Backbone Flexibility in Protein Design Theory and Experiment

Thesis by

Alyce Su

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1998

(Submitted May 18, 1998)

Acknowledgment

I am greatly indebted to many people for making my stay at Caltech fun and fruitful.

I would like to thank Professor Henry Lester, Professor Sela Mager, and Professor Steve Mayo, for seeing me through my first scientific project. I thank Henry for his wonderful ideas and unflagging promotional effort of our paper, Sela for his shrewd judgement, and Steve for his unconditional trust. Without them, there would be no beginning of my scientific career.

As a physicist wanting to tackle complex biology problems, I thank the following professors for their genuine advice and encouragement. I thank Professor Michael Cross, Scott Fraser, David Anderson, Alex Varshasky, Carl Parker, Howard Lipshitz, for discussing research opportunities in Regulatory Gene Networks Pattern Formation. I also thank CNS faculty members, Professor Carver Mead, John Hopfield, Yaser Abu-Mostafa, Peitro Perona, Demitri Psaltis, and Christof Koch, for discussing research opportunities in Computation and Neural Systems.

As a young scientist wanting to pursue a career in science, I was fortunate enough to be enlightened by some of the most influential mentors in the field. I thank Nobel Laureate Rudolph Marcus for believing in me, MIT Professor Carl Pabo for late night discussions at the Athenaeum and support ever since, Professor William Goddard for being the first professor to adopt me in his group, and for his graciousness to us Chinese students, Professor Tom Tombrello for his insightful questions during my candidacy exam, Professor Steve Frautschi for always being there when I needed his guidance, and Professor Henry Lester, profusely, for generously introducing young scientists to the scientific community.

As a member of the Mayo Lab, I thank my advisor Professor Steve Mayo and all my labmates for their patience and help. If it weren't for my advisor's openmindedness, I would have been booted out of the lab a long time ago, and would have never have had the chance to learn any of the biological laboratory techniques I know now. Thank you Steve!!! I thank Bassil Dahiyat for his help throughout the years. Without Bassil, I would never have had the ability to pursue a Ph.D. degree. I thank Jay Luo for his insightful advice and technical knowledge. As the 3rd graduate student to join the Mayo Lab (after Bassil and Jay), I was extremely honored to be able to work side by side with Bassil and Jay, two of the most outstanding young scientists I have ever met. I have learned so much from being in the same room with them. I can never thank Sandy Malakauscas enough for teaching me molecular biology. Without her help, there would be no first systematic backbone flexibility paper in the field. I thank Scott Ross for teaching me how to do NMR, solving the structure, and measuring the relaxation dynamics of a designed mutant protein for me. Scott's constant encouragement and unconditional support have been an extremely valuable gift. I thank Cathy Sarisky for teaching me how to do process NMR data and solving my protein mutant structure for me - after having joined the lab only three months ago. I thank Monica Smith for helping me purifying mg after mg of different mutant proteins. I thank Chantal Morgan for instructing me in the proper use of laboratory equipment, fixing all the equipment I broke, and hosting all the fun parties I went to at Caltech. I thank Ben Gordon and Arthur Street for answering all my computer and programming questions. I thank Dr. Barry Olafson, founder and President of Molecular Engineering Corporation, and Professor Fred Lee, creater of the influential molecular simulation software - POLARIS, for teaching me all about molecular simulations. I thank Professor

Elaine Marzluff for teaching me how to play softball. I thank Dr. Dirk Bokenkamp for teaching me about German beers. I thank Dr. Marie Ary for major rewriting of my thesis. Without her effort, there would be no thesis to submit, and therefore no Ph.D. for me. Thanks Marie!!!

As a Taiwanese immigrant wanting to stay in the U. S., I thank once again all the following recommendation letter writers, for giving me extremely strong and generous evaluations. Among them, Congressman Mr. Randy "Duke" Cunningham, Nobel Laureate Rudolph Marcus, Dr. Barry Olafson, founder and President of Molecular Engineering Corporation, Dr. Newburgh, Executive Officer of the Protein Society, Professor Steve Frautschi, Professor Henry Lester, Professor Sela Mager, Professor Steve Mayo, Dr. Scott Ross, Professor Fred Lee, Professor John Desjarlais, Professor Shin Nan Yang, and Professor Pauchy Hwang. I also thank Attornies Adam Green and Paul Herzog for assisting me with legal matters.

As a female scientist wanting to break into the male-dominant science world, I thank all these male scientists for lending me a helping hand and setting me a benchmark. I thank Dr. Michael Stowell, Dr. Bassil Dahiyat, Professor Charles Musgrave, Dr. Dan Minor, Professor John Desjarlais, Dr. Wyeth Bair, Erik Winfree, Michael Levine, Renny Feldman and Dr. Art Chirino. I especially want to thank Michael Stowell for his advice and support at the most difficult times in my scientific careeer. He has enlightened me in numerous ways and showed me what it takes to succeed in science.

As a Caltech student, I thank all these people for making my Caltech life fun. I thank the gorgeous Dr. Jen Sun and Lavonne Martin for taking me out of the lab, encouraging me to work out in the gym. I thank my ex-roommates Dr. David Hogg and Dr. Salem Fahem for stimulating non-

scientific intellectual discussions. I thank Roian Egnor, Amy Greenwood, Hannah Dvorak, Kate Macleod, Mike Wehr, Dr. Brian Sullivan, Bobby Williams, and Keith Brown, for welcoming a physicist to participate in the Biology Pizza Class. I thank Ben Ramieraz for all the dinner discussions on unconventional ideas in Science and hosting the BI social hour. I thank Professor Buster Bohen for inviting me to the Thanksgiving Dinner and African Music Concert in Santa Monica. I thank Dr. Tobi Delbrok, Dr. Shih-shih Liu, Dr. Rahual Shakespear, for social companionship in my earlier years at Caltech. I thank Dr. Chris Diorio and Eric Bax for helping me with designing VLSI circuits and neural networks. I also thank Dan Fain for taking me to Rave Parties in Hollywood.

As a job seeker fresh out of school, I thank the Caltech Career Development Center for assistance. In particular Counselor Amy Seidel Malak, for her counseling sessions and provision of extra opportunities. I thank recruiters from Merck, Massachussettes General Hospital, Schering Plough, Bristol Myers Squibb, GeneLabs, Mitchell Madison Group, McKinsey, Boston Gonsulting Group, Oliver Wyman & Company, Morgan Stanley, Salomon Smith Barney, Long Term Capital Management, D. E. Shaw, First Quadrant, Pacific Investment Management Company (PIMCO), Group One Trading, Arthur Andersen, Andersen Consulting, KPMG, and Anubis, for taking their time to discuss career opportunities with me, and in certain cases, granting job offers. In particular, Dr. Sid Valluri from McKinsey, who helped in several tangible ways. I also thank Dr. Stephen-wei Chung from Morgan Stanley and Dr. Tom Lee from Symyx, for giving me invaluable job search tips.

As someone who's interested in management consulting, I thank all the 27 participants of the Caltech Case Practice Group I founded. Especially I

wish to thank Jin, Tomislav, Polly, Brian, Russina, Dave, Johan, and Hannah, for all the fun times we shared during and after our practice sessions.

As a resident of Los Angeles, the City of Entertainment and Law, I thank all these people for social opportunities and teaching me about the Jewish culture. I thank Lee Weinberg, J.D. for initiating my search for fun outside of Caltech, Charles Hymowitz for providing me a better understanding of SPICE and funny email jokes, Bruce Singman, J.D. for giving me opportunities for fine dining in Pacific Palisades and Trident videotapes, actor Sam Cohen for introducing the Harvard Alumni Club in Hollywood, Charlie Cohen for movie premieres, post premiere parties and fine dining in Beverly Hills, and possibly an opportunity to dine with my hero Jerry Springer if Jerry does the upcoming MGM production!!!

