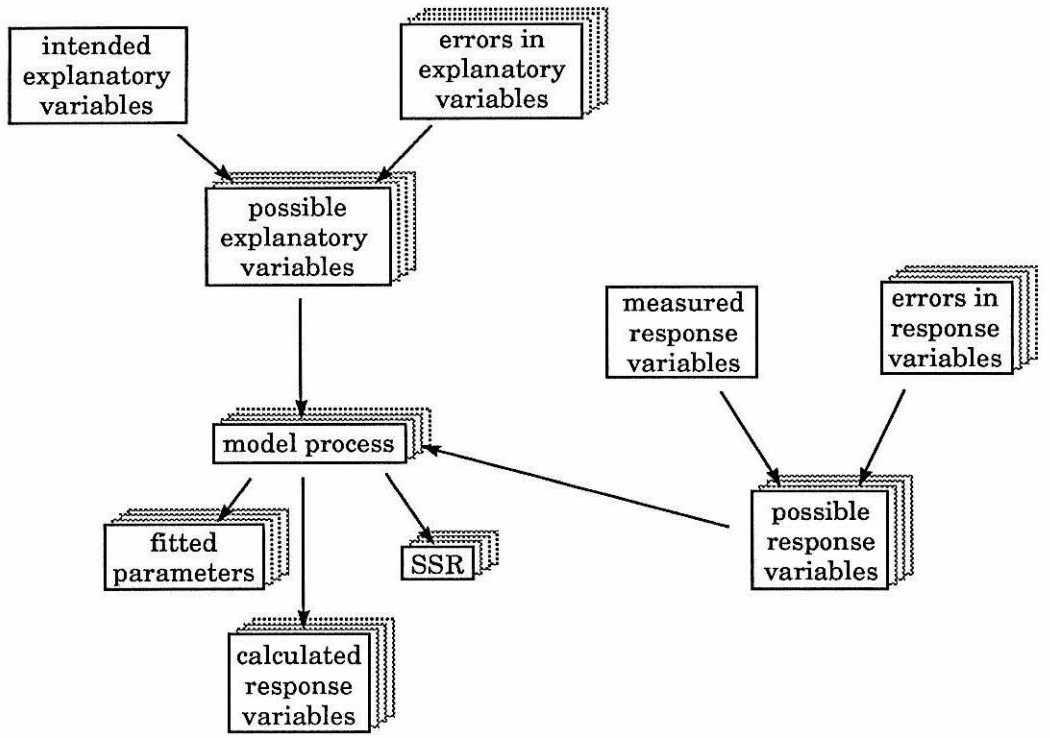


Figure 6. Portia simulations of a binding study. The sampled "possible" explanatory and response variables are values that could have been the true ones in the actual experiment. The model is parametrically adjusted to reconcile these explanatory and response variables. As with Lucius, different sets of simulated measurement errors lead to different SSR scores and fitted parameter sets. These are the parameter sets most consistent with the observed data and the understanding of the measurement errors.

Portia also creates parameter scatter and distribution graph files, which follow exactly the same format as those from Lucius. The empirical cumulative distribution function of the parameter estimates from Portia is $\widehat{C}^{\dagger}$, the Monte Carlo estimate of the confidence function $C^{\dagger}$, which is defined in Chapter 3. Any arbitrary confidence region can be determined from these empirical functions by finding the parameter values corresponding to widely-separated values of i/R. This is illustrated by a plot of such an empirical $\widehat{C}^{\dagger}$ function in Figure 7.

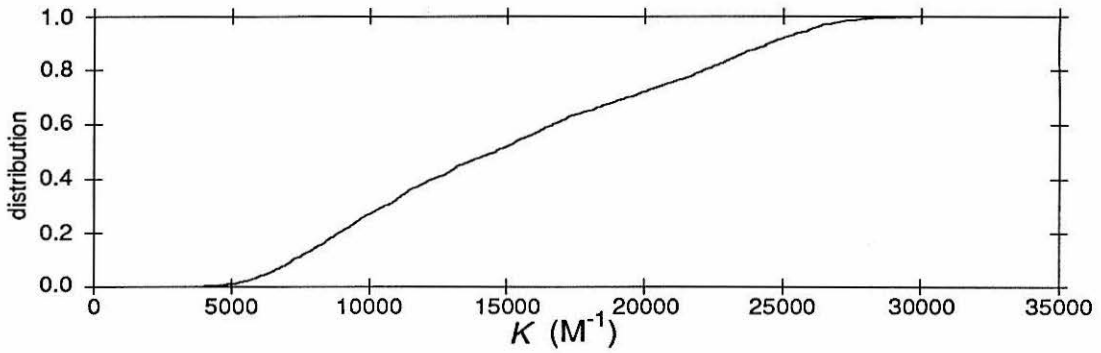


Figure 7. Portia empirical cumulative distribution function for K.

D. Brutus

Brutus, like Portia, finds confidence limits for the fitted parameters, but it does so in a different, slower way. Instead of exploring alternative interpretations of the experimental data set, it explores the range of parameter values to find the ones so extreme that they are inconsistent with any data set that can be fit by the experimental parameters. In other words, it seeks to answer the question: what possible parameter value θ_{H} is so large that the probability of obtaining a best-fit value no larger than $\tilde{\theta}$ is only α if θ_{H} is the true parameter value? Likewise, what possible parameter value θ_{L} is so small that the probability of obtaining a best-fit value no *smaller* than $\tilde{\theta}$ is only α if θ_{L} is the true parameter value? Since the actual experiment *was* best fit by parameter value $\tilde{\theta}$, parameter values that would render obtaining this value extremely improbable are unlikely to have been responsible for the observed data set. The interval of parameter space bounded below by θ_{L} and above by θ_{H} is then the $(1-2\alpha)\times 100\%$ central confidence region for that parameter.

Brutus determines these limiting values by performing a set of simulations, each qualitatively similar to an entire Lucius run. Each simulation pretends that

a different possible parameter value is the true value. The fitted parameter distribution resulting from any such simulation reveals the range of parameter estimates that could be obtained in a binding study like the experimental one if the supposed parameter value were true. The property of interest for each distribution is the probability of obtaining fitted parameter values that are greater than the experimental value. The estimate of this probability, $\widehat{C}^{\star}$, is the fraction of fitted parameter values from the simulation that exceed the experimental best-fit value. If this fraction is very close to zero or very close to one, then the bulk of the fitted values are nowhere near the experimental value. In other words, the observed data set, which *did* give the experimental value, would not have occurred if the supposed model were true. Brutus tries to find the parameter values for which this probability is α and $(1-\alpha)$; these values define the $(1-2\alpha)\times 100\%$ central confidence interval for the parameter estimate.

These values are found by trial and error. Every Monte Carlo study at a different parameter value helps to map out the form of the confidence function $C^{\star}$, as defined in Chapter 3. The empirical measurements of this function at the various sampled parameter values are independent random variables, so estimation of the shape of the $C^{\star}$ from these samples is no mean feat. Every estimate of a single-point value of $C^{\star}$ has associated with it an uncertainty defined by its value and by the number of replications performed in the study. The program aims to sample points near enough to the limiting parameter values to get good estimates of those values, yet far enough apart to avoid downturns in the empirical $\widehat{C}^{\star}$ function. Despite these precautions, downturns *do* occur; they are eliminated by averaging adjacent points until the downturn disappears. This is illustrated in Figure 8.

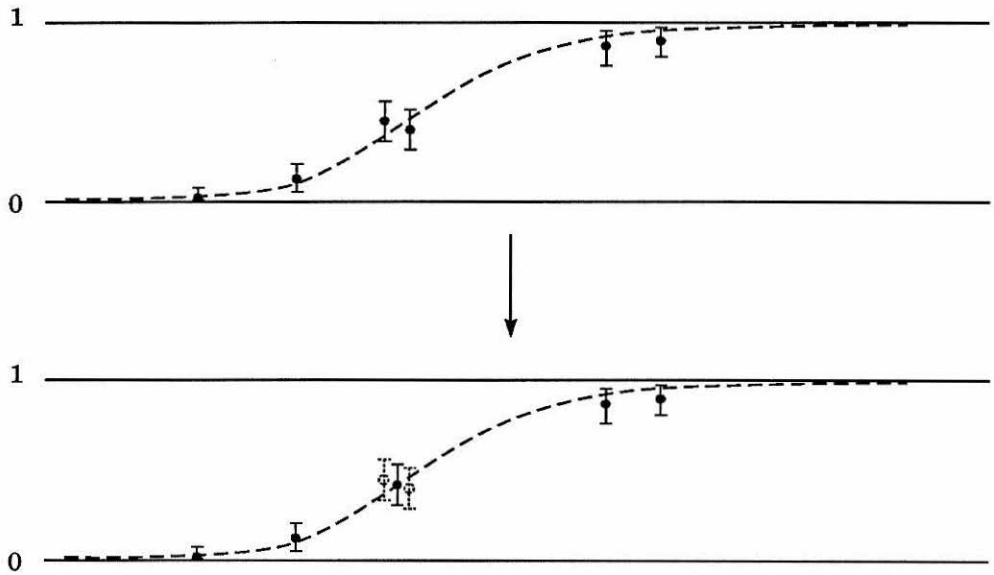


Figure 8. Averaging of downturns in Brutus in order to obtain a set of monotonically increasing "sampled" points for the confidence function $C^{\star}$.

A limiting parameter value is considered "found" if points with sufficiently close $\widehat{C}^{\star}$ values have been sampled that bracket the desired limit; the exact criteria for this bracketing are given in Chapter 3. This is illustrated in Figure 9. Values of the $C^{\star}$ function at parameter values not sampled are estimated by interpolating between sampled points.[2–4] The end result of this process is best described by referring to Figure 10. In this graph, an empirical confidence function for K is plotted against the parameter values. The filled circles "•" are the sampled points; the error bars drawn on them are the 99% confidence limits for the actual value of the $C^{\star}$ function at those points. The values marked "×" are the confidence limits specifically sought

(2) Staniswalis, J. G; Cooper, V. "Kernel estimates of dose response." *Biometrics* **1988**, *44*, 1103–1119.

(3) Priestly, M. B.; Chao, M. T. "Non-parametric function fitting," *J. Roy. Statist. Soc.* **1972**, *34*, 385–392.

(4) Copas, J. B. "Plotting p against x," *Appl. Statist.* **1983**, *32*, 25–31.

by the procedure. Note that all of these points are closely bracketed both above and below by sampled points.

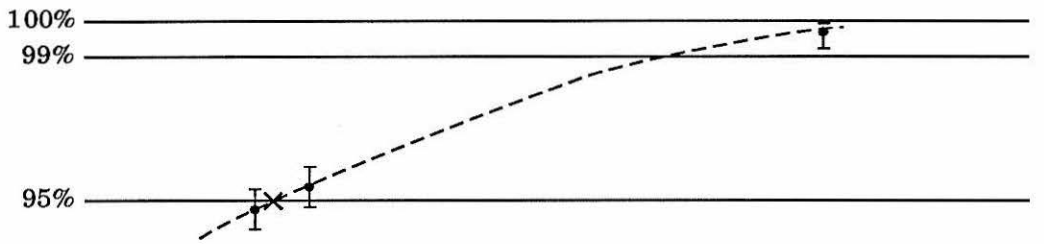


Figure 9. In the process of Brutus's hunt for parameter confidence limits. The 95% confidence value has been closely bracketed by sampled points (•), so the limiting parameter value for this end of the confidence region has been assigned (×). The 99% confidence value is not yet closely bracketed, so the limiting parameter value can not be confidently assigned. The dashed line represents the interpolated confidence function.

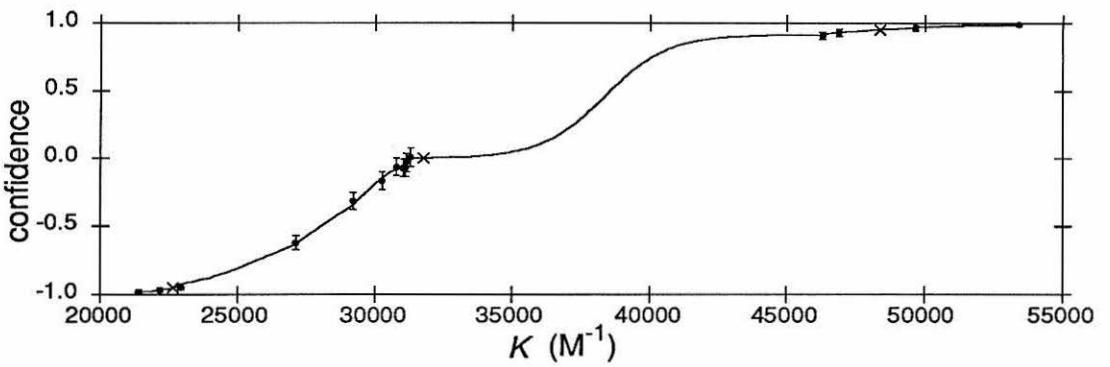


Figure 10. Empirical confidence function C^* for K found by Brutus. The sought limits were 0 and 95%. Note the odd shape of the interpolated function between 32000 and 46000, and also that the 0% limit estimate is not between its bracketing points. These are failings of the interpolated function. Note also that the y-axis ranges from -1 to $+1$ instead of from 0 to 1. This is a simple transformation of the confidence function; here, the 95% confidence region, for example, is in the interval between -0.95 and $+0.95$.

Brutus produces four types of output file. The first, naturally, is a text output file, which summarizes the results of its search for the parameter confidence limits. The first section of this file is identical to the contents of a Lucius text output file.

This is because the first exploration of the $C^\star$ function performed by Brutus is at the best-fit parameter values; this study is identical to a Lucius run. Consequently, a run of Brutus effectively re-runs Lucius. Two of the output files created by Brutus are simply the parameter distribution and parameter scatter graph files from this initial Lucius-like simulation. The remaining output files are also tabular format files for graphing. These files, the *confidence function files*, contain the estimate of the $C^\star$ function for each parameter, as well as all the sampled and averaged points explored in the search for the limiting parameter values. It also contains lower and upper $(1 - \alpha_2) \times 100\%$ confidence limits for the sampled points.[5] A Kaleidagraph macro, "limit bounds to error bars," calculates error bars from these limits so that they can be displayed in a graph. Figure 10 is a graph from such an output file, with error bars.

II. Operation

All of these programs require input directly from the user and from previously-created input files. For the most part, the user input consists of identifying the input and output files. The programs all ask the user for the names of these files; in all cases, a default file name is included in the question. The default name is the name that will be used if the user simply types a carriage return instead of a full file name. Every default name includes an extension; if the user enters a file name without an extension, the programs automatically append the default extension to the specified file name. This reduces the typing required; on the other hand, it is impossible to make the programs read from or write to files that do not have extensions in their names. If ever the user specifies an input file that does not exist, the programs will

(5) The quantity α_2 is explained in the section discussing the preferences file (II.A.2.(a)).

not crash, but will instead report that the file was not found. The user then has the choice to look for a file of a different name, or to cancel the program run.

A. Emul.

1. User interface. Emul is started by simply typing "`emul¬`."[6] The user interface begins by asking for the name of the preferences file. The default name of this file is always "nmr.prf." A preferences file is read by each of the programs in the package. Preference files contain instructions for the operation of all of the programs; a user may customize the package by writing a special preferences file.

The next name Emul asks for is that of the text input file. This is the file containing the description of the experimental design and all of the experimental observations. This request is followed immediately by a request for the name of the output file. The default output file name is taken from the chosen input file name, and is given the extension specified in the preferences file. The user may either accept the default file name by typing a carriage return (¬), or may enter a different name. The user is asked for the name of the text output file only; the names of the other files written by Emul are automatically based on this file name. All of the output files (text, residuals, simulator input, and the optional data summary file) will have the same name, with extensions as specified by the preferences file. *The user is asked only for the name of the text output file at run-time.* The programs do not check if files of the given names already exist; consequently, if they do exist, they will be overwritten without notice.

Emul then asks for the name of the error bars file. This file contains general information about the sizes of the experimental errors. All of this information may

(6) Hereafter, the symbol "¬" will be used to denote a carriage return.

be superceded by specific information in the input file, but an error bars file must nonetheless be specified. The next information requested is a header for the text output file. This is optional; it is merely an opportunity for the user to type a line describing the experiment at the top of the text output file. Emul asks next if it should create a data sumary file. If the answer is yes, then it writes a file detailing its internal representation of the experiment, including all assumed measurement uncertainties and calculations for propagating them.

The program is finally ready for an initial guess of the binding constant. This is the kick it needs to begin adjusting the model parameters to fit the experimental data. No more information from the user is required for the rest of the run. When the run is finished, the program will report the names of all the files it has generated.

2. Input files.

(a) The preferences file. This file contains information, some of it obscure, for the operation of all of the programs in the package. Not every entry is meaningful to every program. I will attempt to explain this file by presenting a listing of a sample preferences file and describing every entry in turn. The discussion will occasionally be technical and incomprehensible; some of these values control inner workings of the program that the user never sees directly. When the entry is a numerical value, the name in the source code of the variable to which the value is assigned is given in italics; in order to fully understand the function of some of these variables, consulting the source code may be necessary. For the most part, however, there should be no need to delve into the inner workings of the procedures. The values presented in this example file should permit satisfactory performance of the programs under most conditions.

`5300` This quantity, *submax*, is the number of replications to perform in a Monte Carlo simulation study of an NMR titration experiment. This number has no meaning for Emul, but it is used by Lucius, Portia, and Brutus. The particular number "5300" is the number of simulations required for Brutus to be able to find 99% parameter confidence limits. *Submax* cannot exceed 100,000.

`51` This quantity, *spread*, is the smoothing factor for determining empirical probability densities from the empirical cumulative distribution functions of the parameter estimates from Lucius, Portia, and the first run of Brutus. These densities are calculated by approximating the derivative of the c.d.f.; the density at the ith parameter value is approximated as the slope of the line drawn between the $(i-l)$th (parameter, c.d.f.) point and the $(i+l)$th (parameter, c.d.f.) point. $Spread = 2l + 1$. It must therefore be a positive odd number.

`nmr.in` This is the default name and extension for the Emul input file.

`nmr.bri` This is the default name for the binary-format simulator input file. The actual file created by Emul will have a file name dictated by the name of the Emul text output file, but the extension will always be the extension given here.

`nmr.lim` This is the default name and extension for the confidence limits file read by Lucius, Portia, and Brutus.

`eml` This is the default extension of the Emul text output file.

`bru` This is the default extension of the Brutus text output file.

`luc` This is the default extension of the Lucius text output file.

`por` This is the default extension of the Portia text output file.

`res` This is the extension of the Emul tabular residuals file.

`lpm` This is the extension of the parameter scatter file from Lucius or Brutus.

`ldn` This is the extension of the parameter distribution file from Lucius or Brutus.

`ppm` This is the extension of the parameter scatter file from Portia.

`pdn` This is the extension of the parameter distribution file from Portia.

`nmr.err` This is the default name and extension of the error bars file read by Emul.

`dsm` This is the extension for the data summary file written by Emul.

`0.6` This quantity, *intwidth*, affects the interpolated confidence function in Brutus. It is a scaling factor for the formula that assigns weights to each point (see the discussion for *share*). This variable is sort of a smoothing factor; if it is small, the value of the interpolated function is determined almost entirely from the closest point; if it is large, the contributions from more distant points become more significant.

`0.1` This quantity, *share*, also is a parameter for the spline interpolation function used by Brutus. The y-value of the interpolated $C^\star$ function is determined at an arbitrary x-value X by a weighted average of all the y-values in the set. Points whose x-values are close to X are weighted more heavily than those that are farther away. The actual weighting factor for each point is proportional to $e^{-d_i^2}$, where d_i is the distance from

x_i to X. It is not a simple arithmetic difference $(X - x_i)$, however. The set of points to be interpolated is unevenly-spaced, and the spline would resemble a step-funcion between points that are far apart. (Figure 10 shows an example of just such behavior, tamed greatly by these parameters.) To correct for this somewhat, an alternative distance measure is also calculated; this is proportional to the number of points between X and point i. Let us call this measure w, and the measure proportional to the arithmetic difference let us call v. The actual weighting factor for point i is $(share) \times v^2 + (1 - share) \times w^2$. Acceptable values are $0 \leq share \leq 1$.

`0.01` This quantity, α_1, is a significance cutoff specifying how close to the desired confidence values (say, 0.05 and 0.95 if the 90% confidence region is sought) the empirical $\widehat{C}^\star$ estimates at the sampled parameter values bracketing the eventual confidence limit estimate must fall. For example, if 5000 replications are performed in a Monte Carlo study, the 99% (that is, $1 - \alpha_1$) confidence limits for the actual probability of success that returns 25 hits (0.500%) are 0.300% and 0.831%. Thus, the estimate of the parameter value that gives a 0.5% probability of success must be bracketed on the low side by a sampled value that experienced between 15 and 25 hits, and on the high side by a sampled value that experienced between 25 and 41 hits. Allowed values are $0 < \alpha_1 < 1$.

`0.01` This quantity, α_2, is the significance cutoff for Brutus to report confidence limits for the parameter confidence limits in its output files. It is strictly for the edification of the user; this number does not affect the program run in any way. Allowed values are the same as for α_1.

`2000` This quantity, *maxit*, is the maximum number of iterations the Levenberg-Marquardt SSR minimization procedures will perform without converging before giving up.

`0.0001` This quantity, *concrit*, is the convergence criterion for the Levenberg-Marquardt minimization procedures. If SSR is improved by less than this fractional amount twice in a row, the routine considers itself converged. Allowed values are $0 < concrit < 1$.

`1.0e100` This quantity, *maxlam*, is the maximum allowable value for λ, a parameter in the Levenberg-Marquardt minimization procedures whose size corresponds to the inability of the quadratic approximation to improve SSR. If λ gets this big, the procedure gives up without converging.

`0.001` This quantity, *upcrit*, is the tolerance for a slightly worse SSR in successive L.-M. steps. If the proportional increase in SSR is less than this value $\times$ the convergence criterion, the procedure does not count that step as an actual worsening. This exists to counteract an observed tendency of these procedures to fail to converge in perfectly acceptable regions of parameter space. Allowed values: $0 \leq upcrit < 1$.

`10` This quantity, *maxconsecups*, is the number of consecutive failures to improve SSR that the L.-M. procedures will tolerate before considering themselves converged. If a parameter guess happens to be very close to the global minimum of the SSR surface, it is possible that the procedure will be unable to ever get any closer. This criterion allows it to consider itself converged in such cases.

10 This quantity, *laminc*, is the factor by which λ is multiplied if SSR increases in a L.-M. step.

0.1 This quantity, *lamdec*, is the factor by which λ is multiplied if SSR *decreases* in a L.-M. step.

44 This is the ASCII code for the column-delimiting character in the tabular data files. 44 is the code for a comma; 9 is the code for a horizontal tab. Since horizontal tabs are transferred to the Macintosh from the IRIS by Versaterm-Pro as spaces, it is necessary to specify some other character. Commas are a good choice, because Kaleidagraph can be told to recognize them as column markers.

15 When Brutus tries to find the parameter value X that gives a probability $C^{\star}(X)$ equal to some desired confidence value, it performs a simulation at the parameter value that is the best current guess of X. This is determined from the interpolated confidence function: the x-value of the interpolated function where y is equal to the desired confidence value is this best guess. Unfortunately, the nonparametric spline is set up so that it is easy to determine the value of y at a given x, but difficult to determine x from a given y. This is the problem of finding the root of an equation. Numerical methods for finding roots involve either iterative narrowing of a region known to contain the root, or iterative improvement of an estimate of the root. The first method employed in Brutus is *bisection*, a member of the former class; after a number of bisection steps have been performed, the method of *false position*, a member of

the second class, is employed.[7] The entry on this line is *bisitmax*, the maximum number of bisection steps to perform.

`30` This is *flspitmax*, the maximum number of false-position iterations.

`1.0e-5` This is the precision required for convergence of the root-finding procedures.

`0.8` This quantity, *squeeze*, affects the root-finding procedures in Brutus by helping to find x-values of the interpolated function that bracket the desired values. If the current guess for one of the brackets (high or low) is on the wrong side of the root, *squeeze* tells the program how far to move it away. If $squeeze = 1$, the guess for the bracketing value will not move at all; if $squeeze = 0$, the guess will be moved all the way out to the most distant sampled point. Intermediate values cause intermediate displacements. Allowable values are $0 \leq squeeze < 1$.

`0.8` This quantity, *limweight*, tells Brutus what y (that is, $C^\star$) value to aim for when it tries to closely bracket a desired limit. The extreme values that the brackets *must* fall within to have "found" the limit are defined by α_1; the *limweight* tells how far inside these extreme values Brutus should attempt to sample. If $limweight = 1$, then Brutus will always shoot for the most extreme value; if $limweight = 0.5$, then it will try to sample an x that will return a y exactly halfway betwen the extreme value and the desired limit. Allowable values are $0.5 \leq limweight \leq 1$.

`0.85` When Brutus tries to sample the parameter values that will bracket some desired confidence limit, it estimates the necessary parameter values from

(7) Press, W. H.; Flannery, D. P.; Teukolsky, S. A.; Vetterling, W.T *Numerical Recipes: the Art of Scientific Computing*; Cambridge University: New York, 1986; pp 243–251.

the current interpolated $C^\star$ function. If the sampled points defining this function are widely or unevenly spaced, the function is often unreliable. Such problems are especially acute when the y-value of one of the bracketing points is much closer to the desired confidence value than is the y-value of the other bracketing point. The empirical solution to this problem is to narrow down the interval. The entry in this line, *maxlopside*, defines how lopsided the bracketing must be to cause Brutus to disregard the interpolated function. In this example, $maxlopside = 0.85$, so corrective action will be called for if the difference between the desired confidence value and the y-value of the most distant point is more than 85% of the total height of the interval. If this happens, the next sampled parameter (x) value will not be determined from the interpolated $C^\star$ function. Instead, it will be the parameter value that is 85% across the interval between the x-values of the bracketing points. This is illustrated in Figure 11. If *maxlopside* is set to 0.5, Brutus will never choose parameter values from the interpolated function, but instead will simply bisect the interval enclosing the desired confidence value. Allowed values are $0.5 \leq maxlopside \leq 1$.

200 This quantity, *morenmr*, is the number of points of the interpolated confidence function that Brutus includes in its confidence function files.

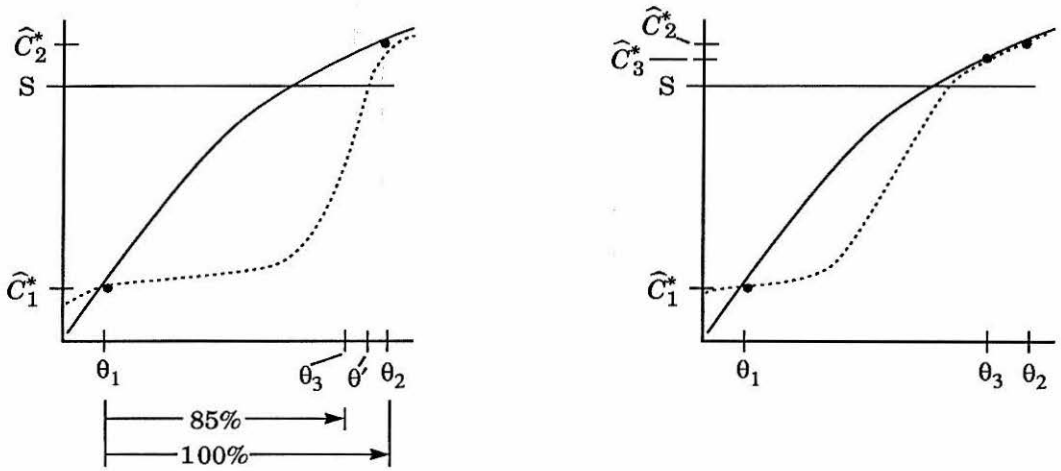


Figure 11. In some cases, the interpolated confidence function $\widehat{C}^\star$ (dashed line) has undesirable behavior between sampled points. In this illustration, the desired confidence limit S is between the sampled $\widehat{C}^\star$ values obtained by simulation about the parameter values θ_1 and θ_2, but it is much closer to $\widehat{C}_2^\star$ than to $\widehat{C}_1^\star$. Instead of using the interpolated parameter value θ', the procedure performs its next simulation about the parameter value θ_3, which is 85% across the interval between the sampled points. This gives a better subsequent interpolated function in the region of interest.

(b) Emul input and error bars files. Before describing the format of these files, it is appropriate to explain how these programs represent a binding study. This will help to justify the input file format.

(i) Internal representation of the binding study. In order for the fitting and simulation programs to properly model a binding titration experiment, they must be able to internally represent the study. Since these experiments may follow a wide range of procedures, the internal representation must be very adaptable. Although it is easy for humans to understand different procedures, computers, which are not as flexible, must follow a specific format. The challenge to the programmer, then, is to develop a format general enough to accomodate any feasible experiment. I hope that the current format fulfils these criteria.

All experiments are assumed to involve NMR observations of at least two samples containing both host and guest. Any or all protons of the host or guest may

be observed in any of the samples. These samples are created by placing solutions containing host and guest into NMR sample tubes. Samples may be made either by adding solution to an existing sample, or by placing solution into an empty tube. Every delivery of a solution aliquot to a sample tube defines a new sample.

There are two types of solution that may be added in such an aliquot. The first type is a stock solution. An experiment may use any number of stock solutions, any of which may contain any concentration of host, guest, both, or neither. The concentration of a stock solution is considered to be a fundamental random variable; its estimated value depends on no other measurements. The second type of solution is a sample solution, which is any solution made from other solutions. The concentration of a sample solution may depend on the volumes and concentrations of all solutions from which it is made. These two types of solution are distinct. Stock solutions cannot be observed, but sample solutions can. The conceptual distinction between stock and sample solutions is so complete that if only a single stock solution is placed in a sample tube, the solution inside the tube is considered a new sample solution.

When creating samples, however, both stock and sample solutions may be added to a sample tube with impunity. The only constraint on adding sample solutions is that they must be created before they are added. This means that the first sample solution must be composed of some quantity of a *stock* solution added to the NMR sample tube. All sample solutions are, fundamentally, made up of nothing but stock solutions, even though they may contain complex mixtures of mixtures.

The actual observations of the NMR spectra complete the experiment. These data are grouped by proton. Associated with each proton is its uncomplexed chemical shift δ_{free}, and a list of chemical shifts δ_{obs} observed in samples containing both

host and guest. It is not necessary for a single proton to be followed in all of the samples, nor is it necessary for a given sample to have any observations associated with it. It *is* necessary, however, for observations to be made only on samples that contain both host and guest. The measurement of the free chemical shift is considered separately from the other NMR observations, and would require illegal mathematical operations if it were not.

All binding studies are represented by these programs in the same way. Associated with each experiment is a list of stock solutions and a list of delivery devices. Samples are in a list of lists. One list contains the different sample tubes used; the sample tubes are lists in their own right of the solutions created in them. If a binding study is conducted so that all aliquots are combined together in a single sample tube, then the list of sample tubes is but one item long. The list of observable samples comprises only the sample solutions that contain both host and guest. The proton chemical shift information is also a list of lists. One list contains all of the observed protons, and each proton has a number of lists, corresponding to its observations, associated with it.

Every measurement has at least two stored quantities associated with it: its value and its uncertainty. The value is self-explanatory; the uncertainty is represented by the expected variance of the measurement. The measurement variances are employed in calculating the weighting factors, as described in Chapter 2. To speed the calculation of the weighting factors, the computer also stores the quantities $\partial[\mathrm{H}]_0/\partial x_j$ and $\partial[\mathrm{G}]_0/\partial x_j$, the derivatives with respect to x_j of the host and guest concentrations, of every sample affected by measurement x_j.

In order for these programs to properly internally represent an experimental data set, the experimental design must be specified to them. The design includes the

general recipe for creating the samples, the values of all measurements performed, and the uncertainties of all measured values.

Describing the recipe for creating samples is straightforward. All delivery devices and stock solutions used in a study are identified and specified, and then the apportionment of solutions into sample tubes is explained. Every solution aliquot is specified by its volume, the device used to deliver it, and the solution being delivered. Successive aliquots are assumed to all be added together unless a new sample tube is indicated.

Proton chemical shift data is reported for each proton in turn. Each list begins with the proton's free chemical shift, followed by its observed chemical shifts in the sample spectra. The sample from which the observation was taken is reported with each observation, so there is no difficulty in not reporting an observation of a given sample.

There are two complementary ways to report experimental uncertainties. The expected error in any specific measurement may be entered after the measurement in the data file; additionally, expected magnitudes of general types of measurement errors may be given in a separate error bars file. An error bars file contains expected uncertainties in NMR measurements, in host and guest concentrations of stock solutions, and in the accuracy and precision of three delivery devices. This is the most convenient way to report measurement uncertainties that are the same for a series of measurements. Any anomalous uncertainties, as might exist, for example, if the signal for a given proton is especially broad for one sample, may be reported along with the measurement in the data file. Uncertainties reported in the data file automatically override the "default" values in the error bars file.

(ii) The error bars file. This file is very short. The first line contains a single entry: this is the uncertainty in NMR shift assignments. A good default value for this quantity is half the spacing between points in the time-domain NMR spectrum. This value should be in the same units in which the chemical shifts are reported in the input file (Hz or ppm). The second line of the file contains two entries. These are the default uncertainties in the determination of host and guest concentrations of the stock solutions. These are reported as proportions: for example, `0.05` means that the uncertainty is 5% of the total concentration. The final three lines of this file refer to three delivery devices. These lines have two entries each: the first is the likely calibration error of the device, and the second is the device precision. The calibration error is a proportion, and the precision is an absolute volume, in liters.

All of these uncertainties are reported as standard deviations. About 68% of the occurrences of a random variable drawn from a normal distribution fall within one standard deviation on both sides of the mean, and 95% fall within two standard deviations. This should give a good idea of the magnitudes to use for these estimates of measurement error.

(iii) The Emul input file. Delivery Devices. The input file begins with a listing of the devices used to deliver volumes of solutions to the sample tube, be they pipets, syringes, graduated cylinders, balances, or whatever. With any such device, there are two measurement uncertainties to consider: accuracy and precision. Accuracy is considered here to be a calibration problem: the device may consistently read or deliver too high or too low. Precision is a matter of reproducibility: it is the amount that two seemingly identical measurements will, in reality, differ.

First choose a name (a string of characters without spaces) to identify each device. No two devices may be given the same name, but device names bear no

relation to any other type of name. Begin the line with the name of the device. If you wish, you may then include accuracy information, as a fractional value, and precision information, as a total volume. For example, entering "`0.008`" for the accuracy means that the presumed standard deviation of the calibration error is 0.8%, and entering "`1e-6`" in the precision slot means that the presumed standard deviation of a series of deliveries that are all nominally the same is 1 μl. If accuracy and precision information are not included on a line, default information will be taken from the error bars file as follows: for the first device, from the third line of the error bars file; for the second device, from the fourth line of the error bars file, for the third device, from the fifth line of the error bars file, and for the fourth and later devices, from the third line of the error bars file.

Stock solutions. A blank line signals that the information for all of the delivery devices has been entered. For stock solutions, as with everything else in this input file, host comes before guest. Choose a name to identify each pipet; start a new line with this name. Then report the stock solution's host concentration, and, if desired, the presumed proportional uncertainty in the measurement of that concentration. Thus, if `0.05` is entered for this uncertainty, the programs will assume that the standard deviation in the measurement is 5% of the measured concentration. If an uncertainty is not reported, it will be taken from the first entry on the second line of the error bars file.

On the next line, enter the guest concentration information for the stock solution in the same way: concentration first, then, if desired, the uncertainty. If no uncertainty is reported, the value of the second entry on the second line of the error bars file will be used. If the concentration in the solution of host or guest, or both, is zero, do not omit the line; that will cause the computer to misunderstand everything

later in the file. Instead, enter "0" for the concentration. Additional stock solutions may be described on succeeding lines.

Creation of Samples. A blank line signals that all stock solution information has been entered, and that the succeeding block of lines will describe how solutions are combined to make the observed samples. Since there are many ways a titration study may be conducted, every solution aliquot must be described thoroughly to avoid ambiguity. As a result, this section is tedious.

The recipe for creating the samples is given in a single block without blank lines. Each line either describes an aliquot added to the sample tube, or calls for a fresh sample tube. Aliquots are always added to the most recent tube.

The description of an aliquot involves merely the name of the new sample solution being created, the name of the device used to deliver the aliquot, the nominal volume of the aliquot, and the name of the solution added. These four data are on one line of the input file. The named delivery device and added solution must have been defined earlier. To indicate that the next aliquot goes into a new empty tube instead of being added to the previous sample, put only one string of characters on a line. The actual string does not matter; it is just a signal to begin a fresh tube. A line with just one entry signals for a new tube, a line with four entries describes an aliquot, and a blank line terminates the sample recipe block. Any other format in this block is an error.

NMR observation data. A blank line indicates that all sample information has been entered, and that the following entries report the pertinent spectra, proton by proton. A proton block begins with a line reporting the proton's free chemical shift. If desired, this line may also report the uncertainty in this measurement. If it

does not contain an uncertainty, the value on the first line of the error bars file will be used. This line may also contain, after the chemical shift and *before* the standard deviation, an "H" or "G" to indicate that the proton belongs to the host or to the guest, respectively. If one of these characters is not present, the proton is assumed to be a guest proton.

The remaining lines in a proton block tell the chemical shifts observed for that proton in the various samples. The first entry in each line is the name of the sample, and the second is the chemical shift observed. If desired, an uncertainty in this chemical shift measurement may also be included as a third entry on the line. If no uncertainty is reported, the default NMR measurement error from the first line of the error bars file will be used. Proton blocks are separated by a blank line. Up to 20 protons may be included. The last proton observation concludes the input file.

Table I. Design and results of an NMR titration binding study.[a]

sample	aliquot volume[b]	solution	pipet	δ^c	$\pm\delta^c$
a	400	buffer	1000λ		
b	25	buffer	200λ		
c	750	host	200λ		
1	3	guest	10λ	2.9544	00.001
2	2	guest	10λ	2.9544	00.001
3	3	guest	10λ	2.9573	00.001
4	9	guest	10λ	2.9750	00.001
5	8	guest	10λ	3.1602	00.001
6	13	guest	50λ	3.6513	00.05
7	13	guest	50λ	3.9292	00.05
8	25	guest	50λ	4.1644	00.001
9	51	guest	250λ	4.3629	00.001
10	130	guest	250λ	4.5173	00.001

[a]Data are from Patrick Kearney. [b]In μl. [c]In ppm.

(iv) Several example input files. Let us examine some hypothetical experiments, and the input files that report them.

Addition of guest. A common experimental design is to begin with a solution of host and guest in an NMR tube and to add aliquots of guest to it. An actual binding study of *N*-methylquinolinium iodide with host P was carried out in the following manner. The host stock solution concentration was 1.99 mM, the guest stock solution concentration was 5.90 mM, and the samples were created by combining aliquots as described in Table I. In this table, the chemical shifts of the *N*-methyl protons of the guest are reported for the samples at which spectra were taken. The free chemical shift of these guest protons is 4.6702 ppm. Four delivery devices were used: 10, 50, 250, and 1000 μl autopipettes. These are designated 10λ, 50λ, 250λ, and 1000λ, respectively. The following file listing is one possible way to report this experiment.

```
10lam 0.05 0.04e-6          (delivery devices)
50lam 0.02 1.0e-6
250lam 0.01 1.0e-6
1000lam 0.006 2.0e-6
                            (blank line: stock solutions follow)
host 1.99e-3                (name and host concentration of stock solution "host")
0                           (guest concentration of stock solution "host")
guest 0
5.90e-3
buffer 0
0
                            (blank line: sample recipe follows)
a 1000lam buffer 400e-6     (samples)
b 200lam buffer 25e-6
c 200lam host 75e-6
1 10lam guest 3e-6
2 10lam guest 9e-6
3 10lam guest 3e-6
4 10lam guest 9e-6
5 10lam guest 8e-6
```

```
6 50lam guest 13e-6
7 50lam guest 13e-6
8 50lam guest 25e-6
9 250lam guest 51e-6
10 250lam guest 130e-6
```

(blank line: proton blocks follow)

```
4.6702
```

(δ_{free})

```
1 2.9544
```

(observations)

```
2 2.9544
3 2.9573
4 2.9750
5 3.1602
6 3.6513 0.05
```

(uncertainty in measurement = 0.05 ppm)

```
7 3.9292 0.05
8 4.1644
9 4.3629
10 4.5173
```

The next example illustrates three points: specifying that a proton belongs to the host, calling for a fresh sample tube, and adding aliquots of solutions that are not stock solutions.

In this experiment, the host stock solution is very concentrated, so it is diluted tenfold and an aliquot of the diluted solution is used to make the observed samples. A spike of the concentrated stock solution is added near the end of the study. In this study, three protons are observed, one of which is a host proton. All chemical shift data are reported in Hz.

```
50lam 0.01 0.2e-6
100lam 0.01 0.5e-6
1000lam 0.006 2e-6

host 10e-3 0.05
0
guest 0
1e-3 0.05
buffer 0
0
```

```
1a 100lam host 100e-6
1b 1000lam buffer 900e-6      (diluting the host solution)
2                             (signaling for a new sample tube)
2a 50lam 1b 50e-6             (50 μl of sample 1b is added to an empty tube)
2b 50lam guest 50e-6
2c 1000lam buffer 300e-6
2d 50lam buffer 10e-6
2e 50lam buffer 20e-6
2f 50lam buffer 40e-6
2g 100lam buffer 80e-6
2h 1000lam buffer 150e-6
2i 1000lam buffer 200e-6
2j 50lam host 10e-6           (now add 10 μl of the host stock solution)
2k 100lam buffer 100e-6

                              (blank line to terminate sample recipe)
1000                          (first proton block)
2c 622.83
2d 624.12
2e 626.83
2f 631.79
2g 640.60
2h 655.35
2i 671.92
2j 544.10
2k 544.10

                              (blank line between proton blocks)
500 H                         (host proton)
2c 575.43
2e 574.63                     (shift in sample 2d not reported for this proton)
2f 573.64
2g 571.88
2h 568.93
2i 565.62
2j 530.63
2k 530.39

2500                          (another guest proton)
2c 1745.66
2d 1748.82
2e 1753.65
2f 1763.58
2g 1781.20
2h 1810.71
2i 1843.84
2j 1580.44
```

2k 1588.21

An experiment Emul can't handle. Although the input format was designed to be flexible, it cannot accommodate all conceivable types of experiment. Since all sample solutions are handled in sequence, there is no way to affect sample solutions in two sample tubes alternately. For example, it is possible to model making up samples in two tubes, A and B, and adding solution from tube A to tube B. However, it is *not* possible to then model taking an aliquot out of tube B and adding it tube A. Since such an experimental design is so bizarre, this prohibition should not pose any serious difficulties.

B. Lucius.

1. User interface. Lucius is executed by simply typing "`lucius¬`." The Lucius user interface is similar to that for Emul. It first asks for the name of the preferences file; the default is always "nmr.prf." Then it asks for the names of the input file and the text output file. The default input file name is specified by the preferences file, and the default name of the text output file is the name of the input file with the default output file extension (from the preferences file) appended to it. The name of the text output file is automatically applied to the names of the parameter distribution and scatter graph files.

Lucius will also ask for the name of a confidence limits file. This file specifies the regions of the fitted parameter and SSR distributions for which limiting values will be reported in the text output file. For example, if this file contains the values 95 and 99, the boundaries of the lower 95% and 99% regions of the SSR distribution will be reported, as will the lower and upper boundaries of the central 95% and 99% regions of the fitted parameter distributions. These numbers do not affect the course of the simulations in any way.

The confidence values reported in the text output file may be slightly different from the numbers in the file. This is because the desired limit may not be an even divisor of the number of replications. For example, the empirical boundaries of the central 99% central region of a distribution are those values below or above which 0.5% of the sampled values fall. If 5300 replications are performed, this fraction corresponds to 26.5 counts. Since it is impossible to record non-integer counts, Lucius looks for the nearest boundary it *can* record. The central region excluding 27 counts on each side is the 98.98% region, and that is the region reported if the 99% region is requested.

It is not necessary to supply initial parameter guesses or an error bars file name to Lucius. The simulator input file already contains the best-fit parameters from Emul, as well as all of the measurement uncertainties from the experiment.

After the input and output files have been named, Lucius begins its set of Monte Carlo simulations. It announces this fact by typing "`simulating`" on the screen. This job may take several minutes; as long as it is running in the foreground, it is impossible to do anything else from your shell. The job may be sent to the background, however, by typing "^Z"[8] followed by "bg¬." These commands halt the foreground job and set it running again in the background, respectively. When the job is finished, it displays a message to that effect, and reiterates the names of its output files.

2. Lucius input files.

Preferences file: this file is thoroughly described in the Emul section.

---

(8) Hereafter the symbol "^" will be used to denote the control key. Striking "^X" means to press the control key, and to then strike key X while the control key is still depressed.

Simulator Input file: This file is created by Emul, so the user has no need to learn its format.

Confidence limits file: This file is very simple. Each line contains a number between 0 and 100 (0 is acceptable; 100 is not), which specifies a percentile region of the SSR and parameter distributions to be reported in the text output file. These numbers do not need to be in order.

C. Portia

Portia is executed by typing "`portia¬`." Portia's user interface and input files are identical to those for Lucius. Output files have exactly the same format as those from Lucius as well, so nothing new needs to be specified here.

D. Brutus

Brutus is executed by typing "`brutus¬`." It asks the user the same questions as Lucius and Portia, and uses exactly the same input files. However, the confidence limits file has a more profound influence on Brutus than on the other programs. For Brutus, this file specifies the parameter confidence regions that must be actively sought. Each boundary of a confidence region must be closely bracketed by the results of two full simulations about different parameter values. Each of these simulations is equivalent to a full run of Lucius. Thus, it is prudent to include only those confidence limits that are most important, and not to bother mapping out the entire confidence function in small increments.

Brutus makes two sets of confidence limits from the values in the file: one for the Lucius-like initial simulation, and one for the methodical search. The requirements for the limits in the Lucius-like run are wholly equivalent to the requirements

in Lucius itself. Different criteria determine the acceptability of limits for the methodical search. In particular, confidence limits that are not whole-number divisors of the number of replications are not a problem, because the interpolated confidence function is continuous. On the other hand, the number of replications *does* dictate the most ambitious confidence limit that will be sought. This condition is discussed more completely in the Appendix to Chapter 3. Basically, as desired confidence limits get closer and closer to 100%, it becomes more and more difficult to bracket them on both sides. Conversely, as the number of replications performed per study increases, the probability of obtaining a reliable bracketing value for any given limit increases. It is possible for the number of replications to be too low for a requested confidence limit to be reliably found. In such a case, Brutus simply substitutes in place of the overly ambitious limit the most ambitious allowed limit. Instead of then simply launching into a long search that will not return the limits the user requested, Brutus reports its updated limits list and gives the user an opportunity to terminate the run. If more ambitious limits than those allowed are desired, the user should run Brutus again, using a preferences file that specifies more replications per study (this is *submax*, the first entry in the preferences file). The Appendix to Chapter 3 contains the formula relating the desired limit to the minimum number of replications necessary.

Running in background. Since Brutus takes so much longer to run than Lucius or Portia, it is especially valuable to run it in the background. This is accomplished in the same manner as running Lucius or Portia in the background: halt the foreground execution by typing "^Z" and then resume execution in the background by typing "bg¬." This is vital if you want to log off the computer while Brutus is running. If you log off while it is running in the foreground, execution will

be terminated. If you log off while it is running in the *background*, the program *will keep on running.* You can check for its completion at any later time by using the "ps" command.

Input files. The input files are exactly the same as those used in Lucius and Portia.

III. Binding Study Design

The key to getting good parameter estimates is to design informative binding studies. An "informative" study is one that minimizes the influence of random measurement errors on the eventual best-fit parameters. Wilcox has devoted a portion of a recent article to emphasizing the importance of maintaining the *minor component* in a sample between 20% and 80% bound.[9] The reason for this stipulation is that, outside of this range, the chemical shifts can be modeled equally well by very different binding constants. If the minor component is bound to near saturation in all samples, for instance, the experiment does not distinguish between binding constants that are large and those that are enormous. Conversely, if the minor component experiences very little binding, all that can be said about the binding constant is that it is small. Table II summarizes the extent of binding at a variety of concentrations for a number of different equilibrium constants. Numbers in the desirable range are set in **bold** type.

---

(9) Wilcox, Craig S. "Design, synthesis, and evaluation of an efficacious functional group dyad. Methods and limitations in the use of NMR for measuring host-guest interactions," In *Frontiers in Supramolecular Organic Chemistry and Photochemistry*, Schneider, H.-J.; Dürr, H., Eds.; VCH: Weinheim, 1990.

Table II. Extent of Complexation.

		% minor component bound when K^c =								
[Maj][a]	[min][b]	3×10^2	1×10^3	3×10^3	1×10^4	3×10^4	1×10^5	3×10^5	1×10^6	3×10^6
500	200	12	**31**	**54**	**78**	91	97	99	100	100
500	100	13	**32**	**57**	81	92	98	99	100	100
500	50	13	**33**	**59**	82	93	98	99	100	100
500	20	13	**33**	**59**	83	94	98	99	100	100
500	10	13	**33**	**60**	83	94	98	99	100	100
200	200	5	15	**30**	**50**	**66**	**80**	88	93	96
200	100	5	16	**33**	**59**	**78**	91	97	99	100
200	50	6	16	**35**	**63**	82	94	98	99	100
200	20	6	16	**37**	**65**	84	95	98	99	100
200	10	6	17	**37**	**66**	85	95	98	99	100
100	100	3	8	19	**38**	**57**	**73**	83	90	94
100	50	3	9	**21**	**44**	**66**	85	94	98	99
100	20	3	9	**22**	**48**	**72**	89	96	99	100
100	10	3	9	**23**	**49**	**73**	90	96	99	100
50	50	1	5	12	**27**	**45**	**64**	**77**	87	92
50	20	1	5	12	**31**	**54**	**77**	91	97	99
50	10	1	5	13	**32**	**57**	81	92	98	99
20	20	1	2	5	15	**30**	**50**	**67**	**80**	88
20	10	0	5	6	16	**33**	**59**	**78**	92	96
10	10	0	1	3	8	19	**38**	**57**	**73**	83

[a]Concentration of major component, in μM. [b]Concentration of minor component, in μM. [c]In M^{-1}.

From this table, it is clear that no single set of host and guest concentrations is ideal for all possible equilibrium constants. In order to get a reliable estimate of a binding constant, it is necessary to run a binding study with appropriate sample concentrations. An initial estimate of the binding constant should be used in choosing the concentrations to include in the binding study. If no initial estimate is available, a quick binding study using just a few samples should first be run in order to get this initial estimate; a more thorough study at the right concentrations will then return the best estimate possible.