I thank Paul Bloom, J.D., for major rewriting of all my job application documents, cooking me dinners, taking care of our kitties - Floyd and Bob!!!, finding creative (yet totally legal) ways to spend money and have a good time, teaching me about the American culture and "common sense(!!!)", educating me about classical films and stars, drafting my greencard application package, and picking me up after I got lost after my McKinsey interviews in downtown Los Angeles. I hope one day Paul will direct my script with Floyd and Bob as cast!

Finally, I want to thank everyone in my family. I thank Dad for his creativity genes, Mom for being my life-long role model, my sister Brenda for her sense of humor, and my brother Charles for his kindness.

Backbone Flexibility in Protein Design - Theory and Experiment

Abstract

The role of backbone flexibility in protein design was studied. First, the effect of explicit backbone motion on the selection of amino acids in protein design was assessed in the core of the streptococcal protein G β1 domain (Gβ1). Concerted backbone motion was introduced by varying Gβ1's supersecondary structure parameter values. The stability and structural flexibility of seven of the redesigned proteins were determined experimentally. Core variants containing as many as six of ten possible mutations retained native-like properties. This result demonstrates that backbone flexibility can be combined with amino acid side-chain selection and that the selection algorithm is sufficiently robust to tolerate perturbations as large as 15% of the native parameter values.

Second, a general, quantitative design method for computing de novo backbone templates was developed. The method had to compute atomic resolution backbones compatible with the atomistic sequence selection algorithm we were using and it had to be applicable to all protein motifs. We again developed a method that uses super-secondary structure parameters to determine the orientation among secondary structural elements, given a target protein fold. Possible backbone arrangements were screened using a cost function which evaluates core packing, hydrogen bonding, loop closure, and backbone torsional geometry. Given a specified number of residues for each secondary structral element, a family of optimal configurations was found. We chose three motifs to test our method (ββα, βαβ, and αα) since their combination could be used to approximate most possible backbone fold. The best structure found for the ββα motif is similar to a zinc finger, and the

best structure for the $\beta\beta\alpha$ motif is similar to a segment of a β-barrel. The backbone obtained for the $\alpha\alpha$ motif resembles minimized protein A.

Last, our backbone design method was evaluated by testing the thermal stability and structural properties of the designed peptides using circular dichroism and 1D nuclear magnetic resonance. From these results, a set of heuristic rules was derived. Taken together, these studies suggest that de novo backbones assembled using our backbone design method may serve as adequate input templates for atomistic sequence selection algorithms.

Table of Contents

Chapter 1

Introduction

Introduction

This thesis contained two independent biophysical projects. First, the role of backbone flexibility in protein design was assessed. Second, a multi-substrate single-file model for ion-coupled transporters was developed. Although the focal biological objects were different, the quantitative physical approach remained the same. In the introductory chapter, I describe background information for the two projects.

I. Backbone Flexibility in Protein Design

In this section, we introduce the following concepts:

(a) proteins as physical systems;

(b) the protein design problem;

(c) approaches to solving the protein design problem;

(d) the role of backbone flexibility in protein design;

(e) incorporating backbone flexibility in protein design.

a) Proteins as Physical Systems

As with any other physical system, we characterize the protein's components, degrees of freedom, and the interacting forces (Creighton, 1993).

Mathematically, a protein is a sequence of amino acids. Just as quarks come in three different flavors, amino acids come in 20 different types, and are abbreviated as A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y (the 26 letters of the alphabet minus B, J, O, U, X, Z, see Table 1). These 20 amino acids are made of N, C, O, H and S atoms, and have different physical properties. They serve as building blocks for proteins and are connected together by peptide bonds. A protein's sequence length can range anywhere

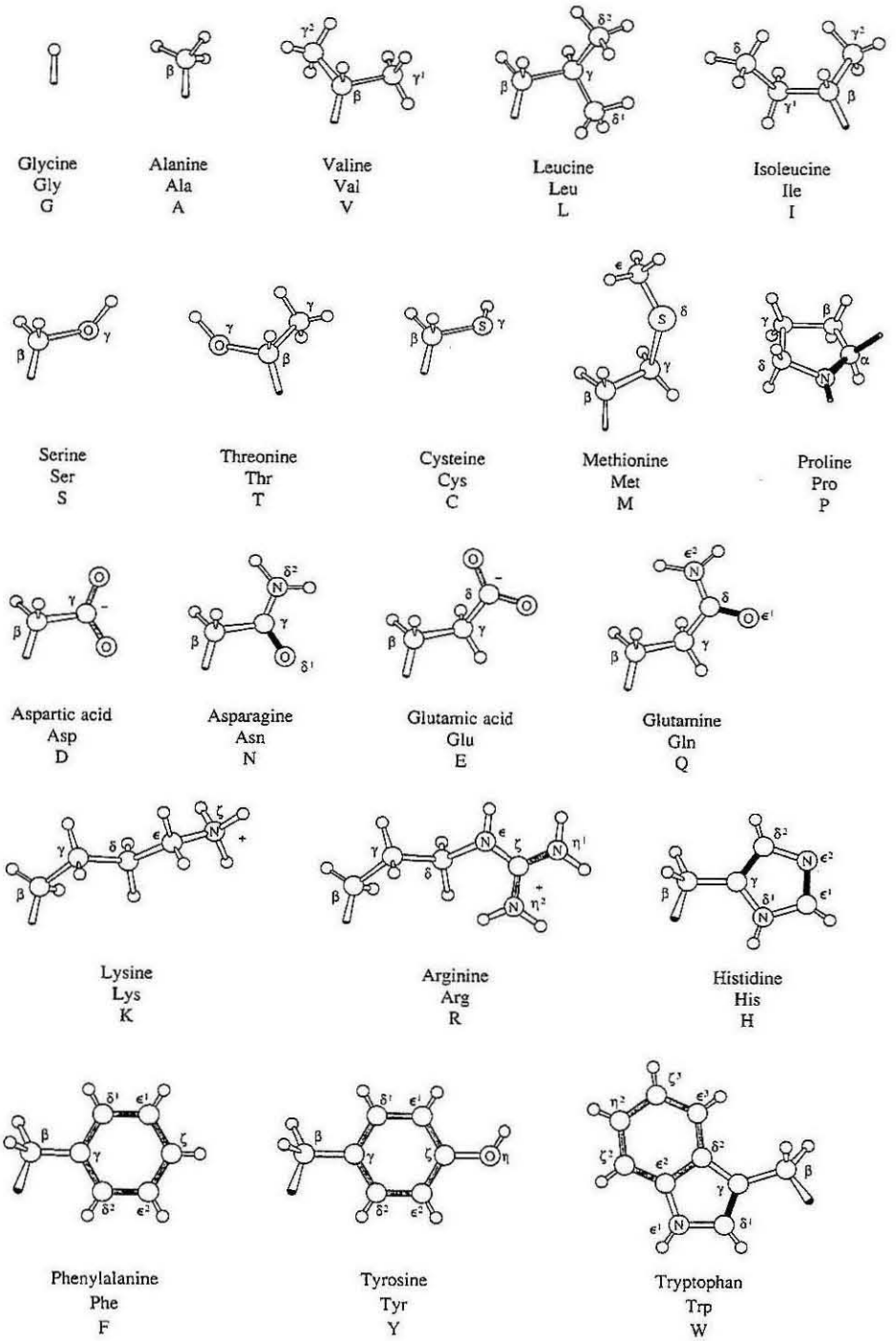


Table 1 Just as quarks come in three different flavors, amino acids come in 20 different types (reprinted from Creighton, Proteins, 2nd edition, 1993).

from 20 to 1000 or more. However, a protein of shorter sequence length, say less than 50, is usually called a "peptide."

Geometrically, a protein is a linear chain of polymer. What this means is that a protein looks like a fishbone, with many side-chains branching off a main-chain. The main-chain is also called the "backbone," consisting of repetitive peptide units $(N\text{-}C\alpha\text{-}C)_n$. The side-chains can take on identities of any of the 20 amino acids. The ordering of a protein's side-chain identities coincides with its sequence (see Figure 1). The protein's sequence is called its "primary structure."

Figure 1 also exhibits the angles completely spanning the protein's degrees of freedom. These degrees of freedom are partitioned between the protein's main-chain (backbone) and side-chain. Main-chain angles include ϕ and ψ. ϕ is the dihedral angle between C-N-Cα-C; ψ is the dihedral angle between N-Cα-C-N. Side-chain χ-n (n = 1, 2, 3, 4) angles are counted along N-Cα-Cβ-Y, and down the side-chain branch, where Y is any non-hydrogen atom on the side-chain. For example, the D in Figure 1 has χ-1 representing the dihedral angle between N-Cα-Cβ-Cγ and χ-2 representing the dihedral angle between Cα-Cβ-Cγ-Oδ. A protein's structure is fully specified once the values of all the ϕ, ψ, and χ angles are specified.

The protein's main-chain (backbone) forms three classes of distinctive patterns, called the "secondary structures." The first class, called the "α-helix", looks like a spiraling spring (see Figure 2a). The second class, called the "β-strand," looks like a series of mountains (see Figure 2b). Both "α-helix" and "β-strand" are periodic, regular structures. The third class, called the "loop", is irregular and is mainly responsible for connecting the regular structures (Figure 2c).

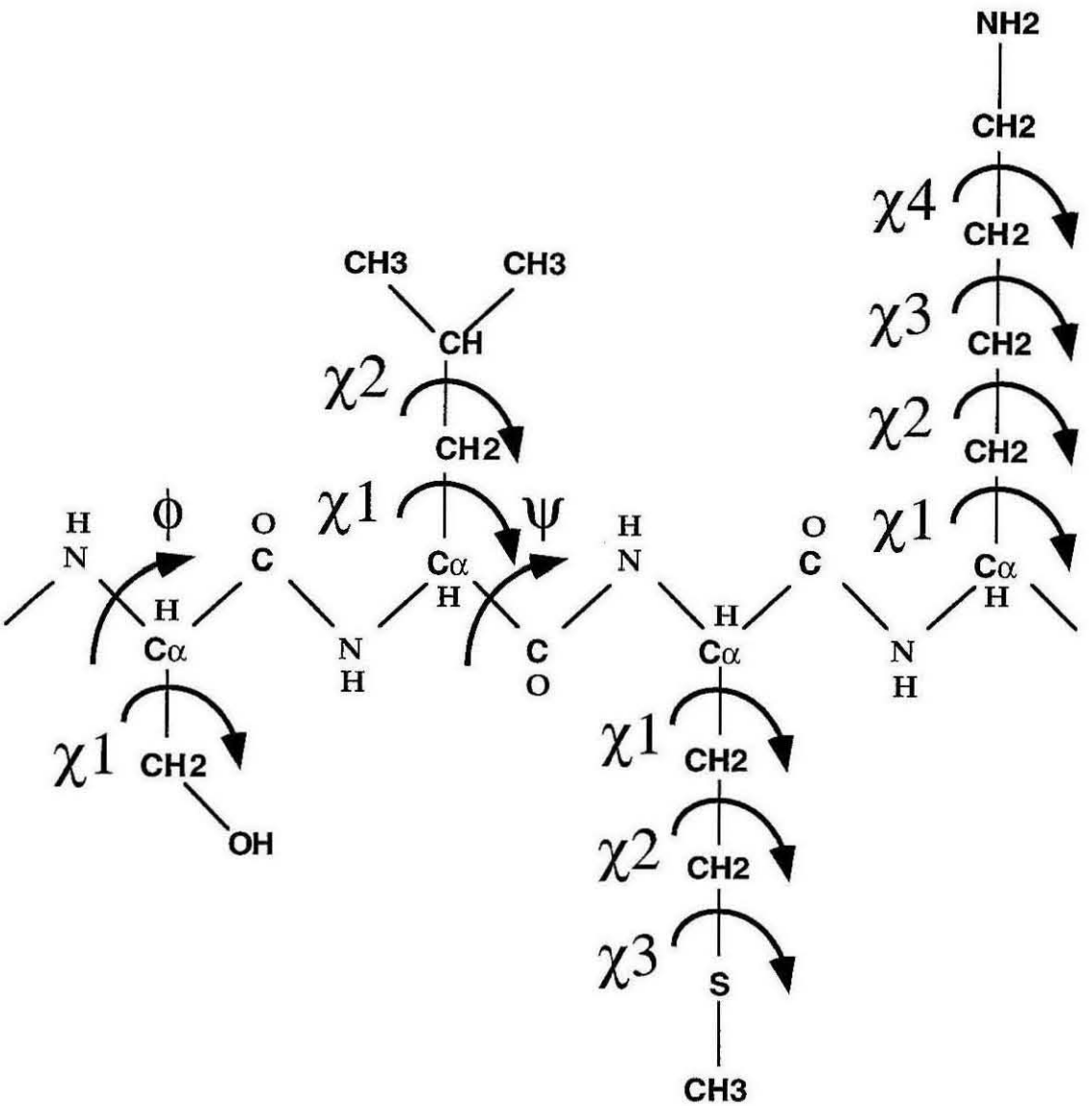
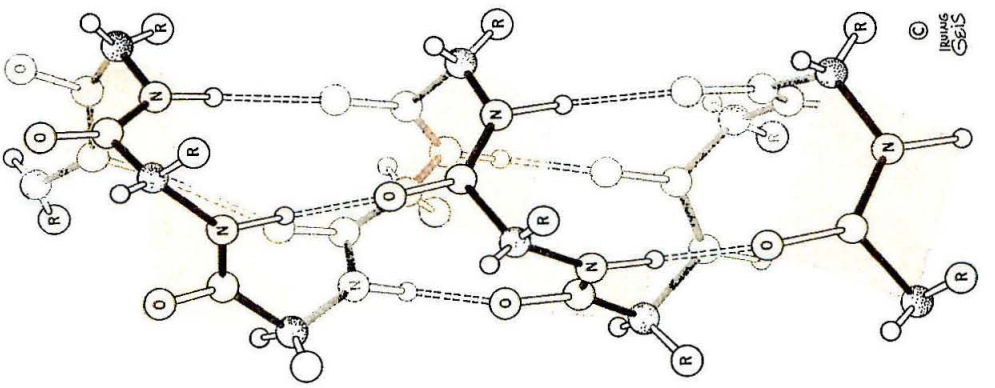
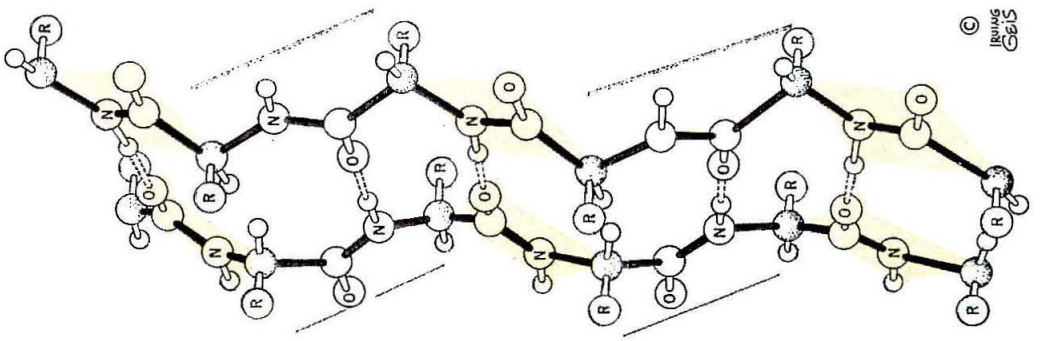


Figure 1 Primary sequence = "SLMK". A protein's structure is fully specified once the values of all the ϕ, ψ, and χ angles are specified.