As an example, consider the case of a host/guest pair with an association constant of 10^5 M^{-1}. The graphs of Figure 12 show the distributions of parameter estimates determined by Monte Carlo simulations of three different experimental

designs. All of these cases represent a system in which two protons, one from host and one from guest, are followed in fifteen samples created by addition of guest to the sample tube. D of the host proton is -100 Hz, and of the guest proton is $+500$ Hz. The three experimental designs followed are G4, G5, and G6 from Chapter 2; G4 was designed to provide a good concentration range for an association constant of 10^4 M^{-1}, G5 for $K = 10^5$ M^{-1}, and G6 for $K = 10^6$ M^{-1}. The program Lucius performed 1500 replications of each experiment to generate these empirical distributions.

Table III. Extent of binding with three experimental designs when $K = 10^5$ M^{-1}.

design	%H bound	%G bound
G4	92–97	20–61
G5	14–94	22–86
G6	33–83	16–50

Table III summarizes the extent of binding of host and guest under each of these three experimental designs. G5 and G6 both appear to be good designs on the basis of the "20–80 rule"; every sample in G4, however, involves near-saturation binding of the host. It is thus not surprising that the distribution of fitted parameters from the Monte Carlo study of this design is very broad. Designs G5 and G6 appear more equal by the "20–80 rule"; G5 covers a wider range of fraction bound, but G6 never strays from the 20–80% region. Both of these experimental designs perform well, especially in comparison to design G4. Surprisingly, design G4 is not even the best for estimating D of the saturated proton. These graphs confirm what is intuitively obvious: *covering the entire range of 20–80% minor component bound gives good parameter estimates.*

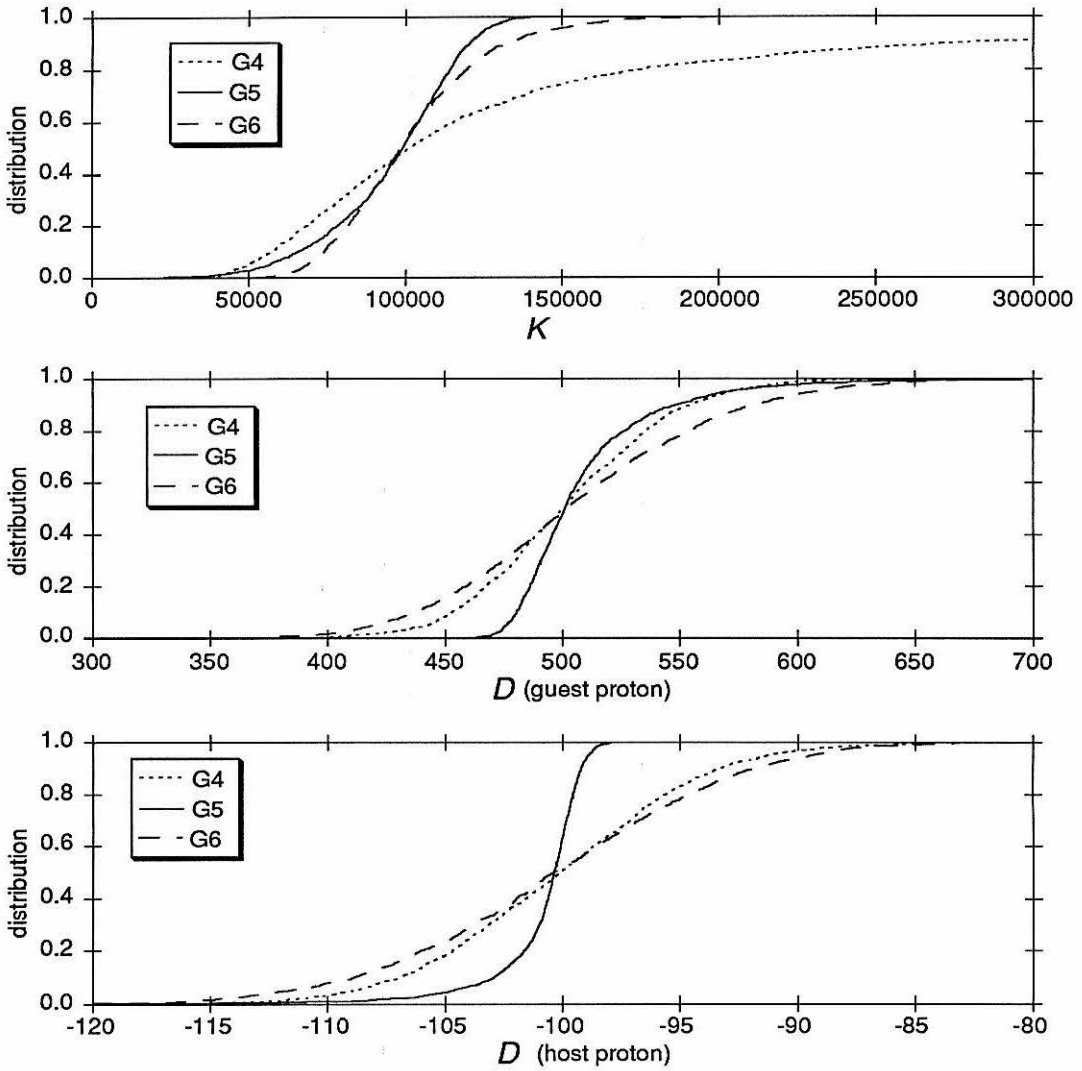


Figure 12. Parameter estimate distributions using the host and guest concentrations in the sets G4, G5, and G6 when $K = 10^5$ M^{-1}.

A Monte Carlo study illustrating another intuitively obvious point is summarized in Figure 13. Two very similar experimental designs are compared here. One of the sets is G5, the best set from Figure 12. The other set uses the same concentrations as G5, but records only five spectra. In order to give this design the best possible statistical advantage, its samples are not made in exactly the same

Table IV. Recipes for samples in set G5 and its trimmed counterpart.

Design G5			Trimmed design		
#	V_a^a	Solution	#	V_a^a	Solution
a	360	buffer	a	360	buffer
b	30	host	b	30	host
1	10	guest	1	10	guest
2	5	guest			
3	5	guest			
4	5	guest			
5	5	guest	2	20	guest
6	10	guest			
7	10	guest			
8	10	guest	3	30	guest
9	15	guest			
10	15	guest			
11	15	guest			
12	20	guest	4	65	guest
13	30	guest			
14	40	guest			
15	55	guest	5	125	guest

[a]In μl.

way as those in G5. Instead of adding separate aliquots for all the samples created in G5, it combines the volumes of aliquots that define unobserved samples into larger aliquots, so that it does not amplify the delivery device imprecision. The exact recipe is summarized in Table IV. Its concentrations are thus more accurately known than those in design G5. Nonetheless, G5 clearly gives better parameter estimates.

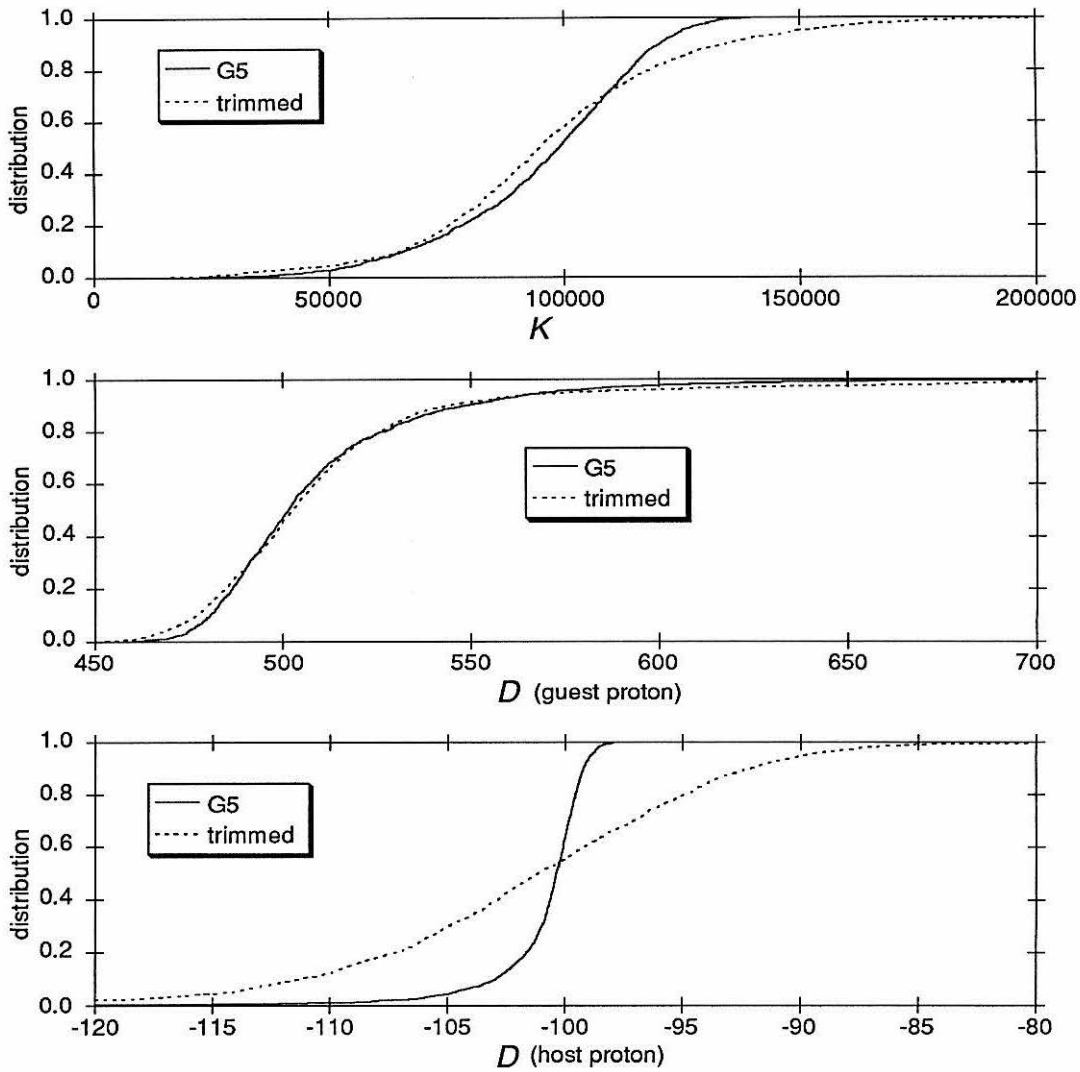


Figure 13. Parameter estimate distributions using data set G5 and a parallel set that covers the same concentration range, but only samples one-third as many points.

Chapter 5

Thermodynamics of Molecular Recognition

Abstract: The thermodynamics of complexation reactions of ethenoanthracene hosts in both borate-*d* and chloroform-*d* have been investigated by variable-temperature binding studies. In many cases, these reactions show significant heat capacity effects. Modeling of the variable-temperature data by a constant ΔC_p° value is convincingly superior to the naïve van't Hoff approach. A simple model of the origin of the heat capacity effect is discussed.

I. Introduction

A. The Importance of Thermodynamic Parameters.

The field of molecular recognition chemistry is inspired by biological receptors. Biological molecules are too large to study in detail, so physical-organic chemists create smaller organic molecules that retain, it is hoped, key features of biological binding interactions. These analogs can be scrutinized in more detail than the biological systems, and can be perturbed in ways more profound than are currently possible with biomolecules. In this way, the forces and interactions important in the larger systems can eventually be understood and perhaps exploited.

As described in chapter 1, a number of forces have been implicated in the complexation processes of ethenoanthracene-based hosts. Donor-acceptor effects appear significant in complexation of aromatic guests that form π-stacked sandwiches inside the rhomboid conformers of the hosts. The cation-π effect explains the tendency of cationic guests to prefer the host cavity even to water, and to occupy the binding site of a completely uncharged tetraester host in chloroform. In addition, the hydrophobic effect is probably a major contributor to the reactions occurring in water.

Discrimination of the individual contributions of these effects to the total energy of the binding event can be achieved by analyzing changes in the binding energy

in response to variations in host or guest structure. Indeed, it was by just such analyses that the donor-acceptor and cation-π effects were first recognized. Further insight into the complexation process could be obtained by dissecting the free energy of complexation into its thermodynamic components, the enthalpy and entropy of complexation.

The Gibbs free energy ΔG of a reaction is a sort of "virtual" energy. It is not directly related to the total heat or work associated with a process, as might be expected of a quantity with the word "energy" in its name. Instead, it is merely a convenient way, expressed in the universal currency of energy, to identify the spontaneous direction and extent of any process occurring at constant pressure.[1] The customary statement of the second law of thermodynamics is

$$\Delta S_{\text{sys}} + \Delta S_{\text{surr}} \geq 0\,.$$

That is, a spontaneous process will act to increase the total entropy of the universe (system + surroundings). In order to predict the spontaneous direction of a given process, it is necessary to know the changes in entropy experienced by both the system (ΔS_{sys}) and the surroundings (ΔS_{surr}). It is inconvenient to consider the surroundings, so this term is profitably replaced by a term depending only on properties of the system. By definition, the infinitesimal entropy change of the surroundings in a small step is

$$dS_{\text{surr}} = \frac{dq_{\text{rev}}}{T}\,,$$

where q_{rev} is the heat absorbed by the surroundings in a reversible pathway with the same initial and final states as the process of interest. If the process occurs

(1) A similar derivation can be found in Atkins, Peter William *Physical Chemistry*; 2 ed.; W. H. Freeman: San Francisco, 1982; pp 145–146.

at constant pressure, the differential element of heat absorbed, regardless of the reversibility of the pathway, is dH. Thus,

$$dS_{\text{surr}} = \frac{dH_{\text{surr}}}{T}.$$

Since the only possible source of the heat absorbed by the surroundings is the system, $dH_{\text{surr}} = -dH_{\text{sys}}$. Thus,

$$dS_{\text{surr}} = -\frac{dH_{sys}}{T},$$

and the thermodynamic criterion of spontaneity is

$$dS_{\text{sys}} - \frac{dH_{\text{sys}}}{T} \geq 0.$$

If we define a state function G, $G \equiv H - TS$, then, employing thermodynamic properties of only the system,

$$dG = dH - TdS \leq 0.$$

If the temperatures of the initial and final states are the same, the second law of thermodynamics may be expressed as equation 1.

$$\Delta G \equiv \Delta H - T\Delta S \leq 0 \tag{1}$$

This state function ΔG, the Gibbs free energy change of the transformation, provides a thermodynamic criterion for spontaneity that depends only upon properties of the system. No explicit knowledge of the surroundings is required.[2]

The two terms in the definition of ΔG are customarily considered individually. A favorable (negative) ΔH, meaning that heat is evolved by the reaction, is interpreted as an intrinsic driving force of the reaction. For a complexation reaction, for example, a negative ΔH would suggest that there is an inherent attraction between

(2) Benzinger has questioned the utility of this particular partitioning of the criterion for spontaneity. Benzinger, T. H. "Thermodynamics, chemical reactions and molecular biology," *Nature* **1971**, *229*, 100–102.

the two complexing species. A favorable (positive) ΔS, on the other hand, is viewed as a relaxation of constraints on the system. If ΔS for a reaction is positive, the system has more conformational, rotational, or translational freedom in its final state than in its initial state. Since a complexation reaction involves bringing two species together to form one single species, molecular mobility must decrease. A positive entropy change for such a reaction would require some other process, such as solvent structure disruption, to occur at the same time. The relative contributions of ΔH and ΔS to a reaction free energy, then, reveal something of the nature of the forces responsible for the reaction.

B. Thermodynamic Properties of Binding Interactions.

Two of the forces implicated in the complexation reactions of our hosts, the cation-π and donor-acceptor effects,[3,4] should be principally enthalpic in nature. A complexation reaction driven by either of these effects would show a large, negative enthalpy change sufficient to overcome an inherently unfavorable entropy change. This enthalpy change would result from the intrinsic affinity of positively-charged or electron-deficient guests toward the binding site. If such an attraction exists, the potential energy of the separated partners is higher than that of the complex. Upon complexation, this potential energy will be released to the surroundings as heat.

The hydrophobic effect would leave a qualitatively different thermodynamic signature. Water is profoundly different from other solvents in the way it behaves toward sparingly soluble solutes. The entropy changes for most dissolution processes

(3) Shepodd, Timothy J.; Petti, Michael A.; Dougherty, Dennis A. "Molecular recognition in aqueous media: donor-acceptor and ion-dipole interactions produce tight binding for highly soluble guests," *J. Am. Chem. Soc.* **1988**, *110*, 1983–1985. Stauffer, David A.; Dougherty, Dennis A. "Ion-dipole effect as a force for molecular recognition in organic media," *Tetrahedron Lett.* **1988**, *29*, 6039–6042.

(4) Petti, Michael A.; Shepodd, Timothy J.; Barrans, Richard E. Jr.; Dougherty, Dennis A. "'Hydrophobic' binding of water-soluble guests by high-symmetry, chiral hosts. An electron-rich receptor site with a general affinity for quaternary ammonium compounds and electron-deficient π systems," *J. Am. Chem. Soc.* **1988**, *110*, 6835–6840.

are favorable, reflecting the enormous number of new configurations available to a system of N solvent molecules when n solute molecules are introduced. The enthalpy change of a dissolution process can be either positive or negative, depending on the relative magnitudes of intermolecular adhesive and cohesive forces. If the attractions between molecules of the same species are stronger than the attractions between molecules of different species, then the enthalpy of mixing is positive. This positive enthalpy results in poor solubility. Nonpolar solutes in water, on the other hand, tend to exhibit *negative* entropies of mixing, that is, conformational mobility somehow *decreases* upon dissolution. In further contrast to other solvents, the enthalpies of such processes tend to be small but favorable. In this case, $-T\Delta S$ is the dominant term in the free energy expression, overwhelming the opposing but ineffectual ΔH contribution.[5]

This "hydrophobic hydration" is conventionally interpreted as a change in the structure of water immediately surrounding the solute. Specifically, the hydration shell about a nonpolar solute is thought to be more similar to ice than to liquid water.[6,7] When a hydrophobic solute is placed in water, the resulting change in the water structure is similar to a freezing process. Enthalpy decreases because the water molecules in the hydration shell are locked into favorable hydrogen-bonding configurations, but this restriction of conformational mobility also reduces the entropy. Overall, this unfavorable solvent entropy change makes hydrocarbons and other

---

(5) Tanford, Charles *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*; Wiley-Interscience: New York, 1980.

(6) Frank, H. S.; Evans, M. W. "Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes," *J. Chem. Phys.* **1956**, *13*, 507.

(7) Némethy, George; Scheraga, Harold A. "Structure of water and hydrophobic bonding in proteins. I. A model for the thermodynamic properties of liquid water," *J. Chem. Phys.* **1962**, *36*, 3382–3400. Némethy, George; Scheraga, Harold A. "Structure of water and hydrophobic bonding in proteins. II. Model for the thermdynamic properties of aqueous solutions of hydrocarbons," *J. Chem. Phys.* **1962**, *36*, 3401–3417.

nonpolar compounds poorly soluble in water. The precise nature of hydrophobic hydration is under active investigation.[8–17]

It is easy to see how the hydrophobic effect could provide the driving force for an association reaction between a nonpolar guest and a host with a hydrophobic binding site. When these species are separated, each is encased in an ice-like hydration shell. The total number of water molecules present in these shells is proportional to the hydrophobic surface area exposed to the solvent. A hydration shell about the host/guest complex employs fewer solvent molecules than the shells about the separated reactants, as illustrated in Figure 1. When the guest enters the host cavity, some of the hydration shell waters contacting the guest exterior and host interior are displaced. Effectively, some of the nonpolar solute is removed from the aqueous environment. The released waters have much more mobility than in the

(8) Shinoda, Kōzō "'Iceberg' formation and stability," *J. Phys. Chem.* **1977**, *81*, 1300–1302. Shinoda, Kōzō; Kobayashi, Makoto; Yamaguchi, Nobuyoshi "Effect of 'iceberg' formation of water on the enthalpy and entropy of solution of paraffin chain compounds: the effect of temperature on the critical micelle concentration of lithium perfluorooctanesulfonate," *J. Phys. Chem.* **1987**, *91*, 5292–5294.

(9) Patterson, Donald; Barbe, M. "Enthalpy-entropy compensation and order in alkane and aqueous systems," *J. Phys. Chem.* **1976**, *80*, 2435–2436. Costas, Miguel; Patterson, Donald "Heat capacities of water + organic-solvent mixtures," J. Chem. Soc. Faraday Trans. 1 **1985**, *81*, 2381–2398.

(10) Mirejovsky, Dorla; Arnett, Edward M. "Heat capacities of solution for alcohols in polar solvents and the new view of hydrophobic effects," *J. Am. Chem. Soc.* **1983**, *105*, 1112–1117.

(11) Ramadan, Mohamed S.; Evans, D. Fennell; Lumry, R. "Why micelles form in water and hydrazine. A reexamination of the origins of hydrophobicity," *J. Phys. Chem.* **1983**, *87*, 4538–4543.

(12) Gill, S. J.; Dec, S. F.; Olofsson, G.; Wadsö, I. "Anomalous heat capacity of hydrophobic solvation," *J. Phys. Chem.* **1985**, *89*, 3758–3761.

(13) Privalov, Peter L.; Gill, Stanley J. "The hydrophobic effect: a reappraisal," *Pure Appl. Chem.* **1989**, *61*, 1097–1104.

(14) Muller, Norbert "Search for a realistic view of hydrophobic effects," Acc. Chem. Res. **1990**, *23*, 23–28.

(15) Pratt, Lawrence R. "Theory of hydrophobic effects," *Ann. Rev. Phys. Chem.* **1985**, *36*, 433–449.

(16) Privalov, Peter L.; Gill,Stanley J. "Stability of protein structure and hydrophobic interaction," *Adv. Prot. Chem.* **1988**, *39*, 191–234.

(17) Hobza, Pavel; Zahradník, Rudolf *Intermolecular Complexes*; Elsevier: New York, 1988.

ice-like state, at the cost of a modest loss of favorable hydrogen bonding interactions. Consequently, the association of two nonpolar molecules in water causes a large entropy increase and a small enthalpy increase. If no other interactions are operating, this is a thermodynamically spontaneous process.

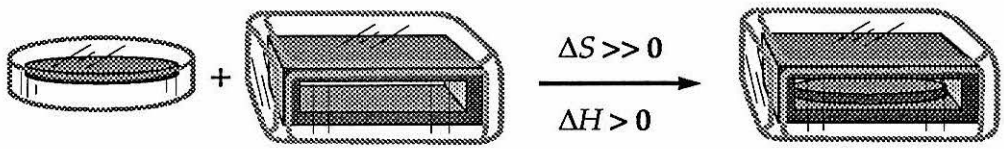


Figure 1. Complexation of a hydrophobic host and guest in water. There is less "iceberg" water associated with the complex than with the separated reactants.

C. Determining Thermodynamic Parameters.

ΔH and ΔS for a reaction can be determined by measuring the equilibrium constant K at a variety of temperatures. This is because ΔG is related by the Boltzmann distribution to the equilibrium populations of any number of chemical states. In the limit of only two states, this distribution can be expressed as

$$\Delta G^\circ = -RT \ln K \,. \tag{2}$$

ΔG° in this equation is the standard free energy of the reaction, and K is the equilibrium constant. This ratio in the case of a simple host/guest complexation reaction is as shown in equation 3.

$$K = \frac{[\mathrm{H{\cdot}G}]}{[\mathrm{H}][\mathrm{G}]} \tag{3}$$

Combination of equations 1 and 2 produces 4.

$$-RT \ln K = \Delta H^\circ - T\Delta S^\circ \tag{4}$$

This equation can be further manipulated to give 5.

$$R \ln K = \Delta S^\circ - \frac{\Delta H^\circ}{T} \tag{5}$$

The two observable quantities in this equation are K and T. R is independently known as a physical constant (1.98720 cal mol^{-1} K^{-1}), so the only unknowns in this equation are $\Delta S°$ and $\Delta H°$. If these are invariant with temperature, this equation is of the form

$$y = a + bx \tag{6}$$

in which x is the independent variable, y is the dependent variable, and a and b are constants. If this model is rigorously correct and there is no experimental error, a plot of $R\ln K$ versus $1/T$ will yield a perfectly straight line with slope $-\Delta H°$ and y-intercept $\Delta S°$. In the more plausible situation of normally-distributed independent measurement errors in $R\ln K$, the $(1/T, R\ln K)$ points will be randomly scattered about this straight line. $\Delta H°$ and $\Delta S°$ can then be estimated by fitting a best straight line to the data by linear least-squares regression.

All the tools necessary to study the thermodynamics of complexation reactions of our hosts are now in place. The precedent provided by the work of other investigators with cyclodextrins[18] and cationic cyclophanes[19] led us to expect to see evidence of a "non-classical" hydrophobic effect.[20] The signature of this effect is what one would expect from a weak hydrophobic interaction and a strong intrinsic attraction operating concurrently. The enthalpy change is large and negative, accounting for most of the free energy change of the reaction, and the entropy change is usually near zero.

---

(18) Harrison, John C.; Eftink, Maurice R. "Cyclodextrin-adamantanecarboxylate inclusion complexes: a model system for the hydrophobic effect," *Biopolumers* **1982**, *21*, 1153–1166. Cromwell, Milliam C.; Byström, Katarina; Eftink, Maurice R. "Cyclodextrin-adamantanecarboxylate inclusion complexes: studies of the variation of cavity size," *J. Chem. Phys.* **1985**, *89*, 326–332. Eftink, Maurice R.; Andy, M. L.; Byström, K.; Perlmutter, H. D.; Kristol, D. S. *J. Am. Chem. Soc.* **1989**, *111*, 6765–6772. Saenger, W. *Angew. Chem. Int. Ed. Engl.* **1980**, *19*, 344–362.

(19) Fergusen, S. B.; Seward, E. M.; Diederich, F.; Sanford, E. M.; Chou, A.; Inocencio-Szweda, P.; Knobler, C. B. "Strong enthalpically driven complexation of neutral benzene guests in aqueous solution," *J. Org. Chem.* **1988**, *53*, 5593–5595.

(20) Jencks, William P. *Catalysis in Chemistry and Enzymology*; McGraw-Hill: New York, 1969; p 427.

II. Studies.

A. Genesis.

Our first foray into the variable-temperature binding study arena was led by Michael Petti. He looked at the binding of host V_{meso} with ATMA in a pD $9.5-10$ phosphate buffer. This laborious experiment bore a surface resemblance to an ordinary NMR titration study; guest was added in increments to host in an NMR sample tube. After each addition, however, a spectrum was taken of the sample at each of five temperatures.

The van't Hoff plot from this experiment is shown in figure 2. The $\Delta H°$ and $\Delta S°$ of this reaction as determined from the best-fit line to the data show favorable entropic and enthalpic contributions. At 298 K, the contribution from $\Delta H°$ is about twice that from $T\Delta S°$. This behavior is in line with a non-classical hydrophobic effect. Comparable results were obtained from a companion study of IV_{meso} with ATMA in the same medium.[21,4]

The data points in Figure 2 show noticeable curvature. Although the fitted straight line is comfortably within the stated 95% confidence limits for the $R\ln K$ values,[22] the residuals from the fit, which are the differences between the fitted and observed values, show a pattern that cannot be reasonably attributed to random error. In fact, the error bars shown for the data are misleading. They show the absolute confidence limits for $R\ln K$ at each temperature, but they do not reflect the fact that the same samples were used in the determinations of K at all temperatures. The only measurement errors that are not shared by all of these determinations are spectrometer peak position and probe temperature fluctuations, which are negligible. The remaining measurement errors are manifest in the sample concentrations. If the

---

(21) Petti, Michael A. Ph.D. Thesis, California Institute of Technology, November 1988, pp 59–68.

(22) Confidence limits were determined by the program Brutus.

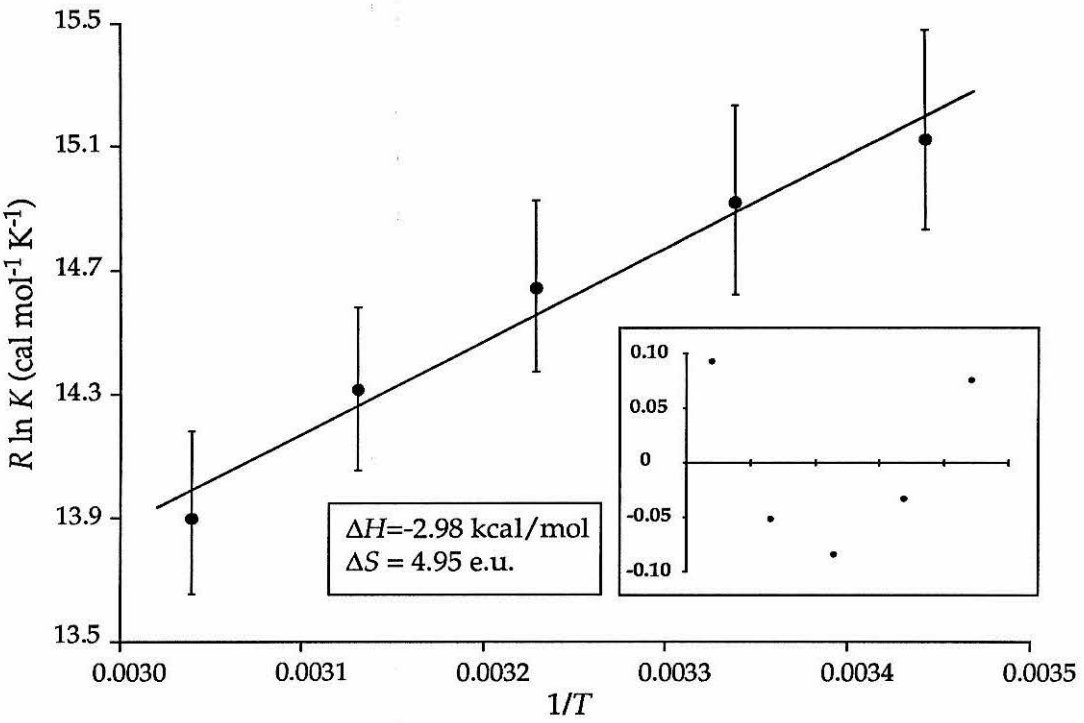


Figure 2. Van't Hoff plot for the complexation of ATMA with V_{meso} in pH 9.5–10 phosphate buffer. Inset: residuals from the fit.

sample concentrations are too high or too low at one of the temperatures, they are also miscalculated in the same manner at all other temperatures. Any biasing effect on the fitted binding constants K will be in the same direction at all temperatures.

This means that the significant errors in the $R\ln K$ values at each temperature are not independent. Furthermore, the curvature of the residuals plot is too consistent to have resulted from random error. Clearly, something not accounted for by the assumed model was taking place, but the nature of this effect was not understood. It potentially could have been a failure of the model, or perhaps a systematic error. This equivocal result, coupled with the laborious nature of the variable-temperature NMR binding studies and the fact that Petti was nearing the end of his graduate studies, forestalled further attempts to catalog $\Delta H°$ and $\Delta S°$ of complexation reactions of our hosts.

A year later, David Stauffer began studies on quaternary ammonium and immonium guests with host P_E in chloroform.[23,24] In these studies, he employed the constant D assumption, an approximation that enables one to obtain variable-temperature binding data without performing a complete binding study at each of many temperatures. Van't Hoff plots of the resulting data again showed curvature, which again was concave downward. Correcting for the thermal expansion of the solvent did not remove this curvature. This indicated emphatically that the curvature resulted from some physical effect not included in the simple model.

B. The Constant D Assumption.

Before proceeding further, it is worthwhile to describe the use and implications of the constant D assumption. This is the belief that the chemical shift of a proton in the host/guest complex is the same at all temperatures. This is expected if the geometry of the complex is not affected by temperature. Similar behavior had been previously seen by Diederich.[25] He found that D, the change of a proton's resonance upon binding, was constant for a given host/guest pair even in different solvents. Changing the temperature should be a much less significant perturbation to the system than changing the solvent, so the constant D assumption is not unreasonable for variable-temperature studies.

This assumption greatly simplifies the determination of K at a variety of temperatures. If D and δ_{free} are known, then measuring the spectrum of a single sample will tell the equilibrium constant. Because the complexation reactions

$$H + G \rightleftharpoons H{\cdot}G$$

(23) Stauffer, David A.; Barrans, Richard E. Jr.; Dougherty, Dennis A. "Concerning the thermodynamics of molecular recognition in aqueous and organic media. Evidence for significant heat capacity effects," *J. Org. Chem.* **1990**, *55*, 2762–2767.

(24) Stauffer, David A. Ph.D. Thesis, California Institute of technology, May 1989, Chapter 2.

(25) Diederich, François; Dick, Klaus; Griebel, Dieter "Complexation of arenes by macrocyclic hosts in aqueous organic solutions," *J. Am. Chem. Soc.* **1986**, *108*, 2273-2286.

are fast on the NMR timescale, the observed signal of a host or guest proton in a sample containing both host and guest is given by equation 7.

$$\delta_{\text{obs}} = \delta_{\text{free}} - DF \tag{7}$$

Here, δ_{free} is the resonance of the proton in uncomplexed host or guest, and F is the fraction of the observed species bound. If $[\text{H}]_0$ and $[\text{G}]_0$ are defined as the total host and guest concentrations, respectively, in the sample, then $F = [\text{H}\cdot\text{G}]/[\text{G}]_0$ if the observed proton is a guest proton, and $F = [\text{H}\cdot\text{G}]/[\text{H}]_0$ if it is a host proton. It is convenient to universally define this fraction bound as $F = [\text{H}\cdot\text{G}]/[\text{S}]_0$, where S denotes the species bearing the proton under consideration. F can be determined from a single observation by solving equation 7.

$$F = \frac{\delta_{\text{free}} - \delta_{\text{obs}}}{D} \tag{8}$$

All that remains is to determine K from this F. The first step is to remove the factors of [H] and [G] in the equilibrium relation (equation 3). [H] and [G] can be expressed in terms of [H·G] and $[\text{H}]_0$ or $[\text{G}]_0$, since $[\text{H}]_0 = [\text{H}] + [\text{H}\cdot\text{G}]$ and $[\text{G}]_0 = [\text{G}] + [\text{H}\cdot\text{G}]$. Thus, the equilibrium relation can be rewritten as

$$K = \frac{[\text{H}\cdot\text{G}]}{([\text{H}]_0 - [\text{H}\cdot\text{G}])([\text{G}]_0 - [\text{H}\cdot\text{G}])}\,. \tag{9}$$

If the species *not* bearing the observed proton is denoted T, this relation becomes

$$K = \frac{[\text{H}\cdot\text{G}]}{([\text{T}]_0 - [\text{H}\cdot\text{G}])([\text{S}]_0 - [\text{H}\cdot\text{G}])} \tag{10};$$

substitution of $[\text{S}]_0F$ for [H·G] gives

$$K = \frac{[\text{S}]_0F}{([\text{T}]_0 - [\text{S}]_0F)([\text{S}]_0 - [\text{S}]_0F)} \tag{11}.$$

After simplification, this becomes equation 12.

$$K = \frac{F}{([\text{T}]_0 - [\text{S}]_0F)(1 - F)} \tag{12}$$

In this manner, every proton observed in this spectrum gives an estimate of K, based on its measured chemical shift and its independently-fitted value of D. The estimates of K obtained in this way from observations of different protons will, because of random measurement errors and the uncertainty in the assigned value of D, necessarily be different.

In principle, it is possible to find the value of K that minimizes

$$\mathrm{SSR} = \sum_{p=1}^{P} (\delta_{\mathrm{calc}\,p} - \delta_{\mathrm{obs}\,p})^2 \tag{13}$$

for the spectrum, where p is an index over all of the P protons observed. If this were desired, however, none of the described estimates of F from the observed spectrum would have any value, except as initial guesses for an iterative fitting procedure. It turns out that it is not possible to analytically solve equation 13 for the value of K that minimizes SSR. Using an iterative procedure to find this best-fit K is certainly possible, but it is not justified if the estimates of K from the different protons are close to each other. If the estimates from the protons are *not* close to each other, using an iterative procedure is still not justified, because such discrepancies would signal that either the model does not adequately describe the data or that the estimated values of D for some or all of the protons are in error. Neither problem would be ameliorated by minimizing SSR.

In practice, then, it is best to determine an estimate of K from observations of each proton, compare the estimates given by different protons, and average them if they are close. If experimental constraints are such that only one proton is followed, the single-proton estimate is the same as the minimum-SSR estimate.

To verify the supposition of constant D, Stauffer performed full binding studies at several temperatures with several different host/guest pairs. The values of D returned for a given complex at each temperature were compared; typically, they

showed a variability of less than 10% over the temperatures studied. Furthermore, the variation did not show a consistent pattern with temperature. This demonstrated that assuming a single value of D at all temperatures would not systematically bias the resulting estimates of K.

In fact, it is quite probable that employing the constant D assumption gives better variable-temperature $R\ln K$ data than does performing full binding titrations at each temperarture. The reason for this is the behavior of the goodness of fit score of a binding study, SSR, as a function of K and D. An example (SSR, K, D) surface is shown in Figure 3. This contour plot shows how K and D can compensate each other to give a low value of SSR. Over a wide range of K values, a value for D can be found that gives an SSR quite close to the global minimum, and vice versa. It is possible for different measurement errors causing only slight changes in the observed data set to ultimately lead to substantially different "best" values of K and D. This problem would be especially acute in experiments in which only a few samples are observed. Full variable-temperature binding studies are just such experiments, because every sample must be observed at every temperature. This is time-consuming and tedious, so an experimenter will want to use as few samples as possible. Even if D is the same at all temperatures, the few experimental random errors that are not shared by all spectra could perturb the observations enough for D to appear different at the different temperatures. This corresponds to scatter in the $R\ln K$ values as well. On the other hand, an incorrect assumed value of D leads to a change in the curvature of $R\ln K$ vs. $1/T$ plots when K is determined according to equations 8 and 12.[26]

---

(26) Reference 24, pp 86–96.

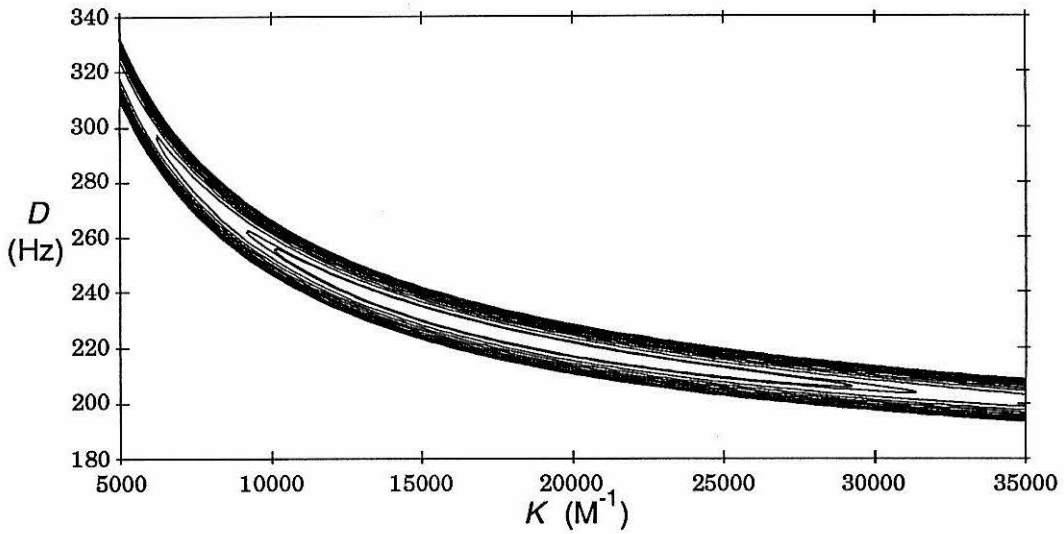


Figure 3. Contour plot of SSR (unweighted) as a function of K and D for an actual data set. Only the contours up to 45 Hz^2 are shown. The innermost contour is 4 Hz^2; the other contours are multiples of 5 Hz^2. The global minimum on this surface is between 3 and 4 Hz^2.

If D is accurately determined in a binding study comprising many samples at a single temperature, consistent use of this value at the other temperatures studied would eliminate this source of fluctuation in the estimates of K. By concentrating experimental efforts on the careful determination of D at one temperature, estimation of K at other temperatures becomes easier and less subject to experimental vagaries. Thus, the constant D assumption allows believable variable-temperature binding data to be quickly and conveniently obtained.

C. Heat Capacity.

Using the constant D assumption, Stauffer obtained variable-temperature binding data for a number of host/guest systems in both borate-*d* and chloroform-*d*. Van't Hoff plots of $R\ln K$ vs. $1/T$ were consistently and unmistakably curved. It was at this point that Greg Simcik, another graduate student in our group, suggested that we were seeing a manifestation of a heat capacity change. Our further analysis

of the thermodynamics of complexation showed that heat capacity was indeed a good additional effect to include in our model.

Since our binding studies are carried out at constant pressure, the heat capacity of interest is C_p, the heat capacity at constant pressure. This is defined as the derivative of enthalpy with respect to temperature, taken at constant pressure.

$$C_p \equiv \left(\frac{\partial H}{\partial T}\right)_p \tag{14}$$

This describes how the enthalpy H of a substance changes with temperature. Clearly, this is the sort of effect needed to account for our van't Hoff plots. If ΔH° of complexation were constant, then the slope of the van't Hoff plots would never change. In reality, however, the van't Hoff plots curve noticeably, so some way to model a change in ΔH° with temperature is desirable. Heat capacity is a logical first choice.

Conveniently, C_p contains information about the change of entropy as well as enthalpy with temperature. The thermodynamic definition of entropy

$$dS = \frac{dq_{\text{rev}}}{T} \tag{15}$$

invokes the heat absorbed in a reversible process. As long as the discussion is restricted to processes occurring at constant pressure, the differential element of heat absorbed, dq, can be identified with the differential element of enthalpy, dH.

$$dS = \frac{dH}{T} \tag{16}$$

Differentiation of both sides of equation 16 with respect to temperature yields equation 17.

$$\frac{dS}{dT} = \frac{1}{T}\frac{dH}{dT} = \frac{1}{T}C_p \tag{17}$$

We are interested not in the absolute values of thermodynamic variables of substances, but in the *changes* in thermodynamic variables wrought by reactions.

Thus, the quantities of interest are ΔG, ΔH, ΔS, and ΔC_p instead of G, H, S, and C_p. ΔG, ΔH, and ΔS are familiar quantities; ΔC_p is not normally discussed. It can be thought of in two ways that are mathematically equivalent but intuitively different. The first, in analogy to the ordinary interpretations of the other quantities, is that ΔC_p is the difference between the heat capacities C_p of the initial and final states of a reaction. The other, in analogy to the definition of C_p, is that ΔC_p is the change with temperature of ΔH.

$$\Delta C_p = \left(\frac{\partial \Delta H}{\partial T}\right)_p \tag{18}$$

The proof that these two definitions are equivalent is left as an exercise to the reader.

Curvature of the van't Hoff plots showed that the naïve assumption of temperature-invariance of $\Delta H°$ and $\Delta S°$ was untenable. The next simplest assumption was that $\Delta C_p°$ itself is independent of temperature. The behavior of van't Hoff plots according to this model can easily be derived.

In order to predict $\Delta G°$ at any temperature, it is first necessary to find $\Delta H°$ and $\Delta S°$. The expression for $\Delta H°$ arises from solving the differential equation 18.

$$\frac{\partial \Delta H}{\partial T} = \Delta C_p$$

$$\partial \Delta H = \Delta C_p \partial T$$

It is most convenient to evaluate this differential equation by a definite integral. Here T is the temperature of interest, and T_0 is a reference temperature at which ΔH is known.

$$\int_{\Delta H_{T_0}}^{\Delta H_T} \partial \Delta H = \int_{T_0}^{T} \Delta C_p \partial T$$

Since ΔC_p is independent of T, it can be taken outside of the integral.

$$\int_{\Delta H_{T_0}}^{\Delta H_T} \partial \Delta H = \Delta C_p \int_{T_0}^{T} \partial T$$

$$\Delta H_T - \Delta H_{T_0} = \Delta C_p (T - T_0)$$

$$\Delta H_T = \Delta H_{T_0} + \Delta C_p (T - T_0) \tag{19}$$

The determination of ΔS° is similar. The differential equation to solve in this case is 20, which is analogous to equation 17.

$$\frac{\partial \Delta S}{\partial T} = \frac{1}{T}\Delta C_p \tag{20}$$

$$\partial \Delta S = \frac{1}{T}\Delta C_p \, \partial T$$

$$\int_{\Delta S_{T_1}}^{\Delta S_T} \partial \Delta S = \Delta C_p \int_{T_1}^{T} \frac{1}{T} \partial T$$

$$\Delta S_T - \Delta S_{T_1} = \Delta C_p \ln\left(\frac{T}{T_1}\right)$$

$$\Delta S_T = \Delta S_{T_1} + \Delta C_p \ln\left(\frac{T}{T_1}\right) \tag{21}$$

Here, T is again the temperature of interest, and T_1 is a reference temperature at which ΔS is known. It is not necessary for T_0 and T_1 to be the same. In fact, it is arithmetically convenient to set $T_0 = 0$ K and $T_1 = 1$ K. Equations 19 and 21 then become 22 and 23.

$$\Delta H_T = \Delta H_0 + \Delta C_p(T - 0) = \Delta H_0 + T\Delta C_p \tag{22}$$

$$\Delta S_T = \Delta S + \Delta C_p \ln\left(\frac{T}{1}\right) = \Delta S_1 + \Delta C_p \ln T \tag{23}$$

These expressions can then be substituted into the definition of ΔG.

$$\Delta G_T^\circ = \Delta H_T^\circ - T\Delta S_T^\circ = \Delta H_0^\circ + T\Delta C_p^\circ - T(\Delta S_1^\circ + \Delta C_p^\circ \ln T)$$

$$\Delta G_T^\circ = \Delta H_0^\circ + T\Delta C_p^\circ - T\Delta S_1^\circ - T\Delta C_p^\circ \ln T \tag{24}$$

A van't Hoff plot of $R\ln K$ should then show behavior described by

$$R\ln K = -\frac{\Delta G^\circ}{T} = -\frac{\Delta H_0^\circ}{T} - \Delta C_p^\circ + \Delta S_1^\circ + \Delta C_p^\circ \ln T$$

$$= -\frac{\Delta H_0^\circ}{T} - \Delta C_p^\circ \ln\left(\frac{1}{T}\right) - \Delta C_p^\circ + \Delta S_1^\circ . \tag{25}$$

Taking the independent variable as $1/T$ and the dependent variable as $R\ln K$, this equation is in the form

$$y = Ax + B\ln x + C\,, \tag{26}$$

where A, B, and C are unknown constants. The model is a linear function of these three parameters.[27,28] The best parameters in a least squares sense can thus be found in one step by simple linear regression. I consequently wrote a Macintosh program (vantHoff) that performs both the linear and log fits to a set of variable-temperature binding data. All of our analyses were carried out by using this program.

D. Regression Analysis.

1. Comparing regression models. In principle and in practice, it is no more difficult to fit a plot of $R\ln K$ vs. $1/T$ with the three parameters $-\Delta H_0^\circ$, ΔC_p°, and $\Delta S_1^\circ - \Delta C_p^\circ$ than with the two parameters ΔH° and $-\Delta S^\circ$. In many cases, the fits dramatically improve when the third parameter is included, as exemplified by Figure 4. This figure shows the data of ATMA + V_{meso} seen earlier and the best-fit curves to this data following equations 5 and 25. These equations will hereafter be referred to as the "linear" and "log" equations, respectively. The log equation clearly fits this data set better than the linear equation, and the residuals from the log fit have no noticeable pattern. This elimination of the pattern to the residuals is strong qualitative evidence that the additional parameter in the log equation accounts for a real effect.

---

(27) This analysis is not novel. See, for example, Everett, D. H.; Wynne-Jones, W. F. K. "The thermodynamics of acid-base equilibria," *Trans. Faraday Soc.* **1939**, *35*, 1380–1401; Clarke, E. C. W.; Glew, D. N. "Evaluation of thermodynamic functions from equilibrium constants," *Trans. Faraday Soc.* **1965**, *62*, 539–547.

(28) Ives, D. J. G.; Marsden, P. B. "The ionization functions of diisopropylcyanoacetic acid in relation to hydration equilibria and the compensation law," *J. Chem. Soc.* **1965**, *62*, 649–676.

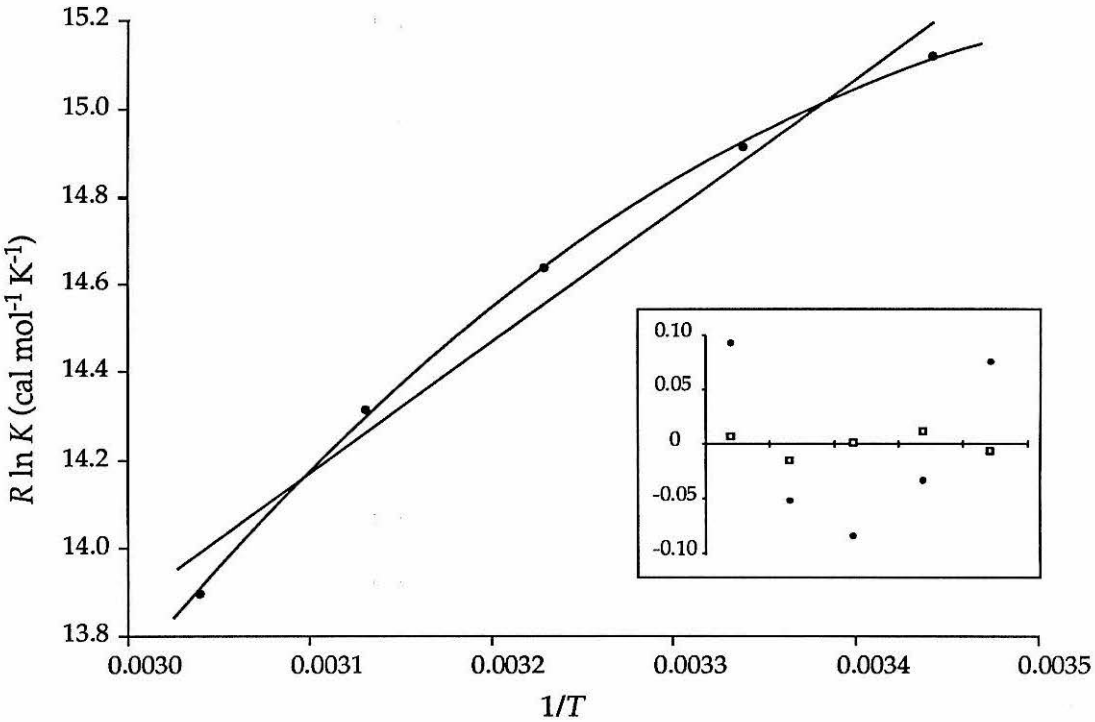


Figure 4. Left: data of Figure 2 with the best-fit curve according to the log equation. Inset: Residuals from the fit by the linear equation (filled circles) and the log equation (open squares).