(a) α-helix



(b) β-sheet



(c) loop

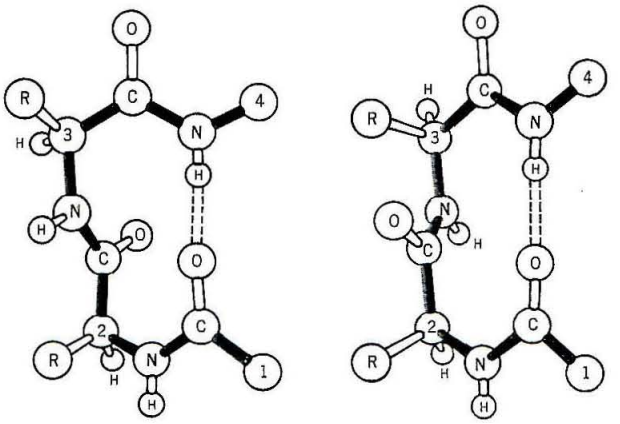


Figure 2 The protein's main-chain (backbone) forms three classes of distinctive patterns, called the "secondary structures." Figure 2a shows the α-helix. Figure 2b shows the β-sheet. Figure 2c shows the irregular loop (reprinted from Cantor and Schimmel, Biophysical Chemistry, 1980).

The protein's side-chain, for a given amino acid, can adopt many different conformations by rotating around its χ-angles. In principle, the rotations can take on any value. But in reality, only a smaller set of discretized values has been observed with statistically higher frequencies around each χ-angle. These preferred rotational states are called "rotamers." As an example, Figure 3 and Table 2 list the more commonly observed rotamers as described by Ponder & Richards in 1987 (Ponder & Richards, 1987).

There are five major physical chemical phenomenon responsible for determining a protein's final structure, also called the protein's "tertiary structure." These factors are (i) the van der Waals potential, (ii) the intrinsic properties of each amino acid, (iii) the hydrogen bonding potential, (iv) the electrostatic potential, and (v) the solvation effect approximated by the protein's surface area. Although these interactions were thought to be better understood than some of the more "fundamental" interactions, say the strong or weak interactions in elementary physics, the exact balance between these interactions becomes more exquisite as proteins are complex systems consisting of thousands of atoms.

Among these interactions, the solvation effect is believed to be the most dominant. The 20 amino acids can be grouped into either "water-loving (hydrophilic)," or "water-hating (hydrophobic)." Once in solvent, the protein wants to bury as many hydrophobic amino acids (forming a hydrophobic core) and expose as many hydrophilic amino acids as possible (forming a hydrophilic surface).

b) The Protein Design Problem

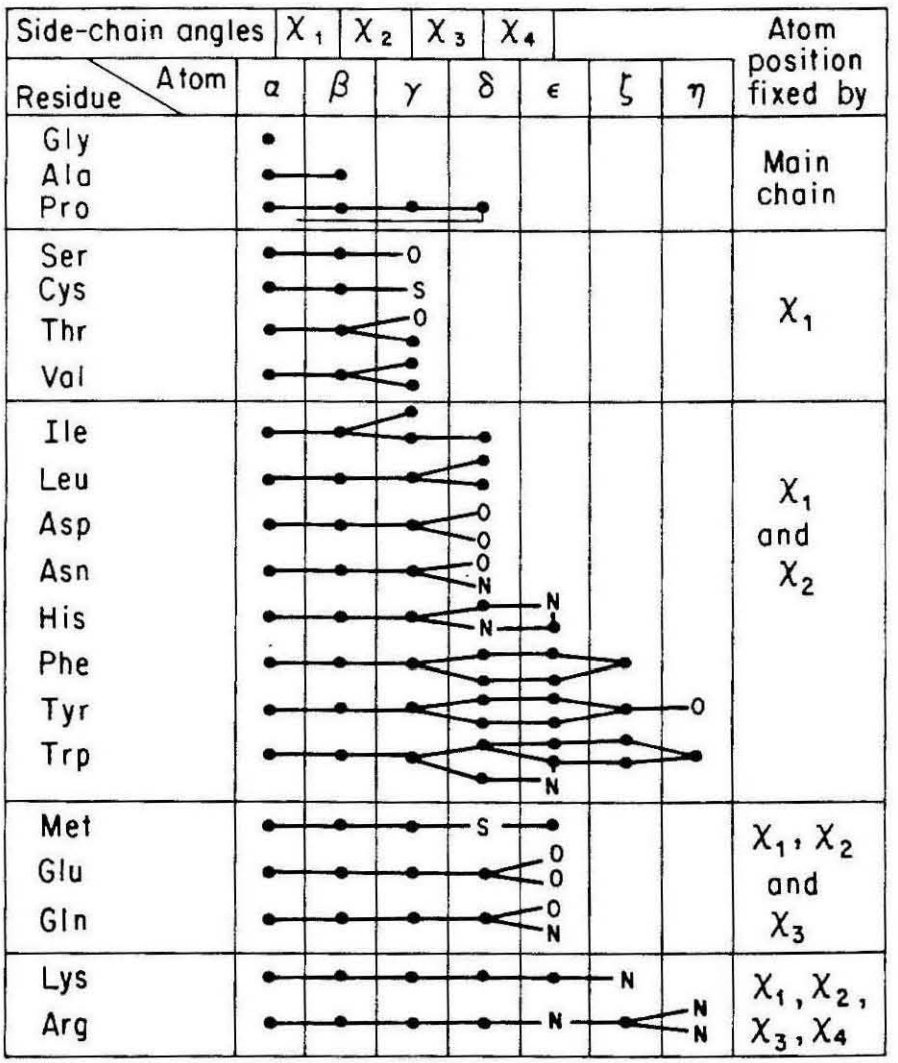


Figure 3 Flexibility of amino acid side-chains. The Figure shows the χ angle values required to fixed the positions of side-chain atoms in each amino acid residue type (reprinted from Ponder & Richards, 1987).

Rotamer	Number	%	Chi 1	Chi 2	Chi 3
Valine					
t	100	67·1	173.5 (9·0)		
−	39	26·2	−63·4 (8·1)		
+	8	5·4	69·3 (9·6)		
Other	2	1·3			
Leucine					
− t	94	63·9	−64·9 (8·2)	176·0 (9·9)	
t +	36	24·5	−176·4 (10·2)	63·1 (8·2)	
t t	7	4·8	−165·3 (10·0)	168·2 (34·2)	
+ +	3	2·0	44·3 (20·0)	60·4 (18·8)	
Other	7	4·8			
Isoleucine					
− t	42	45·2	−60·9 (7·5)	168·7 (11·6)	
− −	17	18·3	−59·6 (9·6)	−64·1 (14·3)	
+ t	15	16·1	61·7 (5·0)	163·8 (16·4)	
t t	12	12·9	−166·6 (10·1)	166·0 (8·9)	
t +	3	3·2	−174·8 (24·9)	72·1 (10·5)	
Other	4	4·3			
Serine					
+	94	48·0	64·7 (16·1)		
−	56	28·6	−69·7 (14·6)		
t	46	23·5	−176·1 (20·2)		
Threonine					
+	81	47·9	62·7 (8·5)		
−	76	45·0	−59·7 (9·4)		
t	8	4·7	−169·5 (6·6)		
Other	4	2·4			
Cysteine					
−	57	60·6	−65·2 (10·1)		
t	23	24·5	−179·6 (9·5)		
+	13	13·8	63·5 (9·6)		
Other	1	1·1			
Proline†					
+	37	39·4	26·9 (7·8)	−29·4 (14·4)	
−	32	34·0	−21·8 (6·4)	31·2 (8·5)	
0	22	23·4	0·3 (6·4)	−0·8 (11·5)	
Other	3	3·2			
Phenylalanine					
−90	37	46·3	−66·3 (10·2)	94·3 (19·5)	
t 90	20	25·0	−179·2 (9·3)	78·9 (8·9)	
+90	17	21·3	66·0 (12·0)	90·7 (9·4)	
− 0	5	6·3	−71·9 (16·3)	−0·4 (26·1)	
Other	1	1·3			
Tyrosine					
−90	52	48·6	−66·5 (11·4)	96·6 (21·8)	
t 90	35	32·7	−179·7 (12·6)	71·9 (13·4)	
+90	16	15·0	63·3 (9·4)	89·1 (13·0)	
− 0	4	3·7	−67·2 (13·2)	−1·0 (20·1)	
Other	2	1·9			
Tryptophan					
− +	11	37·9	−70·4 (7·0)	100·5 (18·2)	
+ −	6	20·7	64·8 (13·0)	−88·9 (5·3)	
t −	4	13·8	−177·3 (7·9)	−95·1 (7·6)	
t +	3	10·3	−179·5 (3·4)	87·5 (3·8)	
− −	2	6·9	−73·3 (6·5)	−87·7 (8·1)	
+ +	1	3·4	62·2	112.5	
Other	2	6·9			
Histidine					
− −	15	34·1	−62·8 (10·0)	−74·3 (17·2)	
t +	11	25·0	−175·2 (15·4)	−87·7 (43·5)	
− +	7	15·9	−69·8 (5·9)	96·1 (32·2)	
+ −	6	13·6	67·9 (17·4)	−80·5 (40·7)	
t −	4	9·1	−177·3 (6·3)	100·5 (14·0)	
+ +	1	2·3	48·0	85·9	
Aspartic acid					
−	51	47·7	−68·3 (9·2)	−25·7 (31·1)	
t	36	33·6	−169·1 (9·5)	3·9 (38·9)	
+	17	15·9	63·7 (9·9)	2·4 (29·4)	
Other	3	2·8			

Table 2 Side-chain angles from the rotamer library χ values and standard deviations (reprinted from Ponder & Richards, 1987).

Rotamer	Number	%	Chi 1	Chi 2	Chi 3
Asparagine					
− −	37	30·3	−68·3 (12·3)	−36·8 (25·2)	
t 0	26	21·3	−177·1 (8·8)	1·3 (34·1)	
− +	16	13·1	−67·2 (10·8)	128·8 (24·2)	
+ 0	14	11·5	63·9 (3·7)	−6·8 (13·5)	
t t	14	11·5	−174·9 (17·9)	−156·8 (58·9)	
+ +	8	6·6	63·6 (6·6)	53·8 (17·1)	
Other	7	5·7			
Glutamic acid					
− t	22	27·2	−69·6 (19·2)	−177·2 (21·7)	−11·4 (44·8)
t t	21	25·9	−176·2 (14·9)	175·4 (10·6)	−6·7 (39·0)
− −	9	11·1	−64·6 (13·5)	−69·1 (17·3)	−33·4 (27·4)
− +	7	8·6	−55·6 (10·6)	77·0 (6·8)	25·3 (32·6)
+ t	7	8·6	69·8 (10·6)	−179·0 (23·7)	6·6 (64·2)
t +	5	6·2	−173·6 (14·6)	70·6 (8·7)	14·0 (37·1)
+ −	4	4·9	63·0 (4·3)	−80·4 (13·9)	16·3 (20·8)
Other	6	7·4			
Glutamine‡					
− t	33	36·7	−66·7 (14·1)	−178·5 (14·9)	—
t t	19	21·1	−174·6 (11·5)	−177·7 (17·1)	—
− − 0	13	14·4	−58·7 (11·2)	−63·8 (16·1)	−46·3 (27·7)
t + 0	8	8·9	−179·4 (21·5)	67·3 (7·9)	26·8 (38·4)
+ t	6	6·7	70·8 (13·0)	−165·6 (9·5)	—
− − t	4	4·4	−51·3 (7·3)	−90·4 (22·8)	165·0 (38·2)
t + t	2	2·2	167·5 (14·8)	70·9 (3·7)	174·2 (7·1)
Other	5	5·6			
Methionine					
− − −	6	37·5	−64·5 (12·7)	−68·5 (6·0)	−75·6 (14·1)
− t	4	25·0	−78·3 (5·4)	−174.7 (15·7)	—
t t	3	18·8	178·9 (8·7)	179·0 (13·4)	—
Other	3	18·8			
Lysine‡					
− t	45	40·9	−68·9 (16·5)	−178·4 (24·7)	—
t t	26	23·6	−172·1 (18·4)	175·3 (23·1)	—
− −	18	16·4	−58·1 (10·5)	−74·9 (23·4)	—
t +	9	8·2	173·4 (9·5)	83·4 (15·5)	—
+ t	4	3·6	71·5 (12·5)	−174·3 (11·1)	—
t −	3	2·7	−175·8 (29·0)	−63·9 (36·5)	—
− +	2	1·8	−104·0 (7·7)	74·6 (35·3)	—
Other	3	2·7			
Arginine‡					
− t	38	46·3	−67·6 (13·3)	176·9 (20·0)	—
t t	19	23·2	−174·1 (17·5)	−178·6 (24·4)	—
+ t	8	9·8	80·0 (20·7)	175·6 (18·0)	—
− −	7	8·5	−67·0 (7·5)	−71·7 (11·8)	—
t +	4	4·9	178·2 (6·7)	69·5 (10·6)	—
+ +	2	2·4	57·1 (2·9)	82·8 (12·7)	—
− +	2	2·4	−76·9 (1·5)	54·2 (20·9)	—

The rotamer abbreviations are as follows: + for a chi value centered in the +60 to +90 range; − for a chi value centered in the −60 to −90 range: t for a value centered near 180; 0 and 90 for chi values centered near 0 and 90, respectively. Note that our definitions of + and − are different from those used by Janin *et al.* (1978), but the same as those of Benedetti *et al.* (1983).

The residues included as "other" had chi values significantly different from any listed library member and represent either extreme outliers of listed library members or rare unlisted library members. In addition, some of the "other" residues may be misinterpreted or poorly defined in the crystallographic structures. All "other" residues and the proteins in which they are found are listed below. The initial 4 letter code is that used by the Protein Data Bank (see Table 1).

1BP2	Val65	1NXB	Val46	2HHB	Leu83	1BP2	Leu118
1NXB	Leu52	2SGA	Leu25	2RHE	Leu55	1MBO	Leu2
2HHB	Leu109	4CYT	Ile57	2ALP	Leu11	5RXN	Ile12
2SGA	Ile72	1NXB	Cys60	1NXB	Thr13	2SGA	Thr74
2SGA	Thr73	3SGB	Thr41	1NXB	Pro44	1NXB	Pro48
2SGA	Pro123	1BP2	Phe5	2RHE	Tyr50	1PPD	Tyr86
1PPD	Trp177	1MBO	Trp14	5PTI	Asp50	2SGA	Asp71
2RHE	Asp97	1SN3	Asn33	1BP2	Asn24	1INS	Asn21
1NXB	Asn61	3SGB	Asn30	1SN3	Asn62	1PPD	Asn84
2RHE	Glu1	1NXB	Glu21	1NXB	Glu56	1SN3	Glu2
2RHE	Glu98	5RXN	Glu54	1NXB	Gln6	2APP	Gln111
1INS	Gln4	4CYT	Gln16	2APP	Gln133	1MBO	Met131
2ALP	Met138	2ALP	Met158	1LZ1	Lys69	1NXB	Lys51
4CYT	Lys72	2SGA	Arg148	1LZ1	Arg115		

† The distribution of Pro residues in our data set with respect to chi 1 does not show distinct clusters corresponding to library members. Thus, we have artificially partitioned the sample to provide the 3 library members listed.