An improved fit to the data brought about by modeling with an equation containing additional adjustable parameters should always be greeted with grave suspicion. Adding a parameter to a model will *never* make the agreement between the data and model any worse, and is nearly certain to improve the agreement. In general, a data set with N points can be modeled exactly by an equation with N adjustable parameters. It is thus no surprise that the log equation always fits experimental van't Hoff plots better than the linear equation does. The ability of the log equation to remove the gross curvature of the plotted residuals strongly argues that modeling the effect of ΔC_p° is justified, but it provides no quantitative demonstration of this fact.

2. F-test for the significance of regression. There is a standard statistical test, however, that furnishes a means for a quantitative comparison between the

linear and log models.[29] It is based, as might be expected, on the sums of the squared residuals from both fits. This F-test for the significance of regression will be intuitively explained here.

Consider a model containing k parameters. The equation describing this model is

$$y^{\star}(\mathbf{x}_i) = a_1^{\star} x_{i1} + a_2^{\star} x_{i2} + \cdots + a_k^{\star} x_{ik} \,,$$

where $\mathbf{x}_i$ is the ith vector of independent variables, $\mathbf{a}^{\star}$ is the model's parameter vector, and $y^{\star}(\mathbf{x}_i)$ is the value of the dependent variable y predicted by the model at the given x-values $\mathbf{x}_i$. Let there also be another model containing $k+l$ parameters. The equation describing this model is

$$y(\mathbf{x}_i) = a_1 x_{i1} + a_2 xi2 + \cdots + a_{k+l} x_{i\,k+l} \,,$$

where the parameter vector is $\mathbf{a}$ and the prediction is $y(\mathbf{x}_i)$. It is possible to construct a statistic to evaluate the null hypothesis that the model $y(\mathbf{x}_i)$ is no more statistically significant than the less complicated model $y^{\star}(\mathbf{x})$.

A data set with N points can be fit by these two models, giving the goodness-of-fit scores SSR and SSR*.

$$\mathrm{SSR}^{\star} = \sum_{i=1}^{N} \left(y^{\star}(\mathbf{x}_i) - y_i \right)^2 \tag{27}$$

$$\mathrm{SSR} = \sum_{i=1}^{N} \left(y(\mathbf{x}_i) - y_i \right)^2 \tag{28}$$

If the null hypothesis is true, then the behavior of the system is described by

$$y_i = y^{\star}(\mathbf{x}_i) + \varepsilon_i \,, \tag{29}$$

where the errors ε are independent and identically distributed normal variables with a mean of zero and a variance of σ^2. The statistic $\mathrm{SSR}^{\star}/\sigma^2$ will follow the χ^2

(29) Hoel, Paul G.; Port, Sidney C.; Stone, Charles J. *Introduction to Statistical Theory*; Houghton Mifflin: Boston, 1971; p 142–147.

distribution with $N-k$ degrees of freedom. A χ^2 variable with n degrees of freedom is the sum of squares of n independent standard normal variates. The mean value of such a χ^2_n variable is n. Including l additional parameters in the model will make the model fit a little better, but the improvement will be small; the extra parameters are fitting only random scatter, not a true functional relation. Thus, the fit score SSR from this more complicated model will still follow a χ^2 distribution multiplied by σ^2, but with l fewer degrees of freedom. The specific expected distribution of SSR is $\sigma^2\chi^2_{N-k-l}$.

Another class of random variable that must be introduced to construct this test is F. An F random variable with m degrees of freedom in the numerator and n degrees of freedom in the denominator is defined as the ratio of a normalized χ^2 variable with m degrees of freedom and a normalized χ^2 variable with n degrees of freedom.

$$F_{m,n} \equiv \frac{\chi^2_m/m}{\chi^2_n/n} \tag{30}$$

A normalized χ^2 variable is a χ^2 variable divided by its number of degrees of freedom, so that its mean value is 1. Since both the numerator and denominator of an F variable have an expectation of 1, all F variables likewise have expectation 1.

It is possible to build several F-statistics based on SSR* and SSR. An especially powerful such statistic is given in equation 31.

$$F = \frac{(\mathrm{SSR}^\star - \mathrm{SSR})/l}{\mathrm{SSR}/(N-k-l)} \tag{31}$$

First, it is necessary to explain why this statistic should follow an F distribution. Recall that under the null hypothesis of no significance of regression, SSR* and SSR behave as $\sigma^2\chi^2_{N-k}$ and $\sigma^2\chi^2_{N-k-l}$ variables, respectively. Thus,

$$\begin{aligned}\mathrm{SSR}^\star - \mathrm{SSR} &\sim \sigma^2\chi^2_{N-k} - \sigma^2\chi^2_{N-k-l} \\ &\sim \sigma^2(\chi^2_{N-k} - \chi^2_{N-k-l}).\end{aligned}$$

When two χ^2 variables are added or subtracted, the resulting variable also follows a χ^2 distribution. Its number of degrees of freedom is just the sum or difference of the degrees of freedom of the initial variables. Consequently, the statistic F is distributed as

$$F \sim \frac{\sigma^2 \chi_l^2 / l}{\sigma^2 \chi_{N-k-l}^2 / (N-k-l)}$$

eliminating the factor of σ^2/σ^2 gives

$$F \sim \frac{\chi_l^2 / l}{\chi_{N-k-l}^2 / (N-k-l)} \sim F_{l,N-k-l}\,.$$

This statistic F has been shown to be the ratio of two normalized χ^2 statistics; if the null hypothesis is correct, it will follow the known $F_{l,N-k-l}$ distribution. In the present case, $y^\star(\mathbf{x}_i)$ is the linear equation, for which $\mathbf{a}^\star = (-\Delta S^\circ, \Delta H^\circ)$, $\mathbf{x}_i = (1, 1/T_i)$, and $k = 2$, and $y(\mathbf{x}_i)$ is the log equation, for which $\mathbf{a} = (\Delta S_1^\circ - \Delta C_p^\circ, -\Delta C_p^\circ, -\Delta H_0^\circ)$ and $\mathbf{x}_i = (1, \ln(1/T_i), 1/T_i)$. The specific test statistic then is

$$F = \frac{\mathrm{SSR}^\star - \mathrm{SSR}}{\mathrm{SSR}/(N-3)}\,, \tag{32}$$

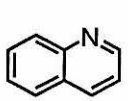
and the appropriate theoretical distribution for comparison is the $F_{1,N-3}$ distribution.

The numerator of this F-statistic is the improvement in the fit caused by adding a heat capacity term to the linear equation. The denominator is the normalized sum of squared residuals remaining after the log fit. If the ΔC_p° term is physically meaningful, one would expect the numerator to be large and the denominator to be small. This is in contrast to the null hypothesis, which states that the test statistic should follow an F distribution, which has an average value of 1. Adding one adjustable parameter to the model is expected to reduce the SSR by only a factor of $(N-2)/(N-3)$. A greater reduction indicates that the null hypothesis is untenable, that is, that adding the extra parameter is statistically justified. The significance

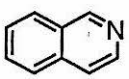
of this effect is quantified by comparing F to the theoretical $F_{1,N-3}$ distribution. If a variable following this distribution is expected to exceed the experimental value of F only some small fraction of the time α, then, to $(1-\alpha)\times 100\%$ confidence, the assumption of the null hypothesis is incorrect. For convenience, I shall call this confidence "p." Using this F-test, it is possible to determine if the improvement in the fit brought about by including ΔC_p° in the model is statistically significant. The confidence p is the probability that the improvement is not due to random chance.[30]

E. Results.

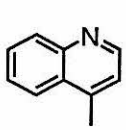
Table I reports values of ΔC_p°, ΔH_0°, and ΔS_1° of complexation for a number of host/guest pairs in chloroform-*d* and in borate-*d*. In addition, Figures 5 through 11 show the van't Hoff plots for some systems unarguably showing a heat capacity effect.



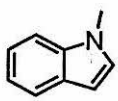
1



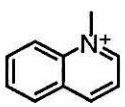
2



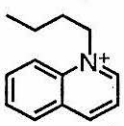
3



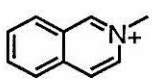
4



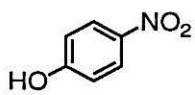
5



6



7



8

These studies clearly demonstrate that ΔH° and ΔS° of these complexation reactions are temperature-dependent, and that the model of the behavior of $R\ln K$ as a function of $1/T$ is substantially improved by including a constant ΔC_p° term. They raise the question, however, of the physical implications of this effect.

---

(30) The application of an F-test to a thermodynamic model including ΔC_p° is not novel; see Reference 28.

Table I. Thermodynamic parameters of association of organic guests with ethenoanthracene hosts in borate-*d* and chloroform-*d*.[a]

host	guest	ΔG°_{298} [b]	ΔH°_{298} [b]	ΔS°_{298} [c]	ΔC_p° [c]
		in borate-*d*			
P	ATMA	-7.3	-4.7	8.6	-100
	1	-6.0	-11	-17	-12
	2	-6.4	-9.8	-11	-25
	3	-7.1	-9.8	-9.1	-130
	4	-4.0	-1.6	8.1	-120
C	ATMA	-5.6	-1.3	14	-110
	1	-6.4	-7.5	-3.8	-39
	2	-6.3	-2.9	11	-61
	3	-6.3	-11.0	24	-190
	4	-5.0	0.3	10	-120
M	ATMA	-6.4	-3.4	1.0	-130
V	ATMA	-5.5	-4.9	2.2	-34
		in $CDCl_3$			
P_E	ATMA	-2.1	-1.5	1.8	-18
	5	-3.5	-3.3	0.4	-24
	6	-2.5	-3.6	-3.8	-24
	7	-2.4	-2.4	0	-19

[a]Data are from reference 23. [b]In kcal mol^{-1}. [c]In cal mol^{-1} K^{-1}.

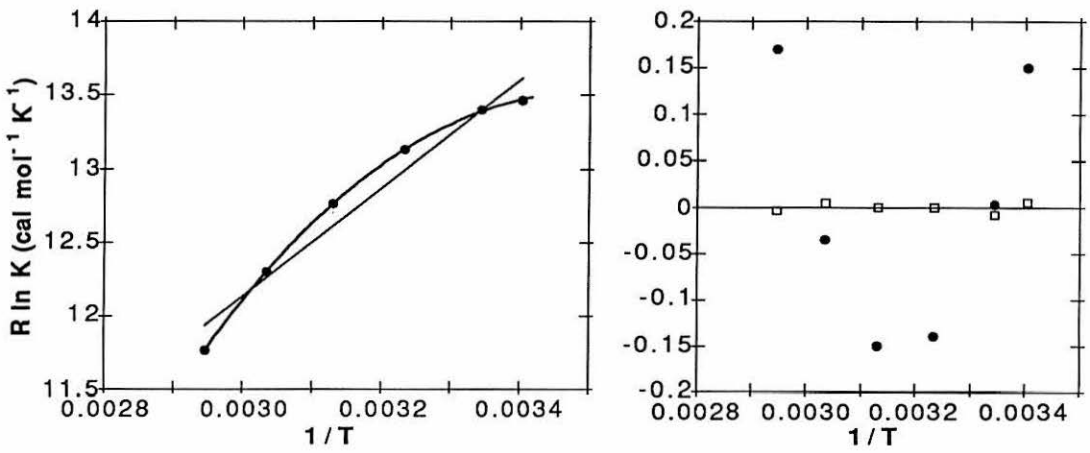


Figure 5. Variable-temperature data of host P_R with guest **4** in borate-*d*. $\Delta H^\circ_{298} = -1.6$ kcal mol^{-1}, $\Delta S^\circ_{298} = +8.1$ cal mol^{-1} K^{-1}, $\Delta C_p^\circ = -120$ cal mol^{-1} K^{-1}, $F_{1,3} = 2334$, $p = 99.998\%$.

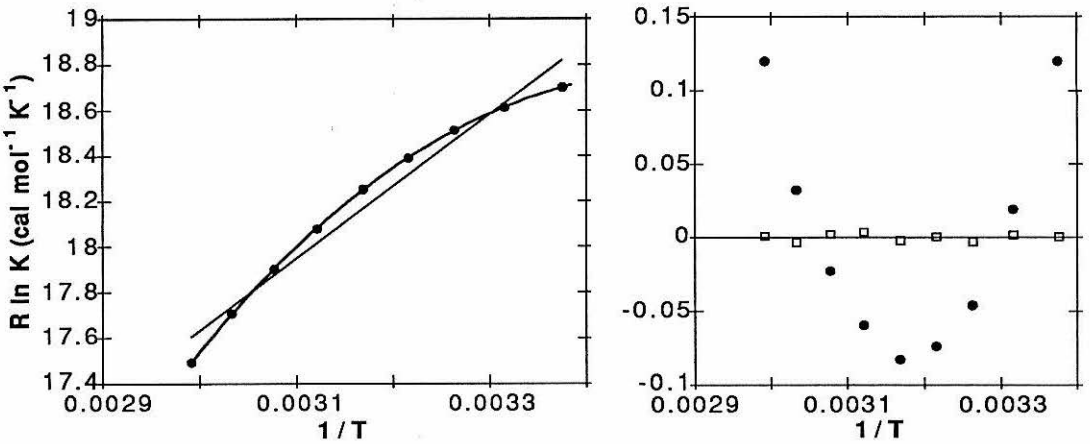


Figure 6. Variable-temperature data of host C_r with guest ATMA in borate-*d*. $\Delta H^\circ_{298} = -1.3$ kcal mol^{-1}, $\Delta S^\circ_{298} = +14$ cal mol^{-1} K^{-1}, $\Delta C^\circ_p = -110$ cal mol^{-1} K^{-1}, $F_{1,6} = 5924$, $p =$ "nine nines."

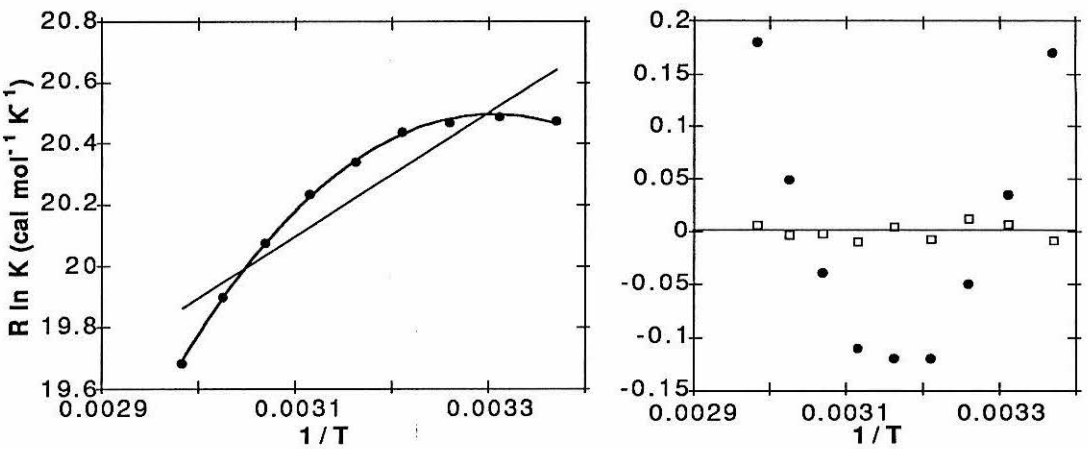


Figure 7. Variable-temperature data of host C_R with guest **3** in borate-*d*. $\Delta H^\circ_{298} = +1.0$ kcal mol^{-1}, $\Delta S^\circ_{298} = +23$ cal mol^{-1} K^{-1}, $\Delta C^\circ_p = -160$ cal mol^{-1} K^{-1}, $F_{1,6} = 1294$, $p =$ "seven nines."

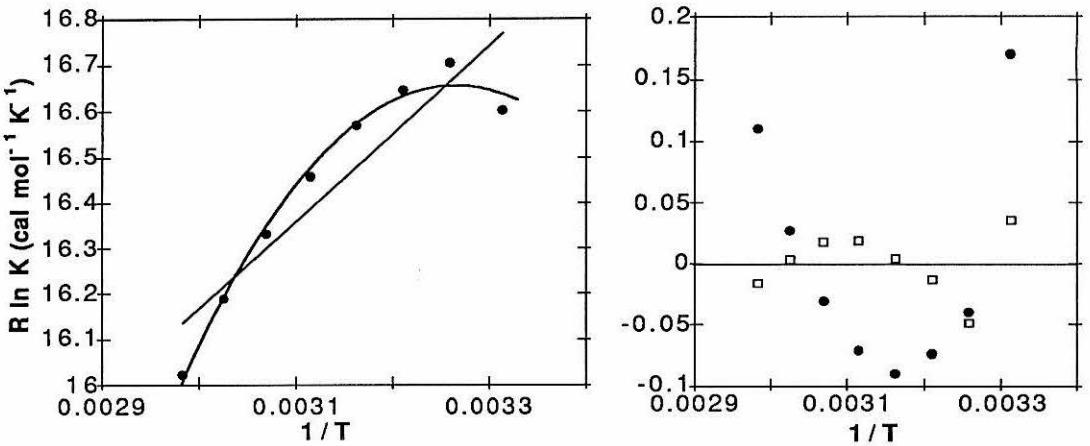


Figure 8. Variable-temperature data of host C_R with guest **4** in borate-*d*. $\Delta H^\circ_{298} = +1.4$ kcal mol^{-1}, $\Delta S^\circ_{298} = +21$ cal mol^{-1} K^{-1}, $\Delta C^\circ_p = -160$ cal mol^{-1} K^{-1}, $F_{1,5} = 59.788$, $p = 99.94\%$.

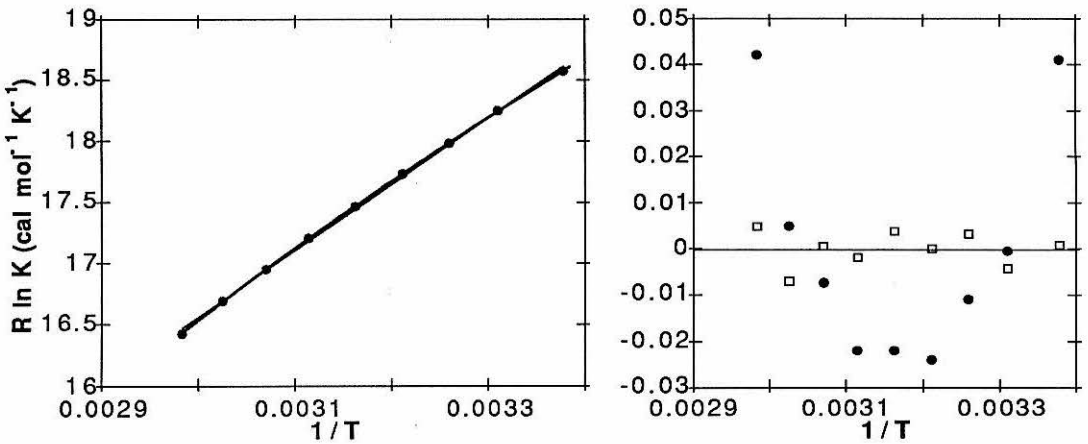


Figure 9. Variable-temperature data of host V_{dl} with guest ATMA in borate-*d*. $\Delta H^\circ_{298} = -4.9$ kcal mol^{-1}, $\Delta S^\circ_{298} = +2.2$ cal mol^{-1} K^{-1}, $\Delta C^\circ_p = -34$ cal mol^{-1} K^{-1}, $F_{1,6} = 955$, $p =$ "seven nines."

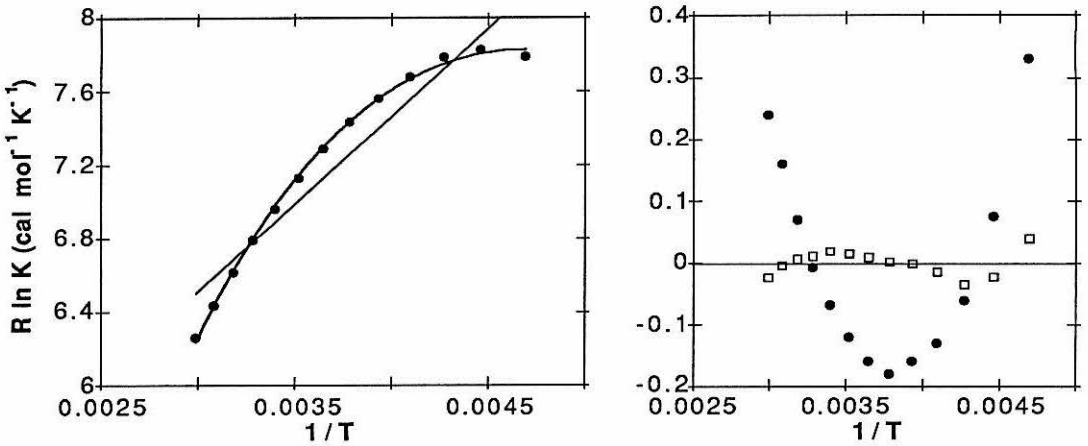


Figure 10. Variable-temperature data of host P_{ER} with guest ATMA in $CDCl_3$. $\Delta H^{\circ}_{298} = -1.5$ kcal mol^{-1}, $\Delta S^{\circ}_{298} = -1.8$ cal mol^{-1} K^{-1}, $\Delta C^{\circ}_{p} = -18$ cal mol^{-1} K^{-1}, $F_{1,10} = 655$, $p =$ "nine nines."

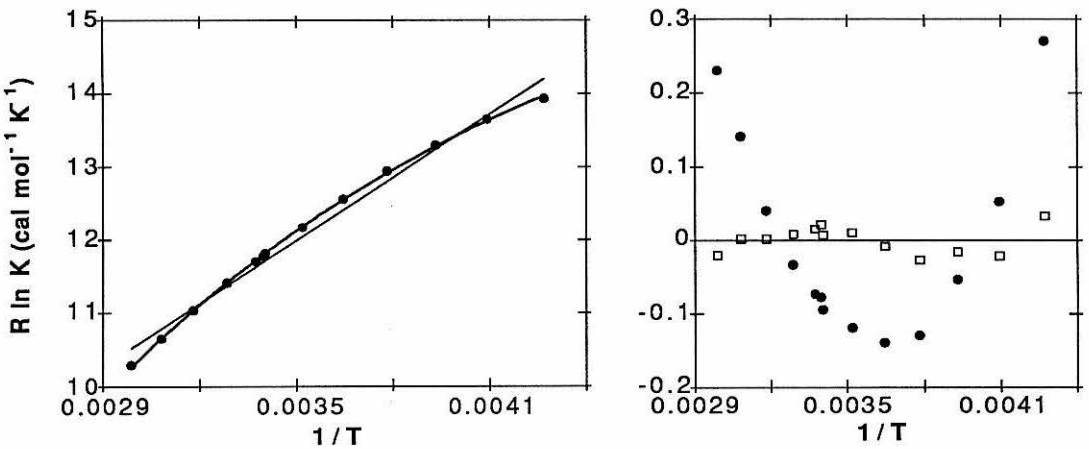


Figure 11. Variable-temperature data of host P_{ER} with guest **5** in $CDCl_3$. $\Delta H^{\circ}_{298} = -3.4$ kcal mol^{-1}, $\Delta S^{\circ}_{298} = 0.4$ cal mol^{-1} K^{-1}, $\Delta C^{\circ}_{p} = -24$ cal mol^{-1} K^{-1}, $F_{1,10} = 565$, $p =$ "nine nines."

III. Interpretation of Variable-Temperature Binding Data.

A. $\Delta H°$ and $\Delta S°$.

Analyses of thermodynamic trends of these aqueous and chloroform complexation reactions have already been published.[23] Customary analysis of the free energy of a reaction according to its enthalpy and entropy components is folly in the face of a significant heat capacity change. A ΔC_p° of 100 entropy units, for example, means that $\Delta H°$ gains 1 kcal mol^{-1} every ten degrees. Similarly, at 300 K, a ten-degree rise in temperature also raises $T\Delta S°$ by about 1 kcal mol^{-1}. The net effect on $\Delta G°$ is nearly zero, but the relative contributions by $\Delta H°$ and $\Delta S°$ change significantly over a narrow temperature range. Figure 12 traces the energetics of a hypothetical system with a $\Delta G°$ of -6 kcal mol^{-1} at 298 K, evenly partitioned at this temperature between $\Delta H°$ and $\Delta S°$. A modest ΔC_p° of -100 cal mol^{-1} K^{-1} transforms this system from being entropically driven at 15 °C to being enthalpically driven at temperatures above 25 °C. In the fifty-degree temperature range between 15 and 65 °C, $\Delta H°$ and $T\Delta S°$ have each changed by about 5 kcal mol^{-1}, while $\Delta G°$ has not varied more than 500 small calories. Interpreting such thermodynamic behavior in terms of intrinsic attractions and changes in conformational freedom implies that the process profoundly and smoothly changes its very mechanism over this narrow temperature range. Such a picture is inconsistent with chemical intuition. Clearly, some significant process associated with the complexation event has not been accounted for.

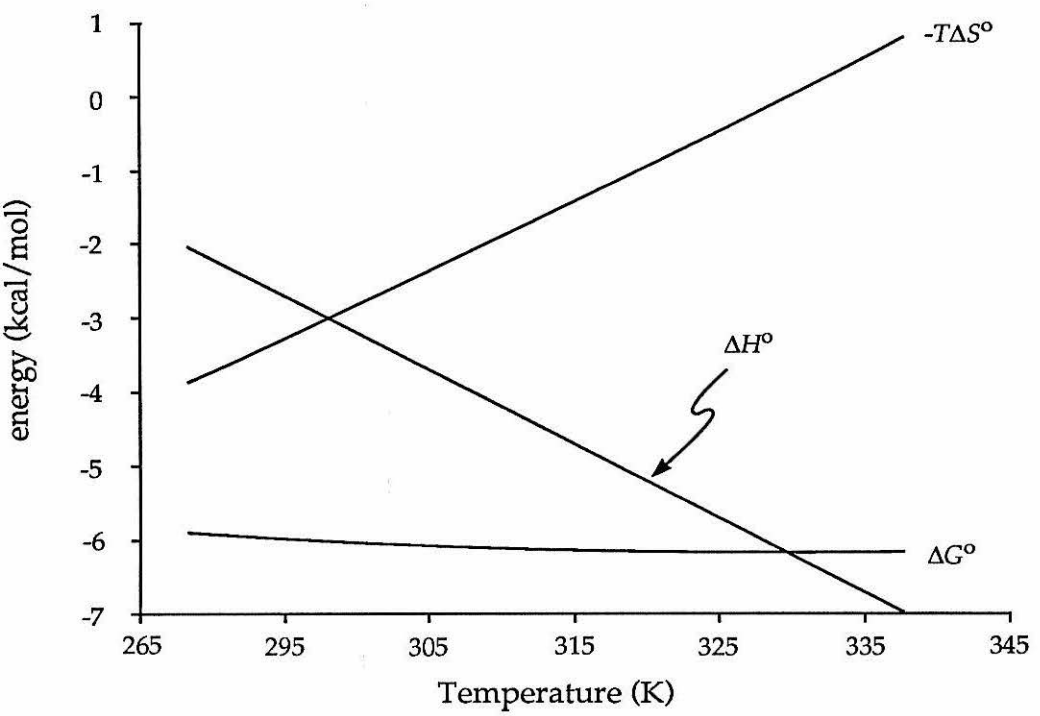


Figure 12. Plot of $\Delta G°$, $\Delta H°$, and $\Delta S°$ as a function of temperature for a system with a $\Delta C_p°$ of -100 cal mol^{-1} K^{-1}. $\Delta G°$ at 298 K is 6 kcal mol^{-1}, half of which is from $\Delta H°$, and half from $-T\Delta S°$.

B. Hydrophobic Hydration.

For the aqueous systems, this missing feature is probably hydrophobic hydration. Purposely omitted from the earlier discussion of the hydrophobic effect was that the dissolution of nonpolar solutes in water is accompanied not only by a slightly favorable enthalpy change and a very unfavorable entropy change, but also by a positive heat capacity change.[5] Since complexation of two nonpolar solutes in water is qualitatively akin to a reversal of a dissolution, one expects hydrophobic association phenomena to be associated with a *negative* heat capacity change. Such changes have indeed long been recognized in biological processes such as protein

denaturation and enzyme-cofactor association.[17,2,13,16,31–33]

The heat capacity change associated with the hydrophobic effect can be attributed to the hydrophobic hydration shell formed around a nonpolar solute. It certainly cannot be a result of internal motions of the solute molecules, because the heat capacity change upon transfer of a nonpolar solute to water is larger than the solute's total heat capacity.[5] There are several current explanations of this hydrophobic heat capacity change, each of which appears to be held with religious conviction by its proponents. All of these theories share the concept that the water molecules of a hydrophobic hydration shell undergo some process for which $\Delta H°$ and $\Delta S°$ oppose each other.[5,8,14,12] The conceptually simplest such picture treats the hydrophobic hydration shell as an iceberg. A hydration shell, like an iceberg, may undergo a "melting" transition in which its rigid clathrate structure gives way to a looser, more fluxional configuration. This melting process absorbs heat but also increases the conformational mobility of the shell: it is enthalpically unfavorable and entropically favorable. The magnitudes of $\Delta H°$ and $\Delta S°$ for this melting process are proportional to the number of water molecules in the hydration shell. An increase in temperature increases the relative importance of the entropic contribution, so a greater fraction of hydration shells will be molten at high temperatures than at low temperatures.

---

(31) Sturdevant, Julian M. "Heat capacity and entropy changes in processes involving proteins," *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 2236–2240.

(32) Kellis, James T. Jr.; Nyberg, Kerstin; Šali, Daša; Fersht, Alan R. "Contribution of hydrophobic interactions to protein stability," *Nature* **1988**, *333*, 784–786.

(33) Eftink, Maurice R.; Anusiem, A. C.; Biltonen, Rodney L. "Enthalpy-entropy compensation and heat capacity changes for protein-ligand interactions: general thermodynamic models and data for the binding of nucleotides to Ribonuclease A," *Biochemistry* **1983**, *22*, 3884–3896.

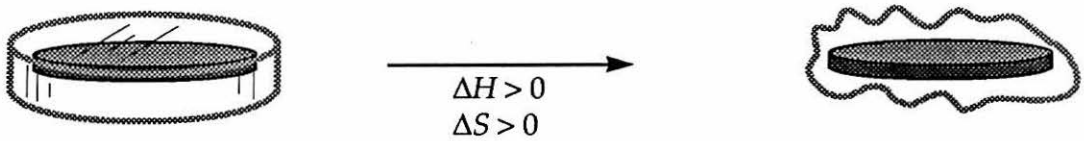


Figure 13. Phase transition (melting) of a hydration shell around a nonpolar solute in water. The presence of ice-like hydration shells increases the heat capacity of the system.

Consequently, a solution of nonpolar molecules in water will absorb an inordinate amount of heat as the temperature rises. The quantity of heat absorbed depends on the number of water molecules in the melting hydration shells. This absorption of heat corresponds to an increase in the enthalpy of the total solution; the increase of enthalpy with temperature is the heat capacity. This can qualitatively be compared to a mixture of ice and water at equilibrium. The heat capacity of such a system is large; in fact, it is infinite. Heat added to the system is consumed by the melting of ice, so that the temperature does not increase. An aqueous solution of a nonpolar substance behaves in a similar manner. Some of the heat added to the solution is absorbed by melting hydrophobic hydration shells.

A hydrophobic association process will involve a decrease in the heat capacity of the solution. Association of hydrophobic solutes reduces the hydrophobic surface area in contact with solvent. The total number of water molecules involved in hydrophobic hydration shells correspondingly drops; the large heat capacity resulting from the order $\rightarrow$ disorder transition of the bound waters will thus no longer be present. This is illustrated in Figure 14. The presence of bound waters in hydration shells increases a solution's heat capacity; releasing waters from hydration shells lowers a solution's heat capacity.

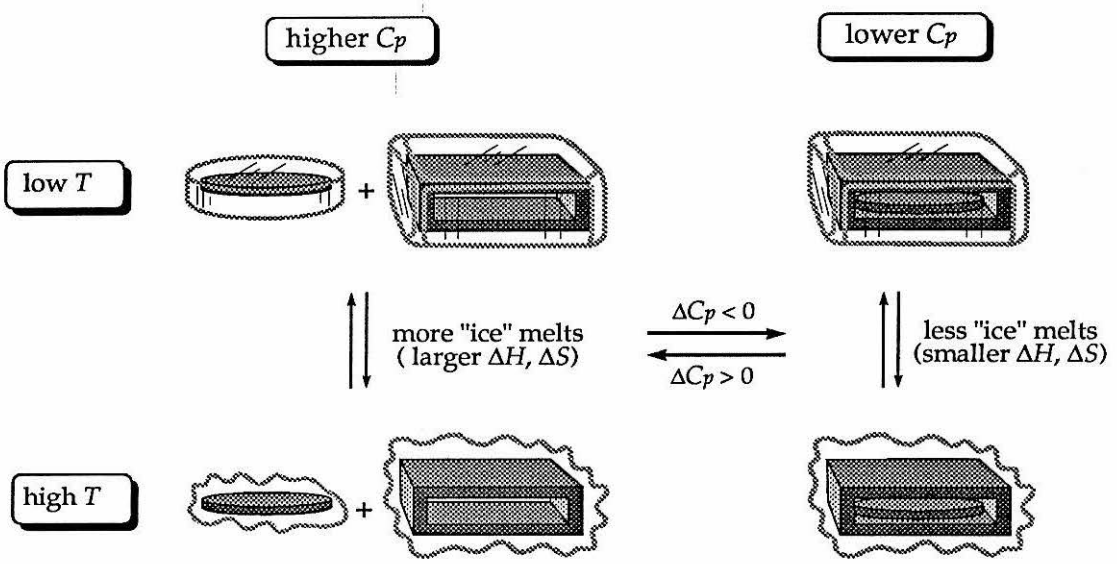


Figure 14. Association of host and guest reduces the number of water molecules involved in hydrophobic hydration shells. This lowers the heat capacity of the system.

C. Behavior of ΔC_p°.

Such an interpretation implies that ΔC_p° is not independent of temperature. Instead, its magnitude should decrease as the temperature increases. As a greater fraction of hydration shells become molten, fewer icy shells remain to melt if the temperature rises further. This aspect of the hydrophobic interaction model is not reflected in the log equation. To investigate the effect a *change* in ΔC_p° would have on our van't Hoff plots, I manufactured a set of (T, K) data that would result from a system for which ΔG° at 298 K was -6 kcal mol^{-1}, with equal contributions from ΔH° and $-T\Delta S^\circ$. ΔC_p° at this temperature was set at -100 cal mol^{-1} K^{-1}, and $d\Delta C_p^\circ/dT$, the change in ΔC_p° with temperature, at 2 cal mol^{-1} K^{-2}. Another hypothetical data set, with the same values at 298 K of ΔG°, ΔH°, ΔS°, and ΔC_p°, but with $d\Delta C_p^\circ/dT$ equal to -2 cal mol^{-1} K^{-2}, was also generated. Subjecting these data sets to analysis according the the linear and log equations gave the fits and residuals plots of Figure 15.

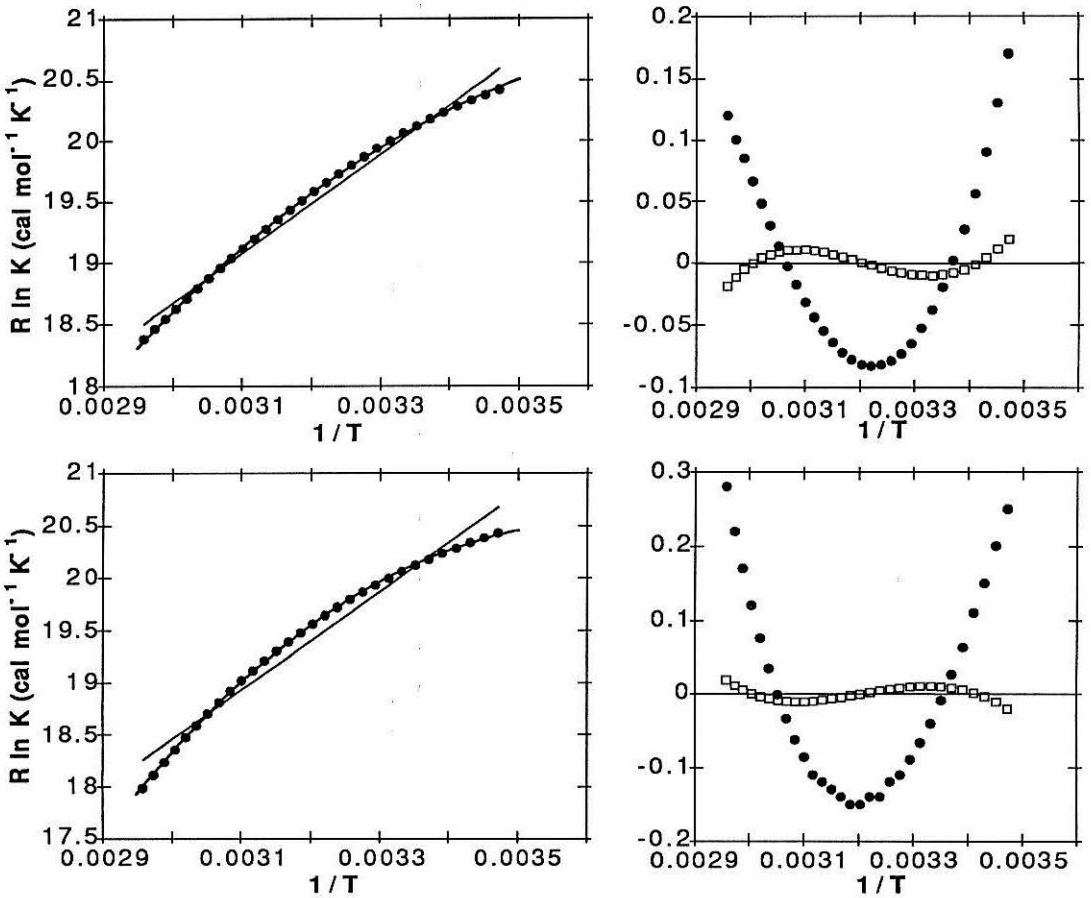


Figure 15. Van't Hoff plots of hypothetical data sets arising if ΔC_p° varies linearly with temperature. Top: $d\Delta C_p^\circ/dT$ is +2 e.u./K. Bottom: $d\Delta C_p^\circ/dT$ is −2 e.u./K. Right: Residuals plots from fitting the linear and log equations to these data sets.

One feature of the residuals plots in this figure is immediately striking: both have an unmistakably sigmoidal pattern. The residuals from the $d\Delta C_p^\circ/dT = -2$ data set have their concave down portion to the right; this portion is on the left in the other data set. Examination of the log residuals plots in Figures 5 through 11 reveals that those plots that do not appear to be random scatter have their concave up portion to the right. This is similar to the residuals plot in Figure 15 for the data set having $\Delta C_p^\circ/dT = +2$. This data set is the most consistent

with the two-state hydration shell model, since the magnitude of ΔC_p° decreases with increasing temperature. Thus, the patterns to the residuals from the log fits suggest, in acordance with the two-state model, that ΔC_p° is in fact not independent of temperature.

D. Further Models.

Nonetheless, I have made no attempt to model a temperature-dependence of ΔC_p° in an analysis of van't Hoff plots. Certainly, adding one more parameter for $d\Delta C_p^\circ/dT$ would not be appropriate; the patterns of the residuals from fits to the actual data shown in Figures 5 to 11 are more complicated than the patterns resulting from a constant $d\Delta C_p^\circ/dT$. Even in the simplistic two-state hydration shell model, the change in heat capacity with temperature is not constant. In order to gain any insight into the binding process, a more realistic model would have to be adopted. Unfortunately, the exact functional dependence of hydrophobic heat capacity upon temperature is a topic of current debate.[12,14,34] Furthermore, it would be inappropriate to fit our van't Hoff data to different models in order to determine which model is best. The quality of our data is insufficient for such analysis, and less complicated systems would be more suited to such a study. The mathematically simple model of a constant ΔC_p° fits our data fairly well, and is sufficient to warn against making rash pronouncements about reaction mechanisms on the basis of enthalpy and entropy changes.

E. Conclusions.

The interpretation of our curved van't Hoff plots has focused on hydrophobic hydration, a phenomenon considered to be unique to water. However, we have also observed indisputable heat capacity changes from complexation reactions in chloroform solution. The only noticeable difference between the variable-temperature

(34) Hearn, R. P.; Richards, F. M.; Sturdevant, J. M.; Watt, G. D. "Thermodynamics of the binding of S-peptide to S-protein to form Ribonuclease S′," *Biochemistry* **1971**, *10*, 806–817.

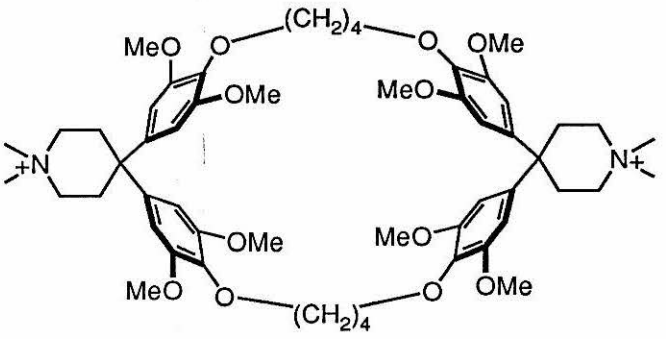
behaviors of complexation reactions in these two dissimilar solvents is that ΔC_p° values are smaller in chloroform. Qualitatively, however, reactions in water and chloroform are indistuinguishable. ΔC_p° is always negative, and the residuals from the log fit, when they have clear patterns, are in the direction consistent with the magnitude of ΔC_p° decreasing with increasing temperature.

It is not clear why complexations are so similar in these two solvents. Our interpretation of the behavior in water hinges on effects that must originate in the solvent rather than in the solute. Similar behavior occurring in chloroform implies that chloroform either has more water-like properties than is commonly supposed, or that the present interpretation of ΔC_p° in terms of hydrophobic interactions is specious. One consideration that makes the former possibility more palatable is that all of the guests that bind to our hosts in chloroform are cationic. It is conceivable that chloroform solvent in the vicinity of the guest/iodide ion pair is highly ordered in a sort of solvophobic solvation shell. Encapsulation of the charged portion of the guest by the macrocyclic host decreases the number of solvent molecules involved in such shells. If these solvophobic solvation shells have heat-absorbing properties similar to those of hydrophobic hydration shells, the complexation process in chloroform solvent would show the same thermodynamic behavior as in aqueous solvent.

IV. Epilogue

A gratifying consequence of our investigations in this field began with a communication we received from François Diederich. His group had previously obtained van't Hoff plots for complexation reactions of their hosts that were perfectly linear.[19] After the present work[23] was published, he sent us about twenty sets of his group's variable-temperature data which, as he claimed, appeared to give linear plots of $R\ln K$ vs. $1/T$. When these data sets were modeled by the log equation, however, one of them, of host **9** with guest **8**, showed a significant heat capacity effect. Figure

16 shows the van't Hoff plot of this data set with its best-fit curve according to the linear equation. The fit appears good upon visual inspection, but examination of the residuals reveals the curved pattern diagnostic of a heat capacity change. Figure 17 shows the best-fit curve from the log equation to this data set, as well as the F statistic and p score for the significance of regression. The best-fit ΔC_p° in this case is -153 cal mol^{-1} K^{-1}, and the inclusion of the heat capacity parameter is valid to 99.5% confidence.



9

Diederich's group subsequently undertook a comprehensive calorimetric study of the binding reactions of their hosts.[35] They found that ΔH° indeed decreased as temperature increased. For the particular system for which our analysis of variable-temperature binding data found a ΔC_p° of -153 cal mol^{-1} K^{-1}, they measured a ΔC_p° of -130 ± 20 cal mol^{-1} K^{-1}. This agreement validates our use of the experimentally simpler log equation.

Although estimating ΔH° and ΔC_p° by using the log equation is not as direct as a calorimetric study, it is a much simpler method to apply. This analysis, combined with the significance-of-regression F test and a critical eye toward the residuals plots, is able to adequately explain the complexation reactions of our hosts.

---

(35) Smithrud, David B.; Wyman, Tara B.; Diederich, François N. "Enthalpically driven cyclophane-arene inclusion complexation: solvent-dependent calorimetric studies," *J. Am. Chem. Soc.* **1991**, *113*, 5420–5426.

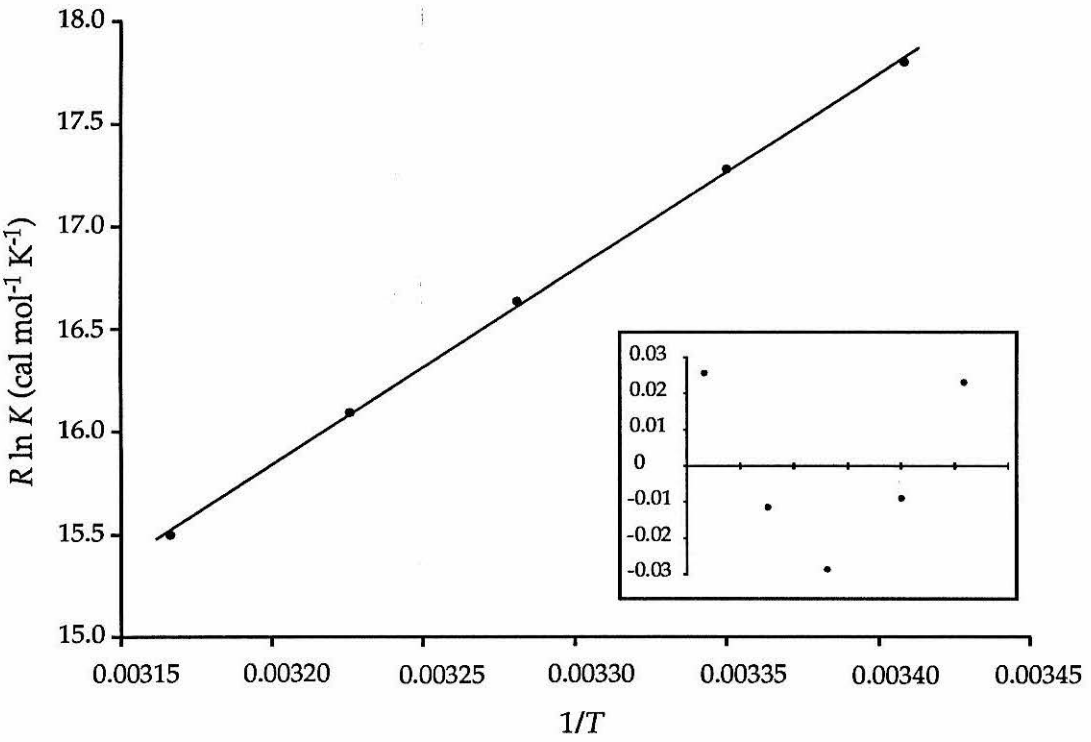


Figure 16. One of François Diederich's data sets fit by the linear equation. Inset: residuals from the fit.

Appendix A. Error Analysis.

One concern about the estimates of ΔC_p° found by the log fit is that they may in fact be far from the true ΔC_p° values. Because ΔC_p° found in this way is a fitted parameter instead of a directly-measured quantity, its accuracy is not immediately obvious.

It is typically possible, when using a linear model, to obtain estimates for the variances of all the fitted parameters. This is accomplished by multiplying the diagonal elements of the inverted normal matrix by the empirical variance from the fit.[36] Such an analysis could easily be carried out for the parameters of the linear and

(36) Dunteman, George H.; *Introduction to Linear Models*; Sage: Beverly Hills, 1984; pp 257–175.

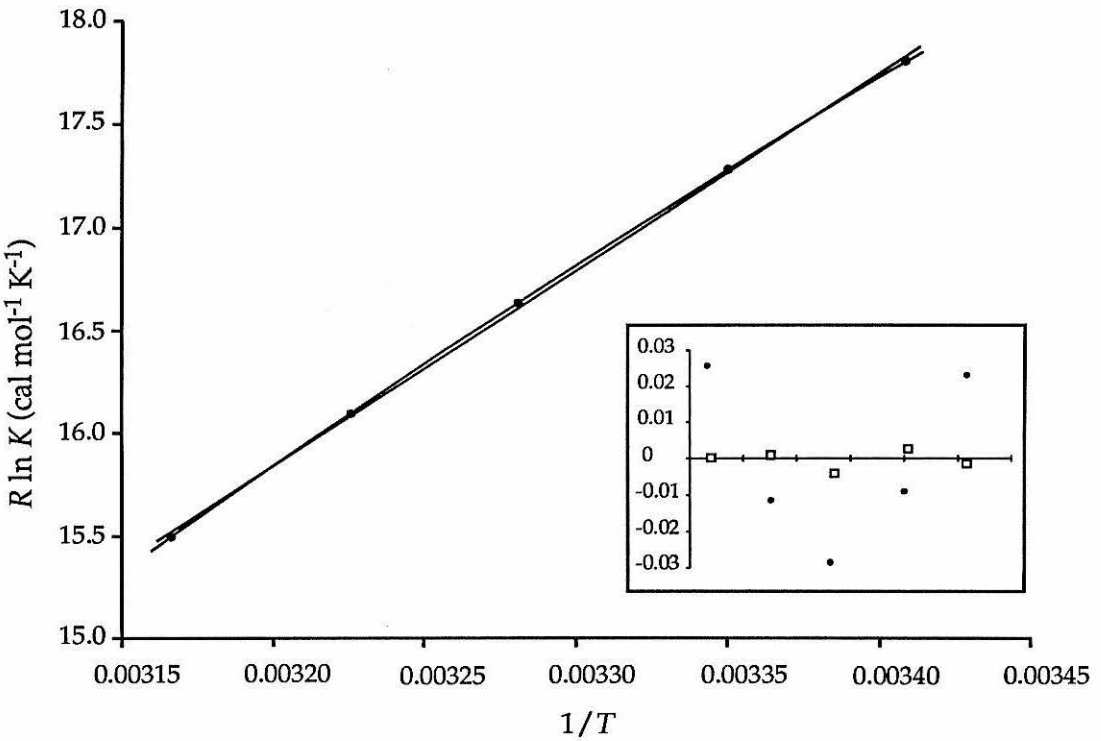


Figure 17. The same set of Diederich's data as shown in Figure 16, fit by both the linear and log equations. Inset: residuals from the fits. Filled circles are from the linear equation; open squares are from the log equation. The F-statistic for significance of regression between these two models is 215; p for this value is 99.5%.

log equations. The fitting procedure already provides the inverted normal matrix *and* the variance from the fit. Assigning uncertainties to the parameters would require about three additional lines of computer code.