‡ Chi 3 of Gln and chi 3 and 4 of Lys and Arg are poorly determined in most structures, and statistics for many of these angles were not used in the derivation of the rotamer library.

Table 2 (continued) Side-chain angles from the rotamer library χ values and standard deviations (reprinted from Ponder & Richards, 1987).

The protein design problem asks the following question: given a three dimensional target protein structure, how can one find an amino acid sequence, that will adopt the desired structure (Pabo, 1983; Yue & Dill, 1992; Bowie & Eisenberg, 1993).

The statement of the problem may be simple, but it is a hard combinatorial problem in nature. For example, consider a protein sequence of length 100. For any target protein structure, there are 20 x 20 x 20 ... = 20^{100} possible sequences to choose from. If an exhaustive search is performed in sequence space, assuming one can sample a billion sequences a second, it will take 10^{105} times of the age of the universe (13 x 10^9 years) to search through all possible sequences.

(c) Approaches to Solving the Protein Design Problem

People have tried to solve the protein design problem for over a decade. Two major approaches have been used, one empirical, the other quantitative. This thesis work employs the quantitative approach to protein design.

The empirical design rules were developed by the pioneers in the field (Regan & DeGrado, 1988; Hecht et al., 1990; Osterhout et al., 1992; DeGrado et al., 1991; Betz et al., 1993; Handel et al., 1993). These scholars attempted to reduce the complexity level associated with the protein design problem by (i) using as the main criterion for sequence-selection the binary patterning of polar and nonpolar amino acids, and (ii) reducing the amino acid set. We illustrate these rules using a four-helix bundle design as an example.

A four-helix bundle is one of the earliest design target (Regan & DeGrado, 1988; Hecht et al., 1990). So it consists of four helices running parallel to each other, forming a "bundle" (see Figure 4). As shown in Figure

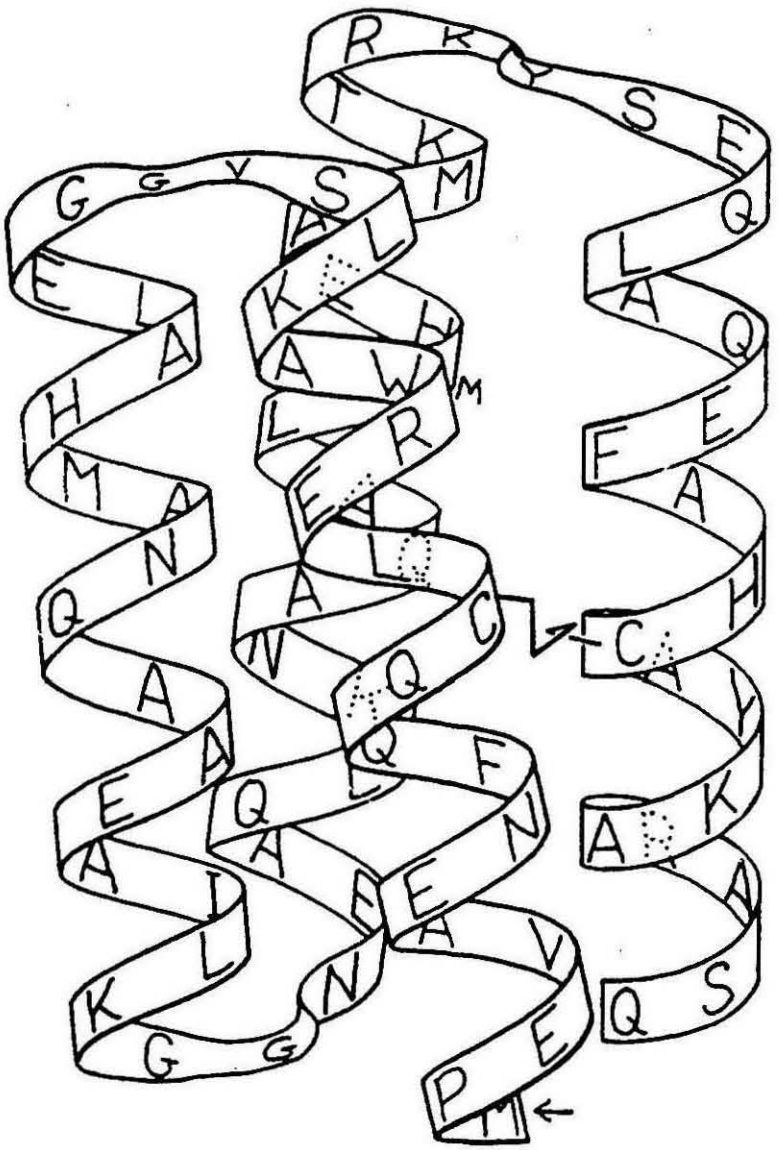


Figure 4 Example of a four-helix bundle. Ribbon drawing of the sequence and proposed three-dimensional structure of Felix, with the disulfide indicated (reprinted from Hecht et al., 1990).

4, there are "interior" and "exterior" faces of the bundle. Considering only the dominant hydrophobic force, the interior and exterior are hypothesized to be formed by nonpolar and polar amino acids, respectively. This is what "binary patterning" means (Hecht et al., 1990). Binary patterning, although crude, is effective in restricting the allowed sequence space. Within the allowed sequence space, the search is further restricted by using a reduced set of amino acids. That is, all the interior hydrophobic positions are assumed to be the amino acid L, all the extrior hydrophilic positions are assumed to be amino acids K or E, and all the loops connecting the helices are assumed to adopt the sequence G-P-R-R-G.

In general, proteins designed using empirical design rules have been shown to have the correct topology, significant secondary structure, and reasonable thermodynamic stabilities (Regan & DeGrado, 1988; Hecht et al., 1990; Osterhout et al., 1992). Even so, their tertiary strucutres appear poorly defined, as demonstrated by several experimental criteria (DeGrado et al., 1991; Betz et al., 1993; Handel et al., 1993). Thus, it is no longer considered sufficient that a given sequence adopt the desired topology; it must also possess the physical properties of well structured natural proteins. This has encouraged the development of more elaborate design strategies to achieve native-like structure and dynamics.

The quantitative design rules were developed after the realization that emperical design rules alone were insufficient. Several groups have begun to develop computational methods to design proteins (Ponder & Richards, 1987; Hellinga et al., 1991; Hurley et al., 1992; Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Harbury et al., 1995; Klemba et al., 1995; Nautiyat et al., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996). At the center of the design approach is the "design cycle," in which theory and experiment alternate.

The starting point is the development of a molecular model, based on rules of protein structure, combined with a computational algorithm for applying these. This is followed by experimental construction and analysis of the properties of the designed protein. If the experimental outcome is failure or partial success, then a next iteration of the design cycle is started in which additional complexity is introduced, rules and parameters are refined, or the algorithms for applying them are modified. The quantitative design approach is therefore a way to test the limits of completeness of understanding experimentally.

Logistically, the quantitative design approach starts with a fixed protein backbone. The backbone is then redecorated with different amino acid sequences that are predicted to be structurally compatible with that fold. This approach is also called the strict "inverse folding" approach (Pabo, 1983) where the backbone conformational degrees of freedom have been removed from the design problem. Key to the strict "inverse folding" approaches are (i) a scoring function based on physical forces to accurately evaluate the compatibility between the sequence and backbone, and (ii) an optimization method that will locate the best-scoring sequence without searching exhaustively. These two daunting tasks have now been overcome by Dahiyat and Mayo (Dahiyat & Mayo, 1997b). In the same publication, they combined a novel forcefield with an optimization algorithm to efficiently search through a sequence space consisting of 1.9×10^{27} sequences, and located a true optimal sequence that successfully adopted the desired target $\beta\beta\alpha$ fold.

(d) The Role of Backbone Flexibility in Protein Design

The role of backbone flexibility in protein design was considered under two contexts: the strict inverse folding approach, and the de novo protein design approach.

In the strict inverse folding approach, the backbone was fixed. However, in several protein core repacking studies, core amino acid sequences optimized on fixed backbones were often very similar to the protein's original sequence (Dahiyat & Mayo, 1996; Dahiyat & Mayo 1997a). This similarity suggested that the fixed backbone can introduce native bias to the designed sequences. On the other hand, it was experimentally observed that the backbone of the target structure could move to alleviate potentially disruptive mutations (Baldwin et al., 1993; Lim et al., 1994). These backbone movements allowed a protein to accomodate amino acids whose total volumes were 10% - 15% bigger than the protein core bounded by the fixed backbone frame. Therefore, sequences that were scored incompatible with the fixed backbone, could turn out to be quite compatible with a relaxed backbone. The role of backbone flexibility in this context is to increase diversity in the designed sequences (Su & Mayo, 1997).

In the de novo protein design approach, a novel optimal backbone has to be constructed before it can be decorated with different amino acid sequences. In order to setup the optimization, the backbone conformational degrees of freedom must be fully characterized. Since the backbone optimization is done in the absense of any sequence information, the scoring function must maximize backbone flexibility (or minimize backbone strain). The role of backbone flexibility in this context is to serve as a criterion for constructing target novel backbones.

(e) Incorporating Backbone Flexibility in Protein Design

After establishing the importance of backbone flexibility in protein design, the challenge now is to incorporate it computationally. There are at least two ways to implement this using a protein's backbone conformational degrees of freedom: one short-range, one long-range. The short-range method relaxes uniformly across every ϕ/ψ angle (Desjarlais, personal communications). The long-range method keeps the ϕ/ψ angles constant across a given secondary structure (an α-helix or a β-strand) but moves the entire secondary structure as a whole (Su & Mayo, 1997). The long-range method is favored over the short-range one for two reasons. First, long-range rather than short-range backbone movements were observed in previous protein design studies (Baldwin et al., 1993; Lim et al., 1994). Second, a protein has fewer long-range degrees of freedom than short-range ones (Chothia et al., 1997; Cohen et al., 1980; Cohen et al., 1981; Chothia & Janin, 1981; Chothia et al., 1981; Chothia & Janin, 1982; Chou et al., 1985; Chou et al., 1986; Cho et al., 1988; Su & Mayo, 1997). As a first order approximation, long-range degrees of freedom are more realistic and computationally efficient.

In chapter 2 of this thesis, we set out to incorporate long-range backbone degrees of freedom into the quantitative protein design approach. The long-range method parsed the protein into a collection of rigid bodies, each rigid body being a piece of secondary structure, such as an α-helix or a β-strand (Chothia et al., 1997; Cohen et al., 1980; Cohen et al., 1981; Chothia & Janin, 1981; Chothia et al., 1981; Chothia & Janin, 1982; Chou et al., 1985; Chou et al., 1986; Cho et al., 1988). Then "supersecondary structure parameters" were used to describe the relative distance and orientation among individual secondary structural elements. Backbone flexibility was introduced by moving the individual secondary structural elements along these distances

and orientations, followed by optimizing the core sequence for each perturbed backbone conformation (Su & Mayo, 1997).

In chapter 3 of this thesis, we sought to expand the range of computational protein design by developing a general, quantitative design method for computing de novo backbone templates. The method had to compute atomic resolution backbones compatible with the atomistic sequence selection algorithm we were using (Dahiyat & Mayo, 1997b) and it had to be applicable to all protein motifs. The algorithm we developed uses supersecondary structure parameters to determine the orientation among secondary structural elements, given a target protein fold. Possible backbone arrangement are screened using a cost function which evaluates core packing (Ponder & Richards, 1987), hydrogen bonding (Ippolito et al., 1991; Sticle et al., 1992; McDonald & Thornton, 1994), loop closure (Bruccoleri, 1993; Donate et al., 1996), and backbone torsional geometry (Salemme, 1983). Given a specified number of residues in each secondary structural element, a family of optimal configurations is found. We chose three motifs to test our method ($\beta\beta\alpha$, $\beta\alpha\beta$, and $\alpha\alpha$) since their combination can be used to approximate most possible backbone fold.

In chapter 4 of this thesis, we evaluate the backbone design method developed in chapter 3 by testing the thermal stability and structural properties of the designed peptides. We also explore relevant issues for integrating computer-generated backbones with the sequence-selection algorithm. Starting with a computer-generated $\beta\beta\alpha$ motif, we optimized five sequences for this backbone and characterized them using circular dichroism and one-dimensional nuclear magnetic resonance. It was found that small differences in the number and location of the hydrophobic residues can significantly change the thermodynamic behaviour of the designed peptides.

This supports the previously acknowledged importance of binary patterning (Hecht, 1990). Based on the results of these five peptides, a set of heuristic rules was derived which could be used to improve computer-generated backbones. Validation of these rules will be the focus of future experimental efforts.

II. A Multi-Substrate Single-File Model for Ion-Coupled Transporters

In this section, we introduce the following concepts:

a) what are ion-coupled transporters;

b) contemporary models for ion-coupled transporters;

c) conflicts between existing models and new data;

d) building the new model using new data;

e) comparing the new model with existing models.

a) What are Ion-Coupled Transporters

Several classes of membrane transport proteins use electrochemical gradients for ions (usually for Na^+ or H^+) to accumulate organic molecules (neurotransmitters, sugars, amino acids, osmolytes) in plant and animal cells (Schultz, 1986; Harvey & Nelson, 1994). The tight flux coupling between these inorganic and organic substrates constitutes a hallmark of ion-coupled transporters and contrasts with properties of ion channels, another major class of membrane transport proteins (Hille, 1992). To explain the mechanism of flux coupling, basically two types of models have been proposed (Hill, 1977; Kanner & Schuldiner, 1987; Rudnick & Clark, 1993). Early models envisioned a recirculating carrier whose motions were largely governed by the binding and dissociation of the substrates (Schultz, 1980; Stein, 1986; Lauger & Jauch, 1986). More recently, sequence analysis of cloned

transporters suggested 6 to 12 putative transmembrane domains, rendering a recirculating carrier less plausible. This is because large membrane proteins don't exhibit such a large scale motion within the membrane.

b) Contemporary Models for Ion-Coupled Transporters

Most contemporary mechanistic concepts of ion-coupled transport employ the alternating-access scheme first enunciated by Jardetzky (1966) and developed in many papers by Lauger (see Lauger, 1991; Wright, 1993; Lester et al., 1994). In this scheme, ion-coupled transporters are viewed as pores or channels that have two gates. While the pore has sites that bind, or perhaps merely accept, all the permeant substrates, the gates have most of the (poorly understood) properties that assure coupled transport. When all the substrates are bound appropriately, the gates undergo conformational changes; and these conformational changes account for the the differences in compartmentalization of the substrates during the transport cycle. Some alternating-access schemes incorporate ordered binding and dissociation of substrates (see for instance Rudnick & Clark, 1993). Now that cloned transporters can be expressed at high densities and studied with good temporal resolution in heterologous expression systems, additional measurements are available on pre-steady state kinetics and charge movements associated with one or a few steps in the transport cycle (Parent et al., 1992a,b; Mager et al., 1993; Mager et al., 1994; Cammack et al., 1994; Wadiche, 1995a). Several studies build on these time-resolved data in the conext of the alternating-access model (Parent et al, 1992b; Mager et al, 1993; Wadiche et al, 1995b).