Such a calculation, however, has not been included in the fitting program. Parameter confidence limits obtained by inverting the normal matrix are valid only if the actual distribution of the dependent variables can be modeled as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon\,, \tag{33}$$

in which $\mathbf{y}$ is the vector $(N \times 1)$ of dependent variables, $\mathbf{X}$ is the matrix $(N \times k)$ of independent variables, β is the vector $(k \times 1)$ of adjustable parameters, and ε is

the vector ($N \times 1$) of measurement errors in the dependent variable. The values of the independent variables are assumed to be known without error. The random errors ε_i are typically assumed to be independent and identically distributed normal variates with an expectation of zero and a variance of σ^2. The formula can be readily adjusted to allow for the different components of ε to have different variances, and even the condition of normality is not crucial to the application of the formula. It is, however, necessary for the measurement errors in the dependent variables to be independent of each other.

In the case of the log equation, the terms in equation 33 are

$$y = \begin{pmatrix} R\ln K_1 \\ R\ln K_2 \\ \vdots \\ R\ln K_N \end{pmatrix}; \; \mathbf{X} = \begin{pmatrix} 1 & -\ln T_1 & 1/T_1 \\ 1 & -\ln T_2 & 1/T_2 \\ \vdots & \vdots & \vdots \\ 1 & -\ln T_N & 1/T_N \end{pmatrix}; \; \beta = \begin{pmatrix} \Delta S_1^\circ - \Delta C_p^\circ \\ -\Delta C_p^\circ \\ -\Delta H_0^\circ \end{pmatrix}$$

so that equation 33 itself becomes

$$\begin{pmatrix} R\ln K_1 \\ R\ln K_2 \\ \vdots \\ R\ln K_N \end{pmatrix} = \begin{pmatrix} \Delta S_1^\circ - \Delta C_p^\circ + \Delta C_p^\circ \ln T_1 - \Delta H_0^\circ/T_1 \\ \Delta S_1^\circ - \Delta C_p^\circ + \Delta C_p^\circ \ln T_2 - \Delta H_0^\circ/T_2 \\ \vdots \\ \Delta S_1^\circ - \Delta C_p^\circ + \Delta C_p^\circ \ln T_N - \Delta H_0^\circ/T_N \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

The elements $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are the errors in determination of $R\ln K$. As was discussed earlier, the errors in K, and consequently in $R\ln K$, are decidedly *not* independent. The design of the variable-temperature binding experiments is such that most of the measurements used in the determination of K are performed only once, and apply to every temperature studied. Since one set of measurement errors eventually affects all experimental K values, the errors in these K values are far from independent. For instance, in a typical variable-temperature binding study, a single sample is observed at a variety of temperatures, and K is determined at each temperature by application of equations 8 and 12. If the true concentrations of host and guest in the sample are lower than the measured values, the fractions of host and guest bound at each temperature will be less than they would have been if the

concentration measurements were correct. Employing equation 12 will then return an erroneously low estimate of K at every temperature. Similarly, an inaccurate estimate of D will bias all measurements of K in the same direction.

The only way to obtain believable confidence limits for parameters obtained from this analysis is to use Monte Carlo sampling. As with the binding titrations, a large number of random simulations of the entire experiment could tell both the fitness of the model and confidence intervals for the parameters. Variable-temperature binding experiments are quite complicated, however, and programming a computer to simulate them would be quite a chore.

The first step of such a Monte Carlo simulation study would be the simulation of a binding titration in order to find the constant value of D. With this D in hand, the NMR spectrum of one sample would be simulated at each of the experimental temperatures. This requires a knowledge of the thermodynamic parameters ΔH_0°, ΔS_1°, and ΔC_p°, and also of the thermal expansion behavior of the solvent.

For testing the fitness of the model, the effect of introducing errors into the experimental measurements would be explored, as in Lucius,[37] by generating data sets that would arise if the fitted model were true. The Monte Carlo binding titration "data" would be iteratively fitted to obtain $\widehat{D}$, a Monte Carlo estimate of D. Monte Carlo NMR spectra would then be generated for each temperature by using the assumed value of D (not $\widehat{D}$) and the assumed parameters of the log equation. These variable-temperature spectra, together with the value $\widehat{D}$, give Monte Carlo estimates $\widehat{K}$ for K at each temperature. The log equation would then be fitted to the plot of $R\ln\widehat{K}$ vs. $1/T$ from these values. The SSR from the actual experiment can be compared to the distribution of this statistic from the simulated experiments. If an

(37) see Chapter 3.

SSR as large as the experimental value turns out to be extremely improbable, then we can confidently declare that the model does not adequately fit the data.

A Monte Carlo protocol for determining confidence intervals for the fitted parameters would be similar to the method used in the program Portia.[37] In this case, alternative values for the experimental measurements, determined from a knowledge of the measurement uncertainties, would form a basis for fitting the log equation to the experimental spectra.

Programs to perform such Monte Carlo experiments would certainly be challenging to create. Their usefulness, however, is less certain. The model of a constant ΔC_p° is almost definitely incorrect. The fitness of this model can for the most part be intuitively rejected on the basis of the sigmoidal patterns to the residuals from the log fits. The log equation allows convenient analysis of variable-temperature data, and is somewhat more refined than the naïve linear equation. If more quantitative conclusions are desired, a more appropriate model should be employed.

The accuracy of the simple estimate of ΔC_p° produced by a log fit depends on many factors, including the magnitude of experimental errors, the accuracy of the estimate of D, and the deviation of the log equation from the actual but unknown behavior of $R \ln K$ as a function of temperature. These factors can be qualitatively assessed by examining the residuals. A good fit indicates that the log estimate of ΔC_p° is valid over the temperature range studied; a sigmoidal residuals pattern indicates that ΔC_p° itself varies over the the sampled temperature range, and widely scattered residuals indicate that the log estimate of ΔC_p° is uncertain. The practice of determining ΔC_p° from a van't Hoff plot can provide compelling evidence that a nonzero ΔC_p° exists, but the actual estimates of ΔC_p° are less certain.

Appendix B. The van't Hoff Fitting Program

The van't Hoff fitting program (vantHoff) is a Macintosh program that operates on a tabular file of variable-temperature binding data to obtain thermodynamic parameters. The results of four fitting procedures are reported in a text output file, and a variety of quantities are written to a tabular file for convenient graphing. Two of the models that are fitted to the data are the linear and log equations described in this chapter; the other two are a quadratic fit and a log fit in which a guess of ΔC_p° is held constant. The quadratic model, like the log model, has three adjustable parameters; the fixed-ΔC_p° model, like the linear model, has only two adjustable parameters. Each three-parameter model is compared to both two-parameter models by the F-test for significance of regression. The program determines the values of F for each of these comparisons, and also reports the confidence score p corresponding to each F.

Two input files are required by the program: a data file and a preferences file. The preferences file must be named "vant-Hoff Preferences." It contains two quantities: the fixed value of ΔC_p° to be used in the fixed-ΔC_p° model, and a reference temperature. The values of ΔG°, ΔH°, and ΔS° according the two models following the log equation will be calculated for this reference temperature. The data file is a text file organized into two columns: the first column contains the absolute temperature (in kelvin), and the second column contains the bimolecular association constant (in M^{-1}). This text file should be written either by a text editing program (such as Word) or by a graphing program (such as Cricketgraph), and saved in text-only format. Best results are obtained by creating the file with Cricketgraph and saving it in tab-delimited text format. Kaleidagraph files, even if saved in text format, must be cleaned up with a text editor before they can function as input. This is because they contain column headings without an indicating marker; vantHoff will try to read the column headings as numerical entries, and consequently crash.

Two output files are generated by vantHoff. The tabular output file is the same file as the input file; the two columns of the input file are merely accompanied by an additional twenty-two columns of output. Columns 3–7 contain simple arithmetic transformations of the two input columns. Column 3 holds $1/T$, column 4 holds $\ln T$, column 5 holds $\ln K$, column 6 holds $R\ln K$, and column 7 holds $RT\ln K$, which is the experimental $-\Delta G^\circ$. The next eight columns hold the calculated $R\ln K$ values and residuals from the fits according to each of the four models. These $R\ln K$ columns allow graphical comparisons betwen the fitted and experimental $R\ln K$ values, and between the residuals from the different models. The remaining nine columns contain the fitted ΔH°, ΔS°, and $-T\Delta S^\circ$ values from the three models that do not require ΔH° and ΔS° to be constant. This enables, among other options, compensation plots of the type seen in Figure 12 of this chapter.

The text output file reports the exact thermodynamic parameters found by each of the models, and some statistics relating to the goodness-of-fit of the models. The first model it describes is the linear model, which fits the VT data by the equation $R\ln K = A/T + B$. The best-fit parameters A and B are reported, along with the fit score SSR (actually SSR*) and the standard deviation (which is $\sqrt{\mathrm{SSR}/(N-2)}$, where N is the number of data points). The model with ΔC_p° fixed at the value in the preferences file is reported next; this model is $R\ln K = A/T+B+\Delta C_p^\circ(\ln T-1)$. The best-fit parameter values A and B are reported, along with SSR and the standard deviation. The values of ΔH_0° and ΔS_1° corresponding to the parameters A and B are also listed. In addition, the calculated values of ΔH°, ΔS°, and ΔG° at the reference temperature specified in the preferences file are reported. The next model summarized is that of the log equation, $R\ln K = A/T + B\ln(1/T) + C$. The parameters A, B, and C, fit scores SSR and standard deviation (which is $\sqrt{\mathrm{SSR}/(N-3)}$ for the three-parameter models), and thermodynamic parameters ΔH_0°, ΔS_1°, ΔC_p° are reported, as are the values of ΔH°, ΔS°, and ΔG° at the

reference temperature. In addition, the results of F-tests for the significance of regression against the linear and fixed ΔC_p° models are reported as well. Finally, the results of the fit to the quadratic model, $R\ln K = A/T^2 + B/T + C$, are summarized. Naturally, the values A, B, C, SSR, and standard deviation are reported, as are the results of the F-tests for the significance of regression against the linear and fixed-ΔC_p° fits. In addition, it reports the results of linear regressions of plots of ΔH° vs. T and of ΔS° vs. $\ln T$. Under the log equation, these plots both give slopes of ΔC_p°, and the y-intercepts are ΔH_0° and ΔS_1°, respectively. These analyses give a sense of the average ΔC_p° found by the quadratic fit over the sampled temperature range; under the quadratic model, ΔC_p° is *not* a constant.

To run the program, simply double-click on the "vantHoff" icon. A Macintosh input file dialog box will appear; select a text input file. An output file dialog box will then appear; choose a name for the text output file. The tabular output file will simply overwrite the input file. The program will perform the fitting analyses and quit on its own. The tabular output file may be opened by any graphing application, and the text output file may be opened by any text editing application.

The More Points program.

There is also a program called More Points, which generates $R\ln K$ points following the log equation. This is useful for plotting a smooth curve in a van't Hoff plot of the fitted log equation. The log equation traces in all of the van't Hoff plots in this chapter were generated in this fashion. This program, and its operation, are very simple. Run the program by double-clicking the "More Points" icon; a text window will appear with instructions to enter "a, b, c, lowT, highT," and "nmr pts." These quantities are, respectively, the A, B, and C from the log equation $R\ln K = A/T + B\ln(1/T) + C$, the low and high ends of the temperature interval for which you want to create the $(1/T, R\ln K)$ points, and the number of such points you wish to create. The program will generate a tabular file called "more," which

contains columns of T, $1/T$, and $R\ln K$ values corresponding to the log equation specified by A, B, and C. These values may be cut and pasted into the tabular output file from vantHoff.

Incidental Details.

Both of these programs are written in THINK Pascal™. The source codes and projects (that is how THINK Pascal™ organizes its jobs) accompany the compiled applications. The linear least-squares minimization procedure in vantHoff was adapted from the Borland Turbo Pascal Numerical Methods Toolbox.

Chapter 6

Catalysis of an S_N2 Reaction by Dynamic Transition-State Stabilization

Abstract: Complexation of pyridine-type nucleophiles by ethenoanthracene-based macrocyclic host molecules accelerates their S_N2 alkylations in water. Analysis of the kinetic data reveals that the transition states of these reactions are bound more strongly than either the reactants or the products. This effect is entropic in origin. A qualitative explanation based on solvent dynamics is offered. The computational procedure used to obtain the rate constant k_c of the catalyzed reaction is described in detail.

I. The Chemical System

A. Background.

Much of the inspiration for synthetic molecular recognition studies comes from the impressive properties of biological receptors. Antibodies, with their high selectivites and strong affinities for their preferred targets, present a notable example. Some of the most impressive biological molecules, however, are the enzymes. In addition to providing a binding site for their substrates, enzymes must arrange for the chemical transformation of bound substrate into product.

These two functions seem at first analysis to be distinct from each other. The receptor site of an enzyme exists to deliver the reactive moiety of the substrate, in the proper orientation, to the catalytic site. The catalytic site then occasions the bond making and breaking. In 1946, however, Linus Pauling[1] theorized that binding and catalytic functions need not be separate. All that a catalyst must do to accelerate a reaction is enable a pathway from the substrate to the desired product that has a lower activation barrier than the other available pathways. This lowering of the activation barrier can be thought of as a stabilization of the transition state. Figures 1 and 2 illustrate this concept.

(1) Pauling, Linus *Nature* **1948**, *161*, 707–709. Pauling, Linus *Chem. Eng. News* **1946**, *24*, 1375–1377.

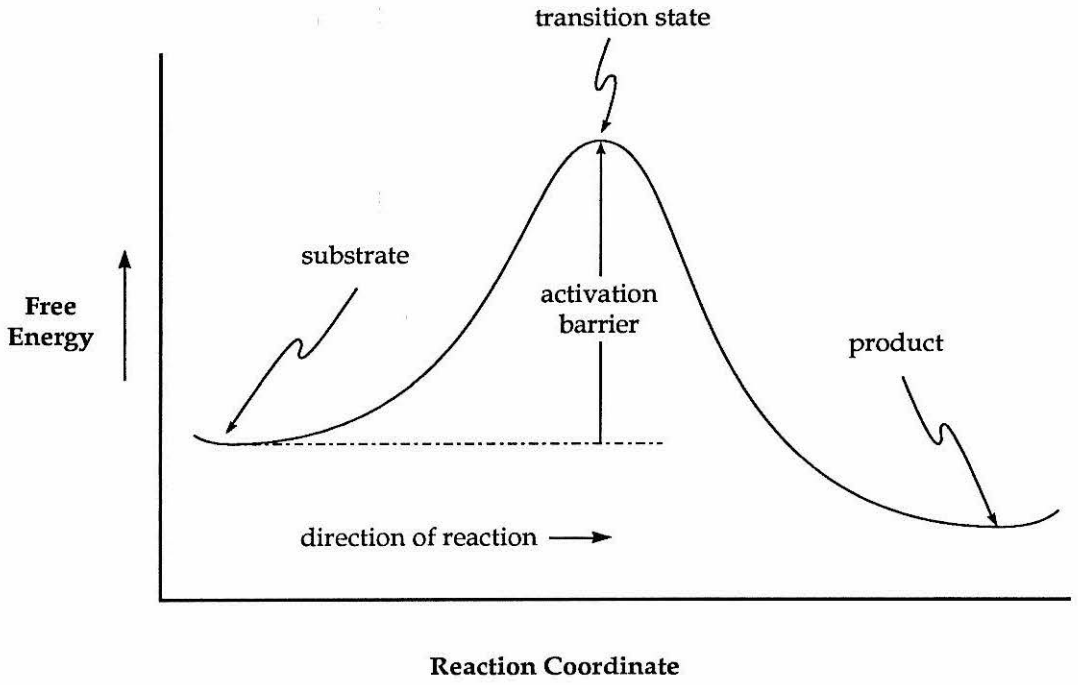


Figure 1. General free energy profile of a one-step reaction.

One possible way to stabilize a molecule is to complex it with another molecule. If it is thermodynamically favorable to bring two separate molecules together, this association process lowers the total free energy of the system. Association of a transition-state species with some other molecule, then, could lower the free energy of the transition state and thus accelerate the reaction.

This means that the only feature required for an enzyme-like catalyst is a binding site for the transition state. In order to bind a short-lived transition state species, however, it is necessary to bind the precusor substrate first. Binding this ground state species introduces the complication that unless the transition state is stabilized more than the substrate, there will be no rate acceleration, as shown in Figure 3. Consequently, this binding site must bind both the substrate and the transition state, with the transition state held the most strongly. Such a situation

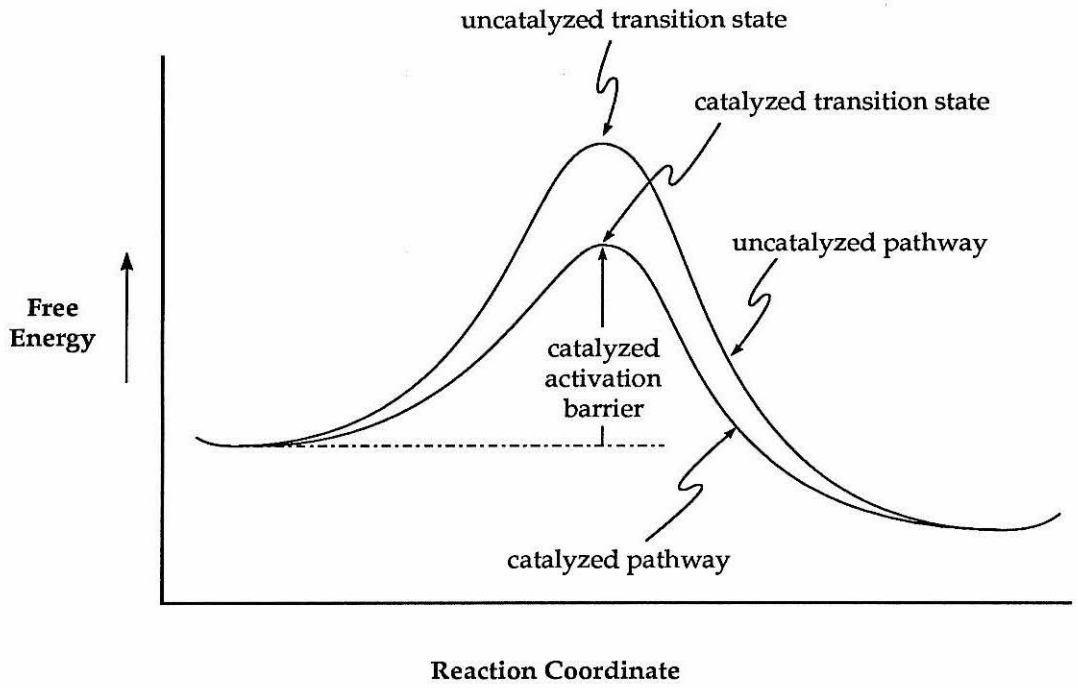


Figure 2. General free energy profile of the reaction of Figure 1 and its catalyzed counterpart. Lowering the free energy of the transition state reduces the activation barrier, accelerating the reaction.

is not at all unimaginable: for any elementary reaction step, the transition state will bear some resemblance to the reactants. A binding site with a strong affinity for the transition state, then, should also bind the reactants, but less tightly. Thus, such a binding site should act as a catalyst for that reaction.

In the time since Pauling's suggestion, circumstantial evidence has arisen from a number of sources to support his claim. Many potent enzyme inhibitors bear a strong resmblance to purported intermediates of the reaction catalyzed by the enzyme.[2] More recently, "catalytic antibodies" have been made by isolating antibodies that

(2) Wolfenden, Richard "Transition state analog inhibitors and enzyme catalysis," *Annu. Rev. Biophys. Bioeng.* **1976**, *5*, 271–306.

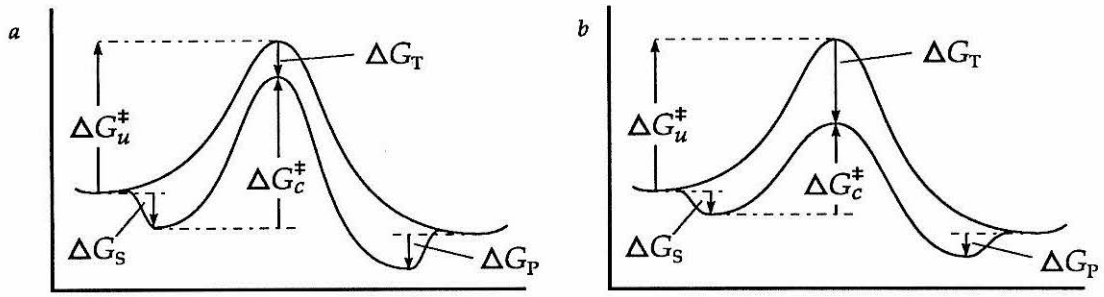


Figure 3. Free energy profiles of one-step reactions showing possible effects of complexation of the reacting species. *a*) Transition state (T) is no more strongly bound than substrate (S), making the "catalyzed" activation barrier as high as the uncatalyzed barrier. *b*) T is more strongly bound than S, making the catalyzed barrier lower than the uncatalyzed barrier. Only in this case is there acceleration of the reaction.

bind to transition-state-analog antigens. In many cases, such antibodies actually catalyze the intended reactions.[3]

B. Studies.

Since our group was actively investigating molecular recognition, we elected to apply our ethenoanthracene-based hosts to the challenge of transition-state stabilization. Accounts of this work have already been published;[4,5] in this chapter I will describe my contributions to and impressions of the project.

1. Approach. We chose, in addition to designing modified hosts to catalyze specific reations, to search for reactions whose transition states would be more

(3) For recent reviews, see Lerner, Richard A.; Benkovic, Stephen J.; Schultz, Peter G. "At the crossroads of chemistry and immunology: catalytic antibodies," *Science* **1991**, *252*, 659–667. Scanlon, Thomas S.; Schultz, Peter G. "Recent advances in catalytic antibodies," *Philos. Trans. R. Soc. London B* **1991**, *332*, 157–164. Shokat, K. M.; Schultz, Peter G. "Catalytic antibodies," *Annu. Rev. Immunol.* **1990**, *8*, 335–363. Mayforth, Ruth D.; Quintáns, José "Designer and catalytic antibodies," *N. Engl. J. Med.* **1990**, *323*, 173–178. Tramontano, Alfonso, Schloeder, Diane "Production of antibodies that mimic enzyme catalytic activity," *Methods Enzymol.* **1989**, *178*, 531–550. Pollack, Scott J.; Nayakama, Grace R.; Schultz, Peter G. "Design of catalytic antibodies," *Methods Enzymol.* **1989**, *178*, 551–568.

(4) Stauffer, David A. Ph.D. Thesis, California Institute of Technology, 1989; Chapter 3.

(5) Stauffer, David A.; Barrans, Richard E. Jr.; Dougherty, Dennis A. "Biomimetic catalysis of an S_N2 reaction resulting from a novel form of transition-state stabilization," *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 915–918.

strongly bound than the reactants *by the hosts already in hand.* This would require a reaction that develops some property in its transition state that is attractive to a host: a property that is absent, or present to a lesser degree, in the reactants. All hosts, especially host P, are efficient receptors for positively charged guests, by virtue of the cation-π effect. Thus, any reaction in which a positive charge develops should be accelerated by host. Practical experimental constraints, however, further limit the reactions that may be studied. Namely, an appreciable fraction of the starting material must be bound, the host must not be destroyed or otherwise inactivated, and the reactants must be soluble in borate-*d*, or some other alkaline aqueous system.

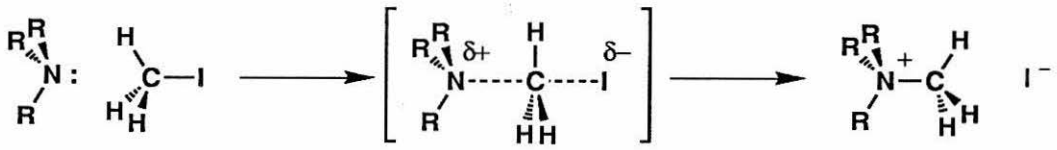


Figure 4. Progress of an S_N2 alkylation of a tertiary amine by iodomethane. The charge on the nitrogen center increases as the reaction proceeds.

One reaction that satisfied these criteria was the S_N2 alkylation of a neutral nucleophile, such as an amine or a sulfide. In such a reaction, a reactant nucleophile smoothly becomes a product cation. If the nucleophile were inside a host cavity, its developing positive charge would become increasingly stabilized as the reaction progresses. Thus, the transition state would enjoy greater stabilization than the reactants. As illustrated in Figure 3*b*, this lowers the overall activation energy and accelerates the reaction.

S_N2 quaternizations of tertiary amines with alkyl halides, known as *Menshutkin reactions*, have been extensively studied. Their rates are known to be sensitive to

the polarity of the solvent.[6] Thus, such reactions were ideal candidates for our investigations.

2. Modeling.

Some Menshutkin reactions were indeed accelerated by host. After trying some more complicated reactions, David Stauffer studied the alkylation of the guest quinoline with iodomethane in borate-*d*. The reaction ran faster with added host P than without, showing that the host, as was hoped, catalyzed the reaction. Nothing more about this effect could be said until the kinetics and energetics of both the catalyzed and uncatalyzed reactions were determined quantitatively.

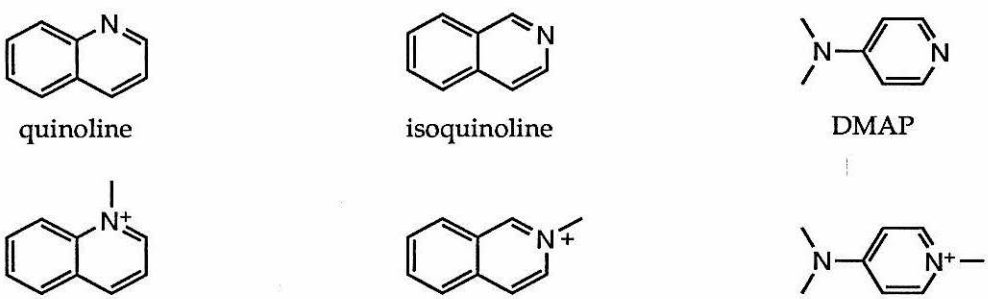


Figure 5. Principal substrates (top) and products (bottom) of studied host-catalyzed alkylations.

Studying the uncatalyzed reaction was straightforward: its kinetic course could be monitored simply by combining nucleophile and alkylating agent in a vessel and measuring the concentration of product as time progressed. The catalyzed reaction, however, could not be followed so easily, because it could not be separated from the uncatalyzed reaction. When host, substrate, and alkylating agent are all present in the same solution, product appears from the action of *both* pathways. My task was

(6) Abraham, M. H. "Solvent effects on the free energies of the reactants and transition states in the Menshutkin reaction of trimethylamine with alkyl halides," *Chem. Commun.* **1969**, 1307–1308.

to develop a way to determine the rate constant of the catalyzed reaction from the data at hand.

The first step was to identify a kinetic model for this process. Then it was necessary to find the rate constant k_c for the catalyzed reaction that best simulated the observed data. To that end, I created an interactive computer program, the Kinetics Simulator, that employs a user's guess for k_c to predict the kinetic behavior. It then compares this predicted behavior to the experimental data. The user varies k_c until a best fit of the predicted to observed behavior is attained.

With k_c in hand, it is easy to determine $\Delta G^{\ddagger}_c$, the free energy of activation of the catalyzed reaction, by plugging k_c into the Eyring equation[7]

$$k = \kappa \frac{k_B T}{h} \exp\left(-\frac{\Delta G^{\ddagger}}{RT}\right) \tag{1}$$

where

κ = transmission coefficient (typically assumed to be unity)

k = rate constant of the reaction (concentration units omitted)

k_B = Boltzmann's constant, 1.380622×10^{-23} JK^{-1}

T = temperature

h = Planck constant, 6.626196×10^{-34} Js

$\Delta G^{\ddagger}$ = free energy of activation of the reaction (per mole)

R = Gas constant = $Nk_B = 8.31441$ J mol^{-1}K^{-1}.

This equation can be rearranged to show $\Delta G^{\ddagger}$ as a function of k:

$$\Delta G^{\ddagger} = -RT \ln\left(\frac{kh}{k_B T}\right). \tag{2}$$

The two free energies of activation, $\Delta G^{\ddagger}_c$ and $\Delta G^{\ddagger}_u$, can now contribute to our understanding of the overall energetic scheme shown in Figure 6.

(7) Wynne-Jones, W. F. K.; Eyring, Henry "The absolute rate of reactions in condensed phases," *J. Chem. Phys.* **1935**, *3*, 493–502.

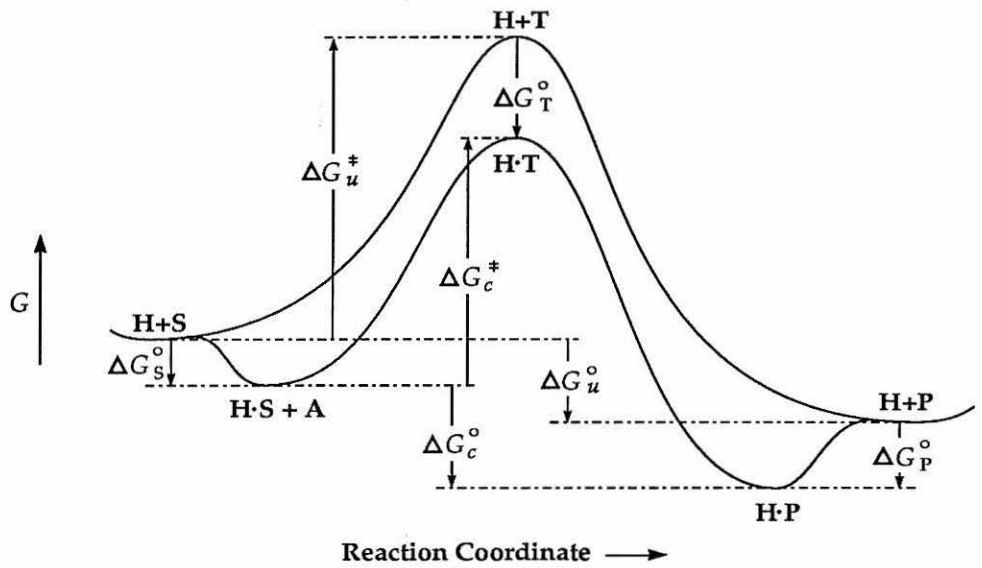


Figure 6. Free energy diagram for catalyzed and uncatalyzed alkylations. Arrows denote the direction of the reaction for which the associated ΔG° is the free energy change.

We are interested here in $\Delta G^\circ_{\mathrm{T}}$, the free energy of binding of the transition state.

$$\mathrm{H} + \mathrm{T} \rightleftharpoons \mathrm{H{\cdot}T} \qquad \Delta G^\circ = \Delta G^\circ_{\mathrm{T}}$$

This binding energy cannot be measured directly in a binding study because the species T is too ephemeral for such an equilibrium to be reached.

Despite this inconvenience, the change in free energy associated with this nonexistent reaction is readily determined by using an equivalent thermodynamic cycle. Every multistep process giving the overall transformation $\mathrm{H} + \mathrm{T} \rightleftharpoons \mathrm{H{\cdot}T}$ has the same overall ΔG°, which is $\Delta G^\circ_{\mathrm{T}}$. This overall ΔG° is the sum of the ΔG° values of the component steps. A convenient hypothetical transformation for finding $\Delta G^\circ_{\mathrm{T}}$ is illustrated in Figure 8.

In this sequence, the transition state (T) first fragments into ground state substrate (S) and alkylating agent(A), $\mathrm{T} \rightarrow \mathrm{S} + \mathrm{A}$. This is the reverse of the

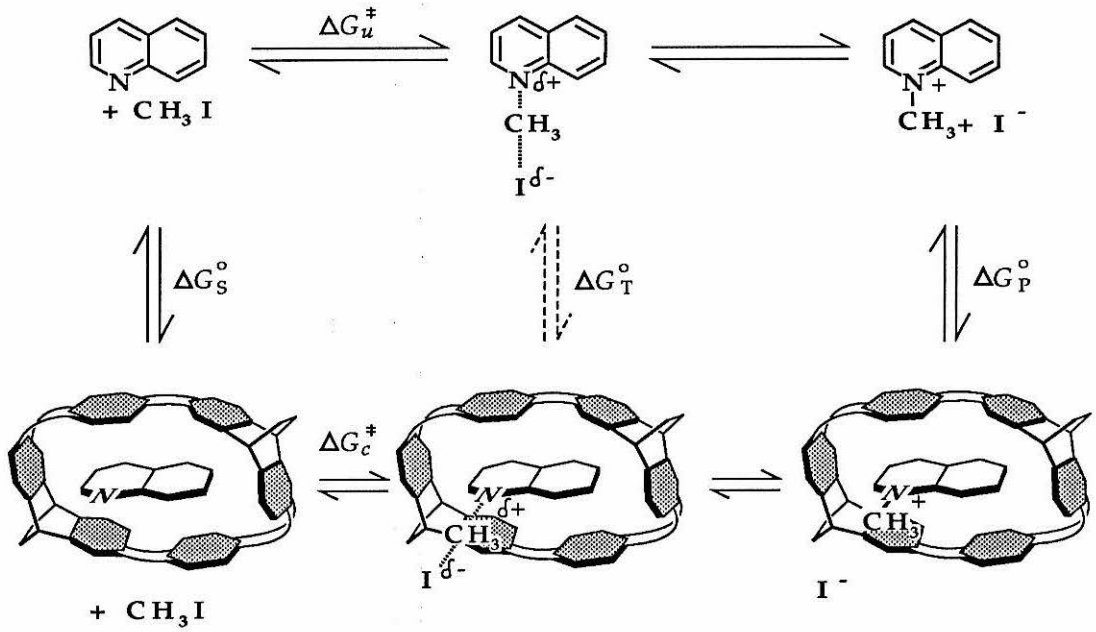


Figure 7. Reaction pathway for catalyzed and uncatalyzed alkylation of quinoline, including the formal complexation reaction of host with the transition state, and its associated free energy $\Delta G^\circ_{\mathrm{T}}$.

transition state forming reaction, and so has a free energy change $-\Delta G^\ddagger{}_u$. Second, the substrate and host come together, H + S → H·S. This gives a free energy change of $\Delta G^\circ_{\mathrm{S}}$. Finally, the bound substrate reacts with the alkylating agent, to give the bound transition state H·S → H·T, $\Delta G = \Delta G^\ddagger{}_c$. Thus, the overall transformation is

$$\begin{array}{rl} \mathrm{T} \rightarrow \mathrm{S} + \mathrm{A} & -\Delta G^\ddagger{}_u \\ \mathrm{H} + \mathrm{S} \rightarrow \mathrm{H{\cdot}S} & \Delta G^\circ_{\mathrm{S}} \\ \mathrm{H{\cdot}S} + \mathrm{A} \rightarrow \mathrm{H{\cdot}T} & \Delta G^\ddagger{}_c \\ \hline \mathrm{H} + \mathrm{T} \rightarrow \mathrm{H{\cdot}T} & \Delta G^\circ_{\mathrm{T}} = \Delta G^\ddagger{}_c - \Delta G^\ddagger{}_u + \Delta G^\circ_{\mathrm{S}}\,. \end{array}$$

By the same reasoning, the change $\Delta Q^\circ_{\mathrm{T}}$ in *any* state function Q associated with the reaction H + T ⇌ H·T can be found by adding together the ΔQ° values of the steps in this alternate multistep process.

$$\Delta Q^{\circ}_{T} = \Delta Q^{\ddagger}{}_{c} - \Delta Q^{\ddagger}{}_{u} + \Delta Q^{\circ}_{S} \qquad (3)$$

This general formula is illustrated, for the specific case of $Q = G$, in Figures 6 and 8.

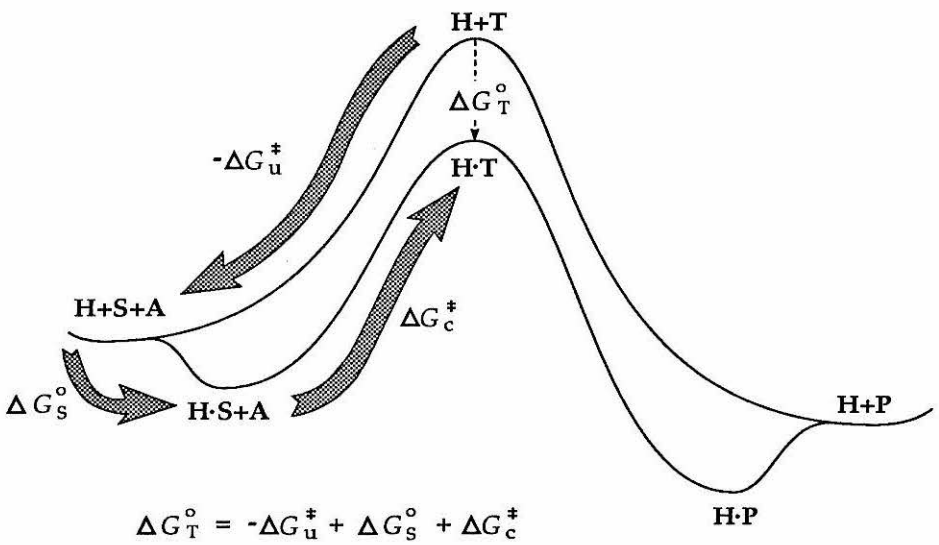


Figure 8. How to determine the free energy of binding of the transition state.

ΔG°_{T} obtained in this way can be directly compared to ΔG°_{S} and ΔG°_{P} to determine the extent of the transition-state stabilization. A ΔG°_{T} that is more negative than ΔG°_{P} would indicate that the transition state is more favorably bound than binding the product. If this occurs, then an artificial "enzyme," in the sense of Pauling's hypothesis, has been prepared.

C. Results and Discussion.

1. Reaction profiles. The data of Table I show that the transition states of the studied alkylations are indeed more stabilized than either starting materials or

products by complexation with host.[8] This was not what we expected. We expected that the host/substrate complex would be stabilized by donor-acceptor interactions and the hydrophobic effect, and that the host/product complex would be further stabilized by the cation-π effect. The transition state, which has a geometry and charge distribution between those of the substrate and the product, was likewise expected to have a free energy of complexation ΔG°_{T} between that of substrate, ΔG°_{S}, and product, ΔG°_{P}. Any stabilizing effect found in the host-transition state complex should also have been present in either the host/substrate complex or the host/product complex. Clearly, something unanticipated was happening.

Table I. Kinetic and thermodynamic parameters of some Menshutkin reactions catalyzed by host P.[a]

Guest	Temp.[b]	k_u^c	k_c^c	$-\Delta G_S^{\circ d}$	$-\Delta G_P^{\circ d}$	$-\Delta G_T^{\circ d}$
quinoline	300.3	0.2402	2.3	5.50	7.70	7.92
	305.3	0.446	5.1	5.55	7.76	8.42
	310.1	0.787	5.8	5.58	7.82	8.22
	315.0	1.47	14.3	5.60	7.88	8.52
	320.0	2.55	22.6	5.60	7.95	8.46
	324.9	4.50	36	5.58	8.01	8.40
	329.8	7.30	123	5.54	8.09	8.89
isoquinoline	300.3	3.38	21	6.42	7.28	7.85
	305.3	5.48	41	6.49	7.38	8.01
	310.1	9.87	97	6.54	7.48	8.32
	315.0	17.2	122	6.59	7.57	8.17
	320.0	27.1	380	6.63	7.66	8.67
DMAP	308	3.37	18	5.75	6.48	6.70

[a]Alkylating agent was iodomethane. Data are from Reference 4. [b]Temperature in K. [c]10^{-4} $M^{-1}s^{-1}$. [d]kcal mol^{-1}.

The thermodynamic parameters of activation of the catalyzed and uncatalyzed reactions tell quite different stories. Table II shows the activation parameters for these reactions of quinoline and isoquinoline, determined by Eyring plots of the

(8) Under different experimental conditions, this is not always the case. Work at higher pH (Leslie Jimenez, unpublished results) has revealed acceleration of alkylation without the transition state being bound more strongly than the products.

Table II. Activation parameters for catalyzed and uncatalyzed Menshutkin reactions in borate-*d*.[a]

substrate	$\Delta G_u^{\ddagger\,b,c}$	$\Delta G_c^{\ddagger\,b,c}$	$\Delta H_u^{\ddagger\,c}$	$\Delta H_c^{\ddagger\,c}$	$\Delta S_u^{\ddagger\,d}$	$\Delta S_c^{\ddagger\,d}$
quinoline	23.9	21.2 (21.1)[e]	22.4	23.9 (20.9)[e]	−5.1	8.9 (−0.8)[e]
isoquinoline	22.7	21.4 (21.3)[e]	20.5	26.2 (23.1)[e]	−7.4	16 (6)[e]

[a]Values are from Reference 4. The alkylating agent was iodomethane. [b]$\Delta G^{\ddagger}$ at 298 K. [c]kcal mol^{-1}. [d]cal mol^{-1} K^{-1}. [e]The values in parentheses are from the Eyring plot omitting the highest-temperature measurement.

catalyzed and uncatalyzed kinetic data from different temperatures. Two sets of numbers appear for the catalyzed data because Stauffer elected to analyze the Eyring plots with and without the measurement at the highest temperature. The change upon including or omitting this value is substantial, but both analyses give $\Delta S_{\mathrm{T}}^{\circ} > \Delta S_{\mathrm{P}}^{\circ}$, and $\Delta H_{\mathrm{T}}^{\circ} \geq \Delta H_{\mathrm{S}}^{\circ}$.

The catalyzed reactions, of course, have lower $\Delta G^{\ddagger}$ than the uncatalyzed reactions. The enthalpies do not reflect this trend; if anything, the activation enthalpies of the catlyzed reactions are *less* favorable than those for the uncatalyzed reactions. The activation *entropies* are where the differences between the catalyzed and uncatalyzed reactions are most manifest. The uncatalyzed reactions, as may be expected for reactions in which two parts are brought together in a specific orientation, have negative entropies of activation. The catalyzed reactions, on the other hand, show a *zero to slightly positive* entropy of activation. Although the exact values of these activation entropies are uncertain because of scatter in the Eyring plots, it is clear that the activation entropies of the catalyzed reactions are much more favorable (more positive) than the activation entropies of the uncatalyzed reactions.

These activation parameters show that the principal catalytic activity of the host is *entropic* in consequence. It is more entropically favorable to reach the bound transition state from the bound substrate than to reach the free transition state from the free substrate. Since it is difficult to imagine that encirclement of the substrate

by host widens the range of its allowed attack angles on the alkylating agent, this differential entropic effect must be a property of the *solvent*.

The negative values of $\Delta S^{\ddagger}$ for the uncatalyzed reactions indicate that the solvent is not able to adopt as many configurations about the transition state as it is able to adopt about the substrate. The more positive $\Delta S^{\ddagger}$ values for the catalyzed reactions show that this bottleneck is removed by the host. This can be intuitively rationalized by noting that the participants in the uncatalyzed reaction are in direct contact with water, but the participants in the catalyzed reaction are insulated from the solvent by the intervening host molecule. Thus, entropic restrictions imposed upon the solvent by the transition state should be less severe in the catalyzed case.

Table III. Thermodynamic parameters of binding of reactants, transition states, and products of Menshutkin reactions to host P in borate-*d* at 298 K.[a]

substrate	ΔG_S° [b,d]	ΔG_T° [c,d]	ΔG_P° [d]	ΔH_S° [b,d]	ΔH_T° [d]	ΔS_S° [b,e]	ΔS_T° [e]
quinoline	−6.0	−8.7 (−8.8)	−7.6	−11	−9.5 (−12.5)	−17	−3.0 (−12.7)
isoquinoline	−6.4	−7.7 (−7.8)	−7.2	−9.8	−4.1 (−7.2)	−11	12.4 (2.4)

[a]Alkylating agent was iodomethane. ΔG_S° and ΔG_P° were obtained from binding studies. ΔH_S° and ΔS_S° were obtained from variable-temperature studies; see reference 9 and chapter 5. ΔG_T°, ΔH_T°, and ΔS_T° were determined by using equation 3; see Figure 8. The values in parentheses are from the Eyring plot omitting the highest-temperature measurement. [b]From reference 9. [c]From reference 5. [d]kcal mol^{-1}. [e]cal mol^{-1} K^{-1}.

The stabilization of the transition state by interaction with host is perhaps best seen by dissecting ΔG_T° into its components ΔH_T° and ΔS_T°. This can be accomplished by using equation 3 in conjunction with the ΔH_S° and ΔS_S° determined from a variable-temperature binding study.[9] The thermodynamic parameters of binding of substrate, transition state, and products for two reactions are shown in Table III. Some of these values differ from those in Table II, because they were

(9) Stauffer, David A.; Barrans, Richard E. Jr.; Dougherty, Dennis A. "Concerning the thermodynamics of molecular recognition in aqueous and organic media. Evidence for significant heat capacity effects," *J. Org. Chem.* **1990**, *55*, 2762–2767.

derived from different measurements. However, it is still clear that the principal difference between substrate and transition-state binding is entropic.

2. Interpretation. We would like a consistent, intuitive physical interpretation of these observations. The effect enabling catalysis must result from some feature of the transition state that is not present in either the substrate or the products, and that is better stabilized by the host than by the solvent. Static properties of the transition state, such as geometry and charge distribution, have already been eliminated from consideration. One property of the transition state that *does* fulfill these criteria is its ephemeral nature: the transition state simply is not present for long. The reaction coordinate of the Menshutkin reaction can be identified with an asymmetric stretch of the nucleophile–C(alkylating agent)–X(leaving group) system. The actual substitution reaction is a vibration in this mode, occurring on a vibrational timescale. The rapid course of this reaction prevents the system from being fully stabilized by solvent at all points along its progress.

There are only a few actions water molecules can take to accommodate a Menshutkin reaction's rapidly changing charge distribution. Water is not at all polarizable; rapid solvent electronic polarization cannot provide significant stabilization of the solute. Its O–H stretching and bending vibrations momentarily alter its molecular dipole moment, but the force constants in these modes are high, and the reduced masses low. This prohibits a distortion of any consequence lasting longer than a natural vibrational period. Water's principal mechanism for responding to an electric field, which accounts for its high static dielectric constant, is reorientation of its molecular dipole. This occurs on a *rotational* timescale. A somewhat faster related process involves restricted rotations, such as those about the axes of established hydrogen bonds. Such librational motion has been credited with about

half the charge-solvating ability of water.[10] Still, the vibrational timescale of the Menshutkin reaction is faster than the rotational timescale of the solvent response.

The host, on the other hand, can react more quickly. It responds to electronic moments within its cavity by polarizing its π elctrons. This is a very fast process; electrons can move much faster than nuclei. The π elctrons of the host are sufficiently agile to accomodate the charge distribution of the Menshutkin reaction at every stage of its progress. This process does not exact a heavy entropic price.

In order for the solvent to stabilize the uncatalyzed reaction, however, it must assume a configuration suitable for solvating the transition state before the reaction even occurs. This action is entropically very costly, undercutting any enthalpic stabilization gained. Encapsulation of the substrate by host attenuates the need for a specific solvent configuration in the transition state. This relative solvent relaxation makes the bound transition state entropically easier to reach and makes the binding of the transition state to host entropically favorable.

This interpretation is consistent with previous theoretical and experimental studies. Theoretical studies of the degenerate S$_N$2 reaction of chloride with methyl chloride in water have suggested that the transition-state configuration involves the solvent arranged about the solute in a rigid, almost hydrophobic hydration shell.[11,12]

---

(10) Maroncelli, Mark; Fleming, Graham R. "Computer simulation of the dynamics of aqueous solvation," *J. Chem. Phys.* **1988**, *98*, 5044–5069.

(11) Chandrasekhar, Jayaraman; Smith, Scott F.; Jorgensen, William L. "Theoretical examination of the S$_N$2 reaction involving chloride ion and methyl chloride in the gas phase and aqueous solution," *J. Am. Chem. Soc.* **1985**, *107*, 154–163.

(12) Gertner, Bradley J.; Whitnell, Robert M.; Wilson, Kent R.; Hynes, James T. "Activation to the transition state: reactant and solvent energy flow for a model S$_N$2 reaction in water," *J. Am. Chem. Soc.* **1991**, *113*, 74–87. Gertner, Bradley J.; Wilson, Kent R.; Hynes, James T. "Nonequilibrium solvation effects on reaction rates for model S$_N$2 reactions in water," *J. Chem. Phys.* **1989**, *90*, 3537–3558. Bergsma, John P.; Gertner, Bradley J.; Wilson, Kent R.; Hynes, James T. "Molecular dynamics of a model S$_N$2 reaction in water," *J. Chem. Phys.* **1987**, *86*, 1356–1376. Gertner, Bradley J.; Bergsma, John P.; Wilson, Kent R.; Lee, Sangyoub; Hynes, James T. "Nonadiabatic solvation model for S$_N$2 reactions in polar solvents," *J. Chem. Phys.* **1987**, *86*, 1377–1386.

A statistical free energy perturbation study of a model Menshutkin reaction in water has likewise shown a solvent configuration about its transition state involving entropically unfavorable solvent reorganization.[13] Furthermore, analyses of kinetic and thermodynamic trends of Menshutkin reactions in a variety of solvents have suggested that prior solvent reorganization plays a crucial entropic role.[14]

II. Computational Methods

A. General Concepts.

1. Nomenclature. Because both alkylation and association processes are occurring in this system, a number of different subscripted quantities appear in the following discussion. In this chapter, the total concentration of a species, such as host or product, is denoted by the subscript "tot," as in "$[P]_{tot}$." This is different from the notation adopted elsewhere in this dissertation, in which a total concentration is denoted by the subscript "0," as in "$[G]_0$." The notation is different in this chapter because many concentrations change with time. Any concentration that is specific to the ith simulated time step is given the subscript "i." These indices start with zero, so that initial concentrations have the subscript "0." For instance, $[P]_{tot\,0}$ is the total product concentration at the beginning of the experiment.

2. The kinetic model. The alkylation of substrate by iodomethane could safely be assumed to follow an S_N2 pathway, making the kinetics first order each in substrate and alkylating agent, and second order overall.

$$\text{uncatalyzed reaction: } S + A \rightarrow P$$

---

(13) Gao, Jiali "*A priori* computation of a solvent-enhanced S_N2 reaction profile in water: the Menshutkin reaction," *J. Am. Chem. Soc.* **1991**, *113*, 7796–7797.

(14) Arnett, Edward M.; Reich, Ronald "Electronic effects on the Menshutkin reaction. A complete kinetic and thermodynamic dissection of alkyl transfer to 3- and 4-substituted pyridines," *J. Am. Chem. Soc.* **1980**, *102*, 5892–5902.

$$\text{uncatalyzed rate: } \frac{d[\mathrm{P}]}{dt} = k_u[\mathrm{S}][\mathrm{A}] \tag{4}$$

The catalyzed reaction was likewise identified as an SN2 alkylation; it is hard to envision any action the host would take to change the reaction mechanism.

$$\text{catalyzed reaction: } \mathrm{H{\cdot}S} + \mathrm{A} \rightarrow \mathrm{H{\cdot}P}$$

$$\text{catalyzed rate: } \frac{d[\mathrm{H{\cdot}P}]}{dt} = k_c[\mathrm{H{\cdot}S}][\mathrm{A}] \tag{5}$$

At ordinary temperatures, the reverse reaction does not occur, so it can be neglected. The rate of formation of product then depends on the concentrations of free substrate, host/substrate complex, and alkylating agent.

$$\text{total product concentration: } [\mathrm{P}]_{\mathrm{tot}} = [\mathrm{H{\cdot}P}] + [\mathrm{P}]$$

$$\text{total rate: } \frac{d[\mathrm{P}]_{\mathrm{tot}}}{dt} = \frac{d[\mathrm{P}]}{dt} + \frac{d[\mathrm{H{\cdot}P}]}{dt} = [\mathrm{A}]\big(k_u[\mathrm{S}] + k_c[\mathrm{H{\cdot}S}]\big) \tag{6}$$

Equation 6 shows that the total rate of appearance of product can be predicted from the two rate constants k_c and k_u and the three concentrations [A], [S], and [H·S]. Determination of [S] and [H·S] is, unfortunately, not straightforward, because these concentrations depend on many factors. The first is the host-substrate association constant K_{S}.

$$\mathrm{H} + \mathrm{S} \rightleftharpoons \mathrm{H{\cdot}S};\ K_{\mathrm{S}} = \frac{[\mathrm{H{\cdot}S}]}{[\mathrm{H}][\mathrm{S}]} \tag{7}$$

The fraction of substrate bound depends also on the concentration of host that is available to complex with it. As product forms, it competes with substrate for host.

$$\mathrm{H} + \mathrm{P} \rightleftharpoons \mathrm{H{\cdot}P};\ K_{\mathrm{P}} = \frac{[\mathrm{H{\cdot}P}]}{[\mathrm{H}][\mathrm{P}]} \tag{8}$$

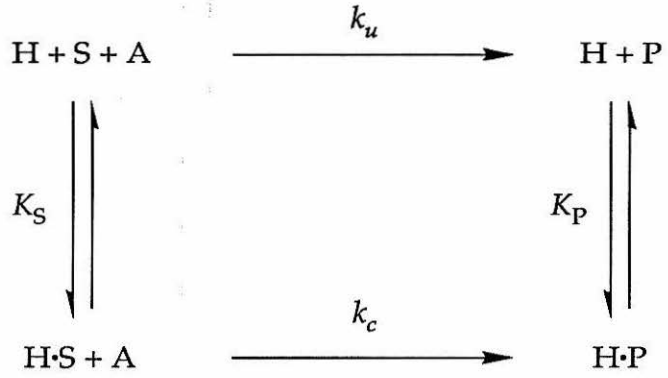


Figure 9. Reactions accounted for in the kinetic model for host-catalyzed alkylations.

These interrelations are summarized in the kinetic scheme of Figure 9. Both substrate and product may complex with host, and either complexed or free substrate may react with alkylating agent to yield product. To satisfactorily extract rate constants and activation barriers from experimental measurements of concentrations over time, the kinetic behavior implied by this scheme must be accurately determined.

3. Predicting kinetic behavior. In order to analytically model the kinetic behavior of this system, that is, to determine a mathematical function to directly predict $[P]_{tot}$ as a function of time, it is necessary to solve the first-order differential equation 6. This requires determination of [A], [S], and [H·S] as functions of time (or of $[P]_{tot}$) and integration with respect to the two differential elements. In this case, two equilibrium relations together determine [H], the concentration of uncomplexed host, since that is the common factor in equations 7 and 8.

The analytical solution of this system of equations is very difficult, if not impossible, so a numerical approach has been employed. This enables calculation of the concentrations of all species of interest ([H], [S], [P], [H·S], [H·P]) to any arbitrary degree of accuracy given the total concentrations of substrate, product, and

host. However, it does not allow calculation of the *derivatives* of any of these quantites with respect to time, as is needed to directly solve the differential equation 6. Consequently, the concentration of product as a function of time is approximated by what amounts to very primitive stepwise numerical integration. Details of the computational procedure are given in later sections.

In summary, the Kinetics Simulator determines the concentration of product as a function of time, given the initial concentrations of substrate, product and host, the two equilibrium constants (K_S and K_P), and the two rate constants (k_c and k_u). All of these quantities except for k_c can be measured independently.

4. Determining k_c. Since k_c is the only unknown, obtaining its value is straightforward. The known quantities can be entered into the computer, along with a guess for the unknown k_c. The simulation predicts the course of the kinetic behavior, which is compared to the experimental data. The program shows, in addition to a graphical display of the calculated and experimental kinetic data, the root mean square deviation (RMS), a measure of the goodness of fit. This is functionally related to the residual sum of the squares (SSR),

$$\mathrm{RMS} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\left([\mathrm{P}]_{\mathrm{tot}\,i\,\mathrm{calc}} - [\mathrm{P}]_{\mathrm{tot}\,i\,\mathrm{obs}}\right)^2} = \sqrt{\frac{1}{\mathrm{N}}\,\mathrm{SSR}}, \qquad (9)$$

but is normalized to give an estimate of the geometric average discrepancy at each point. A value of k_c is optimal in the least squares sense if it gives the smallest RMS of any possible value of k_c. The user may determine this value by successively trying different values of k_c until RMS ceases to improve.

B. Concentration Determination in Multiple Association Processes.

1. Substrate and product binding. In the kinetic scheme depicted in Figure 9, two association processes involving the host operate. To determine the concentrations of all species involved in equations 7 and 8 in a single analytical

step, one must solve the two simultaneous equations 7 and 8, and the mass balance equations 10, 11, and 12.

$$[S]_{tot} = [S] + [H{\cdot}S] \tag{10}$$

$$[P]_{tot} = [P] + [H{\cdot}P] \tag{11}$$

$$[H]_{tot} = [H] + [H{\cdot}S] + [H{\cdot}P] \tag{12}$$

I have been unable to combine these equations in a way that yields one of the unknown quantities [H], [S], [P], [H·S], or [H·P] in terms of only the known quantites $[H]_{tot}$, $[S]_{tot}$, $[P]_{tot}$, K_S, or K_P. Even if such a relation were to be found, it would be valid only for an equilibrium system involving no processes but 7 and 8; it would not hold if more equilibria were considered. Consequently, I have choosen an approach that is, in effect, what Nature does in cases of multiple equilibria. Each system (H/P or H/S) is treated as if it were oblivious to the other. Either one can be treated first; in the case of this example let it be the H/S system.

Initially, all host, substrate, and product are considered to be uncomplexed. At this time, neither substrate nor product equilibrium relations are satisfied (Figure 10).

This system will first be perturbed by solving for [H·S] given the initial concentrations [H] and [S] (Figure 11). For now, the equilibrium relation of equation 7 is satisfied.

The next step is to allow the remaining free host and *product* to interact. As far as the product is concerned, host complexed with substrate is not available to it; all it can use is that which is uncomplexed (Figure 12).

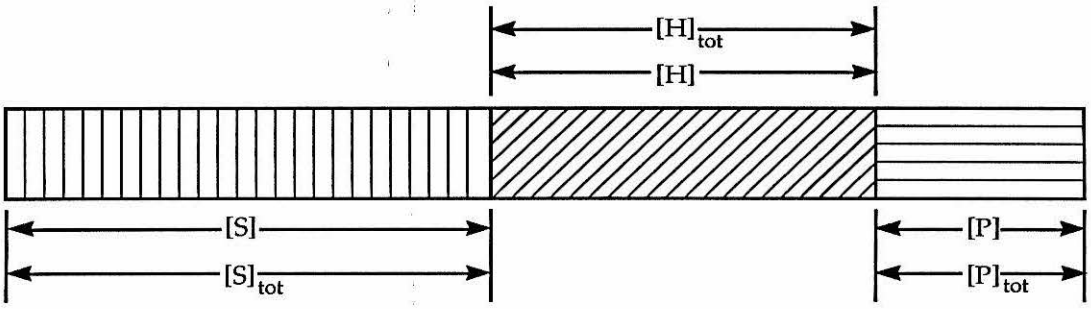


Figure 10. This figure and the four following, 11 through 14, illustrate the method of successive approximations used to determine concentrations in cases of multiple equilibria. These figures arbitrarily show a case in which $K_S = K_P = 20\,000\ M^{-1}$, $[H]_{tot} = 100\ \mu M$, $[S]_{tot} = 126\ \mu M$, and $[P]_{tot} = 54\ \mu M$. In these diagrams, chemical species are represented by hatched boxes; the concentration of a species is proportional to the area of its box. Substrate is denoted by vertical hatches, host by oblique hatches, and product by horizontal hatches. This figure shows the concentrations assumed by the computer before any equilibria are considered. Host, substrate, and product are all uncomplexed.

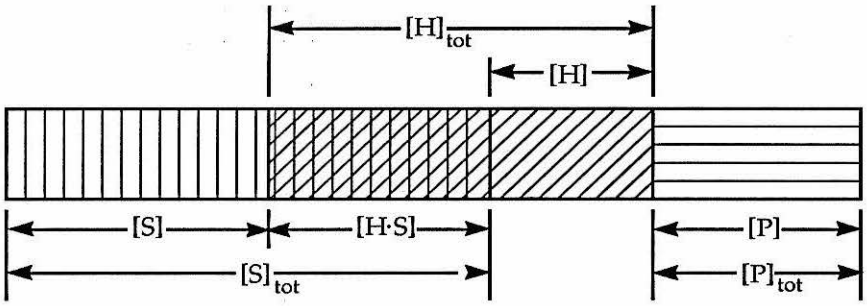


Figure 11. First stage of the first iteration. Some host and substrate have associated; [H·S] = 57.73 μM. The region of intersection between the vertically hatched box (substrate) and the diagonally hatched box (host) represents host/substrate complex.

The first iteration is now complete. The accuracy of all concentrations will be considered adequate if the pertinent equilibrium relations are "close enough" to being satisfied. In this case, these equilibrium relations are equations 7 and 8, for substrate and product binding to host. The degree to which a relation is satisfied is expressed by the fractional error in the relation. For example, for the generalized

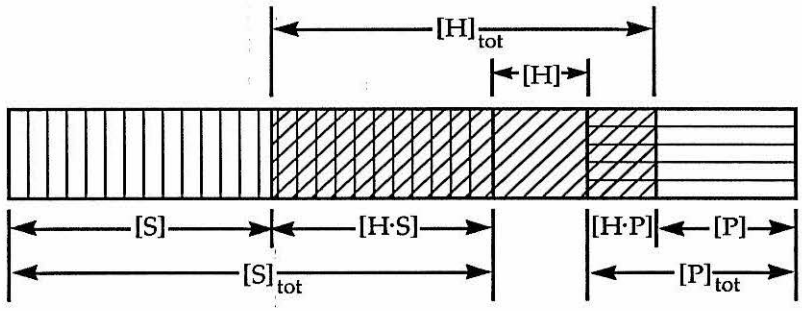


Figure 12. Second step of the first iteration. The product binding reaction has been considered. [H·S] still is 57.73 μM, and now [H·P] is 17.76 μM. Host/product complex is represented by the intersection of the host and product boxes.

reaction A + B $\rightleftharpoons$ C, the equilibrium relation is $K = [C]/([A][B])$. Estimated values of [A], [B], and [C] can be tested by this relation if K is known. The fractional error of these concentrations is given by equation 13.

$$\text{fractional error} = \frac{1}{K}\left|\frac{[C]}{[A][B]} - K\right| \tag{13}$$

This is the absolute value of the difference between the calculated reaction quotient and the actual equilibrium constant, expressed as a fraction of the actual equilibrium constant. The closer [A], [B], and [C] are to the proper equilibrium values, the closer to zero this fractional error will be. If it falls below an arbitrary cutoff value known as the *convergence criterion*, the concentration estimates are considered adequately consistent with the equilibrium constant.

At the end of an iteration, equilibrium relation 8 will always be perfectly satisfied because there has been no change in the concentration estimates since it was last applied. Consequently, it is only necessary to evaluate the concentration estimates with respect to relation 7.

If the equilibrium relations are not closely enough satisfied, another iteration is performed. As before, [H·S] will be estimated first, follwed by [H·P]. Only species

appearing explicitly in the equilibrium relation under consideration are used. To determine [H·S], the host employed as H·P is ignored, and the total host concentration, as far as the substrate is concerned, is [H] + [H·S]. To satisfy the substrate binding relation, [H], [S], and [H·S] will change. At this stage, the change involves dissociation of host/substrate complex into free host and substrate, because the most recent perturbation decreased [H] without affecting [S] or [H·S]. This step is shown in Figure 13.

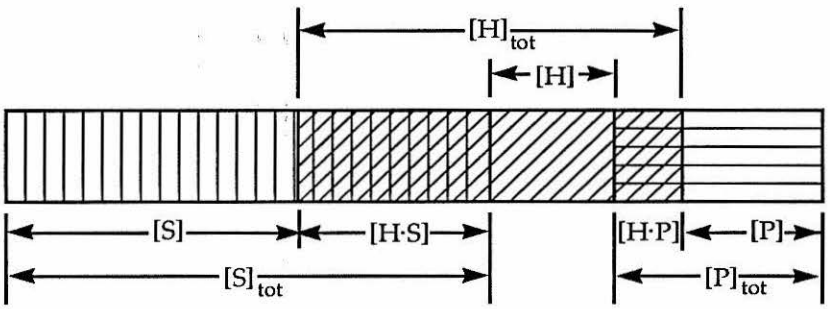


Figure 13. First stage of the second iteration. Host/substrate complex has dissociated in response to removal of free host by product. Now [H·S] is 49.69 μM.

As before, the second stage of the iteration is to satisfy the host/product association relation. As before, host in the form H·S is invisible to the product, so the effective total host concentration is [H] + [H·P]. The concentrations to be altered are [H], [H·P], and [P]. This will involve association of free host and product to form more host/product complex, since the last step generated more free host. Figure 14 illustrates this second step. Because the concentration estimates are improving, the change involved is very small.

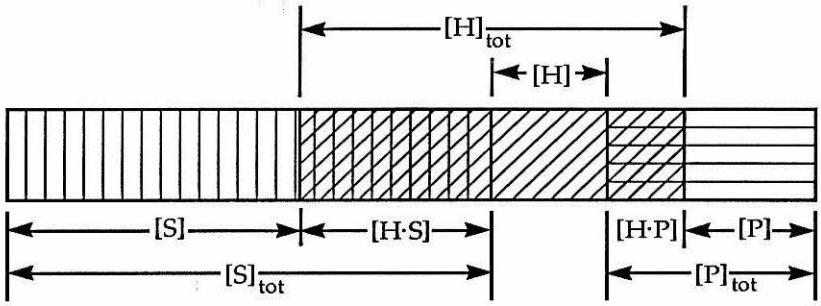


Figure 14. Second stage of the second iteration. Since more host is now available, product binds to some of it. [H·P] now has increased to 20.03 μM.

The second interation is now complete. The fractional error in relation 7 is again evaluated; it will be lower than in the previous iteration. Further iterations are performed until the fractional error is less than the convergence criterion. All following iterations are qualitatively identical to the second.

2. Other association equilibria. The simulator is also set up to model, in a similar manner, two additional association processes: competitive binding of alkylating agent to host (equation 14), and competitive binding of an additional, nonnucleophilic guest to the host (equation 15).

$$\mathrm{H} + \mathrm{A} \rightleftharpoons \mathrm{H{\cdot}A};\ K_\mathrm{A} = \frac{[\mathrm{H{\cdot}A}]}{[\mathrm{H}][\mathrm{A}]} \tag{14}$$

$$\mathrm{H} + \mathrm{I} \rightleftharpoons \mathrm{H{\cdot}I};\ K_\mathrm{I} = \frac{[\mathrm{H{\cdot}I}]}{[\mathrm{H}][\mathrm{I}]} \tag{15}$$

The host/alkylating agent interaction was introduced because David Stauffer noticed that host and substrate appear to associate less in the presence of large concentrations of iodomethane. Competitive binding of host by iodomethane was hoped to provide a means to simulate this effect without introducing a new K_S and K_P whenever the concentration of iodomethane changed. Most likely, the reduced affinity of host for substrate when the concentration of iodomethane is large is a

result of a change in the solvent structure, and not of competition by iodomethane. *No* evidence of association of host with iodomethane has been observed in direct binding studies. Nonetheless, binding of iodomethane can be modeled, if desired.

The capability to simulate competitive inhibition was added to the program because such a process can be mechanistically informative. If a competitor of known binding constant K_I is added to a host-catalyzed alkylation, observation of the expected inhibition of the alkylation would confirm that association of the host and substrate is indeed responsible for the observed rate acceleration, and that the inhibitor associates with the same region of the host as the substrate.

All that was required to introduce these processes to the simulation was to add a step in the equilibrium iteration in which the concentrations of free host, free alkylating agent, and host/alkylating agent complex were made consistent with the value of K_A, and another step in which the concentrations of free host, free inhibitor, and host/inhibitor complex were made consistent with the value of K_I. Of course, after these steps are executed, it is necessary to check the concentrations of all species involved in other processes for convergence with respect to their own equilibrium constants. The sequence of association steps is repeated until such adequate convergence is reached.

C. Computation of Reaction Progress.

Once the concentrations of free and bound substrate are known, determining the instantaneous reaction rate is trivial. This rate is given by equation 6, repeated here for easy reference.

$$\text{total rate: } \frac{d[\mathrm{P}]_{\mathrm{tot}}}{dt} = [\mathrm{A}]\left(k_u[\mathrm{S}] + k_c[\mathrm{H{\cdot}S}]\right) \tag{6}$$

The reaction is followed in a series of steps. In each step, $d[\mathrm{P}]_{\mathrm{tot}}/dt$ is determined from [A], [S], and [H·S]. Multiplication of this rate by a time increment Δt gives a

first-order approximation to Δ[P], the amount of product formed during that time increment. This means that there will now be new values of the total concentrations of substrate, product, and alkylating agent.

$$\Delta[\mathrm{P}] \simeq \Delta t \, \frac{d[\mathrm{P}]_{\mathrm{tot}}}{dt} \tag{16}$$

$$[\mathrm{P}]_{\mathrm{tot}\,i} = [\mathrm{P}]_{\mathrm{tot}\,i-1} + \Delta[\mathrm{P}] \tag{17}$$

$$[\mathrm{S}]_{\mathrm{tot}\,i} = [\mathrm{S}]_{\mathrm{tot}\,i-1} - \Delta[\mathrm{P}] \tag{18}$$

$$[\mathrm{A}]_{\mathrm{tot}\,i} = [\mathrm{A}]_{\mathrm{tot}\,i-1} - \Delta[\mathrm{P}] \tag{19}$$

When these new total values are in hand, the association equilibria can again be solved, leading to a new instantaneous reaction rate. Multiplication of this new rate by the time increment gives a new Δ[P]. The appearance of product with the passage of time is calculated in a sequence of such steps.

This method always overestimates the reaction rate. The rate calculated by applying equation 6 to [S] and [H·S] is the rate at the *beginning* of the time interval; throughout the rest of the interval, as substrate is consumed, the rate continuously decreases. Thus, it is important to use small time intervals Δt so that the calculated Δ[P] for that interval does not overshoot the correct value by too much.

The size of the time interval Δt reflects the reaction rate $d[\mathrm{P}]_{\mathrm{tot}}/dt$. At the start of the reaction, Δt is set so that Δ[P] is equal to some arbitrary fraction of the initial substrate concentration. This fraction, the *increment*, is chosen by the user. As substrate is consumed, the rate decreases, so that Δ[P] becomes smaller and smaller. If Δ[P] falls below another arbitrary user-determined parameter known as the *floor*, the time increment is recalculated to bring Δ[P] back to its initial size. In this manner, short time steps are taken in the initial stages of the simulation when the rate is fastest, and longer time steps are taken later. This improves the accuracy obtainable by a simulation with a set number of time steps. Because the

change in rate is greatest when the rate is fastest, the first-order approximation is its least valid here. Accuracy is improved in this critical region by using smaller time intervals. As the rate changes more slowly in the latter part of the reaction, the time interval can be lengthened with no loss in accuracy.

D. Comparison with Experimental Data.

In order to determine the RMS deviation (equation 9), one must know the predicted total product concentration at the times the data points were taken. These times are not likely to be the same as the times encountered in the step-by-step reaction simulation. It is necessary then to interpolate between the simulated points. Fortunately, this is as easy as can be imagined: the first-order approximation *assumes* that product appearance with time is linear between steps. Thus, determination of the calculated total product concentration at any arbitrary time requires only linear interpolation between the bracketing points.

E. Experimental Section

Details of the procedures of the kinetic experiments can be found in earlier reports.[4,5] This section reports the verification of the Kinetics Simulator program.

The Kinetics Simulator exists to model kinetic behavior which, in general, cannot be modeled analytically. There are special cases, however, in which the kinetics *can* be analytically modeled, or at least very closely approximated. In such cases, the findings of the Kinetics Simulator can be compared to the analytical predictions.

1. No association, no catalysis. The first and simplest case to be verified was the simple uncatalyzed reaction under pseudo-first-order conditions. If $K_S = 0$, there is no reaction by the catalyzed pathway, and the rate is described by equation 20.

$$\text{rate} = \frac{d[\text{P}]}{dt} = k_u[\text{A}][\text{S}] = k_u[\text{A}]_{\text{tot}}\,[\text{S}]_{\text{tot}} \tag{20}$$

If $[A]_{tot\,0} \gg [S]_{tot\,0}$, then $[A]_{tot}$ will be effectively constant throughout the course of the reaction. Thus, $[A]_{tot} \simeq [A]_{tot\,0}$, and

$$\frac{d[P]_{tot}}{dt} \simeq k_u [A]_{tot\,0} [S]_{tot} . \tag{21}$$

This differential equation is readily solved to yield

$$[P]_{tot} = [P]_{tot\,0} + [S]_{tot\,0}\left\{1 - \exp(-[A]_{tot\,0}\, k_u t)\right\} . \tag{22}$$

Initial conditions for this test were arbitrarily chosen as $K_S = K_P = 0$, $k_u = 10^{-5}$ $M^{-1}s^{-1}$, $k_c = 0$, $[H]_{tot} = 10$ μM, $[A]_{tot\,0} = 10$ mM, $[S]_{tot\,0} = 100$ μM, and $[P]_{tot\,0} = 0$. These values were substituted into equation 22, and also used as input for the Kinetics Simulator. Three different preference sets, as described in the introduction to the Kinetics Simulator User's Manual, were used in the simulations. The results are shown in Table IV, and the preference sets in Table V.

Table IV. Comparison of results from the Kinetics Simulator with predictions from the pseudo-first-order approximation for an uncatalyzed alkylation.

time[a]	$[P]^{b,c}_{calc}$	$[P]^{b,d}_{sim}$	$[P]^{b,e}_{sim}$	$[P]^{b,f}_{sim}$
1 000	0.598	0.6	0.6	0.598
10 000	5.824	6.0	5.851	5.825
50 000	25.918	27.077	26.003	25.901
100 000	45.119	47.214	45.203	45.059
300 000	83.470	86.833	83.619	83.351
500 000	95.021	99.923	95.234	94.965

[a]time in minutes. [b]Total product concentration in μmol l^{-1}. [c]Calculated using equation 22. [d]Simulated using preference set 1. [e]Simulated using preference set 2. [f]Simulated using preference set 3.

Table V. Preference sets used for verification of the Kinetics Simulator.

set	increment	floor	con. crit.	tail
1 ("screaming")	0.1	0.95	0.001	0.01
2 ("default")	0.01	0.5	0.0001	0.05
3 ("careful")	0.001	0.95	0.0001	0.01

Note that the product concentrations simulated using preference sets 1 and 2 are higher than the values calculated with equation 22. This is to be expected; as discussed in subsection C, the simulator overestimates the rate. The product concentrations simulated using preference set 3, however, are actually *lower* than the calculated values. This parameter set requires the simulator to take a very short time step, making its predictions the most accurate. Although simulations using this parameter set still overestimate the reaction rate, the predictions in this case are even better than those from equation 22, whose first-order approximation overestimates the kinetics even worse.

2. Association. In general, the concentration of free host cannot be analytically determined when host binds to both substrate and product. Furthermore, since total host concentration is constant, but total substrate concentration is not, the fraction of substrate bound will change as substrate is consumed. If, however, substrate and product have the same binding constant, the fraction of free and bound host will always remain the same, as will the fractions of free and bound substrate and product. In this special case of $K_S = K_P$, the reaction kinetics can be modeled analytically. The formula for the kinetics in this situation can be derived as follows.

$$\begin{aligned}
\text{rate} = -\frac{d[\mathrm{S}]_{\mathrm{tot}}}{dt} &= k_u[\mathrm{A}]_{\mathrm{tot}}\,[\mathrm{S}] + k_c[\mathrm{A}]_{\mathrm{tot}}\,[\mathrm{H{\cdot}S}] \\
&= k_u[\mathrm{A}]_{\mathrm{tot}}\,[\mathrm{S}] + k_c[\mathrm{A}]_{\mathrm{tot}}\,K_S[\mathrm{H}][\mathrm{S}] \\
&= [\mathrm{A}]_{\mathrm{tot}}\,[\mathrm{S}]\big(k_u + k_c K_S[\mathrm{H}]\big) \\
&= [\mathrm{A}]_{\mathrm{tot}}\,\frac{[\mathrm{S}]_{\mathrm{tot}}}{1 + K_S[\mathrm{H}]}\big(k_u + k_c K_S[\mathrm{H}]\big) \\
\frac{d[\mathrm{S}]_{\mathrm{tot}}}{[\mathrm{S}]_{\mathrm{tot}}} &= -\frac{[\mathrm{A}]_{\mathrm{tot}}\big(k_u + k_c K_S[\mathrm{H}]\big)}{1 + K_S[\mathrm{H}]}\,dt
\end{aligned} \tag{23}$$

We know that [H] is constant with respect to time. $[\mathrm{A}]_{\mathrm{tot}}$ is not, but it is approximately so when alkylating agent is present in great excess, and can be represented by $[\mathrm{A}]_{\mathrm{tot}\,0}$.

$$\frac{d[\mathrm{S}]_{\mathrm{tot}}}{[\mathrm{S}]_{\mathrm{tot}}} \simeq -\frac{[\mathrm{A}]_{\mathrm{tot}\,0}\left(k_u + k_c K_{\mathrm{S}}[\mathrm{H}]\right)}{1 + K_{\mathrm{S}}[\mathrm{H}]}\, dt$$

$$\int_{[\mathrm{S}]_{\mathrm{tot}\,0}}^{[\mathrm{S}]_{\mathrm{tot}\,t}} \frac{d[\mathrm{S}]_{\mathrm{tot}}}{[\mathrm{S}]_{\mathrm{tot}}} \simeq -\frac{[\mathrm{A}]_{\mathrm{tot}\,0}\left(k_u + k_c K_{\mathrm{S}}[\mathrm{H}]\right)}{1 + K_{\mathrm{S}}[\mathrm{H}]} \int_0^t dt$$

$$\ln\frac{[\mathrm{S}]_{\mathrm{tot}\,t}}{[\mathrm{S}]_{\mathrm{tot}\,0}} \simeq -\frac{[\mathrm{A}]_{\mathrm{tot}\,0}\left(k_u + k_c K_{\mathrm{S}}[\mathrm{H}]\right)}{1 + K_{\mathrm{S}}[\mathrm{H}]}\, t$$

$$[\mathrm{S}]_{\mathrm{tot}} \simeq [\mathrm{S}]_{\mathrm{tot}\,0}\, \exp\left\{-t\, \frac{[\mathrm{A}]_{\mathrm{tot}\,0}\left(k_u + k_c K_{\mathrm{S}}[\mathrm{H}]\right)}{1 + K_{\mathrm{S}}[\mathrm{H}]}\right\}$$

$$[\mathrm{P}]_{\mathrm{tot}} \simeq [\mathrm{P}]_{\mathrm{tot}\,0} + [\mathrm{S}]_{\mathrm{tot}\,0}\left[1 - \exp\left\{-t\, \frac{[\mathrm{A}]_{\mathrm{tot}\,0}\left(k_u + k_c K_{\mathrm{S}}[\mathrm{H}]\right)}{1 + K_{\mathrm{S}}[\mathrm{H}]}\right\}\right] \qquad (24)$$

[H], which is constant throughout the reaction, can be found from the usual association equilibrium expression

$$[\mathrm{H{\cdot}G}] = \frac{1}{2}\left\{[\mathrm{H}]_{\mathrm{tot}} + [\mathrm{G}]_{\mathrm{tot}} + 1/K - \sqrt{\left([\mathrm{H}]_{\mathrm{tot}} + [\mathrm{G}]_{\mathrm{tot}} + 1/K\right)^2 - 4\,[\mathrm{H}]_{\mathrm{tot}}[\mathrm{G}]_{\mathrm{tot}}}\right\} \qquad (25)$$

substituting $\left([\mathrm{S}]_{\mathrm{tot}} + [\mathrm{P}]_{\mathrm{tot}}\right)$ for $[\mathrm{G}]_{\mathrm{tot}}$.

No catalysis. The next case considered was one in which the substrate would bind to host, but would not be alkylated in the bound state. Under these conditions, host would actually inhibit the reaction. The initial conditions chosen for this test case were $K_{\mathrm{S}} = K_{\mathrm{P}} = 0$, $k_u = 10^{-5}\ \mathrm{M^{-1}s^{-1}}$, $k_c = 0$, $[\mathrm{H}]_{\mathrm{tot}\,0} = 100\ \mu\mathrm{M}$, $[\mathrm{A}]_{\mathrm{tot}\,0} = 10\ \mathrm{mM}$, $[\mathrm{S}]_{\mathrm{tot}\,0} = 100\ \mu\mathrm{M}$, and $[\mathrm{P}]_{\mathrm{tot}\,0} = 0$. The product concentration is given by equation 24, which in this case reduces to equation 26.

$$[\mathrm{P}]_{\mathrm{tot}} = 10^{-4}\,\mathrm{M}\,\left[1 - \exp(-t\ 4.3923 \times 10^{-6}\,\mathrm{min^{-1}})\right] \qquad (26)$$

The predictions of equation 26 and the Kinetics Simulator are compared in Table VI.

Table VI. Comparison of results from the Kinetics Simulator with predictions from the pseudo-first-order approximation for a host-inhibited alkylation.

time[a]	$[P]^{b,c}_{calc}$	$[P]^{b,d}_{sim}$	$[P]^{b,e}_{sim}$	$[P]^{b,f}_{sim}$
1 000	0.438	0.439	0.439	0.438
10 000	4.297	4.392	4.316	4.298
50 000	19.717	20.576	19.788	19.709
100 000	35.547	37.171	35.635	35.510
300 000	73.224	73.165	73.329	73.103
500 000	88.877	91.969	89.033	88.774

[a]time in minutes. [b]Total product concentration in μmol l^{-1}. [c]Calculated using equation 26. [d]Simulated using preference set 1. [e]Simulated using preference set 2. [f]Simulated using preference set 3.

Catalysis. The initial conditions chosen were identical to those in the previous system, except that $k_c = 10^{-4}\,M^{-1}s^{-1}$. This makes the concentration prediction of equation 24 equal to

$$[P]_{tot} = 10^{-4}\,M\left[1 - \exp(-t\ 2.04692 \times 10^{-5}\,min^{-1})\right]. \qquad (27)$$

The predictions of equation 27 and of the Kinetics Simulator are given in Table VII.

Table VII. Comparison of results from the Kinetics Simulator with predictions from the pseudo-first-order approximation for a host-catalyzed alkylation.

time[a]	$[P]^{b,c}_{calc}$	$[P]^{b,d}_{sim}$	$[P]^{b,e}_{sim}$	$[P]^{b,f}_{sim}$
1 000	2.026	2.047	2.036	2.027
3 000	5.956	6.141	5.983	5.957
10 000	18.510	19.371	18.578	18.503
20 000	33.359	33.062	33.683	33.562
50 000	64.065	66.954	64.172	63.959
100 000	87.087	90.493	87.243	86.977

[a]time in minutes. [b]Total product concentration in micromolar. [c]Calculated using equation 27. [d]Simulated using preference set 1. [e]Simulated using preference set 2. [f]Simulated using preference set 3.

III. Kinetics Simulator User's Manual

A. Introduction.

1. Philosophy. The Kinetics Simulator is a Macintosh program written in Microsoft QuickBasic. It has many features familiar to users of standard Macintosh applications: pull-down menus, edit windows, and dialog boxes. This similarity to standard applications, though intentional, unfortunately does not run very deep. When edit windows or dialog boxes are open, for instance, nothing but the forward windows can be made active, and all menus are disabled. The principal edit window does *not* employ the standard, familiar Macintosh text editing package TextEdit. Only one file may be opened at any time. Window sizes and and captions are *not* contained in a separate resource file to allow alterations without recompiling. These deviations from standard performance all exist to simplify the programming. Making a more ordinary-feeling Macintosh application would have required much longer and more complex code, a knowledge of the inner workings of the Macintosh that could only come from a thorough study of the many large and expensive *Inside Macintosh* volumes, and quite probably the use of a programming environment other than QuickBasic. Since the Kinetics Simulator has the modest purpose of determining k_c from the kinetic data for a specific set of reactions studied in the Dougherty group, it seemed wise not to expend the inordinate effort required to make it into a full-fledged application.

2. User interface. Despite these deficiences, there are still a number of things the Kinetic Simulator *can* do. It allows interactive determination of the best fit k_c to the experimental data. The data are plotted on the screen, along with the kinetic behavior expected from the parameters input by the user. Thus, the user may evaluate the similarity of the predicted behavior to the data by visual inspection as well as by the reported RMS deviation. The data and the predicted curve can be exported in two ways. A tabular file of the experimental and predicted

$(t, [\mathrm{P}]_{\mathrm{tot}})$ points, which can be read by such graphing applications as Cricketgraph and Kaleidagraph, is one option. The other is a text output file more readable by humans, which reports all parameters used in the simulation as well as the simulated and experimental time points.

3. Input. Input to the program is handled in several ways. The principal measured data ($[\mathrm{H}]_{\mathrm{tot}\,0}$, $[\mathrm{S}]_{\mathrm{tot}\,0}$, $[\mathrm{P}]_{\mathrm{tot}\,0}$, $[\mathrm{A}]_{\mathrm{tot}\,0}$, K_{S}, K_{P}, k_u, and the observed concentrations of product at later times) are entered in an edit window that appears upon selecting "Edit" from the "Data" menu, or striking ⌘E. This information can also be saved to an external file which can be recalled as input later. The rate constant k_c of the catalyzed reaction is entered in its own edit window, which appears upon selecting "New k(cat)" from the "Data" menu, or striking ⌘K. The equilibrium constant K_{A} for association between host and alkylating agent also is entered in its own edit window. This window is accessed by selecting "New Ka" from the "Data" menu or striking ⌘A. Similarly, there is also a separate edit window for entering the equlibrium constant for association between host and a competitive inhibitor, and the total concentration of this competitive inhibitor. This window appears when "New Ki" is selected from the "Data"menu, or when ⌘I is typed.

4. Association equilibria. The kinetic behavior can be predicted from models involving a number of association equilibria. The default model involves association of the host with both substrate and product, but the user may elect to consider association of the host with the alkylating agent or with an added competitive inhibitor as well. In addition, if any other association equilibrium is desired, it could be included with only minor modification of the source code.

5. Preferences. The user may set four preferences that govern the simulation. Up to ten groups of preferences may be saved to an external data file and altered or recalled at will by the user. These preferences are the *convergence criterion*, the

tail, the *increment*, and the *floor*. Modification of these preferences allows the user to tailor the simulation to his own needs for speed, accuracy, and memory usage.

The *convergence criterion* tells the maximum acceptable fractional error in the concentrations of associating species. It must be greater than zero. The fractional error for an equilibrium relation is defined in equation 13. If the fractional errors of *all* pertinent association equilibria are below this criterion, then the concentrations are accepted. No further iterations for refining the concentrations need to be performed.

The *tail* specifies the minimum fraction of substrate remaining at the end of a simulation. For example, if the tail is 0.05, the simulation will not proceed beyond 95% completion. As long as the value of this quantity is small, it will have no effect on a simulation. This parameter exists primarily for historical reasons. If the simulated reaction rate is so fast that the reaction would proceed to nearly total completion in the time specified by the observations, simulating the entire time course of the reaction may require more steps than will fit in the arrays. Halting the simulation at a set maximum completion is one way to prevent such an overflow.

The *increment* is the fraction of the initial concentration of substrate that is converted into product in the first simulation step. It must be between zero and one, exclusive. The value of the increment is used to calculate the length of the time step. If the increment is b, for instance, the concentration of product formed in the first simulation step is $\Delta[\mathrm{P}] = b[\mathrm{S}]_{\mathrm{tot}\,0}$. The time increment Δt is obtained by rearranging the approximate relation (equation 16), to give equation 28.

$$\Delta t = \frac{\Delta[\mathrm{P}]}{d[\mathrm{P}]_{\mathrm{tot}}/dt} = \frac{d[\mathrm{S}]_{\mathrm{tot}\,0}}{[\mathrm{A}](k_c[\mathrm{H{\cdot}S}] + k_u[\mathrm{S}])} \tag{28}$$

The final preference is the *floor*. It may be any number from zero to one, inclusive. This allows the simulator to gracefully account for decreasing reaction

rates caused by scarce substrate or by product inhibition. If the concentration of product formed in a time step is less than the floor multiplied by the amount of product formed in the first time step, $\Delta[\mathrm{P}] < \text{floor} \times b[\mathrm{S}]_{\mathrm{tot}\,0}$, a new value for the time increment Δt will be obtained by again employing equation 28.

B. Tutorial.

Suppose that you have studied the kinetics of alkylation of the newly-discovered alkaloid beemeramine with iodomethane in borate-*d*, both uncatalyzed and in the presence of host P. You have indepedently determined that the binding constant of beemeramine with host P is 20 000 M^{-1}, that the binding constant of N-methylbeemerammonium iodine with host P is 100 000 M^{-1}, and that the kinetics of the uncatalyzed reaction have revealed a rate constant k_u of 5×10^{-4} $\mathrm{M}^{-1}\mathrm{s}^{-1}$. An experiment in which host was 100 μM, iodomethane 10 μM and beemeramine 150 μM in borate-*d* gave the "without inhibitor" kinetic data in Table VIII. An additional experiment with the same initial host, substrate, and iodomethane concentrations, but with the reaction mixture also 100 μM in the competitive inhibitor ATMA, $K_{\mathrm{I}} = 80\,000$, gave the "with inhibitor" data in the same table.

To start the program, double-click on the Kinetics Simulator icon. The menu bar will change, but no new windows will appear on the desktop. To enter your kinetic data, bring up the data edit window by typing ⌘N or selecting "New" from the "Data" menu. Once this edit window is active, you may enter data into any of the edit fields by clicking on the edit field, clicking on the button adjacent to the edit field, or hitting "¬"[15] or "Enter" from the previous edit field. Note that you must enter k_u in units of $\mathrm{M}^{-1}\mathrm{s}^{-1}$, and concentrations in units of M. This means, for instance, that "`100e-6`" should be entered for $[\mathrm{H}]_{\mathrm{tot}}$. Try your hand at entering impossible values in some of the fields, such as negative concentrations or non-numeric entries such as "`Dennis`," and see how the program responds.

(15) As in Chapter 4, the symbol "¬" here denotes a carriage return.

Table VIII. Kinetic data for the Kinetics Simulator tutorial.

without inhibitor		with inhibitor	
time[a]	[P][b]	time[a]	[P][b]
0	0	0	0
30	23	30	12
60	42	60	24
90	57	90	33
120	68	120	41
150	77	150	49
180	86	180	56
210	93	240	68
240	98	300	78
270	103	360	86
300	108	420	93
330	112	480	100
360	115		
390	119		
420	120		
450	123		
480	125		

[a]Time in minutes. [b]Total product concentration in μM.

Once you have entered the appropriate data into this window, bring up the kinetic data edit window by clicking on the "kinetic data" button or by striking "¬" or "Enter" from the P(init) edit field. This kinetic data edit window has edit fields in two columns: the first for time, and the second for total concentration of product. Enter the "without inhibitor" data from Table VIII into the proper columns. Note that in *this* window, product concentration must be entered in units of μM, not M. Striking "¬" or "Enter" from an edit field will send you to the field directly below while striking "Tab" will send you to the edit field to the right. Advancing from the last row in the window will cause all the edit fields to scroll down. You may also scroll the window one row at a time by clicking on one of the two buttons at the right of the window; the upper button will scroll up, and the lower button will scroll down.

When you are finished entering the kinetic data, close the kinetic data edit window and the data edit window by clicking in their close boxes (upper left-hand

corner). The kinetic data will now be plotted in a large display window. Save the data you just entered by typing ⌘S or selecting "Save" from the "Data" menu. A file dialog box will appear, asking you to name the data. Call it "`Busywork`" or whatever you prefer. Note that when you do this, the title on the graph reflects the new data name.

Now it is time to find a best-fit k_c. Bring up the k_c edit window by typing ⌘K or selecting "New k(cat)" from the "Data" menu, and enter an intial guess for k_c; start with "`0.001`." As soon as you close the k_c edit window, the computer will begin calculating the resulting kinetic behavior. When it is finished, it will plot this curve on the graph with the data. Visual inspection of the curve, as well as the RMS deviation, will reveal that this initial guess is far too low.

Before entering any other guesses, however, bring up the "Preferences" menu. It will show a check next to "Default." If there are other options available, such as "Faster" or "Screaming," select one of these. This will allow faster calculations at the cost of some accuracy. Re-calculate the kinetic behavior under this new preferences set by typing ⌘C or selecting "Calculate" from the "Data" menu. Note that the curve plotted is more angular, and shows slightly faster product formation, than that obtained using the default preferences.

Now refine your estimate of k_c by entering values to make the simulated curve conform more closely to the data. When you have gotten fairly close, change the preference set to something more accurate, such as "Slower" or "Careful." Perform your last few refinements using these preferences, which minimize the errors inherent in the simulation.

Finally, save the best-fit run for closer inspection and for plotting with a graphing program. For a file that can be read by Cricketgraph or Kaleidagraph, type ⌘G or select "Tabular" from the "Output" menu; for a file that makes more sense to

a human, type ⌘R or select "Save Run" from the "Output" menu. In both cases, you will encounter a file dialog box asking what to name the file and where to put it. When these output files have been created, quit the program by typing ⌘Q or selecting "Quit" from the "Output" menu.

To examine the human-readable output file, open it from within any application that edits text files, such as Microsoft Word. If you want a hard copy of this file, print it from within your application.

The tabular output file can be opened directly from within Cricketgraph. Just be sure the "Show all TEXT files" box is checked in the Cricketgraph file dialog box. To open this file from within Kaliedagraph, select (one) "Tab" as the column delimiter, skip one line, and select the "Read Titles" box. Making graphs from this data file is accomplished by normal means within the graphing application. In Kaleidagraph, however, if you wish to draw a line graph of the simulated kinetic behavior, the data must first be sorted by the time column, or stray lines will apppear.

Now that you are familiar with the output files that the Kinetics Simulator generates, run the program again to explore some additional features. (Do this by again double-clicking on the Kinetics Simulator icon.) Open the file you saved earlier by typing ⌘O or selecting "Open" from the "Data" menu, and picking the file's name from the file dialog box that comes up. Call up the k_c edit window (⌘K) and enter the best-fit k_c you found earlier. (If you can't remember it, make a guess and refine it the way you did before.) This time try to find the binding constant of the host with the alkylating agent that gives the best fit. To alter this K_A, call up its edit window by typing ⌘A or selecting "New Ka" from the "Data" menu. Enter a guess in the edit field (5 M^{-1} is a good starting value), close the edit window, and the computer will calculate and plot the resulting kinetic behavior. Notice that this

calculation process is incredibly slow; no, the computer has not "hung up." This is just a property of the method for calculating equilibrium concentrations for some sets of initial conditions. And no, you cannot use some other Macintosh application under the Multifinder while you are waiting, because all of the menus are disabled. Sorry. To the extent that your patience endures, optimize k_c in a series of trials, and then make another guess at K_A. Continue varying both K_A and k_c until you can no longer improve the fit.

Next, look at the "with inhibitor" data from Table VIII. This is the data from the catalyzed reaction in the presence of a competitive inhibitor. You could enter this data exactly as you did before by typing ⌘N or selecting "New" from the "Data" menu and typing the numbers into the empty edit fields, but, since the initial conditions of this reaction are so similar to those in the previous reaction, you can save a little time by doing something different. Select "Save as..." from the "Data" menu, and give a new name, such as "`Inhibited Busywork`," to your data. Then edit the kinetic data associated with this file by typing ⌘E or selecting "Edit" from the "Data" menu, and clicking on the "Kinetic Data" button in the data edit window. Then enter the "with inhibitor" data from Table VIII into the kinetic data edit window, overwriting the previous values. Remove any left-over values from the previous data set. When these windows are closed, the new kinetic data will be plotted on the screen. Before setting k_c, however, enter the concentration and binding constant of the added inhibitor. Call up the K_I edit window by typing ⌘I or selcting "New Ki" from the "Data" menu, and enter "`80000`" for K_I and "`100`" for $[I]_{tot}$. Note that $[I]_{tot}$ is entered in units of μM. When this K_I edit window is closed, the computer will *not* automatically calculate the kinetic behavior; it is waiting for a value of k_c or K_A. First, bring up the K_A edit window and re-set K_A to zero. Then, enter a guess for k_c and refine it as before.

Congratulations! You are now familiar with most of the features of the Kinetics Simulator!

C. Kinetics Simulator Reference.

1. Menus.

(a) The Output menu.

Print Graph(⌘P): This command performs a screen dump to the Laser Writer, allowing the plot on the screen to be printed out directly from the program. It does not work on our Macintosh IIfx, however, because the screen is too big. As a result, this menu option is always disabled. If a hard copy of the plot of the experimental and simulated kinetic data is desired, make a "Tabular" output file and work on it within a graphing program.

Save Run(⌘R): This command writes information about the experminental and simulated kinetic behavior to an ordinary text file. It reports the preferences used in the simulation, as well as the physical quantites $[H]_{tot\,0}$, $[S]_{tot\,0}$, $[P]_{tot\,0}$, $[A]_{tot\,0}$, $[I]_{tot}$, K_S, K_P, K_A, K_I, k_c, k_u, and the RMS deviation of the simulated run from the data. Then it lists time, observed product concentration, and simulated product concentration for each data point reported.

Tabular(⌘T): This command writes the simulated and experimental kinetic data to a tab-delimited text file that can be read by graphing programs such as Cricketgraph and Kaleidagraph. The first line of this file is an asterisk, which is a signal to Cricketgraph that the next line contains column headings. Naturally, the second line of this file contains the column headings. The subsequent lines contain the data. This is arranged into three columns: time, simulated total product concentration, and experimental total product concentration. The first lines report the experimental data: time in the first column, the interpolated value of the simulated total product concentration in the second column, and the experimentally observed total product concentration in the third column. The remaining lines report the simulated kinetic behavior. Since there are no experimental observations for these times, the third column is always empty. The lines reporting the experimental and

interpolated numbers are in order with respect to time, as are the lines reporting only the simulated numbers, but because these lists are in successive blocks, the overall list is *not* in order with respect to time. This is not a problem in Cricketgraph, but it *is* a problem in Kaleidagraph if the simulated kinetic behavior is plotted as a line graph. In such a case, the program draws a line from each point to the one listed after it, giving a graph like that shown in Figure 15.

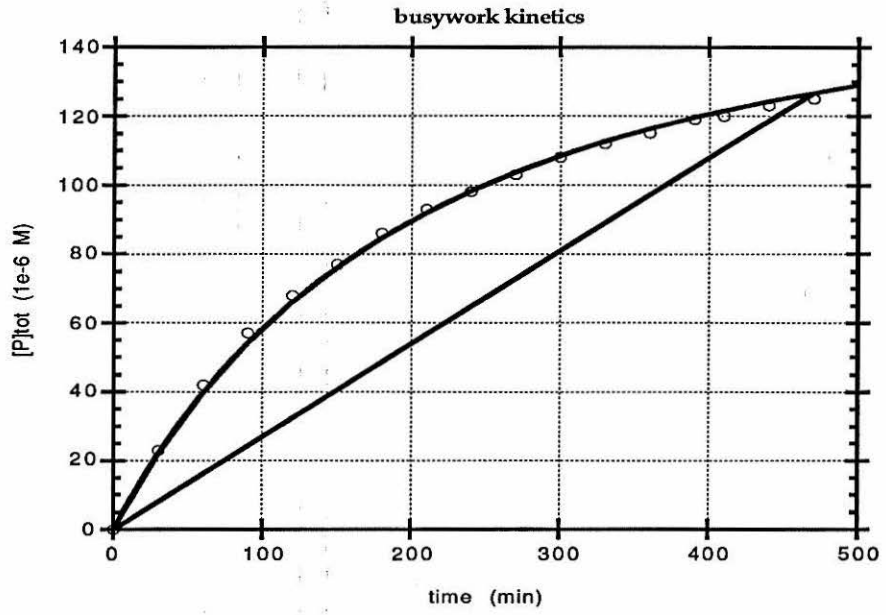


Figure 15. Kaleidagraph rendering of an unsorted tabular output file.

This difficulty can be corrected in several ways. The interpolated data (the elements in the second column of lines that have three columns) can be moved to a forth column, masked, or removed, or the entire file can be sorted by time.

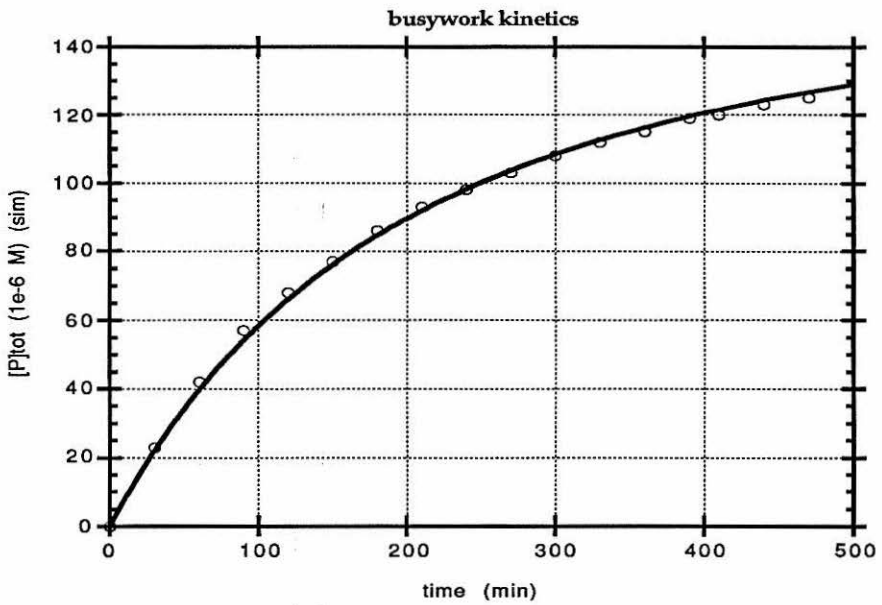


Figure 16. Kaleidagraph rendering of a tabular output file after sorting.

Quit(⌘Q): This command has the same effect in the Kinetics Simulator as in any standard application: it terminates the program. If the data currently in use have not been saved, the user will be asked if a data file should be created before quitting.

(b) The Data menu.

Open (⌘O): This command reads data from a previously saved file. If the data currently in use have not been saved, the user will be asked if the current data should be saved before opening a new file. This is because the program does not allow more than one set of data to be open at a time. When this command is selected, a file dialog box will prompt the user for the input file. After the file is read, the kinetic data are plotted on the screen.

New (⌘N): This command brings up the data edit window for a new data set to be entered. If the current data set has not been saved, the user will be asked if it should be saved before creating the new data set. After the new data set is satisfactorily entered and the data edit windows are closed, the kinetic data are plotted on the screen.

Edit (⌘E): This command brings up the data edit window with the values of the current data set. This gives the user an opportunity to change the current data. When the window is closed, the kinetic data are plotted on the screen.

New k(cat) (⌘K): This command brings up the k_c edit window. The current value of k_c, the rate constant of the catalyzed reaction, appears in the edit field. This value may be changed by selecting it with the cursor and entering a new value. When the widow is closed by clicking its close box or hitting "¬,"the program will simulate the kinetic behavior resulting from this k_c and the initial conditions, and plot its findings on the screen.

New Ka (⌘A): This command brings up the K_{A} edit window. The current value of K_{A}, the association constant of host with alkylating agent, appears in the edit field. This value may be changed by selecting it with the cursor and entering a new value. When the window is closed, the program calculates the expected kinetic behavior, and plots the result on the screen.

New Ki(⌘I): This command brings up the K_{I} edit window. This window has two edit fields: K_{I}, the association constant of host with the competitive inhibitor, and $[\mathrm{I}]_{\mathrm{tot}}$, the total concentration of competitive inhibitor. Either of these quantities may be changed in the usual way. When the window is closed, no further action is taken by the program.

Calculate(⌘C): This command prompts the computer to begin calculating the kinetic behavior from the current data. When it is finished, the results are plotted

on the screen. This command can be used whenever a parameter has changed but simulation did not automatically execute, such as after changing to a new preferences set or editing the initial concentrations.

ReDraw(⌘D): This command causes the plot of experimental and simulated kinetic data to be drawn again. This redrawing is usually done automatically, so it is typically not necessary to call this command. It is sometimes convenient for restoring the drawing in the plot window after running other applications under the Multifinder or using desk accessories. Since the Kinetics Simulator is ignorant of such activity, the user sometimes needs to help it along.

Save(⌘S): This command writes some of the important variables to an external data file, which is readable only by the Kinetics Simulator. If the current data set has not been altered since the last time it was saved, this command is disabled. If the current data set is new, this command is equivalent to "Save as..." (below). The saved variables are the ones contained in the data edit windows: K_{S}, K_{P}, k_u, $[\mathrm{H}]_{\mathrm{tot}}$, $[\mathrm{A}]_{\mathrm{tot}\,0}$, $[\mathrm{S}]_{\mathrm{tot}\,0}$, $[\mathrm{P}]_{\mathrm{tot}\,0}$, and the experimental kinetic data. The variables *not* saved are k_c, K_{A}, K_{I}, and $[\mathrm{I}]_{\mathrm{tot}}$. Why are these not saved? The quantity k_c is not saved because it isn't data; it's *fitted* to the data. The same reasoning applies to K_{A}. So what about K_{I} and $[\mathrm{I}]_{\mathrm{tot}}$? These values indeed are experimentally measured. The reason they are not included in the file is that they were added to the program after it had met with some use. I did not want to change the data file format, because that would have made older files unreadable. Consequently, when you need to simulate a catalyzed alkylation in the presence of a competitive inhibitor, you must input K_{I} and $[\mathrm{I}]_{\mathrm{tot}}$ every time you load the file.

Save as...: This command is similar to "Save:" it writes the current data to an external file. The difference between these two commands is that "Save as..." prompts the user for a name for the data file. As long as there is a current data set, the "Save as..." command is never disabled.

(c) The Labels menu.

Title(⌘T): This command brings up the graph label edit window. This window enables the user to change the labels of the kinetic behavior graph in the plot window. The labels that can be changed are the graph title, the x-axis label, and the y-axis label. When the graph label edit window is summoned by this command, the graph title edit field is initially selected.

X Axis(⌘X): This command also brings up the graph label edit window, but with the x-axis label edit field selected. Note that in this program ⌘X is not the command-key equivalent of "Cut." The Kinetics Simulator *has* no "Cut."

Y Axis(⌘Y): This command, like "Title" and "X Axis," brings up the graph label edit window. In this case, as might be expected, the y-axis label edit field is selected.

(d) The Preferences menu.

This menu is the only menu in this program that can be altered by the user. It is here that a preference set is created, modified, or selected. The functions of these preferences are explained in the introduction to section III. The user can add preference sets to the list, up to a total of ten, and change the values of any preferences except those in the default set. It is not possible, however, to remove a preference set from within the program. Since undesirable preference sets can be altered and renamed, this drawback should not detract from the utility of the program. If it becomes necessary to start the preferences file over again, run the "pref startup" program by clicking on its icon. This will generate a new preferences file, containing only the default preferences.

The first elements of this menu are names of preference sets that can be selected. A check appears by the set that is currently in use. The last element, "Edit," brings

up the preference select window. This window has buttons for each of the saved preference sets, and an additional button for adding a new preference set.

Clicking on any of the buttons in the preference select window brings up the preference edit window. This window has edit fields for each of the four preferences (convergence criterion, tail, increment, and floor), and an additional field for the preference set's name. If the default preference set is selected, the numerical values of the preferences can be looked at but not changed. The other preference sets, however, may be changed in any way.

2. Windows.

Most of the windows used in the Kinetics Simulator program are simple edit windows, which contain some text and a few buttons and edit fields. For the most part, these windows may be resized, but no advantage is gained from resizing, since the contents will not readjust their positions to fit their new surroundings. By the same token, these edit windows may also be repositioned, but the program does not keep track of their new positions. If one of these windows is repositioned and then closed, it will come up in its original place when it is summoned again, instead of in the place to which it was repositioned. Two windows in the program have special features: the kinetic data edit window and the plot window. These windows will be addressed in turn.

(a) The kinetic data edit window.

This window looks a bit like a spreadsheet window, which, in a more advanced program, it would actually be. The data are arranged in rows and columns, just like a spreadsheet. Unlike a spreadsheet, however, resizing the window will not reveal more data, and the data cannot be smoothly scrolled in the window with familiar Macintosh scroll bars.

The data portion of the window is divided into two columns of ten rows each. Each row represents a measurement of total product concentration at some time, with the entry in the first column telling the time in minutes and the second column telling the product concentration in micromoles/liter. Although there are only ten rows in this edit window, the program can accomodate more than ten concentration measurements. There are several ways that the kinetic data edit window can view all parts of the kinetic data set.

If "¬" or "Enter" are typed from within any of the edit fields in this window, the edit field directly below becomes selected. If this new edit field would be beyond the end of the column, then all of the entries are moved one edit field upward, so that the newly-selected entry is on the last line in the window. Pressing "Tab" from within an edit field has a similar effect. Instead of selecting the edit field directly below, however, the result of this action depends on the position of the current edit field. If the current field is in the first column, the next field in the same row is selected. If the current field is in the second column, the first edit field in the next row is selected. Again, if this requires selecting an edit field in a row beyond the tenth in the window, all entries will be moved to the edit field directly above, and the new selected field will be in the last row.

Furthermore, there are two oval buttons at the right-hand side of the window, one near the top and one near the bottom. These can move the text in the window up or down. Clicking on the top button will move the window text down one row, so that earlier rows can be viewed. Conversely, clicking on the bottom button moves the window up one row. In this way, all of the kinetic data can eventually be displayed in this window.

(b) The plot window.

This is the only window that keeps track of its size. If the user resizes this window, the program re-plots the kinetic data curve to fit within its new dimensions.

If the window is made too small to contain a plot (because the title, reported parameters, and axis labels take up some space), then the plot is omitted. In this way, the plot window can be sized to take up only part of the screen, which is sometimes convenient when running several applications under the Multifinder.

3. Limits. There are some size limits that the program as written will not exceed. There may only be nine user-defined preference sets (in addition to the default set), there may be no more than thirty kinetic data points, and the simulated reaction may not contain more than four thousand steps. It is clear how to avoid exceeding these first two limits; how to remain within the third is less obvious. If the program presents an alert box stating "`Arrays too short for full simulation`," then the current simulation conditions require too many steps. If this happens, change the parameters to take a larger increment, or to have a higher (0.9 or above) floor.

Chapter 7

Laboratory Projects

This chapter is organized into four divisions, each one of which is devoted to a single project. There is an experimental section for each project. Footnotes and compound numbers are continuous through the chapter.

General Experimentals

Flash chromatography[1] was performed with Silica Gel 60, 230–400 mesh (EM Science) and thin-layer chromatography (TLC) on glass-backed 250 μm Silica Gel F-254 (EM Science) plates stored in a dessicator. NMR spectra were recorded on Varian EM-390 and Bruker AM-500 spectrometers.

Purification of solvents. When solvents are described as "dry" or "freshly-distilled," they were purified in the following manner. *N,N*-Dimethylformamide (DMF): Benzene or cyclohexane (100 ml) was added to the pot of the DMF still containing CaO and distilled away at atmospheric pressure. Vacuum distillation apparatus was then set up, and a fore-cut was removed before the actual sample was obtained. Acetonitrile, benzene, toluene, and dichloromethane were distilled from calcium hydride. Tetrahydrofuran and diethyl ether were distilled from benzophenone ketyl. Morpholine and pyridine were refluxed overnight with KOH, then distilled from calcium hydride or barium oxide.

Notational conventions. Flash chromatography column dimensions are specified by length (cm) "×" diameter (cm). Thus, a flash column 15 cm long and 2 cm in diameter is reported as "15 × 2." When a chromatographic solvent ramp was used, that is, when an increasing proportion of a second solvent is added to the initial solvent as the elution progresses, this is indicated by (initial solvent) "+" (second solvent). Extractive workups described as "dried over anhydrous $MgSO_4$" were always gravitationally filtered through fluted paper.

(1) Still, W. Clark; Kahn, Michael; Mitra, Abhijt "Rapid chromatographic technique for preparative separations with moderate resolution," *J. Org. Chem.* **1978**, *43*, 2923–2925.

I. 9,10-Disubstituted Ethenoanthracenes

Abstract: A scheme for functionalizing ethenoanthracenes at the 9 and 10 (bridgehead) positions was explored. Aldehydes at these positions were to have been substituted by Wittig olefinations to allow ready access to hosts bearing substituents rigorously excluded from the binding cavity. The olefination step proved to be destructive to the ethenoanthracene.

A. Purpose.

The first project I worked on at Caltech was aimed at functionalizing the ethenoanthracene building block from which hosts were constructed. By placing substituents at the bridgehead positions of this unit, we hoped to achieve several aims. The first was to make the hosts more water-soluble. If the substituents sported hydrophilic groups, the resulting hosts would have more water-solubilizing functionality and about the same amount of hydrophobic surface area as the unsubstituted parents. Furthermore, substituents at these positions were expected to sterically disrupt micellar packing, destabilizing the aggregate phase even in the absence of any additional hydrophilic interactions.[2] These steric interactions were also expected to inhibit the collapse of flexible hosts into the "bowl" conformation,[5] thus enforcing a preorganized binding cavity. Another aim was to furnish a scaffold from which catalytic groups could be appended. Furthermore, such substituents were expected

(2) Although we have not investigated the nature of aggregates formed by our hosts, they probably involve tubular arrays of host rings stacked face-to-face. Diederich has implicated such structures in the aggregation of his hosts,[3,4] which have structures similar to ours. Substituents pointing parallel to the cavity axis of the hosts would certainly impede the formation of such aggregates.

(3) Diederich, François; Dick, Klaus " New water-soluble macrocycles of the paracyclophane type: aggregation behavior and host-guest-interaction with hydrophobic substrates," *Tetrahedron Lett.* **1982**, *23*, 3167–3170.

(4) François Diederich, personal communication.

(5) Shepodd, Timothy J.; Petti, Michael A.; Dougherty, Dennis A. "Tight, oriented binding of an aliphatic guest by a new class of water-soluble molecules with hydrophobic binding sites," *J. Am. Chem. Soc.* **1986**, *108*, 6085–6087.

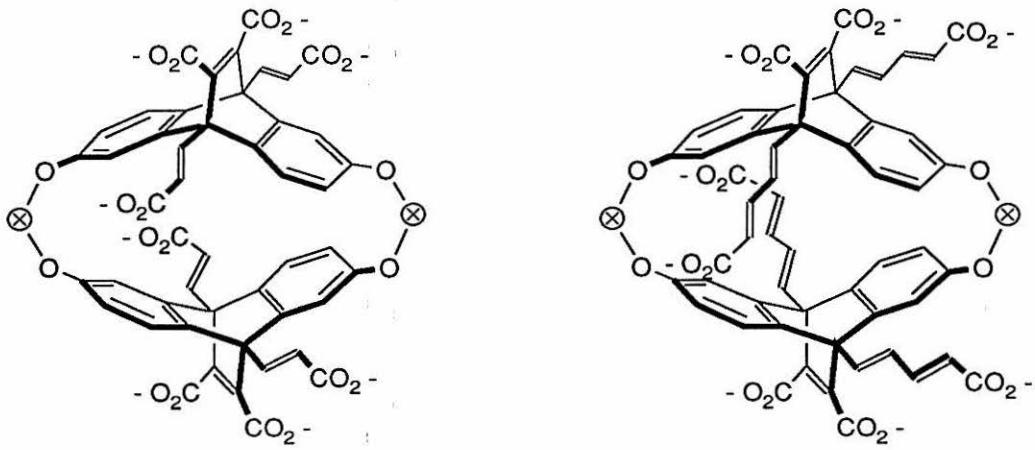


Figure 1. Desirable ethenoanthracene host molecules with carboxylate-bearing substitiuents at the 9- and 10-positions.

to partially cover the openings of the binding cavity, providing a more complete hydrophobic environment for the guests. Finally, we hoped that these substituents would hinder both the approach of guests to the cavity and their departure from it, so that distinct signals for the separated species and the host/guest complex could be observed by NMR.

B. Design.

A scheme for functionalizing these positions with olefinic substituents had already been developed,[6] but we felt that more versatile moieties were called for to fulfill our ambitious aims. To that end, we envisioned aldehyde substituents at the bridgehead positions, which could be elaborated by Wittig olefination technology to a wide variety of final products. The syntheses of these different products would diverge only after the creation of the common dialdehyde intermediate. If a large quantity of the dialdehyde were on hand, there would be easy access to hosts with different substituents. The most immediately attractive such products were those with carboxylic acid groups held rigorously away from the binding site by *trans* linkages across double bonds. These targets are depicted in Figure 1.

(6) Rider, Michael A. M.S. Thesis, California Institute of Technology, 1985.

C. Execution.

The proposed synthetic path is depicted in Scheme I. The solid arrows indicate the extent of our success in this scheme. Commercially-available anthraflavic acid **1**, the starting material for the synthesis of the parent hosts, was also to be the basis of this second host family as well. Protection of the phenols as methyl ethers gave 2,6-dimethoxyanthraquinone **2**, which received two carbon atoms from sulfur ylide additions to its carbonyl groups.[7] The resulting dispiro dioxirane **3** was isomerized by lithium bromide to the conjugated aldehyde–alcohol **4**,[7] which was reduced to the diol **5** and protected as the bis-*t*-butyldimethylsilyl ether **6**. This soluble anthracene readily engaged in a Diels–Alder reaction with dimethyl acetylenedicarboxylate, and the silyl ethers were cleanly cleaved by boron trifluoride etherate[8] to yield the diol **8**. Oxidizing this diol to the dialdehyde proved to be problematic, but not nearly as problematic as the subsequent Horner–Emmons–Wadsworth olefination.

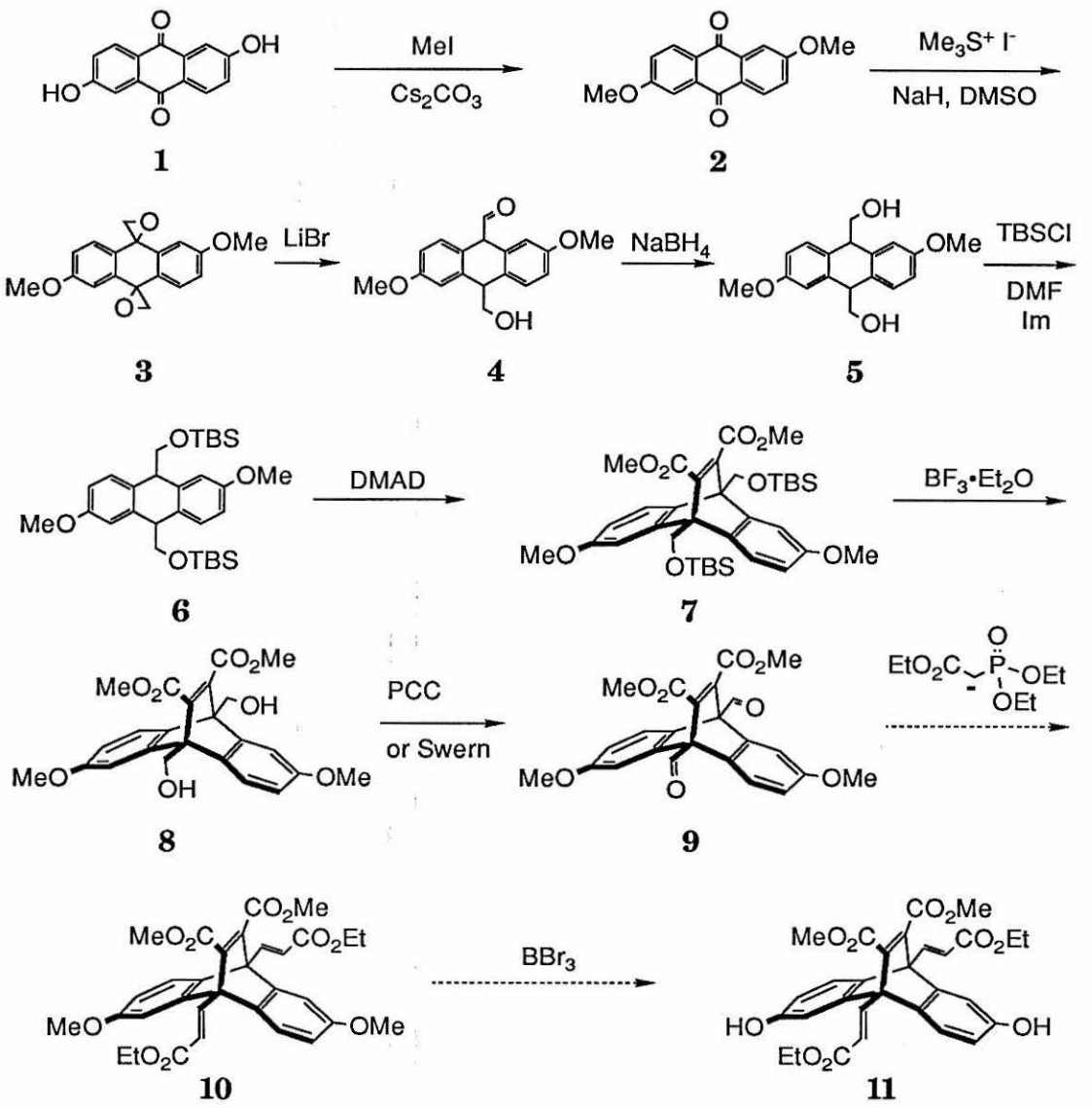
The olefination reaction worked splendidly with model compounds such as trimethylacetaldehyde (pivalaldehyde) **12** or 9,10-anthracenedicarboxaldehyde **13**, but attempting the reaction on the more relevant model **14** resulted in the destruction of all starting materials (Scheme II). The cause of this unpleasant behavior was never fully discerned, but our suspicion fell on the proximity to the targeted aldehyde groups of the methyl esters on the ethenoanthracene bridge.

---

(7) Lin, Yang-i; Lang, S. A. Jr.; Seifert, Christina M.; Child, R. G.; Morton, G. O.; Fabio, P. F. "Aldehyde syntheses. Study of the preparation of 9,10-anthracenedicarboxaldehyde," *J. Org. Chem.* **1979**, *44*, 4701–4703.

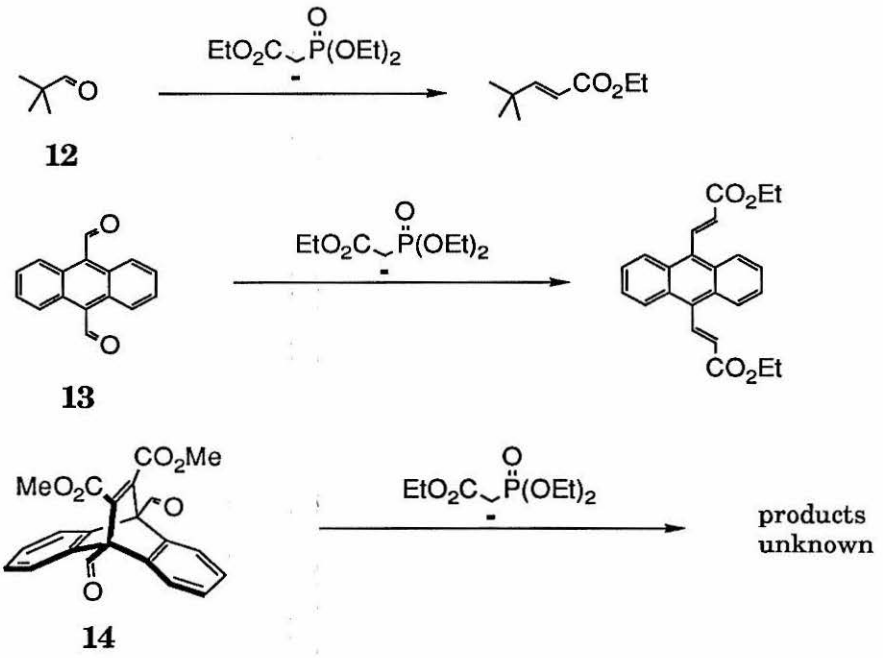
(8) Kelly, David R.; Roberts, Stanley M.; Newton, Roger F. "The cleavage of *t*-butyldimethylsilyl ethers with boron trifluoride etherate," *Synth. Commun.* **1979**, *9*, 295–299.

Scheme I



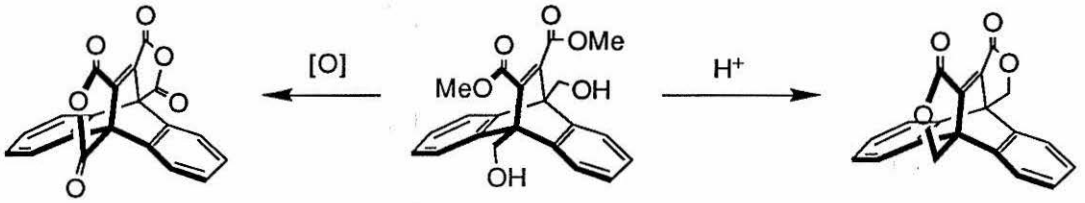
Attack of the phosphorus-stabilized carbanion of the Horner–Emmons–Wadsworth reagent on the aldehyde presumably would create an alkoxide pointing directly at the carbonyl group of the nearby methyl ester. Other reactions of this skeleton had demonstrated the facility of interactions between the bridgehead substituents and the carbonyl groups of the bridge. For example, an analogue of the diol **8** readily formed the bis-lactone upon mild acid catalysis (silica gel chromatography was

Scheme II



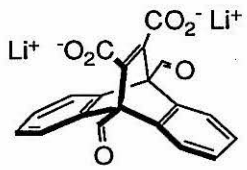
sufficient), and over-oxidation of the diol to the dicarboxylic acid with pyridinium chlorochromate generated the corresponding bis-anhydride, as shown in Scheme III.

Scheme III

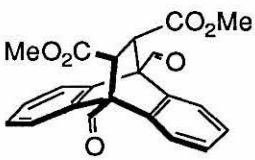


It was not clear how any such reaction could demolish the entire carbocyclic framework of the substrate, but no alternative explanations presented themselves. Consequently, we sought to perform the olefination on analogues of **9** with compromised carbonyl substituents. The two such models we investigated were the lithium

soap **15**, and the Diels–Alder adduct between 2,6-di-(*t*-butyldimethylsiloxy)anthracene and dimethyl fumarate, **14**. Formation of the lithium salts of the carboxylates was expected to remove their electrophilic character, and saturation of the bridge was expected both to remove the two carboxylic ester groups from conjugation and to move them slightly away from the bridgehead substituents. In both cases, the outcomes of the attempted Horner–Emmons–Wadsworth reactions were in line with the outcomes of all previous attempts with bridged anthracenes: the substrates were wholly obliterated. As a result, this project was terminated.



15



16

D. Conclusions.

In retrospect, it appears that this project was too ambitious. The dialdehyde **11** bears a hefty complement of chemical functionality on a small, rigid, unsaturated framework. All of the different substituents probably interact, so the system's unusual chemical behavior should have come as no surprise. If the targeted hosts could be made, they would probably be very interesting. Horner–Emmons–Wadsworth olefination of the 9,10-dicarboxaldehyde, however, is not the way to make them.

E. Experimental Section.

Trimethylsulfonium iodide was prepared according to Fieser and Fieser.[9] Other compounds were purchased as noted or prepared as described below.

---

(9) Fieser, L. F.; Fieser M.; *Reagents for Organic Synthesis*; John Wiley and Sons: New York, 1967; p 1236.

Preparation of **2,6-Dimethoxyanthraquinone 2:** 97% Anthraflavic acid **1** (Aldrich, 30.00 g, 0.1249 mol) and 700 ml DMF were magnetically stirred to give partial solution. Anhydrous K_2CO_3 (Baker, 69.04 g, 0.4495 mol) and iodomethane (Aldrich, 42 ml, 0.50 mol) were added sequentially, and the mixture was stirred under nitrogen at 60°C. After 6 h, 35 ml conc. aqueous ammonia was added to destroy the remaining iodometheane and the reaction mixture was poured into 500 ml 2 M NaOH, filtered, and washed with water. The residue was crystallized in several crops from chloroform to give fluffy yellow needles. Yield: 22.05 g (75%). ^{1}H NMR ($CDCl_3$): 8.13 (d, J = 9 Hz, 2H), 7.61 (d, J = 3 Hz, 2H), 7.11 (dd, J = 3 Hz, J = 9 Hz), 3.84 (s, 6H).

Preparation of **bis-oxirane 3:** To a solution of sodium hydride, 60% oil dispersion (Aldrich, 8.96 g, 0.224 mmol) in 750 ml dry DMSO was added anthraquinone **2** (25.05 g, 0.093 mmol). In the dark and under nitrogen, a solution of trimethylsulfonium iodide (41.65 g, 0.204 mmol) in 350 ml dry DMSO was added dropwise from a pressure-equalized addition funnel over a 30-min period. The reaction was stirred under nitrogen an additional 2 h and the clear red solution was poured onto 2 l crushed ice.The resulting yellow precipitate was filtered, washed with water, and dried under vacuum. Yield: 27.60 g (99.7%); ^{1}H NMR ($CDCl_3$): 7.25 (d, J = 10 Hz, 2H), 6.94–6.77 (m, 4H), 3.7 (s, 6H), 3.15 (s, 4H).

Preparation of **10-hydroxymethyl-2,6-dimethoxy-9-anthraldehyde 4:** To a light yellow suspension of oxirane **3** (26.41 g,0.089 mol) in 1600 ml dry acetonitrile was added anhydrous lithium bromide (Alfa, 39.70 g). The reaction mixture was stirred under nitrogen for 23 h. The deep orange suspension was allowed to settle, and the supernatant was decanted away. The residue was filtered, washed with water, and dried under vacuum to a light orange solid. Yield: 20.02 g (76%); ^{1}H NMR ($CDCl_3$): 11.31 (s, 1H), 8.80 (d, J = 10 Hz, 1H), 8.35 (d, J = 10 Hz, 1 H),

8.24 (d, J = 3 Hz, 1H), 7.64 (d, J = 3 Hz, 1H), 7.40–7.11 (m, 2H), 5.27 (s, 3H), 3.86 (s, 6H).

Preparation of **9,10-bis(hydroxymethyl)-2,6-dimethoxyanthracene 5:** To a magnetically-stirred suspension of aldehyde **4** (18.83 g, 0.064 mol) in 190 ml dry DMSO was added 5 ml absolute methanol and sodium borohydride (Aldrich, 0.665g, 0.0176 mmol). TLC (1:1 EtOAc/PE) indicated the presence of fluorescent yellow starting material as well as fluorescent blue product, so an additional 170 nm DMSO, 1 g sodium borohydride, and 150 ml methanol were added. Extended reaction times did not eliminate the trace of starting material. Approximately 100 ml MeOH was removed by rotary evaporation, and the solid product in the flask was allowed to settle. The supernatant was decanted away, and the residue was filtered, washed with water, dried under vacuum, and recrystallized from nitrobenzene. Starting material was still visible by TLC but could not be detected by NMR. Yield: 14.64 g (77.2%); ^{1}H NMR (DMSO-d_6): 8.00 (d, J = 9.6 Hz, 2H), 7.30 (d, J = 3.0 Hz, 2H), 6.87 (dd, J = 3.0 Hz, J = 9.6 Hz, 2 H), 5.12 (m, 6H), 3.56 (s, 6H).

Preparation of **9,10-bis(*t*-butyldimethylsilyloxymethyl)-2,6-dimethoxy anthracene 6:** Diol **5** (1.999g, 6.70 mmol), *t*-butyldimethylchlorosilane (TBSCl) (Aldrich, 5.081 g, 33.7 mmol), imidazole (Aldrich, 2.734 g), and 18 ml dry DMF in a 100-ml snrb flask were heated to 100 °C, affording a clear solution, and stirred for 4 h. Solvent and TBSCl were removed under reduced pressure, and the solid residue was partitioned between ether and water. The ether phase was extracted copiously with water and dried over $MgSO_4$. A flash column (CH_2Cl_2, 8×6) separated most of the product satisfactorily; the impure fractions were again chromatographed (CH_2Cl_2, 25×1). Yield: 3.166 g (90%); ^{1}H NMR ($CDCl_3$): 8.26 (d, J = 10 Hz, 2H), 7.60 (d, J = 3 Hz, 2H), 7.18 (dd, J = 10 Hz, J = 3 Hz, 3 H), 5.53 (s, 4H), 3.91 (s, 6H), 0.90 (s, 18H), 0.10 (s, 12H).

Preparation of **9,10-bis (*t*-butyldimethylsilyloxymethyl)-11,12-dicarbomethoxy-9,10-etheno-9,10-dihydro-2,6-dimethoxyanthracene 7:** Anthracene **6** (1.200 g, 2.278 mmol), DMAD (Lancaster, 2.8 ml, 22.79 mmol), and 9.0 ml toluene were refluxed for 2 d. Toluene was removed by rotary evaporation, and the dark, viscous residue was passed through a 1 cm pad of silica gel with dichloromethane. Solvent was evaporated, and the resulting yellow oil was heated to 80 °C for 30 min. A yellow oil was obtained after flash chromatography (CH_2Cl_2, 16 × 2), which gave a yellow solid upon heating to 100 °C under vacuum. The product was crystallized from methanol to give colorless needles. Yield: 1.1031 g (72%); ^{1}H NMR ($CDCl_3$): 7.22 (d, J = 10 Hz, 2 H), 6.91 (d, J = 3 Hz, 2H), 6.51 (dd, J = 10 Hz, J = 3 Hz, 2H), 5.00 (s, 4H), 3.72 (s, 6H), 0.95 (s, 18H), 0.26 (s, 12H).

Preparation of **11,12-dicarbomethoxy-9,10-etheno-9,10-dihydro-9,10-bis(hydroxymethyl)-2,6-dimethoxyanthracene 8:** Silyl ether **7** (99.95 mg, 0.145 mmol) was dissolved in 1.64 ml dry chloroform. Boron trifluoride etherate (Aldrich, 117 μl, 0.953 mmol) was added; after stirring under nitrogen for 1 h, solvent was removed by rotary evaporation, leaving a yellow oil that was quenched with 1 M $NaHCO_3$. The product was partitioned between CH_2Cl_2 and water, and the organic phase was dried over onhydrous $MgSO_4$. Yield: 43.46 mg (100%). ^{1}H NMR (DMSO-d_6): 7.24 (d, J = 7.5 Hz, 2H), 6.96 (d, J = 3.0 Hz, 2H), 6.55 (dd, J = 7.5 Hz, J = 3.0 Hz, 2 H), 5.31 (t, J = 3.0 Hz, 2H), 4.75 (d, J = 3.0 Hz, 4H), 3.68 (s, 6H), 3.67 (s,6H).

Preparation of **11,12-dicarbomethoxy-9,10-etheno-9,10-dihydro-2,6-dimethoxy-9,10-anthracenedicarboxaldehyde 9:** To a stirred suspension of pyridinium chlorochromate (Aldrich, 68.1 mg, 0.316 mmol) in 0.42 ml dry dichloromethane was added diol **8** (46.43 mg, 0.105 mmol) and 1.39 ml dry dichloromethane. After 7.5 h, the reaction was diluted to fivefold volume with dichloromethane and forced through a 1 cm silica gel pad to remove chromium species. Solvent was

removed by rotary evaporation to leave a brown oil, 33.02 mg (72%). ^{1}H NMR ($CDCl_3$): 10.58 (s, 2H), 7.36 (d, J = 8.5 Hz, 2H), 7.12 (d, J = 3.0 Hz, 2H), 6.45 (dd, J = 3.0 Hz, J = 8.5 Hz, 2H), 3.67 (s, 6H), 3.61 (s, 6H).

Acid-catalyzed lactone formation from a 9,10-bis(hydroxymethyl)-ethenoanthracene: 9,10-etheno-9,10-dihydro-9,10-bis(*t*-butyldimethylsilyloxymethyl)-11,12-dicarbomethoxyanthracene (49.96 mg, 0.08205 mmol, 1 eq) was dissolved in 1 ml chloroform in a 5-ml snps flask. The flask was evacuated and flushed with dry nitrogen twice, and $BF_3{\cdot}Et_2O$ complex (Aldrich, 44 μl, 48 mg, 0.338 mmol, 2.06 eq) was added *via* syringe. The reaction was stirred under nitrogen for 40 min, and TLC (3:7 EtOAC/PE) indicated that extensive cleavage had occurred. The reaction was quenched by addition of triethylamine (50 μl, 37 mg, 0.361 mmol, 2.2 eq). The reaction mixture was extracted with water, sat. $NaHCO_3$, and brine, and dried over anhydrous $MgSO_4$. A crude NMR spectrum showed the desired diol. Purification by flash chromatography (1:1 EtOAc/PE, 20 × 1) was attempted; the eluted product was the bis-lactone depicted in Scheme III. ^{1}H NMR ($CDCl_3$): 7.30(m, 4H), 7.10 (m, 4H), 5.56 (s, 4H).

Intramolecular bis-anhydride of 9,10-dihydro-9,10-ethenoanthracene-9,10,11,12-tetracarboxylic acid: 9,10-etheno-9,10-dihydro-9,10-bis(hydroxymethyl)-11,12-dicarbomethoxyanthracene (255.11 mg, 0.6710 mmol, 1 eq) was suspended in 3.6 ml dry dichloromethane. Pyridinium chlorochromate (Aldrich, 578.66 mg, 2.684 mmol, 2.00 eq) was added to the suspension, and sonicated. The reaction was stirred under nitrogen for 12 h. The reaction mixture was forced through a silica gel column (CH_2Cl_2, 2 × 6) to remove the inorganic species, and then purified by flash chromatography (15:85 EtOAc/PE, 17 × 2.5) to yield a very high-R_f product, which was identified as the anhydride depicted in Scheme III. ^{1}H NMR ($CDCl_3$): 7.86(m, 4H), 7.31(m, 4H).

Attempted preparation of alkene 10: Lithium chloride (9 mg, 0.21 mmol, 1.3 eq)) and dry acetonitrile (100 μl) were placed in an oven-dried 5-ml snps flask. The flask was evacuated and flushed with nitrogen twice, and triethyl phosphonoacetate (Aldrich, 40 μl, 45 mg, 0.20 mmol, 1.2 eq) and 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) (Aldrich, 26 μl, 26 mg, 0.17 mmol, 1.0 eq) were added to the mixture, followed by a solution of dialdehyde **9** (35.82 mg, 0.08208 mmol, 1 eq) in 280 μl acetonitrile. This was allowed to stir for 2.75 h, and the reaction was quenched by addition of sat. NH_4Cl. The mixture was partitioned between ether and water; the aqueous phase was extracted 4 times with ether, and the combined organic phases were dried over anhydrous $MgSO_4$. The product, a brown goo, had no discrete peaks in its NMR spectrum.

Preparation of anthracene-9,10-bis(3-ethyl acrylate): Sodium hydride, 60% mineral oil dispersion (Aldrich, 36 mg, 0.90 mmol, 2.1 eq) was weighed out in an oven-dried 100-ml 3-neck rb flask and rinsed twice with dry petroleum ether. Freshly-distilled THF (0.6 ml) was added to it, and a solution of triethyl phosphonoacetate (Aldrich, 100 μl, 143 mg, 0.64 mmol, 1.5 eq) and 9,10-anthracenedicarboxaldehyde **13** (Kodak, 50.05 mg, 0.2137 mmol, 1 eq) in 11 ml THF was placed in a pressure-equalized addition funnel. The reaction flask was cooled in an ice bath, and the contents of the addition funnel were added to the NaH suspension with stirring over a period of 5 min. The reaction was stirred at 0 °C for an additional hour, and then it was quenched by the addition of sat. NH_4Cl. The product from aqueous extractive workup was purified by flash chromatography (CH_2Cl_2, 15 × 2) to give 30.27 mg (44%) highly fluorescent yellow material. ^1H NMR ($CDCl_3$): 8.67 (d, J = 16 Hz, 2H), 8.29 (m, 4H), 7.55 (m, 4H), 6.42 (d, J = 16 Hz, 2H), 4.41 (qua, J = 8 Hz, 4H), 1.44 (t, J = 8 Hz, 6H).

Attempted Horner–Emmons–Wadsworth reaction of lithium salt 15: To an NMR sample of dilithium 9,10-etheno-9,10-dihydro-anthracene-9,10-dicarboxaldehyde-11,12-dicarboxylate **15** (0.0322 mmol) in 400 μl DMSO-d_6 was added triethyl phosphonoacetate (Aldrich, 16 μl, 18 mg, 0.081 mmol, 1.3 eq) and lithium amide (Aldrich, 6.16 mg, 0.27 mmol, 4.2 eq). No reaction was seen after 3 h, so a solution of methyllithium was added. As the aldehyde peak in the NMR spectrum disappeared, so did the aryl resonances of the substrate.

Attempted Horner–Emmons–Wadsworth reaction of **fumarate adduct 16:** Sodium hydride 60% mineral oil dispersion (approximately 5 mg, 0.1 mmol, 4 eq) was placed in a dry micro-reactor and rinsed twice with dry petroleum ether. THF (225 μl) was added to it, and the mixture was cooled in an ice bath. A solution of triethylphosphonoacetate (Aldrich, 10 μl, 14 mg, 0.062 mmol, 2.7 eq) and ethanoanthracene **16** (4.35 mg, 0.0115 mmol, 1 eq) in 450 μl THF was added to it over a 15-min period. The reaction was stirred at 0 °C for an additional 10 min, and the solvent was removed by flash chromatography. A crude NMR spectrum of the yellow product showed no discernable peaks.

II. Macrocyclic Cyclophanes Bearing Chelating Diphosphine Ligands

Abstract: In an effort to produce a homogeneous, chiral hydrogenation catalyst able to discriminate between the faces of prochiral olefins lacking a second metal-coordinating site, a scheme to create a macrocycle containing both an ethenoanthracene unit and a diphosphine ligand was explored. Reduction of the corresponding phosphine oxide macrocycle was not achieved.

A. Purpose.

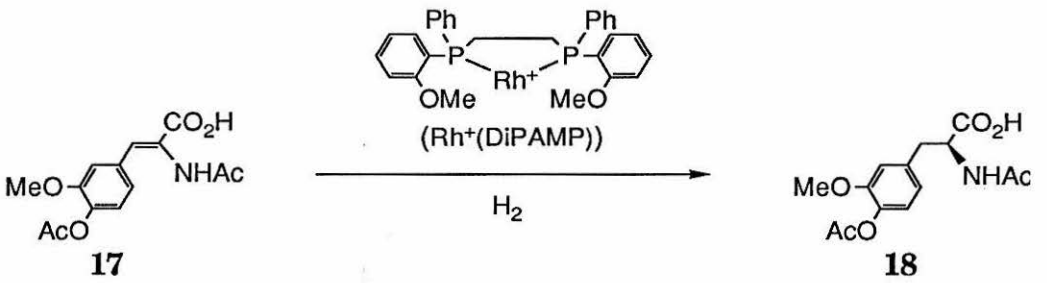
Some of the most impressive achievements in the field of chiral synthesis have been in asymmetric catalysis. In principle, a single molecule of a chiral catalyst can transform an army of chiral substrate molecules each into a single enantiomer of chiral product. Such a process is an elegant way to amplify the chirality of a small quantity of catalyst.

The most accomplished synthetic catalysts producing asymetric induction are the chiral hydrogenation catalysts. These catalysts are related to the homogenous Wilkenson's catalyst, containing phosphine ligands complexed to a platinum group metal with open coordination sities. Many chiral ligands containing a chelating diphosphine group have been developed.[10]

---

(10) (a) Ojima, Iwao; Clos, Núria, Bastos; Cecilia "Recent advances in catalytic asymmetric reactions promoted by transition metal complexes," *Tetrahedron* **1989**, *45*, 6901–6939. (b) Brunner, Henri "Enantioselective synthesis of organic compounds wth optically active transition metal catalysis in substoichiometric quantities," *Top. Stereochem.* **1988**, *18*, 129–247. (c) *Asymmetric Catalysis*; NATO ASI Series E 103; Bosnich, B., Ed.; Martinus Nijhoff: Dordrecht, 1986. (d) *Asymmetric Synthesis*; Vol. 5, "Chiral Catalysis," Morrison, J. D., Ed.; Academic: Orlando, FL, 1985. (e) Pino, P.; Consiglio, G. "Organometallic catalysis in asymmetric synthesis," *Pure. Appl. Chem.* **1983**, *55*, 1781–1790. (f) Kagan, H. B. "Asymmetric synthesis using organometallic catalysis," In *Comprehensive Organometallic Chemistry*; Wilkinson, G.; Stone, R. F. G.; Abel, E. W., Eds.; Pergamon: Oxford, 1982; Vol. 8, p 463. (g) Halpern, Jack "Mechanism and stereoselectivity of asymmetric hydrogenation," *Science* **1982**, *217*, 401–407. (h) Bosnich, B. Fryzuk, Michael D. "Asymmetric synthesis mediated by transition metal complexes," *Top. Stereochem.* **1981**, *12*, 119–154.

Perhaps the most commercially important of these catalysts is DiAMP,[11] used by Monsanto to make the anti-Parknison's drug *l*-DOPA. In the asymmetric induction step of this synthesis, the achiral enamide **17** is transformed, with 95% e.e., into the chiral amino acid **18**.[12]



17 **18**

The enamide functionality of the substrate is necessary in order to obtain such a high stereoselectivity. Mechanistic studies have shown that both the amide and the olefin coordinate to the metal center, and that the two diastereometric complexes of enamide with catalyst have vastly different rates of hydrogen addition to the double bond. Ironically, the *minor* diastereomer has the faster hydrogenation rate.[13]

B. Design.

It was our intent to incorporate a chiral diphosphine moiety into a macrocyclic binding site related to our hosts. This would provide, upon introduction of a metal center, a host containing a hydrogenation catalyst. A prochiral olefin complexed to the catalyst would be held in place by two interactions: coordination of the

---

(11) Vineyard, B. D.; Knowles, W. S.; Sabacky, M. J.; Bachman, G. L.; Weinkauff, D. J. "Asymmetric hydrogenation. Rhodium chiral bisphosphine catalyst," *J. Am. Chem. Soc.* **1977**, *99*, 5946–5952.

(12) Knowles, William S. "Asymmetric Hydrogenation," *Acc. Chem. Res.* **1983**, *16*, 106–112.

(13) Landis, Clark R.; Halpern, Jack "Asymmetric hydrogenation of methyl-(*z*)-α-acetamidocinnamate catalyzed by {1,2-bis((phenyl-*o*-anisoyl)phosphino)ethano}rhodium(I): kinetics, mechanism, and origin of enantioselection," *J. Am. Chem. Soc.* **1987**, *109*, 1746–1754. Chan, A. S. C.; Pluth, J. J.; Halpern, Jack "Identification of the enantioselective step in the asymmetric catalytic hydrogenation of a prochiral olefin," *J. Am. Chem. Soc.* **1980**, *102*, 5952–5954.

carbon-carbon double bond to the metal, and encapsulation of the hydrocarbon inside the binding cavity of the host. We hoped in this way to extend the scope of the dissymetric hydrogenation to alkenes lacking a second metal-coordinating site.

Our design involved a "heterodimer" host. The host macrocycle was to include one ethenoanthracene unit, one DiPAMP unit, and two linkers. The planned synthesis is sketched in Scheme IV.

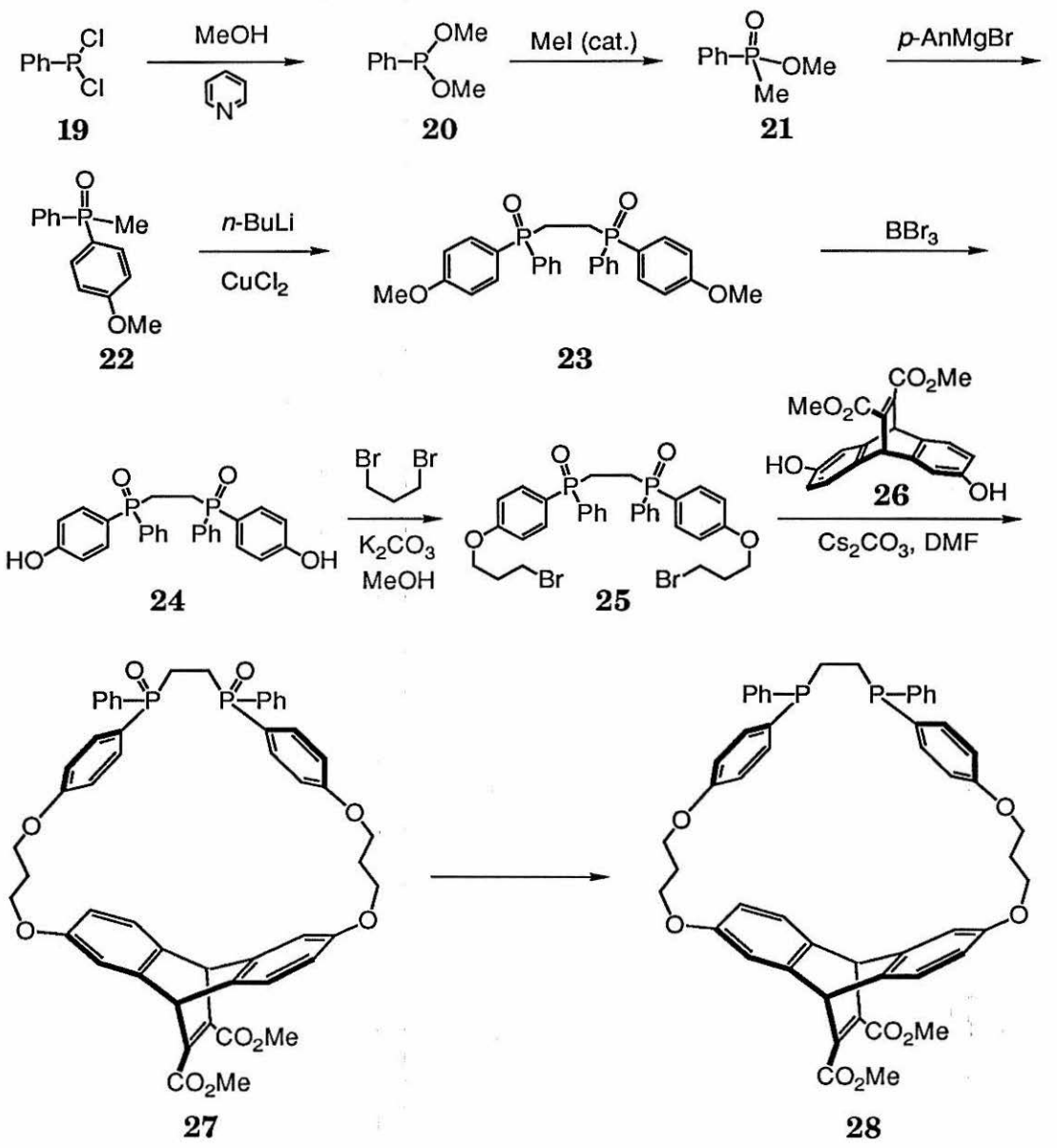
Most of this scheme is devoted to the synthesis of the diphosphine portion of the molecule. The key step in this convergent synthesis is the macrocyclization, in which the complete host skeleton is assembled from the separate diphosphine and ethenoanthracene parts. After this step, only final touches are required, namely, reduction of the phosphine oxides and saponification of the esters.

C. Execution.

The reduction of the phosphine oxide **27** was expected to be the most uncertain step of this sequence. Because the macrocycle contains reducible functional groups in addition to the phosphine oxides, it was necessary to find a procedure that reduced phosphine oxides without affecting the α,β-unsaturated esters of the ethenoanthracene. It was also important that the reduction procedure be stereospecific. The configurations of phosphines and phosphine oxides are stably chiral at ordinary temperatures;[14] indisciminate alteration of configurations at the two phosphine centers in the presence of the chiral ehtenoanthracene could lead to three diastereomers of the macrocycle, each of which would possess different catalytic and complexation properties. Although the synthesis sketched in Scheme IV employs no means to control the stereochemistry of the diphosphine ligand, such control could

(14) Horner, L. "Darstellung und eigenschaften optisch aktiver, tertiärer phosphine," *Pure. Appl. Chem.* **1965**, *9*, 225–244. Horner, L.; Winkler, H. "Phosphororganische verbindungen XLI die aktivierung senergien der racemiserung optisch aktiver tertiärer phosphine," *Tetrahedron Lett.*, **1964**, 461–462.

Scheme IV

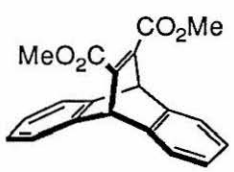


be introduced in a few additional standard steps.[15] Once it was demonstrated that the macrocycle could be made, specific stereoisomers could be produced by such

(15) (a) Korpiun, Olaf; Mislow, Kurt "A new route to the preparation and configurational correlation of optically active phosphine oxides," *J. Am. Chem. Soc.* **1967**, *89*, 4784–4786. (b) Korpiun, Olaf; Lewis, Robert A.; Chickos, James; Mislow, Kurt "Synthesis and absolute configuration of optically active phosphine oxides and phosphinates," *J. Am. Chem. Soc.* **1968**, *90*, 4842–4846.

longer pathways.

The model ethenoanthracene **29** was chosen to evaluate methods of reducing phosphine oxides. If it could withstand reaction conditions that reduced phosphine oxides, then these conditions could be expected to safely and effectively perform the reduction of the macrocycle **27**. Two reduction procedures, both employing chlorides of silicon, were tested in this way. The first, which is reported to reduce phosphine oxides with inversion of configuration, used trichlorosilane with diethylcyclohexylamine in acetonitrile.[16] The second, which also proceeds with inversion and is reportedly stereospecific if the reaction is brief, used hexachlorodisilane in benzene or chloroform.[17] These two procedures were tested on the model ethenoanthracene **29** to assess their compatibility with the complete macrocycle **27**. The procedure employing trichlorosilane caused some degradation of the substrate, but the scheme using hexachlorodisilane passed this test with flying colors. Under conditions in which triphenylphophine oxide, diphenylmethylphosphine oxide, (p-anisyl)phenylmethylphosphine oxide, and 1,2-bis(diphenylphosphino)ethane were fully reduced, **29** was unaffected.



29

(16) Johnson, C. R.; Imamoto, T. "Synthesis of polydentate ligands with homochiral phosphine centers," *J. Org. Chem.* **1987**, *52*, 2170–2174.

(17) Naumann, Klaus; Zon, Gerald; Mislow, Kurt "The use of hexachlorodisilane as a reducing agent. Stereospecific deoxygenation of acyclic phosphine oxides," *J. Am. Chem. Soc.* **1969**, *91*, 7102–7023.

The actual production of the phosphine oxide macrocycle **27** was uneventful. Commercially-available dichlorophenylphosphine **19** was converted to the malodorous, water-sensitive dimethyl phenylphosphonite **20** with methanol and pyridine in petroleum ether. This isomerized, by an iodomethane-catalyzed Arbuzov rearrangement of capricious violence, to the much stabler, odorless methyl methylphenylphosphinate **21**. This compound added, with displacement of methanol, the Grinard reagent made from 4-bromoanisole to give p-anisylmethylphenylphosphine oxide **22**. Oxidative coupling of the methyl substituents on phophorous gave the ethano bridged bis(phosphine oxide) **23**. The two anisole nuclei were readily demethylated by boron tribromide at room temperature to yield the insoluble diol **24**, which added two equivalents of linker 1,3-dibromopropane. The resulting "3/4 molecule" **25** was dissolved with an equimolar amount of 2,6-dihydroxyethenoanthracene **26** in dry DMF and slowly added to an excess of cesium carbonate in dry DMF to form the macrocycle **27** in low yield.

The first portent of unanticipated difficulty with phosphine oxide reduction appeared in further model studies with 1,2-bis(p-anisylphenylphosphino)ethane **23**. Reduction conditions which had previously proven to be sufficient for other phosphine oxides failed when tried on this substrate. For instance, no reaction occurred with hexachlorodisilane in refluxing benzene or acetonitrile, even after thirty minutes. In refluxing 1,2-dichloroethane, however, the reduction went to partial conversion when a threefold excess of hexachlorodisilane and a 25 minute reaction time were used. More reductant and longer times did not improve the conversion. Any conversion at all, however, was heartening. These reduction conditions were consequently tested on the macrocycle **27**: no evidence of deoxygenation was seen, even after 24 hours. The starting material did, however, begin to degrade in this time; additional TLC spots that did not correspond to the high-R_f behavior of a phosphine had appeared.

This route to deoxygenationing phosphine oxides did not appear likely to bear fruit, so the project was suspended for a while until another path to **28** could be found. An ideal path would involve a protected phosphine in the macrocyclization step, which would later be unmasked under mild conditions to yield the free phosphine. The difficulty lay in finding a protecting group for phosphines.

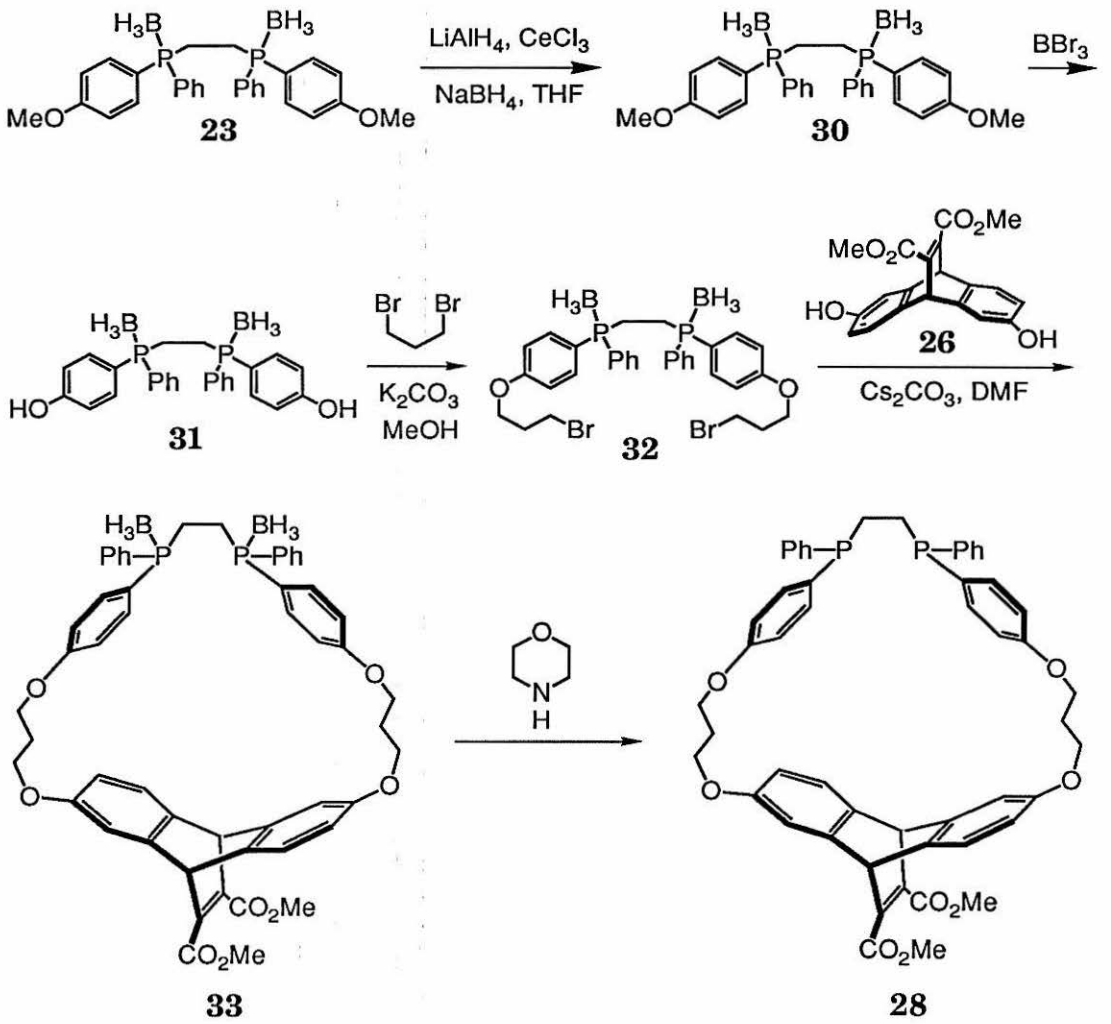
An attractive candidate came to my attention in a paper by Imamoto *et al.* describing some of the chemistry of phosphine-borane complexes.[18] It reported that these $R_3P–BH_3$ complexes could be formed in the room-temperature reaction of a phosphine oxide with sodium borohydride, lithium aluminum hydride, and anhydrous cerium trichloride. They were able to withstand the strongly basic conditions of oxidative coupling, but could be cleaved to yield the free phosphine by simply heating them in morpholine or diethylamine.

This introduced the option of the synthesis shown in Scheme V. After coupling to give the diphosphinoethane, the phosphine oxides could be converted to phosphine-borane complexes and then demethylated. The resulting diol was to be alkylated with 1,3-dibromopropane, macrocyclized with **26** by cesium carbonate in DMF, and then deprotected with morpholine.

Following this pathway did not prove to be beneficial, however. The phosphine-borane complex **30** could indeed be made, but the product of its demethylylation had undesirable solubility properties reminiscent of its phosphine oxide analogue **24**. Consequently, it was difficult to characterize. Attempts to alkylate this product with an excess of 1,3-dibromopropane were wholly unsuccessful. Thin-layer chromotagraphy showed only baseline spots; this indicates either a failure to alkylate

(18) Imamoto, Tsuneo, Oshiki, Toshiyuki; Onozawa, Takashi; Kusumoto, Tetsuo; Sato, Kazuhiko "Synthesis and reactions of phosphine–boranes. Synthesis of new bidentate ligands with homochiral phosphine centers via optically pure phosphine–boranes," *J. Am. Chem. Soc.* **1990**, *112*, 5244–5252. Imamoto, Tsuneo; Kusumoto, Tetsuo, Suzuki, Nobuyo; Sato, Kazuhiko "Phosphine oxides and $LiAlH_4$–$NaBH_4$–$CeCl_3$: synthesis and reactions of phosphine–boranes," *J. Am. Chem. Soc.* **1985**, *107*, 5301–5303.

Scheme V



or quaternization of deprotected phosphines. This project was consequently terminated.

D. Experimental Section

Dimethyl phenylphosphonite (**20**) and methyl methylphenylphosphinate (**21**) were prepared by the published procedure of Mislow *et al.*[15a]

Preparation of ***p*-anisylmethylphenylphosphine oxide 22**: All glassware was oven-dried prior to use. A solution of *p*-bromoanisole (Aldrich, 14.0 ml, 20.9 g,

0.112 mol, 1.20 eq) in 50 ml freshly-distilled ether was placed in a pressure-equalized addition funnel over a 250-ml 3-neck rb flask containing magnesium turnings (3.971 g, 0.163 mol, 1.75 eq). The reaction was maintained under a dry nitrogen atmosphere. About 1 ml of the solution was added to the turnings, and the mixture was stirred magnetically. When an exothermic reaction developed, the flow of bromide solution was resumed at a rate to maintain the reaction. After the addition was complete, the reaction mixture was refluxed for 1 h. A solution of methyl methylphenylphosphinate **21** (15.8 g, 0.093 mmol, 1 eq) in 50 ml benzene was placed into the addition funnel, and was added dropwise to the refluxing Grignard reagent. After the addition, the mixture was refluxed overnight. The reaction was quenched by placing dry ice in the addition funnel and waiting several hours. The pasty mixture was treated with 2 M sulfuric acid, which broke up the magnesium salts and consumed the unreacted magnesium metal. The organic phase was extracted twice with 2 M NaOH, water, and brine, dried over anhydrous $MgSO_4$, and the solvent was removed by rotary evaporation. The product was purified by crystallization from amyl acetate. Yield (2 crops): 9.051 g (40%). ^{1}H NMR ($CDCl_3$): 7.87–7.40 (br, 7H), 7.00 (dd, J = 3 Hz, J = 9 Hz, 2H), 3.86 (s, 3H), 1.98 (d, J = 13, 3H).

Preparation of **1,2-bis(*p*-anisylphenylphosphino)ethane 23**: The general procedure of Mislow *et al.*[19] was followed. All glassware was oven-dried prior to use. A 400-ml 3-neck rb flask containing *p*-anisylmethylphenylphosphine oxide **22** (7.385 g, 0.30 mmol, 1 eq) was charged with freshly-distilled tetrahydrofuran (200 ml) and cooled in a dry ice/acetone bath. *n*-Butyllithium (1.0 M in hexanes, 45 ml, 1.5 eq) was slowly added to the reaction mixture; after ten minutes, anhydrous cupric chloride (Aldrich, 6.5 g, 0.48 mmol, 1.6 eq) was added all at once. The reaction

(19) Maryanoff, Cynthia A.; Maryanoff, Bruce E.; Tang, Reginald; Mislow, Kurt "One-step synthesis of optically pure 1,2-ethano bis sulfoxides and phosphine oxides *via* the copper-promoted oxidative dimerization of chiral sulfinyl and phosphinyl carbanions," *J. Am. Chem. Soc.* **1973**, *95*, 5839–5840.

mixture was stirred under dry nitrogen as the mixture warmed to room temperature, about 5 h. Approximately 1 l of oxygen was bubbled through the mixture, and dilute sulfuric acid was added until pH neutrality was reached. Solvent was removed under reduced pressure, and the product was partitioned between dichloromethane and 1 M NH_4OH. The organic phase was extracted with water 3×, 2 M HCl, 6 M H_2SO_4, water again, 1 M NH_4OH, and brine, and dried over anhydrous $MgSO_4$. The product was purified by two recrystallizations from amyl acetate. Yield: 3.217 g (44%). ^1H NMR ($CDCl_3$): 7.83–7.30 (m, 14H), 6.94 (d, J = 7 Hz, 4H), 3.80 (s, 6H), 2.42 (d, J = 3 Hz, 4H).

Preparation of **1,2-bis((*p*-hydroxyphenyl)phenylphosphino)ethane 24**: 1,2-bis(*p*-anisylphenylphosphino)ethane **23** (903 mg, 1.84 mmol, 1 eq) was dissolved in 40 ml dry dichloromethane in a dry 100-ml snrb flask. The solution was cooled in a dry ice/acetone bath with stirring under dry nitrogen, and then boron tribromide (Aldrich, 0.87 ml, 2.3 g, 9.2 mmol, 2.5 eq) was added to it. The reaction mixture was allowed to warm to room temperature, and, after 19 h, was quenched by addition of 5 ml anhydrous ether. Solvent was removed by rotary evaporation, and the product was neutralized with saturated $NaHCO_3$. The resulting aqueous suspension was filtered by suction through a medium glass frit, and the residue was washed copiously with water. The solid was vacuum pumped overnight. This compound was too insoluble to record an NMR spectrum.

Preparation of **1,2-bis((*p*-(3-bromopropoxy)phenyl)phenylphosphino)-ethane 25**: Diol 24 (501 mg, 1.071 mmol, 1 eq), anhydrous potassium carbonate (Baker, 362 mg, 2.62 mmol, 1.22 eq), and 1,3-dibromopropane (Aldrich, 1.100 ml, 2.190 g, 5.06 eq) were combined in a 50-ml snrb flask. Methanol (12.5 ml) was added, and the reaction mixture was refluxed with magnetic stirring for 24 h. Methanol was removed by rotary evaporation, and the remaining solids were partitioned between dichloromethane and water. The organic phase was dried over anhydrous $MgSO_4$,

and purified by flash chromatography (1:19 MeOH/CH_2Cl_2, 30 × 2). Yield: 477 mg (63%). ^{1}H NMR($CDCl_3$): 7.90–7.26 (m, 14H), 6.98 (d, J = 9 Hz, 4H), 4.10 (t, J = 6 Hz, 4H), 3.55 (t, J = 6 Hz, 4H), 2.50 (d, J = 2 Hz, 4H), 2.27 (m, J = 6 Hz, 4H).

Preparation of **phosphine oxide macrocycle 27**: An oven-dried 100-ml snrb flask was charged with **25** (604 mg, 0.858 mmol, 1 eq), **26** (323 mg, 0.916 mmol, 1.07 eq), and 50 ml dry DMF. The solids dissolved, and were taken up into a 60-ml disposable polypropylene syringe, along with two 10-ml rinses of the flask. Cesium carbonate (Aldrich, 1.260 g, 3.867 mmol, 2.25 eq) and 30 ml dry DMF were placed in the now-empty flask. The flask and syringe were wrapped in aluminum foil to exclude light, and the solution of diol and dibromide was added to the stirred cesium carbonate suspension by a syringe pump over a period of 3 h. The mixture was allowed to stand for 3 days. The solvent was removed by rotary evaporation, and the residue was partitioned between dichloromethane and water. The aqueous phase was extracted twice with CH_2Cl_2, and the combined organic phases were extracted 5 times with water, once with brine, and dried over anhydrous $MgSO_4$. The product was purified by flash chromatography (1:19 EtOH/CH_2Cl_2, 29 × 2). The highest-R_f fraction was isolated. ^{1}H NMR ($CDCl_3$): 7.60 (m) 7.44 (m), 7.18 (d, J = 7 Hz), 6.84 (d, J = 2 Hz), 6.82 (d, J = 10 Hz), 6.44 (dd, J = 7 Hz, J = 2 Hz), 5.26 (s), 4.00 (m), 3.79 (s), 2.36 (m), 2.26 (m). HRMS: calculated for $C_{53}H_{4910}P_2$: 895.2801002; measured: 895.2758.

Preparation of **11,12-dicarbomethoxy-9,10-etheno-9,10-dihydroanthracene 29**: Anthracene, blue-violet fluorescence (MCB, 3.000 g, 16.832 mmol, 1 eq), dimethyl acetylenedicarboxylate (Aldrich, 8.0 ml, 9.2 g, 65 mmol, 3.9 eq), and xylenes (20 ml) were combined in a 100-ml snrb flask. The mixture was heated with stirring under dry nitrogen, and refluxed overnight. The xylenes were removed by short-path distillation at water aspirator vacuum, and much of the DMAD was removed by steam distillation at atmospheric pressure. The residue was dissolved in

dichloromethane, extracted with brine, and dried over anhydrous $MgSO_4$. Activated charcoal and diatomaceous earth were added to the mixture, and the slurry was filtered. The filtrate was purified by flash chromatography (CH_2Cl_2, 15 × 4). Most fractions were white crystalline solids; some fractions were yellow oils from which colorless crystals eventually formed. These crystals were digested with boiling isooctane and filtered while warm. Total yield: 3.927 g (73%). ^1H NMR ($CDCl_3$): 7.33 (m, 4H), 6.98 (m, 4H), 5.44 (s, 2H), 3.75 (s, 6H).

Reduction of *p*-anisylmethylphenylphosphine oxide: *p*-Anisylmethylphenylphosphine oxide **22** (29.94 mg, 0.1216 mmol, 1 eq) was dissolved in 0.500 ml dry 1,2-dichloroethane in an oven-dried 10-ml snrb flask. The solution was heated to reflux, and hexachlorodisilane (Aldrich, 96%, 27 μl, 42 mg, 0.16 mmol, 1.3 eq) was added to the solution. After 5 min, the reaction flask was placed in an ice bath for 1 min, and the reaction was quenched by addition of sat. $NaHCO_3$. The mixture was extracted with CH_2Cl_2 twice, and the CH_2Cl_2 phase was dried over anhydrous $MgSO_4$. TLC (1:19 EtOH/CH_2Cl_2) showed nearly quantitative conversion to a high R_f compound.

Reduction of 1,2-bis(*p*-anisylphenylphosphino)ethane 23: 1,2-bis(*p*-Anisylphenylphosphino)ethane **23** (9.29 mg, 0.0189 mmol, 1 eq) and dry 1,2-dichloroethane (400 μl) were placed in an oven-dried micro-reactor and heated to 90 °C. The solid dissolved with heating, and hexachlorodisilane (Aldrich, 20 μl, 32 mg, 0.12 mmol, 3.1 eq) was added. Reduction was incomplete after 25 min, so an additional 10 μl Si_2Cl_6 was added (total 4.6 eq). The reduction was nearly complete 20 min after this second addition. The reaction was removed from heat, quenched by addition of 1 M K_2CO_3, and partitioned between dichloromethane and water. The organic phase was dried over anhydrous $MgSO_4$, and purified by flash chromatography (CH_2Cl_2, 18 × 0.8). Yield: 4.38 mg (55%). ^1H NMR ($CDCl_3$): 7.29 (m, 14H), 6.85 (d, J = 9 Hz, 4H), 3.79 (s, 6H), 2.03 (t, J = 4 Hz, 4H).

Attempted preparation of phosphine macrocycle 28: Phosphine oxide macrocycle **27** (2.52 mg, 2.8 μmol, 1 eq) was dissolved in 400 μl dry 1,2-dichloroethane in an oven-dried micro-reactor and heated to reflux. Hexachlorodisilane (Aldrich, 96%, 10 μl, 16 mg, 58 μmol, 10 eq) was added, and the reaction was followed by TLC (CH_2Cl_2). No compound with a nonzero R_f value was seen in 24 h. The reaction was removed from heat, and quenched by addition of 1 M K_2CO_3. The mixture was subjected to an extractive workup, and the product was analyzed by TLC. There were no fractions that were mobile in CH_2Cl_2 (phosphine oxides are very retentive, while phosphines are very mobile in this solvent), but there were a multitude of different spots when the eluent was 3:97 EtOH/CH_2Cl_2. The reduction was thus judged a failure.

Preparation of ***p*-anisylmethylphenylphosphine–borane:** Cerium trichloride heptahydrate (Aldrich, 1.1g, 2.95 mmol, 2.95 eq) was placed in a 25 ml Shlenk flask, and the flask was evacuated and heated to 140 °C for 2 h, then cooled under vacuum and placed under argon atmosphere. 5 ml freshly-distilled THF was added to the flask, followed by sodium borohydride (Aldrich, 114 mg, 3.01 mmol, 3.01 eq), and this slurry was magnetically stirred for 1 h. *p*-Anisylmethylphosphine oxide **22** (247 mg, 1.003 mmol, 1 eq) was added to this slurry and stirred for 30 min. Lithium aluminum hydride (50 mg, 1.317 mmol, 1.3 eq) was then added and allowed to stir for 2 days. The reaction was quenched by addition of water, followed by 1M HCl after the foaming died down. Toluene (10 ml) was added to the mixture, and the aqueous and toluene phases were separated. The aqueous layer was extracted thrice with CH_2Cl_2, and the combined organic phases were extracted with brine, dried over anhydrous $MgSO_4$, and purified by flash chromatography (CH_2Cl_2 + 8:92 MeOH/CH_2Cl_2, 2 × 2). Yield: 112.06 mg (46%). NMR ($CDCl_3$): (#390)

Preparation of **bis-phosphine–borane 30:** An oven-dried 100 ml snrb flask was loaded with anhydrous cerium trichloride (378 mg, 1.533 mmol, 3.08 eq) in a

drybox. The flask was removed from the drybox, 5 ml freshly-distilled THF was added *via* syringe, and sodium borohydride (Aldrich, 65 mg, 1.718 mmol, 3.46 eq) was added as well. The mixture was stirred for 30 min, then 1,2-bis(*p*-anisylmethylphenylphosphino)ethane **23** (122 mg, 0.249 mmol, 1 eq) was added in 20 ml freshly-distilled THF and stirred for an additional 30 min. Lithium aluminum hydride (20 mg, 0.53 mmol, 1.1 eq) was added, and the mixture was stirred for 1 day. Little reduction had occurred at this time, so an additional 50 mg $LiAlH_4$ was added, and the reaction mixture was heated to reflux for 3 h. The reaction was allowed to cool, and was quenched by addition of water. Dichloromethane and sat. Na, K tartrate were added to the solution and allowed to stand for 1 day. The aqueous phase was extracted with CH_2Cl_2, and the combined organic phases were extracted with sat. Na, K tartrate, water, and brine, and dried over anhydrous $MgSO_4$. Yield: 71.80 mg (59%). ^{1}H NMR ($CDCl_3$): 7.83–7.30 (br, 7H), 6.95 (d, J = 7 Hz, 2H), 3.82 (s, 3H), 1.81 (d, J = 10 Hz, 3H).

Preparation of **1,2-bis(*p*-hydroxyphenylphosphine-borane)ethane 31:** An oven-dried 10-ml snrb flask was charged with the bis-phosphine–borane **30** (18.63 mg, 0.0383 mmol, 1 eq), 500 μl dichloromethane, and boron tribromide (Aldrich, 8 μl, 21 mg, 0.08 mmol, 1.1 eq). The reaction was stirred under nitrogen for 4 h. Cleavage did not appear to be progressing toward completion, so an additional 8 μl BBr_3 was added. After 20 h, a third 8-μl sample of BBr_3 was added to the reaction mixture, and the demethylation appeared to be complete 1 h later. The reaction was quenched by addition of anhydrous ether, and solvent was removed by rotary evaporation. The remainder was dissolved in CH_2Cl_2, extracted twice with water, and dried over anhydrous $MgSO_4$. Yield: > 20 mg (> 100%). ^{1}H NMR ($CDCl_3$): 7.8–7.4 (br), 6.95 (m), 2.85–2.60 (br).

Attempted alkylation of **methyl(hydroxyphenyl)phenylphosphine–borane complex**: A 10-ml snrb flask was charged with methyl(hydroxyphenyl)phenylphosphine–borane complex (21.89 mg, 0.0383 mmol, 1 eq), anhydrous K_2CO_3 (Mallinckrodt, 50 mg, 0.36 mmol, 4.7 eq), 1,3-dibromopropane (Aldrich, 0.33 ml, 65 mg, 0.32 mmol, 4.2 eq), and 3 ml methanol. After 2 days of reflux, only a substance with a baseline R_f (solvent for TLC: 3:97 $MeOH/CH_2Cl_2$) was present.

Preparation of ***p*-anisylmethylphenylphosphine**: *p*-Anisylmethylphenylphosphine–borane complex (112.06 mg, 0.459 mmol) was dissolved in 1.00 ml dry morpholine, and the solution was freeze–pump–thaw degassed three times. The mixture was heated to 60 °C for 8 h. Some morpholine was removed by short-path distillation under reduced pressure; the remainder was removed by passage with dichloromethane through a (9×1) column of flash silica gel. Yield: 91.55 mg (87%). ^{1}H NMR ($CDCl_3$): 7.52–7.14 (bf, 7H), 6.85 (d, 9 Hz, 2H), 3.76 (s, 3H), 1.58 (d, 4 Hz, 3H).

Exposure of an ethenoanthracene dimethyl ester to trichlorosilane: An oven-dried 10-ml snrb flask was charged with 11, 12-Bis(carbomethoxy)-9,10-etheno-9,10-dihydroanthracene **29** (20.03 mg, 0.0625 mmol) and evacuated and flushed with dry nitrogen four times. Trichlorosilane (Aldrich, 87 μl, 117 mg, 0.87 mmol), diethylcyclohexylamine (Aldrich, 205 μl, 174 mg, 1.12 mmol), and dry acetonitrile (1 ml) were added, and the mixture was heated to 70 °C for 2 h. The reaction was quenched by addition of 2 ml 1 M K_2CO_3. The mixture was allowed to stand for 2 h, then was diluted with CH_2Cl_2 and extracted twice with with water, once each with 1 M K_2CO_3 and brine, and dried over anhydrous $MgSO_4$. TLC (CH_2Cl_2) and NMR indicated that **29** was still present, but that other compounds were present as well.

Exposure of an ethenoanthracene dimethyl ester to hexachlorodisilane: 11, 12-Bis(carbomethoxy)-9,10-etheno-9,10-dihydroanthracene **29** (14.76 mg,

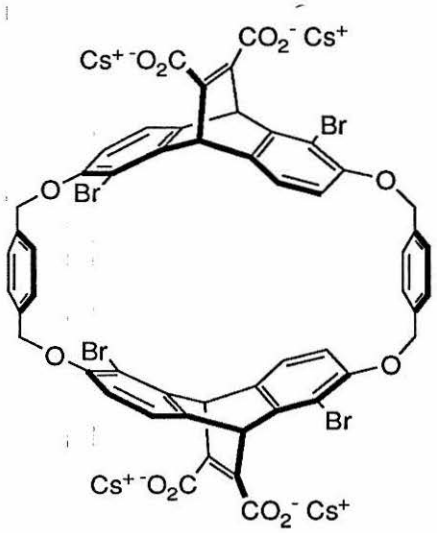
0.046 mmol) was dissolved in 0.625 ml dry and degassed benzene in a micro-reactor. The mixture was heated to reflux, and hexachlorodisilane (Aldrich, 21 μl, 33 mg, 0.12 mmol) was added to the refluxing solution. After 5 min, the reactor was placed in an ice bath, and the reaction was quenched by addition of sat. $NaHCO_3$. The mixture was extracted with CH_2Cl_2, and the organic phase was dried over anhydrous $MgSO_4$. TLC (1:4 EtOAc/PE) showed only one spot, and the NMR spectrum of the product was identical to that of the starting material.

III. Amino Diacids as Water-Solubilizing Groups

Abstract: Ethenoanthracenes linked by peptide bonds at the 11- and 12- positions to amino diacids promise to make water-soluble hosts readily accessible.

A. Purpose.

Discontinuation of the efforts to synthesize hosts with rigid 9,10-substituents did not remove the need for more water-soluble hosts. Instead, in the time since that project was halted, the need for hosts with greater solubility has only intensified. Hosts of more recent concern have even more hydrophobic surface area, such as **34** (TBP), or fewer water-solubilizing groups, such as the elusive **28**. Host **34** (TBP) is so hydrophobic that it aggregates at *all* concentrations observable by NMR.[20] In order to make this and other hosts soluble enough to study conveniently, some other way needs to be found to append a sufficient number of water-solubilizing groups to the ethenoathracene unit.



34

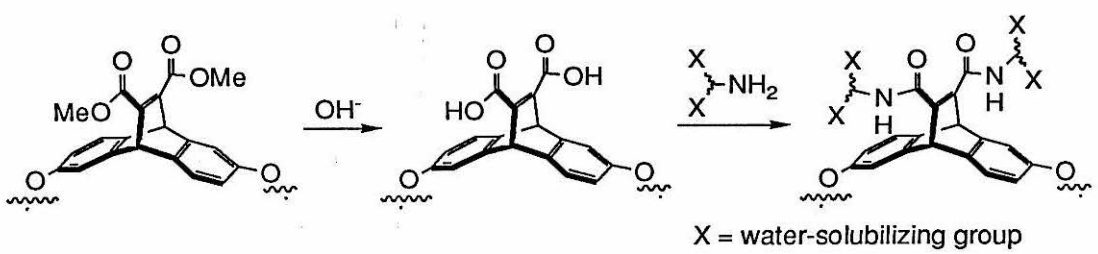
(20) Patrick Kearney, unpublished results.

After our experience with 9,10-substituents, we realized that the most important characteristic of any functionalization scheme was feasibility. Convenience was only slightly less important. The host synthesis was fairly long and difficult as it stood: a modest improvement in solubility as a result of an additional long and low-yielding series of steps would come at too high a cost. Additional considerations, such as synthetic flexibility and rigorous separation of the hydrophobic and hydrophilic regions of the host, were still significant, but could not be permitted to make the synthesis long or risky.

B. Design.

All of these conditions, including synthetic flexibility and exclusion of the water-solublizing groups from the binding site, were met in the synthesis outlined in Scheme VI. In this plan, the carboxylic acid groups of the ethenoanthracene form amides with amines possessing branching functionality. This branching increases the number of hydrophilic groups in the molecule, and the carboxylic acids conveniently provide the point of attachment on the skeleton that is the farthest removed from the cavity. The technology of such a transformation is highly developed because of its importance in polypeptide synthesis. Consequently, we believed that, one way or another, this step could be made to work.

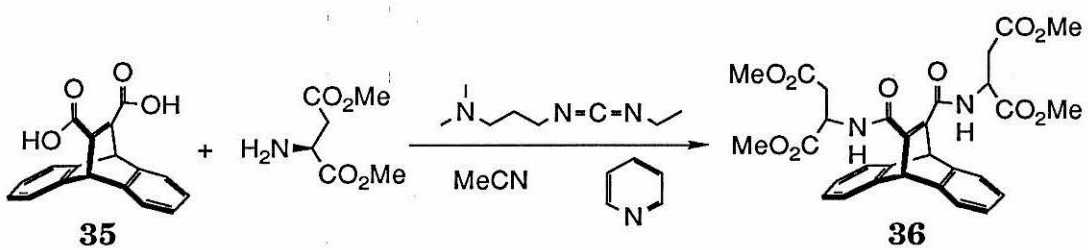
Scheme VI



C. Execution.

Aaron Clements, an undergraduate working in the group, was the first to work on this project and formed the amide **36** between 9,10-etheno-9,10-dihydroanthracene-11,12-dicarboxylic acid **35** and L-aspartic acid dimethyl ester. He used as a promoter the water-soluble diimide 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide, which has the considerable advantage over dicyclohexylcarbodiimide (DCC) that it and its urea product are easily separated from the amide reaction product by silica gel chromatography. I took over the project at this point, and began by optimizing the solvent conditions for this transformation.

Scheme VII

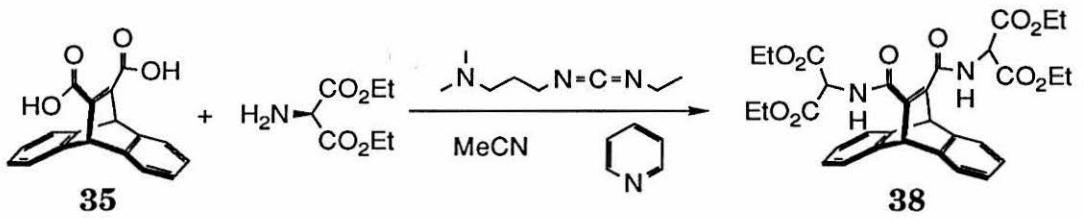


This compound was easily saponified to form the tetracarboxylic acid **37** in high yield. One disadvantage of the aspartate system, however, was obvious in the NMR spectra of these compounds. In the parent host P, all of the protons, which are either aryl or benzyl, are fairly far downfield in an NMR spectrum, leaving a wide window open for observing guest resonances. The aspartates, on the other hand, have the additional protons of the amino acid prominently appearing in this window. Even when the amido NH proton is removed by exchange with D_2O, the remaining three alkyl protons form an ABM system, whose complex splitting obscures a broad region of the spectrum. Furthermore, functionalization with chiral aspartic acid introduces stereochemical issues into the system. If, for instance, some of the aspartate substituents are racemized during saponification, still more

resonances would appear in the NMR spectrum. *Seven* diastereomers can arise from randomization of the aspartic acid moieties bonded to the four carboxylic acid sites of an otherwise homochiral host.

The simple solution to this problem was to functionalize the carboxylic acid positions with an achiral amino acid. Conveniently, the protected amino acid diethyl aminomalonate is commercially available, and it indeed is capable of the same chemistry as L-aspartic acid dimethyl ester (Scheme VIII). The NMR spectrum of its diamide with an ethenoanthracene, however, is immensely simpler, and there is no danger of additional diastereomers forming as a result of epimerization.

Scheme VIII



No rigorous solubility studies have been undertaken with either of these amino diacid functionalized ethenoanthracenes. Strong qualitative evidence that these molecules will have exemplary water solubilities is provided, however, by the NMR spectrum of tetraacid **39** which is reproduced in Figure 2. This spectrum was taken in neat D_2O. The concentration of the sample was not determined, but it is clear from the relative intensities of the solute and solvent peaks that it is fairly high. The solubility of this molecule in a neutral or slightly alkaline buffered medium should be even higher.

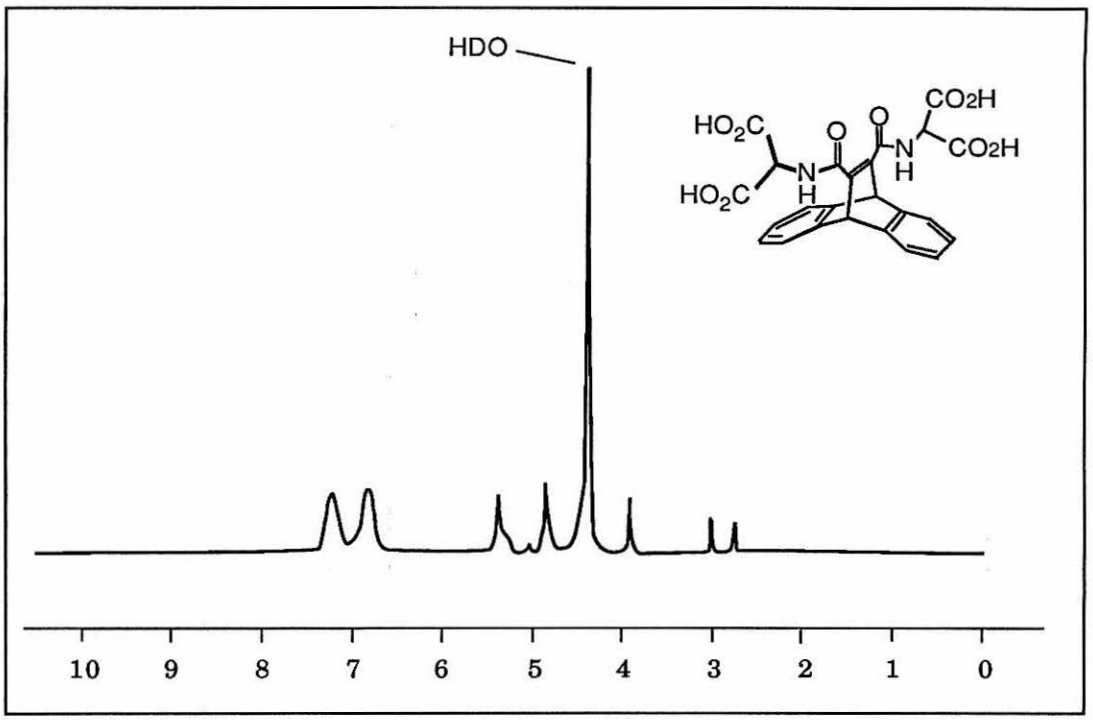
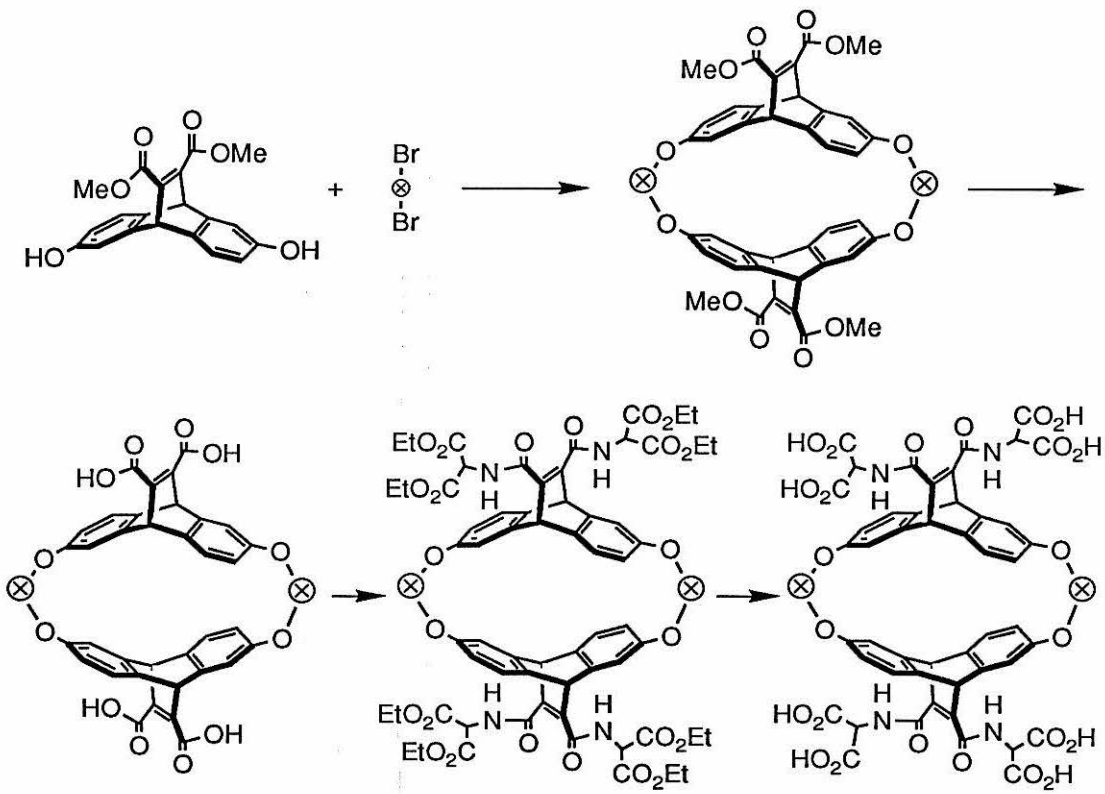


Figure 2. Artist's rendition of the NMR spectrum of **39** in D_2O.

This encouraging result directed our thoughts toward a scheme for incorporating this functionality into a host. The key consideration for such a scheme is whether the peptide-forming step should occur before or after the macrocyclization step. These two possibilities are shown in Schemes IX and X. Both involve the same steps, but in different sequence. The Scheme IX has several steps, including a fourfold functionalization, after the macrocyclization step. Scheme X, in contrast, involves only one step after the macrocyclization, and the peptide-forming step is a difunctionalization rather than a tetrafunctionalization.

The yield of a tetrafunctionalization step will be lower than for difunctionalization; statistically, the tetrafunctionalization yield will be the square of the difunctionalization yield. Furthermore, macrocycle is quite precious: these macrocyclizations have at best moderate yields, and separation of macrocyclization reaction

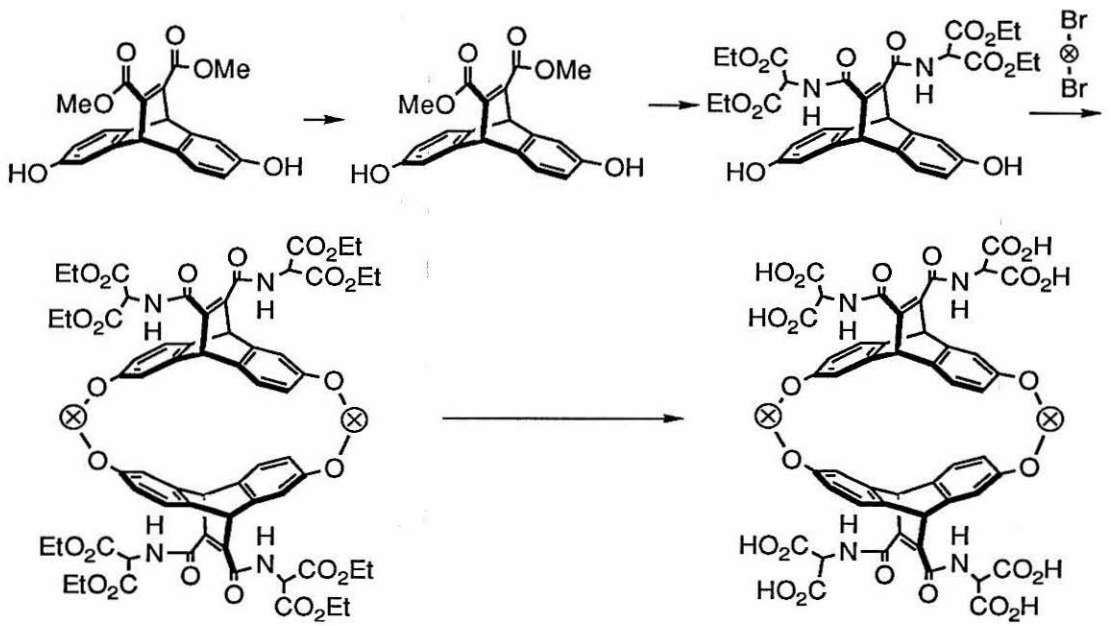
Scheme IX



products is tedious. Once macrocycle is formed, as much as possible should be retained. This argues strongly against using macrocycle as the starting material for any reaction but the most high-yielding. The second sequence (Scheme X) is thus the synthesis of choice.

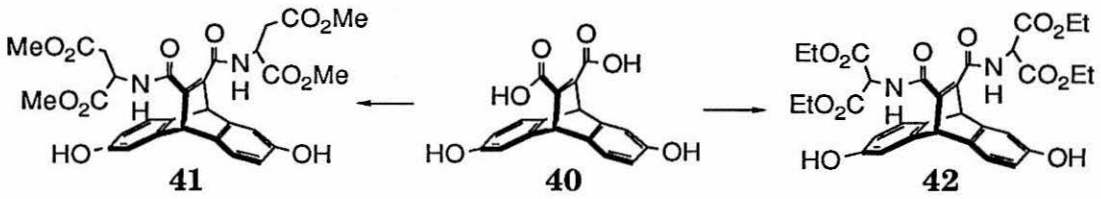
The only uncertainty in this second sequence was in the peptide-forming step. The free phenols of the ethenoanthracene could conceivably compete with the amino ester for the carboxylic acid groups, forming esters and thus lowering the amide yield. If this turned out to be the case, a protecting group for the phenols would have to be used, adding at least two more steps to the scheme. Since amines are much better nucleophiles than phenols, this was not expected to be a serious problem, but the peptide-forming reaction was tested on **40** to be certain. Both L-aspartatic acid

Scheme X



dimethyl ester and diethyl 2-aminomalonate formed amides with **40** (Scheme XI), confirming that the lower synthetic route is feasible.

Scheme XI



D. Conclusions.

Although a host with diacid substituents has not yet been synthesized, these studies indicate that such a synthesis is workable. The synthesis shown in Scheme IX compares well to the general host synthesis. It is two steps longer, but these steps, a saponification and a peptide formation, are clean and high-yielding. The only anticipated disadvantage is the inconvenient chromatographic behavior of the

amides themselves. This should be only a minor inconvenience, however, because this problem is corrected by using a more polar eluting solvent. The expected improvement in water solubility should make physical studies of the hosts easier, compensating for the additional synthetic difficulties.

E. Experimental Section.

Preparation of **9,10-etheno-9,10-dihydroanthracene-11,12-dicarboxylic acid 35**: Diester **29** (3.927 g, 12.25 mmol, 1 eq) was suspended in 250 ml methanol in a 500-ml snrb flask. Potassium hydroxide (7.0 g, 125 mmmol, 5.1 eq) was added, and the reaction curdled. The slurry was refluxed for 1 h, and the solvent was removed by rotary evaporation. The residue was dissolved in water, and precipitated by addition of conc. HCl. The precipitate was filtered, and the residue was rinsed with dilute HCl. Water was removed from the filter cake by reduced pressure. Yield: 3.531 g (98.6%).

Preparation of **ethenoanthracene bis-(dimethylaspartate) 36**: Dicarboxylic acid **35** (1.000 g, 3.42 mmol, 1 eq), L–aspartic acid dimethyl ester hydrochloride (Sigma, 1.365 g, 6.91 mmol, 1.01 eq), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (Aldrich, 1.330 g, 6.94 mmol, 1.01 eq), dry acetonitrile (20 ml), and pyridine (1.00 ml, 9,78 mg, 12.4 mmol, 1.81 mmol) were combined in a 100-ml snrb flask, making a cloudy yellow solution. This mixture was allowed to stand without stirring overnight; solvent was removed by rotary evaporation. The remaining yellow goo was partitioned between water and dichloromethane. The aqueous phase was extracted twice with dichloromethane, and the combined organic phases were extracted with 2 M HCl 2×, brine, and dried over anhydrous $MgSO_4$. Yield: 1.576 g (80%) white foam. ^{1}H NMR ($CDCl_3$): 7.93 (d, J = 6.7 Hz, 2H), 7.39 (m, 4H), 6.99 (m, 4H), 5.65 (s, 2H), 4.88 (m, 2H), 3.71 (s, 6H), 3.62 (s, 6H), 2.97 (dd, J = 4.8 Hz, J = 17 Hz, 2 H), 2.86 (dd, J = 4.8 Hz, J = 17 Hz, 2H).

Preparation of **ethenoanthracene bis-(diethylmalonate) 38**: A 10-ml snrb flask was charged with dicarboxylic acid **35** (100 mg, 0.342 mmol, 1 eq), diethyl 2-aminomalonate hydrochloride (Sigma, 148 mg, 0.699 mmol, 1.02 eq), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (Aldrich, 135 mg, 0.704 mmol, 1.03 eq), dry pyridine (100 μl, 97.8 mg, 1.24 mmol, 1.81 eq), and acetonitrile (2.000 ml). The yellow solution was stirred overnight. The reaction mixture was diluted with water and extracted with 6 × 5 ml portions of CH_2Cl_2 .The combined organic phases were extracted with 6 × 5 ml portions of 2 M HCl, once with brine, and dried over anhydrous $MgSO_4$. The product was purified by flash chromatography (8:92 $MeOH/CH_2Cl_2$, 5 × 1). Yield: 150.76 mg (72%). ^{1}H NMR ($CDCl_3$): 7.79 (d, 2H, J = 7.5 Hz), 7.37 (m, 4H), 6.98 (m, 4H), 5.61 (s, 2H), 5.14 (d, 2H, J = 7.5 Hz) 4.20 (qua, 8H, J = 7.5 Hz), 1.20 (t, 12H, J = 7.5 Hz).

Preparation of **ethenoanthracene bis-malonic acid 39**: Tetraester **38** (79.62 mg, 0.1308 mmol, 1 eq) was dissolved in 2.0 ml methanol in a 25-ml snps flask. A solution of KOH (39 mg, 0.70 mmol, 1.32 eq) in 0.3 ml methanol was added to it. The reaction mixture turned an orange-red color that gave way to a bright yellow. After 30 min, the solvent was removed by rotary evaporation. The residue was dissolved in Milli-Q purified water, and ion exchanged for NH_4^+. The free tetraacid was recovered by lyophilization. Yield: 69.93 mg (107%). ^{1}H NMR (D_2O): 7.21 (m, 4H), 6.80 (m, 4H), 5.38 (s, 2H). The methine proton at the malonate 2-position was not seen.

Preparation of **2,6-dihydroxyethenoanthracene bis-(dimethylaspartate) 41**: A 10-ml snrb flask was charged with racemic 9,10-etheno-9,10-dihydroanthracene-11,12-dicarboxylic acid **40** (28.18 mg, 0.0869 mmol, 1 eq), L–aspartic acid dimethyl ester (Sigma, 35 mg, 0.18 mmol, 1.0 eq), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (Aldrich, 37 mg, 0.19 mmol, 1.1 eq), dry pyridine (60 μl, 59 mg, 0.74 mmol, 4.3 eq), acetonitrile (2 ml), and DMF (1 ml). Solids never

fully dissolved. After 2 days, the acetonitrile was removed by rotary evaporation, and the remainder was partitioned between dichloromethane and 2 M HCl. The aqueous layer was extracted twice with CH_2Cl_2, and the combined organic phases were extracted with 6×20 ml 0.05 M HCl, once with brine, and dried over anhydrous $MgSO_4$. The product was purified by flash chromatography (8:92 MeOH/CH_2Cl_2, 15×1). ^{1}H NMR ($CDCl_3$): Peaks are doubled because there are two diastereomers present. ^{1}H NMR ($CDCl_3$): 8.10 (m), 7.07 (br), 7.01 (d, J = 3), 6.98 (d, J = 3), 6.82 (s), 6.76 (s), 6.28 (d, 4 Hz), 6.25 (d, 4 Hz), 5.39 (s), 5.33 (s), 4.90 (m), 4.60 (m), 3.62 (s), 3.57 (s), 2.90 (m).

Preparation of **2,6-dihydroxyethenoanthracene bis-diethylmalonate 42**: A 25-ml snrb flask was charged with diol diacid **40** (13.81 mg, 0.0427 mmol, 1 eq), diethyl 2-aminomalonate (Sigma, 20 mg, 0.085 mmol, 1.1 eq), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (Aldrich, 19 mg, 0.099 mmol, 1.2 eq), dry pyridine (14 μl, 14 mg, 0.18 mmol, 2.1 eq), and acetonitrile (0.500 ml). The yellow suspension was allowed to stand overnight, and solvent was removed by rotary evaporation. The residue was partitioned between CH_2Cl_2 and dilute HCl. The aqueous phase was extracted thrice with CH_2Cl_2, and the combined organic phases were extracted thrice with dilute HCl, once with brine, and dried over anhydrous $MgSO_4$. Yield: 19.29 mg (75%). ^{1}H NMR ($CDCl_3$): 8.03 (d, J = 7 Hz, 2H), 7.00 (d, J = 7 Hz, 2H), 6.78 (br), 6.27 (br. d, J = 8 Hz, 2H), 5.40 (s, 2H), 5.17 (d, J = 6 Hz, 2H), 4.18 (qua, J = 8 Hz, 8H), 1.20 (t, J = 8 Hz, 12 H).

IV. Does the Cation-π Effect Apply to Alkali Metals?

Abstract: Cyclophanes with small cavities were prepared as possible hosts for alkali metal cations. None were observed to associate with alkali metals in acetonitrile or chloroform.

A. Purpose.

After the publication by Stauffer and Dougherty implicating the cation-π effect in biological acetylcholine binding,[21] Roderick MacKinnon, an investigator studying the voltage-gated potassium channel, found evidence of similar effects operating in that system. Namely, a threonine residue near the mouth of the channel is required for blockage by tetraethylammonium (TEA). Replacement of this residue with lysine, arginine, glutamine, or valine renders the channel insensitive to TEA; replacement by tyrosine increases the blocking affinity by a factor of fifty.[22] Furthermore, the portion of the protein thought to be the actual transmembrane channel does not support a model of recognition of the K^+ by negative charges. The channel is composed of four identical subunits; the "pore" region of each subunit is a segment of about forty residues connecting two α-helices. None of the twenty or so residues making up the pore are anionic; instead, four rigorously conserved residues are aromatic.

These findings suggested that occupation of the channel by alkali metal cations was made more favorable by cation-π interactions between the metal and the channel lining. If cation-π interactions between aromatic rings and alkali metals are a significant force in biological systems, then they should also be detectable in smaller,

---

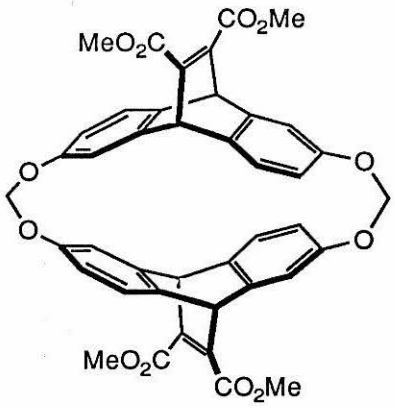
(21) Dougherty, D. A.; Stauffer, D. A. "Acetylcholine binding by a synthetic receptor: implications for biological recognition," *Science* **1990**, *250*, 1558–1560.

(22) Miller, Christopher "1990: *Annus mirabilis* of potassium channels," *Science* **1991**, *252*, 1092–1096.

more tractable synthetic systems. This raised the tantalyzing prospect of artificial cyclophane hosts complexing alkali metals.

B. Design and Execution.

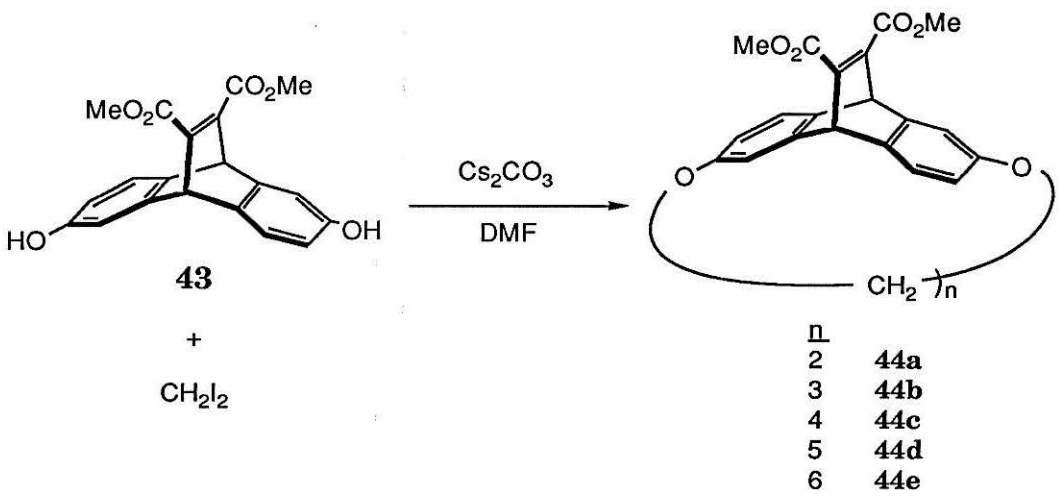
Consequently, we decided to make some cyclophane hosts, patterned after our successful ethenoanthracene macrocycles, with cavities of an appropriate size to bind alkali metal ions. The first target was **44**, which is a member of our general ethenoanthracene family. It contains two ethenoanthracene units and two linker groups, just as larger and more familiar hosts such as P and V, but the linker group is only a single methylene unit. This gives the molecule a much smaller cavity: small enough, we hoped, to provide a favorable alkali metal binding site.



44

The obvious way to make this molecule is to perform a macrocyclization under high dilution conditions between equimolar amounts of 2,6-dihydroxyethenoanthracene **43** and diiodomethane, as shown in Scheme XII. This reaction was carried out using racemic **43**, yielding a number of products which were tediously separated from each other by centrifugal chromatography. These spots all had similar NMR spectra, which were qualitatively similar to the expected spectrum of **43**. All had signals of the proper intensities in the aromatic, methyl ester, and methylene regions of the spectrum, but all were far more complex than warranted by the expected highly symmetrical structure of **44**.

Scheme XII



Several effects could have been responsible for this behavior. One was the number of diastereomers possible for each cyclic oligomer of a 2,6-disubstituted ethenoanthracene. Since the starting material of the macrocyclization was racemic, a "dimer" macrocycle could comprise ethenoanthracene units either of the same or opposite absolute configurations. There are also two possible diastereomers of a "trimer" macrocycle; its ethenoanthracene units could be either all the same, or one could be different from the other two. A "tetramer" macrocycle has four possible diastereomers: all ethenoanthracene units the same, one different from the other three, two of each configuration with identical units adjacent to each other, or two each with identical units opposite from each other. Higher oligomers have still more diastereomers.

Both dimer macrocycles should have simple NMR spectra. The trimer with three identical subunits should also have a simple spectrum, but the spectrum of its isomer with one unique subunit should be more complicated. The tetramer with four identical subunits, as well as the one with only subunits of different configuration connected to each other, should also have simple spectra, but the spectra of the

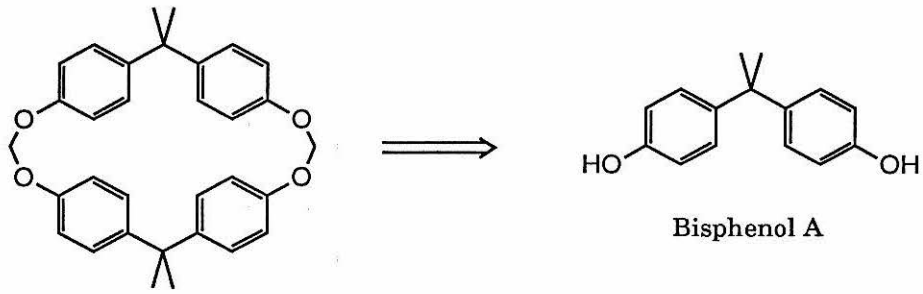
remaining tetramers should be much busier. The presence in a sample of many different diastereomers, or of single diastereomers with inherently messy spectra, would cause behavior as was observed.

On the other hand, a sample of even the two dimer diastereomers together would have had a simpler spectrum than any we observed. To check if this complexity was caused by several conformational isomers which were inhibited from rapidly interconverting by the close contacts and demanding sterics of **44**, two of these compounds were investigated by variable-temperature NMR. If heating the samples caused coalescence of several of the peaks and a simplified spectrum, or even a broadening of some of the peaks, such conformational effects would be implicated. However, when these compounds in nitrobenzene-d_5 solution were heated *even to* $117\,^{\circ}C$, no tendency toward coalescence was observed.

We remained mystified by these compounds until we received the results of their mass spectra. These showed that the spots we separated from the reaction mixture were not individual diastereomers of a few of the lower cyclic oligomers of **44**, but instead were mixtures of all diastereomers of a certain molecular weight. Thus, the first spot isolated contained the trimers **44b**, the second the tetramers **44c**, all the way up to the hexamers **44e**. Furthermore, a still-later spot failed to give a mass spectrum. The complexity of the NMR spectra was due to the presence of different diastereomers in each sample, and to the inherent complexity of the spectra of the lower-symmetry diastereomers. These stereochemical complications could be avoided by using enantiomerically pure **43** in the macrocyclization reaction. There was no point to performing such an experiment, however, because the macrocyclization using the more readily-available racemic **43** adequately demonstrated that the desired dimer **44a** was not formed.

Moreover, those stereochemical issues were avoided altogether in another family of cyclic compounds, **45**. In this series, the function of the diol **43** is assumed by

the much less dear diol Bisphenol A, which is a commercially-available compound of industrial importance. Since Bisphenol A has no stereogenic centers, its cyclic oligomers each have only one stereoisomer.

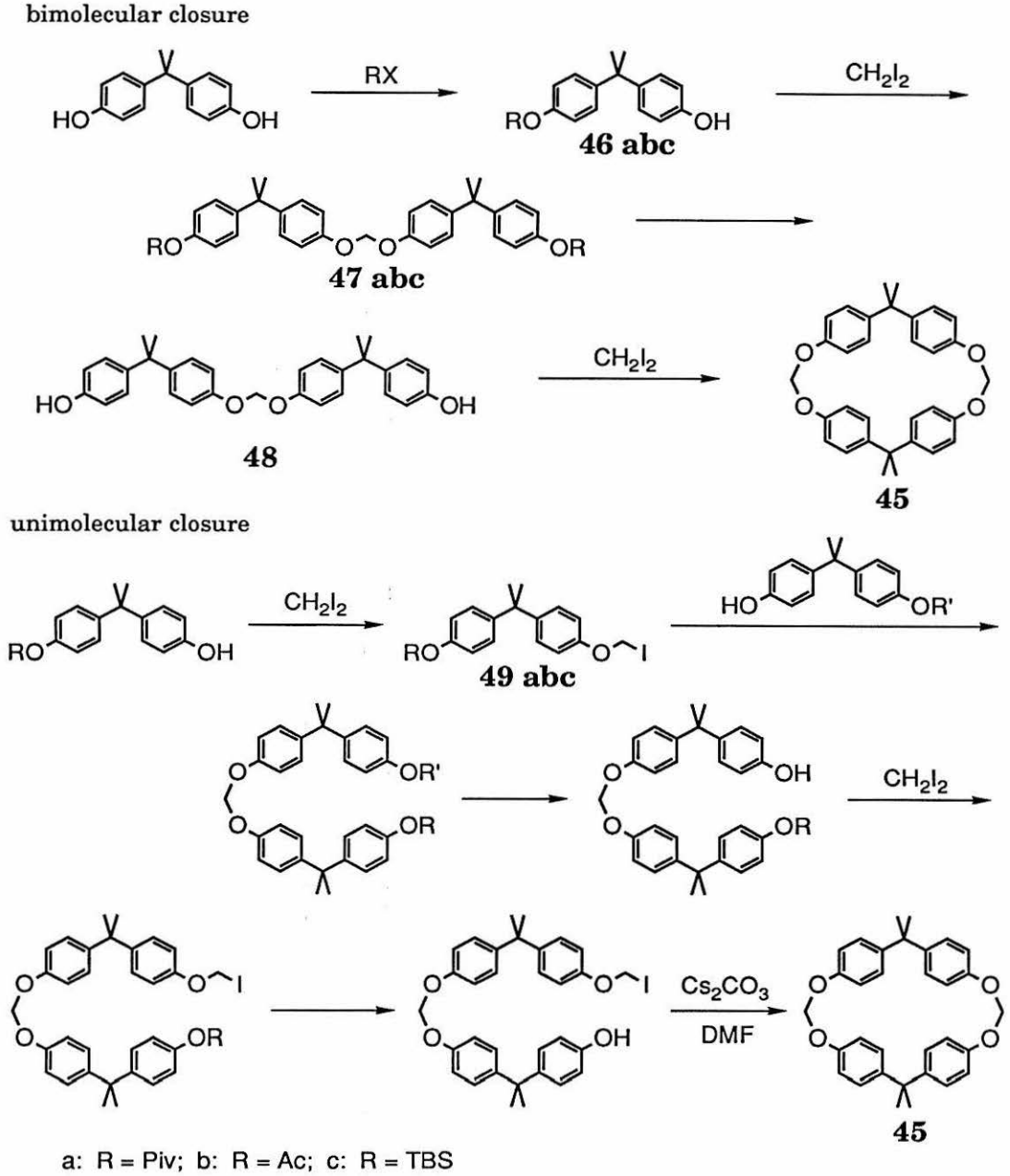


45 **Bisphenol A**

These compounds were made much like **44**, in a high-dilution reaction of the diol and diiodomethane in dry DMF in the presence of excess cesium carbonate. There was no reaction at room temperature, however, so the cyclization was carried out at 60 °C. This produced, as expected, a number of compounds with closely-spaced R_f values by TLC. The major spot could be obtained in pure form by two recrystallizations from toluene, and its NMR spectrum was as expected. This compound showed, however, no ability to extract any of the alkali metal picrates from water to chloroform solution. Subsequent mass spectral results indicated that this compound was actually the tetramer, with *four* Bisphenol A units per molecule, instead of the dimer, as we had supposed. The cavity of this molecule is obviously far too large to furnish a snug environment for an alkali metal cation.

The results of these macrocyclizations demonstrated that a simple mixture of a diphenol with diiodomethane would not produce the dimeric macrocycles we desired. We therefore sought a longer, stepwise synthesis employing a closure reaction that was bimolecular or unimolecular instead of tetramolecular. Plans for such syntheses are shown in Scheme XIII.

Scheme XIII



The scheme allowing for the bimolecular closure requires a phenol protecting group R that can withstand the conditions of the alkylation by diiodomethane (cesium carbonate, dry DMF, 60 °C), yet can be removed under conditions that do not

destroy a formaldehyde acetal. Acetate and *t*-butyldimethylsilyl were unsatisfactory in the former regard; these protecting groups were removed under the alkylation conditions. Pivalate (2,2,2-trimethylacetate), however, proved satisfactory. This group was unaffected by the alkylation conditions, yet could be removed in the presence of the formaldehyde acetal by two equivalents of *n*-butyllithium. Unfortunately, however, the first alkylation step, **46a** → **47a**, had a poor yield, and its products were contaminated with Bisphenol A dipivalate, which was probably present in small quantities in the starting material. The low reaction yield then ensured that this contaminant made up a siginificant portion of the isolated product fraction.

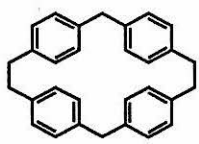
The synthesis employing a unimolecular closure reaction is even more demanding. It requires *two* phenol protecting groups, *both* of which can endure alkylations, but which can be removed in the presence of a formaldehyde acetal. Futhermore, one of these groups must be selectively removed in the presence of the other, and the other must be labile in conditions to which a halomethyl aryl ether is stable. More immediately, it requires a halomethyl aryl ether to be synthetically accessible and isolable. This requirement was not fulfilled: normal alkylation conditions (cesium carbonate, DMF, six-fold excess of diiodomethane) yielded only the formaldehyde acetal **47a**, and no iodomethyl ether **49a**.

Further investigations into these approaches were not pursued, because some even smaller cyclophanes appeared to be both more desirable and more accessible. [2.1.2.1]-Paracyclophane, **50**, has been reported several times in the literature,[23,24] and its oxygenated relative, **51**, was expected to be available from Bisphenol A and an intermediate of the synthesis of **50**.

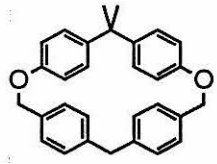
---

(23) Sergheraert, C.; Marcinal, P.; Cuignet, E. "Preparation du [2.1.2.1] paracyclophane," *Tetrahedron Lett.* **1977** *33*, 2879–2880.

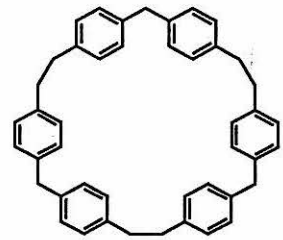
(24) Grützmacher, Hans-Friedrich; Huseman, Wolfram "Synthesis and reactions of some new dithia[3.1.3.1]paracyclophanes and [2.1.2.1]paracyclophanes," *Tetrahedron Lett.* **1987**, *43*, 3205–3211.



50



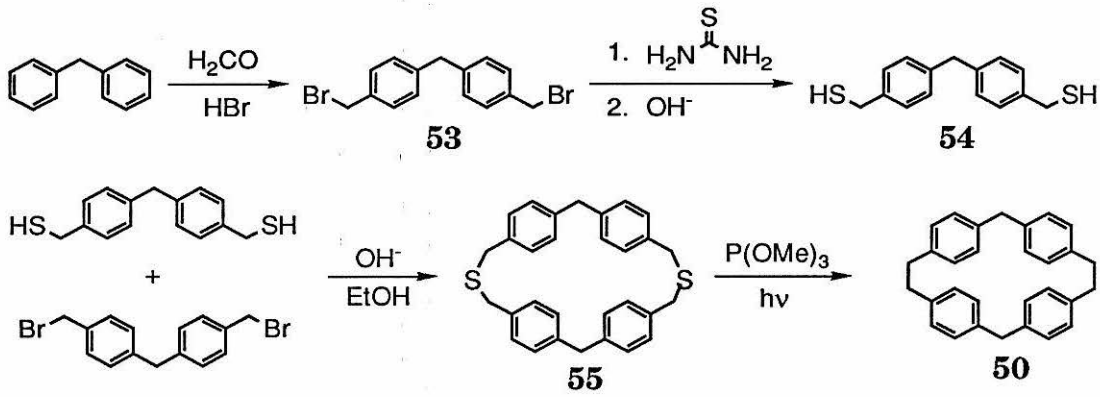
51



52

Two syntheses of [2.1.2.1] paracyclophane have been published. One involves simple reductive Wurtz coupling of 4,4′-bis(bromomethyl) diphenylmethane;[23] the other, longer synthesis involves sulfur extrusion from a dithia [3.1.3.1]-paracyclophane.[24] We elected to follow the longer synthesis because of its reported higher yield and lack of complications from higher oligomers such as [2.1.2.1.2.1] paracyclophane, **52**.

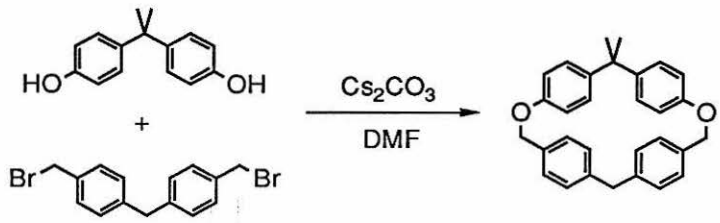
Scheme XIV



This synthesis was carried out as reported by Grützmacher and Huseman[24] and illustrated in Scheme XIV. The first two steps of this process were straightforward, and could be carried out as reported. Accounts of the final steps omitted some key observations, however, so the full details of these steps appear here in the Experimental section. The final product had a satisfactory molecular weight by HRMS, and its NMR spectrum was in accord with the literature report.

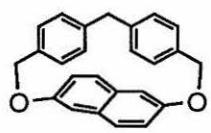
The oxygenated analogue **51** of [2.1.2.1]-paracyclophane was prepared by a high-dilution addition of a 1:1 mixture of 4,4′-bis (bromomethyl) diphenylmethane **53** and Bisphenol A to a large excess of cesium carbonate in dry DMF (Scheme XV). This product also showed a satisfactory NMR spectrum and HRMS.

Scheme XV

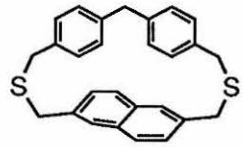


Once these cyclophanes were in hand, their association behavior with each of the five alkali metal picrates was studied by liquid-liquid extraction between water and chloroform. No enhancement of extraction into the chloroform phase was seen. Their association behavior in acetonitrile-d_3 was then studied by NMR; no shifting of host resonances occurred upon addition of picrate, indicating that binding probably did not occur to any appreciable extent. Solid-liquid extraction was used in a last attempt to measure association constants between these cyclophanes and the alkali metals in chloroform. This technique proved to be extremely unreliable and irreproducible despite extensive precautions. I found no evidence that the presence of either of these cyclophanes increased the solubility of any of the alkali metal picrates in chloroform. In contrast, a positive control study with 18-crown-6 showed a dramatic enhancement of the amount of picrate drawn into chloroform. This showed that the solid-liquid extraction technique indeed can detect binding, and that the failure to detect binding in the case of the cyclophanes is a failure of the cyclophanes, and not of the technique.

These results marked the end of my association with this project. Alison McCurdy continued the investigaton with such cyclophanes as **56** and **57**, which have even smaller cavities than **50** and **51**. Preliminary liquid-liquid extraction studies have not shown these molecules to be any more effective at solubilizing metal picrates in chloroform than the molecules I studied.



56



57

C. Conclusions.

These investigations indicate that an array of electron-rich aromatic rings properly situated to surround an alkali cation is, by itself, not effective in stabilizing the cation in chloroform or acetonitrile. It is not clear why a cation-π effect is not present for alkali metal cations and cyclophanes **50** and **51** when such an effect has been so clearly demonstrated for quaternary ammonium ions and host P_E. Perhaps the cation-π effect seen with the ethenoanthracene-based hosts is due to a specific property of the quaternary ammonium guests or their counterions, or of the ethenoanthracene hosts themselves. One clear lesson from these studies is that the cation-π effect is not general enough for one to be confident of the stabilization of any cation by encapsulation within a properly-sized cyclic array of aromatic rings.

D. Experimental Section.

1. Binding Studies. Absorbances of the alkali metal picrates in acetonitrile were determined on a Hewlett–Packard 8451A diode array spectrometer from a series of five-fold serial dilutions of samples of the picrates in spectrophotometric grade acetonitrile. The concentration range covered in this calibration was 5×10^{-3} M

down to 5×10^{-9} M. The spectra were best resolved, and the absorbances at 380 nm obeyed Beer's Law, at concentrations between 10^{-4} and 10^{-6} M. The values of A_{380}/cl (A = absorbance, c = concentration, l = path length) from each of the spectra in this range were averaged to obtain an overall extinction coefficient of 18,000 M^{-1} cm^{-1}. The literature value for the picrate extinction coefficients in acetonitrile is 16,900 M^{-1} cm^{-1}.[25]

All picrates except for lithium appeared to have similar extinction coefficients (for Li, $\varepsilon \simeq 14000$; for the others, $\varepsilon \simeq 18000$). The amomalous lithium result was attributed to the poorer form of the lithium picrate crystals in comparison to the other picrates; apparently, these crystals contained residual ethanol. The experimental extinction coefficient of lithium picrate was therefore disregarded.

(a) Liquid–liquid extraction studies. The basic procedure followed was as described by Cram and co-workers.[26] Solutions of the alkali metal picrates[27] in Milli-Q purified water were made up as follows: Li: 0.01494 M; Na: 0.1498 M; K: 0.1495 M; Rb: 0.00986 M; Cs: 0.009978 M. Polypropylene centrifuge tubes (2-ml size) were charged with 0.500 ml of an ethanol-free chloroform solution of **45c** (bisphenol A tetramer) and one of the pictrate solutions. Of the Li, Na, and K picrate solutions, 0.500 ml were used; 0.750 ml of the Rb and Cs picrate solutions were used. An additional reference tube was made up with host and 0.500 ml purified water. Each tube was capped, agitated at high speed on a vortex mixer for 1 min, and centrifuged at high speed in a Micro-Centaur centrifuge for 10 min. A 100 μl aliquot was removed from each chloroform layer with a Hamilton 100 μl glass

(25) Moore, Stephen S.; Tarnowski, Thomas L.; Newcomb, Martin; Cram, Donald J. "Host-guest complexation. 4. Remote Substituent effects on macrocyclic polyether binding to metal and ammonium ions." *J. Am. Chem. Soc.* **1977**, *99*, 6398–6405.

(26) Koenig, Karl E.; Lein, George M.; Stuckler, Peter; Kaneda, Takahiro; Cram, Donald J. "Host-guest complexation. 16. Synthesis and cation binding characteristics of macrocyclic polyethers containing convergent methoxyaryl groups," *J. Am. Chem. Soc.* **1979**, *101*, 3553–3566.

(27) Silberrad, Oswald; Phillips, Henry Ablett "The metallic picrates," *J. Chem. Soc.* **1908**, *93*, 474–489.

syringe, and diluted to the mark ml with spectrophotometric grade acetonitrile in a 5-ml volumetric flask. The sample made from the tube without picrate was used as the reference for the other samples. Spectra were plotted from 300 to 600 nm, and the absorbance of each sample at 380 nm was recorded. All of the absorbances were below the range from which quantitative concentration information could be obtained, and none of the spectra clearly showed picrate absorption bands.

Subsequent studies showed that, in time, chloroform samples held in polypropylene centrifuge tubes of different colors developed different UV absorbances. None of these absorbances were observed in the extraction samples, but the use of polypropylene centrifuge tubes was nonetheless discontinued.

Similar liquid–liquid extraction studies were carried out with the [2.1.2.1]paracyclophanes **50** and **51**. The same picrate stock solutions as described previously were used. The concentration of the chloroform solution of host **50** was 7.633 mM, and of host **51** was 15.11 mM. The same volumes of solutions were combined as before. The cyclophane and picrate solutions were placed into 13×100 mm disposable test tubes, each vortexed at high speed for 2 min, and then centrifuged at maximum speed in a clinical centrifuge for 1 h. UV samples were prepared from the chloroform fractions in the same way as before. In no cases could picrate absorbances be detected in these samples.

(b) NMR titration binding studies of [2.1.2.1]paracyclophanes with alkali metal picrates in acetonitrile-d_3: NMR samples were made up of each of the alkali metal picrates in acetonitrile with each of the [2.1.2.1]paracyclophanes. The concentrations of the species in the appropriate samples were: **50**: 540 μM; **51**: $\sim$ 400 μM; Li: 5.28 mM; Na: 5.12 mM; K: 4.36 mM; Rb: 2.33 mM; Cs: 2.21 mM. None of the spectra showed any differences from the spectra of the cyclophanes alone.

(c) Solid–liquid extraction studies. All glassware except for pipets used in these extraction studies was soaked overnight in a 10% nitric acid solution, rinsed ten times with deionized water, three times with Milli-Q water, and oven-dried prior to use. Ethanol-free chloroform was prepared by distillation of spectrophotometric grade chloroform from P_2O_5 through a 20 cm long Vigreaux column, stored in an acid-washed, oven-dried brown bottle in the dark inside a nitrogen-flushed dessicator, and used within one week of preparation. Gelman Accrodisc 0.45 μm PTFE syringe end filters were prepared by passing 50 ml spectrophotometric grade chloroform through each, and evacuating in a vacuum dessicator for at least 3 h. The general extraction protocol was as follows. Five 1-ml ampoules were charged with dry crystals of alkali metal picrate, and 1.2 ml of a chloroform solution was added to each. An ampoule of chloroform solution alone was also prepared. Each sample was frozen in liquid nitrogen, and the ampoules were sealed under vacuum. Each ampoule was placed in a 13×100 mm test tube and covered with water. The test tubes were each placed in 50-ml Erlenmeyer flasks, which were filled with water above the level of the chloroform solutions in the ampoules. These were placed in the tray of a Bransonic 2200 ultrasonic cleaner, which was covered with aluminum foil to exclude light, and the samples were sonicated for six successive 25-min sessions. Between each session, the positions of the samples in the sonicator were interchanged, and the sonicator water was replaced. The ampoules were cracked, and their contents forced through prepared filters. An aliquot (0.8 ml) of each filtrate was placed in its own 10-ml snrb flask; solvent was removed by rotary evaporation, and the flasks were evacuated overnight. The residue in each flask was diluted to the mark in a 2-ml volumetric flask with spectrophotometric grade acetonitrile, and the UV spectrum of each sample was recorded between 200 and 600 nm. The cuvette was cleaned after each sample with 2 rinses of acetonitrile, 2 rinses of chloroform, and 10 more rinses of acetonitrile. A blank spectrum of acetonitrile was run between

each sample to be certain that the cuvette was properly cleaned; if necessary, more rinsings were performed until host and picrate absorbances were no longer present. A fresh reference spectrum of acetonitrile was taken whenever the baseline appeared to change. The absorbance at 380 nm of the sample prepared from the ampoule without picrate was subtracted from the absorbance at 380 nm of each of the other samples. The concentrations of picrates in the chloroform solutions in each of the ampoules were calculated from these absorbance values.

Determining the solubilities of the alkali metal picrates in ethanol-free chloroform: The solid-liquid extraction procedure described above was carried out four times. The solubilities obtained in each study are reported in Table I. Although these values are not highly reproducible, they are consistently low, and all numbers for a given picrate are within an order of magnitude of each other.

Table I. Picrate solubilities.[a]

	alkali metal				
trial	Li	Na	K	Rb	Cs
1	5.95×10^{-5}	6.18×10^{-5}	2.53×10^{-5}		7.65×10^{-6}
2	8.85×10^{-5}	3.00×10^{-5}	6.85×10^{-5}	6.75×10^{-6}	7.12×10^{-6}
3	3.55×10^{-5}		3.14×10^{-5}		6.53×10^{-5}
4	5.00×10^{-5}	6.10×10^{-6}	1.70×10^{-5}	6.65×10^{-6}	1.24×10^{-5}

[a]In M. Blank entries are a result of ampoule breakage during sonication.

Binding studies of alkali metal picrates with cyclic hosts in chloroform by solid–liquid extraction: The solid-liquid extraction procedure described above was followed. The 2-ml UV sample solutions were prepared by first adding 0.40 ml spectrophotometric grade chloroform to the dried residue, and then diluting with acetonitrile up to 2.00 ml. This was to combat the poor solubility of the cyclophanes in acetonitrile. Previous control studies established that up to 20% chloroform in a sample affected neither the shape nor the intensity of the picrate

absorption between 300 and 600 nm. The concentrations of [2.1.2.1]paracyclophane **50**, dioxa[2.1.2.1]paracyclophane **51**, and 18-crown-6 in their extraction solutions were 25.86 mM, 32.2 mM, and 33.9 mM, respectively. Although more than the usual amounts of pictrates were used in the study with 18-crown-6, this host brought the entire picrate sample into solution in all cases but potassium. The total concentrations of picrate in the extracted solutions as determined from the measurements of the UV samples are reported in Table II. These results do not suggest that the cyclophanes function as alkali metal binders.

Table II. Picrate solubilities in the presence of cyclophanes.[a]

		alkali metal				
host	[H][b]	Li	Na	K	Rb	Cs
50	25.86	3.43×10^{-5}		6.48×10^{-5}	8.15×10^{-6}	1.77×10^{-5}
51	32.20	4.30×10^{-5}	3.30×10^{-5}	4.95×10^{-5}	3.22×10^{-6}	1.07×10^{-5}
18-crown-6	33.9	1.97×10^{-2}[c]	1.44×10^{-2}[c]	3.17×10^{-2}	1.90×10^{-2}[c]	2.40×10^{-2}[c]

[a] In M. Blank entry is a result of ampoule breakage during sonication. [b] In mM. [c] All picrate dissolved.

2. Synthesis. 1, 1′-methylenebis (4-bromomethylbenzene) **53** was prepared by the method of Steinburg and Cram.[28] The synthesis of [2.1.2.1]paracyclophane was carried out according to the published report of Grützmacher and Huseman;[24] this report is sketchy in its later steps, so the procedures for these steps are reported in more detail here.

Preparation of **ethenoanthracene-methylene macrocycles 44:** A DMF solution 2.18 M in diiodomethane (Aldrich) was prepared by weighing out 0.585 g of the diiodomethane in a 5 ml volumetric flask and diluting to the mark with dry DMF. Diol **43** (347 mg, 0.984 mmol, 1 eq) was placed in an oven-dried snps flask; 2.25 ml of the DMF solution of the diiodide (2.25 ml × 0.585 g / 5.0 ml =

(28) Steinburg, H.; Cram, D. J. "Macro Rings. II. Polynuclear Paracyclophanes," *J. Am. Chem. Soc.* **1952**, *74*, 5388–5391.

0.263 g, 0.982 mmol, 0.999 eq) and 10 ml DMF were added to it. This solution was transferred with additional DMF to a dry 60-ml disposable polypropylene syringe (total volume was about 40 ml). Cesium carbonate (Aldrich, 3.225 g, 9.898 mmol, 5.03 eq) and 100 ml dry DMF were placed in an oven-dried 300 ml snrb flask and capped with a rubber septum. The syringe and flask were wrapped with aluminum foil and stirred under dry nitrogen. The solution of the diol and diiodide was added to the cesium carbonate suspension by syringe pump over a period of 2 days, and the reaction was stirred at room temperature under nitrogen for an additional 4 days. Solvent was removed under reduced pressure, and the residue was partitioned between water and dichloromethane. The aqueous phase was acidified with HCl, and extracted with dichloromethane. The combined organic phases were extracted twice with 1 M K_2CO_3, dried over $MgSO_4$, and rotovapped to a yellow oil. Preliminary purification by flash chromatography (EtOAc, 14 × 2) yielded a white foam, which was purified again by flash chromatography (1:9 ether/dichloromethane, 18 × 4) giving partial separation of six different fractions. Each group of partially-purified fractions was separately purified by Chromatotron (dichloromethane + ether, 1-mm plate) one or more times to eventually isolate all compounds. NMR spectra of all fractions were complex. MS: spot 1: 1094 (trimer + 1), spot 2: 1457 (tetramer + 1), spot 3: 1820 (pentamer), spot 4: 2185 (hexamer + 1), spot 5: 2185 (heptamer +12?), and the final fraction showed no MS.

Methylene-linked cyclic Bisphenol A oligomers 45: All glassware was oven-dried prior to use. Diiodomethane (Aldrich, 1588 mg) was weighed out in a 5 ml volumetric flask and brought to the mark with dry DMF. Bisphenol A (Aldrich, 1143 mg, 5.007 mmol, 1 eq) was weighed out in a 25-ml snps flask, and an aliquot of the diiodomethane solution (4.20 ml, 1334 mg, 4.98 mmol, 1 eq) was added to it. The resulting solution was placed in a 60-ml disposable syringe, along with 2 × 20 ml rinsings of the flask. A 500-ml snrb flask was charged with a stir bar, cesium

carbonate (Aldrich, 8.179 g, 25.1 mmol, 2.5 eq), and 80 ml DMF. The reaction flask was placed in an oil bath at 60 °C, and the solution of diol and diiodide was added by a syringe pump over the course of 2 days. The reaction was stirred at this temperature for an additional 2 days. The DMF was removed under reduced pressure, and the residue was partitioned betweem 1 M NaOH and ether. The ether layer was extracted with water and brine, dried over anhydrous $MgSO_4$, and purified by flash chromatography (dichloromethane, 12 × 4). The high-R_f fraction was recrystallized twice from toluene to yield 68.13 mg (30%) of a single product, which was the second-fastest spot by TLC (CH_2Cl_2). NMR ($CDCl_3$): 7.05 (effective AB quartet, J = 10 Hz, 16H), 5.68 (2, 4H), 1.60 (s, 12H). MS: $m/e = 961$. This is the tetramer. HRMS of fragmentation peak at $m/e = 430$: calculated for $C_{46}H_{44}O_4$: 660.3239603; measured: 660.3268.

Preparation of **Bisphenol A mono-pivaloyl ester 46a:** Bisphenol A (Aldrich, 10.001 g, 43.81 mmol, 2 equiv), trimethylacetyl chloride (Aldrich, 5.2 ml, 5.1 g, 42 mmol, 0.96 equiv), dry pyridine (4.0 ml, 3.9 g, 49 mmol, 1.1 equiv), and anhydrous ether (20 ml) were placed in a 100 ml snrb flask and magnetically stirred under dry nitrogen at room temperature for 2 h. The reaction mixture was partitioned between ether and water; the water layer was acidified to pH 2 with HCl. The ether phase was extracted twice with 1 M HCl, dried over anhydrous $MgSO_4$, and solvent was removed by rotary evaporation. The clear, colorless goo was purified by flash chromatography (CH_2Cl_2 + Et_2O, 22 × 5) to give a white solid. Yield: 7.159 g (52%). ^{1}H NMR ($CDCl_3$): 7.30–6.60 (m, 8H), 4.93 (s, 1H), 1.74 (s, 3H), 1.36 (s, 6H).

Bisphenol A monoacetate 46b: Bisphenol A (Aldrich, 3.003 g, 13.15 mmol, 2 eq) was dissolved in 10 ml anhydrous ether in a 50 ml snrb flask. Dry pyridine (1.020 ml, 998 mg, 12.61 mmol, 1.9 eq) was added to the solution; acetic anhydride (1.24 ml, 1.34 g, 13.1 mmol, 1 eq) was then added dropwise, and the resulting viscous

mixture was allowed to stand overnight. This was combined with a similar mixture prepared from Bisphenol A (3.00 g) and acetyl chloride (1.24 ml, 1.37 g, 17.4 mmol). The mixture was partitioned between ether and dilute HCl; the ether phase was extracted once with dilute HCl, thrice with sat. $NaHCO_3$, dried over anhydrous $MgSO_4$, and purified by flash chromatography (1:9 Et_2O/CH_2Cl_2, 31 × 4). The mixed high- and medium-R_f fractions were chromatographed again (CH_2Cl_2, same column); total yield of medium-R_f fraction: 3.084 g (41%). ^{1}H NMR ($CDCl_3$): 7.21 (m, 2H), 7.08 (m, 2H), 6.79 (m, 2H), 6.71 (m, 2H), 4.85 (s, 1H), 2.30 (s, 3H), 1.65 (s, 6H).

Bisphenol A mono-(*t*-butyldimethylsilyl) ether 46c: Bisphenol A (Aldrich, 5.006 g, 21.93 mmol, 2 eq) was placed in a 100-ml snrb flask and dissolved in 100 ml DMF (from the bottle). *t*-Butyldimethylchlorosilane (Aldrich, 3.323 g, 22.04 mmol, 1.00 eq) was added, and stirred to dissolve. Addition of triethylamine (Fisher, 4.0 ml, 2.9 g, 28 mmol, 1.3 eq) caused immediate precipitation of white solid. After standing overnight, the solvent was removed under reduced pressure. The residue was partitioned between ether and water; the ether phase was extracted 4× with water, once each with 2M HCl and brine, and dried over anhydrous $MgSO_4$. The product was purified by flash chromatography (1:9 Et_2O/CH_2Cl_2, 30 × 3), and the mixed high- and medium-R_f fractions were separated by flash chromatography again (CH_2Cl_2, 22 × 3.5). Yield: 3.109 g (41%). ^{1}H NMR ($CDCl_3$): 7.05 (m, 4H), 6.71 (m, 4H), 4.69 (s, 1H), 1.64 (s, 6H), 1.00 (s, 9H), 0.20 (s, 6H).

Preparation of **methylene-linked di-pivaloyl Bisphenol A dimer 47a:** Bisphenol A mono-pivaloyl ester **46a** (1.001 g, 3.204 mmol, 1 eq), cesium carbonate (Aldrich, 2.544 g, 7.808 mmol, 2.43 eq), and 3 ml dry DMF were placed in an oven-dried 50 ml 3-neck rb flask. This mixture was stirred under dry nitrogen and heated in an oil bath to 50 °C. Diiodomethane (Aldrich, 424 mg, 1.58 mmol, 1 eq) was weighed out in an oven-dried flask, dissolved in 4 ml dry DMF, and

transferred to a syringe. This solution was added to the reaction mixture by a syringe pump over the course of 30 min. After 18 h, the mixture was allowed to cool, and was triturated with ether. A white solid precipitated, which was removed by suction filtration through a medium glass frit. The residue was rinsed with ether, and the solvent was removed from the combined filtrates by rotary evaporation, yielding a pinkish paste. This paste was partitioned between ether and water; the ether phase was extracted twice with 1 M K_2CO_3, dried over anhydrous $MgSO_4$, and purified by flash chromatography (CH_2Cl_2, 10 × 4). This product was further purified by Chromatotron (PE + EtOAc, 2 mm plate). The fast fraction from this chromatography appeared to be a mixture of desired product and Bisphenol A di-pivaloyl ester. This fraction was again purified by Chromatotron (PE + EtOAc, 1 mm plate), yielding 35.86 mg (1.8%) desired product. NMR ($CDCl_3$): 7.20 (m, 4H), 7.13 (m, 4H), 6.99 (m, 4H), 6.94 (m, 4H), 5.69 (s, 2H), 1.64 (s, 4H), 1.34 (s, 12H).

Preparation of **methylene-bis-Bisphenol A 48:** Methylene-linked di-pivaloyl dimer **47a** (13.69 mg, 0.0215 mmol, 2 equiv) was placed in an oven-dried 10-ml snrb flask, and dissolved in 300 μl dry toluene. *n*-Butyllithium (Aldrich, 1.6 M in hexanes, 220 μl, 0.352 mmol, 16 equiv) was added, and the reaction was stirred at room temperature until starting material was gone. The reaction was quenched by addition of 1 M NH_4Cl solution, and partitioned betwen ether and water. The ether phase was dried over anhydrous $MgSO_4$, and the solvent was removed by rotary evaporation. The residue was purified by flash chromatography (1:19 Et_2O/CH_2Cl_2, 17 × 1). Product was not weighed. ^{1}H NMR ($CDCl_3$): 7.12 (m, 4H), 7.06 (m, 4H), 6.97 (m, 4H), 6.70 (m, 4H), 5.67 (s, 2H), 1.60 (s, 12H).

Attempted preparation of diacetate dimer 47b: All glassware was oven-dried prior to use. Bisphenol A monoacetate **46b** (3.084 g, 11.41 mmol, 1 eq) and cesium carbonate (Aldrich, 7.434 g, 22.83 mmol, 2.00 eq) were placed in a 250-ml

snrb flask. 30 ml dry DMF was added, and the mixture was heated to 80 °C in an oil bath. The mixture quickly darkened. A solution of iodomethane (Aldrich, 1.527g, 5.70 mmol, 1 eq) in 30 ml DMF was added to the heated reaction mixture by syringe pump over the course of 2 h, and the reaction mixture was stirred with heating for an additional 4 h. The mixture was filtered by suction, and DMF was removed from the filtrate at reduced pressure. TLC (1:9 Et_2O/CH_2Cl_2) indicated that extensive deprotection of the starting material had occurred.

Attempted preparation of bis (TBS) dimer 47c: This reaction was carried out in the same way as the attempted coupling of the monoacetate **46b**. The monoether **46c** (1.001g, 2.922 mmol, 1.03 eq) and cesium carbonate (Aldrich, 1.940 mg, 5.95 mmol, 2.1 eq) were placed in an oven-dried 50-ml 3-neck rb flask with 3 ml dry DMF. This mixture was heated to 50 °C, and a solution of diiodomethane (Aldrich, 380 mg, 1.42 mmol, 1 eq) in 4 ml dry DMF was added to the reaction mixture by syringe pump over the course of 4 h. Stirring with heating was continued overnight; the reaction mixture was then filtered, and solvent was removed from the filtrate under reduced pressure. TLC (CH_2Cl_2) indicated that extensive deprotection of the starting material had occurred.

Attempted preparation of Bisphenol A mono-pivaloyl ester iodomethyl ether 49a: Diiodomethane (Aldrich, 566 mg, 2.11 mmol, 6.5 eq) was weighed out in an oven-dried 25 ml snrb flask. Bisphenol A mono-pivaloyl ester **46a** (101 mg, 0.323 mmol, 1eq), cesium carbonate (Aldrich, 322 mg, 0.988 mmol, 3.05 eq), and 2.00 ml dry DMF were added, and the mixture was stirred magnetically under nitrogen at 80 °C for 21 h. Solvent was removed under reduced pressure, and the residue was partitioned between dichloromethane and water. The organic phase was extracted with 1 M K_2CO_3 and brine, and dried over anhydrous $MgSO_4$. High-R_f fractions were removed from starting materials by flash chromatography

(CH_2Cl_2, 15 × 2), and the fast fractions were separated from each other by Chromatotron (PE + EtOAc, 1 mm plate). The only products identified were Bisphenol A dipivaloyl ester and the methylene-linked di-pivaloyl dimer **47a**.

Preparation of **[2.1.2.1]-paracyclophane 50:** 382 mg 2,18-dithia [1.3.1.3]-paracyclophane **55** (382 mg, 0.8444 mmol) and trimethyl phosphite (Aldrich, 250 ml) (Caution: stench!) were placed in a 250-ml quartz immersion well, and irradiated by a Vycor-filtered 450 W Hanovia lamp. After 5 h, the suspended solids had dissolved, so the photolysis was halted. The phosphite was removed under reduced pressure, and the residue was partitioned between water and dichloromethane. The organic phase was extracted twice with H_2O, once each with 0.5 M NaOH and brine, and dried over anhydrous $MgSO_4$. Solvent was removed by rotary evaporation, and the resulting yellowish oil was purified by flash chromatography (1:19 EtOAc/PE, 4 × 8.5). The fastest fractions were further purified by Chromatotron (PE + EtOAc, 1 mm plate) to yield the cyclophane (123.73 mg, 37.7%). ^{1}H NMR (CD_3CN): 6.84 (d, J = 9 Hz, 8H), 6.78 (d, J = 9 Hz, 8H), 3.71 (s, 4H), 2.91 (s, 8H).

Dioxa cyclophane 51: Bisphenol A (Aldrich, 1.000 g, 4.380 mmol, 1 eq) and 1,1′-methylenebis(4-bromomethylbenzene) **53** (1.552 g, 4.383 mmol, 1.001 eq) were dissolved in 40 ml dry DMF, and transferred to a 60-ml disposable syringe. A 1000-ml 3-neck rb flask was charged with cesium carbonate (Aldrich, 7.102 g, 21.80 mmol, 2.49 eq) and 460 ml dry DMF. This was heated to 80 °C with stirring under dry nitrogen. The solution of diol and dibromide was added to the heated suspension of base by a syringe pump over the course of three days. At the completion of the addition, the solids were removed by suction filtration, and the DMF was removed under reduced pressure. The product was partitioned between ether and water; the ether solution was dried over anhydrous $MgSO_4$, and the resulting pinkish-brown solid was purified by flash chromatography (dichloromethane + 1:9 ether/dichloromethane, 8 × 2). The high-R_f fractions were recrystallized twice from toluene to

yield 147 mg (8.0%) white crystals. NMR (CD_3CN): 7.1538 (d, 8H), 6.86 (d, 4H, J = 9 Hz), 6.57 (d, 4H, J = 9 Hz), 5.01 (s, 4H), 3.72 (s, 2H), 1.59 (s, 6H). HRMS: calculated for $C_{30}H_{28}O_2$ = 420.2089304; measured: 420.2103.

2,-18-dithia[2.1.2.1]paracyclophane 55: A 5-l 3-neck flask was charged with a large magnetic stir bar, potassium hydroxide (1.468 g, 26.26 mmol, 1.53 eq) and 1.5 l denatured ethanol. The solution was sparged with dry nitrogen for 2 h. A 500 ml solution of 1:1 ethanol/benzene was also sparged with nitrogen for 2 h. 1,1-methylenebis(4-benzylmercaptan) **54** (2.200 g, 8.514 mmol, 1 eq) was dissolved in 250 ml ethanol/benzene and placed in a pressure-equalized addition funnel equipped with a needle valve control; 1,1-methylenebis(4-bromomethylbenzene) **53** (3.014 g, 8.512 mmol, 1 eq) was dissolved in 250 ml THF freshly distilled from benzophenone ketyl and placed in another pressure-equalized addition funnel with a needle valve control. The reaction flask was stirred under dry nitrogen, and the two funnels were carefully (and often!) adjusted to deliver the same flow. A white precipitate immediately formed when the solutions met. The addition of mercaptan and bromide solutions took about 28 h, and the reaction mixture was stirred for an additional 16 h. The reaction mixture was filtered through diatomaceous earth, no desired product was detected in the filtrate. The literature preparation for this compound neglected to mention that it precipitates from solution. The product was recovered from the diatomaceous earth by digestion with dichloromethane and toluene. Yield: 0.6 g (15%).