c) Conflicts Between Existing Models and New Data

However, the newer measurements have also revealed several additional classes of complexities that cannot be explained by straightforward alternating-access models. (1) There are leakage currents -- Na^+ fluxes in the absence of organic susbtrate (Schwarz, 1990; Schwartz & Tachibana, 1990; Umbach et al., 1990; Cammack et al., 1994). (2) There are major departures from accepted stoichiometry, so that transport-associated currents are several times larger than the flux of organic substrate (Mager et al, 1994; Wadiche et al, 1995a). (3) There are actual or inferred quantized current events that exceed by several orders of magnitude the single-charge events expected from the model (Mager et al, 1994; Wadiche et al, 1995b; DeFelice, 1995).

d) Building the New Model Using New Data

Although more complex alternating-access models can be developed to account for some of these new phenomenona, the time seemed ripe for an new class of models. Our formulation is termed the multi-substrate single-file transport model. We borrow heavily from ion channel models that incorporate a pore with several simultaneously bound ions (Hille, 1992). In particular, we do not explicitly allow conformational changes that modify the compartmentalization of the substrates. The gates of the alternating-access model have been de-emphasized. Instead, functional compartmentalization arises because the pore (or lumen or channel) of the transporter mediates multiple substrate bindings and substrate-substrate interactions that favor, albeit only statistically, permeation in fixed ratios of inorganic ions to organic substrate.

In this first discussion on the topic, we test the hypothesis of "multi-substrate single-file transport" in a quantitative, physically realistic fashion. We simulate the function of three ion-coupled transporters for which high-

resolution functional studies have been reported: the GABA transporter GAT1 (Mager et al., 1993), the serotonin transporter 5-HTT (Mager et al., 1994) and the Na^+-glucose transporter SGLT1 (Parent et al., 1992a,b). In each case, the multi-substrate single-file transport model has been found to reproduce available experimental data within experimental error. The model also accounts for newer phenomena such as the leakage currents of all these transporters and the variable stoichiometry of 5-HT have been recapitulated, among other permeation properties.

e) Comparing the New Model with Existing Models

Our approach has certainly been foreshadowed by many previous suggestions that transporters have channel-like mechanisms, for instance, in mediated ionic transport (Frohlich, 1988; Krupka, 1989; Hasegawa et al., 1992), in electrogenic membrane systems (Andersen et al., 1985; Lagnado et al., 1988; Nakamoto et al., 1989; Hilgemann et al., 1991; Lauger, 1991; Gadsby et al., 1993; Rakowski, 1993), in neurotransmitter transporters, (Krupka & Deves, 1988; Schwarztz & Tachibana, 1990), and in facilitative sugar transporters (Barnett et al., 1975; Lowe & Walmsley, 1986; Walmsleym, 1988; Baldwin, 1993; Hernandez and Fischbarg, 1994). Detailed theories have been based on electro-diffusion (Chen & Eisenberg, 1993; Eisenberg, 1994) and have concerned channels that can simultaneously contain two ionic species at once (Franciolini & Nonner, 1994). Although molecular cloning has given us knowledge about the amino-acid sequence of many ion-coupled transporters (Harvey and Nelson, 1994), there is still little relevant structural information at the atomic scale, or even at the level of tertiary structure or membrane topology. Therefore the model is cast in purely formal terms at present.

References

Andersoen OS, Silveira JEN, Steinmetz PR. 1985. Intrinsic characteristics of the proton pump in the lumiamenal membrane of a tight urinary epithelium. The relation between transport rate and DetalhumH. J. Gen. Physiol. 86:215-234.

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science 262*:1715–1718.

Baldwin SA. 1993. Mammalian passive glucose ion-coupled transporters: members of an ubitquitous family of active and passive transport proteins. Biochim. Biophys. Acta. 1165:17-49.

Barnett JEG, Holman GD, Chalkley RA, Munday KA. 1975. Evidneence for two asymmetric conformational states in the human erythrocyte sugar-transport system. Biochem. J. 145:417-429.

Betz SF, Raleigh DP, DeGrado WF. 1993. De novo protein design: from molten globules to native-like states. *Curr Opin Struc Biol 3* 601-610.

Betz SF, DeGrado WF. 1996. Controlling topology and native-like behavior of de novo-designed peptides — Design and characterization of antiparallel 4-stranded coiled coils. *Biochemistry 35*:6955–6962.

Bowie JU, Eisenberg D. 1993. Inverted protein structure prediction. *Curr Opin Struct Biol 3* 437-444.

Bruccoleri RE. 1993. Application of systematic conformational search to protein modeling. *Mol Sim 10* 151-174.

Cammack JN, Rakhilin SV, Schwartz EA. 1994. A GABA ion-coupled transporter operates asymmetrically and with variable stoichiometry. Neuron. 13:1-20.

Canter CR, Shimmel PR. 1980. Biophysical Chemistry. *W. H. Freeman and Company, San Francisco.*

Chen DP, Eisenberg RS. 1993. Flux, coupling, and selectivity in ionic channels of one conformation. Biophys. J. 65:727-746.

Chothia C, Janin J. 1981. Relative orientation of close-packed *b*-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:4146–4150.

Chothia C, Janin J. 1982. Orthogonal packing of *b*-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:3955–3965.

Chothia C, Levitt M, Richardson D. 1977. Structure of proteins: packing of *a*-helices and pleated sheets. *Proc Natl Acad Sci USA 74*:4130–4134.

Chothia C, Levitt M, Richardson D. 1981. Helix to helix packing in proteins. *J Mol Biol 145*:215–250.

Chou K-C, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. 1985. Interactions between an *a*-helix and *b*-sheet energetics of *a/b* packing in proteins. *J Mol Biol 186*:591–609.

Chou K-C, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. 1986. Interactions between two *b*-sheets energetics of *b/b* packing in proteins. *J Mol Biol 188*:641–649.

Chou K-C, Maggiora GM, Nemethy G, Scheraga HA. 1988. Energetics of the structure of the four-*a*-helix bundle in proteins. *Proc Natl Acad Sci USA 85*:4295–4299.

Cohen FE, Sternberg MJE, Taylor WR. 1980. Analysis and prediction of protein *b*-sheet structures by a combinatorial approach. *Nature 285*:378–382.

Cohen FE, Sternberg MJE, Taylor WR. 1981. Analysis of the tertiary structure of protein *b*-sheet sandwiches. *J Mol Biol 148*:253–272.

Cohen FE, Sternberg MJE, Taylor WR. 1982. Analysis and prediction of the packing of *a*-helices against a *b*-sheet in the tertiary structure of globular proteins. *J Mol Biol 156*:821–862.

Creighton TE. 1993. Proteins: structures and molecular properties, 2nd edition. *W. H. Freeman and Company, New York.*

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.

Dahiyat BI, Mayo SL. 1997a. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94* 10172-10177.

Dahiyat BI, Mayo SL. 1997b. De novo protein design - fully automated sequence selection. *Science 278* 82-87.

DeGrado W, Raleigh D, Handel T. 1991. Protein design, what are we learning? *Curr Opin Struct Biol* 1 984-993.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

DeFelice LJ, Galli R, Blakely D. 1995. Current fluctuations in norepinephrine transporters. Biophys. J. 68:A232.

Donate LE, Rufino SD, Canard LHJ, Blundell TL. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database of modeling and prediction. *Protein Sci 5* 2600-2616.

Eisenberg RS. 1994. Atomic biology, electrostatics, and ionic channels. New Developments and Theoretical Studies of Proteins:1-116. World Scientific Publishing, Philadelphia. publisher?

Franciolini F, Nonner W. 1994. A multiion permeation mechanism in neuronal background chloride channels. J. Gen. Physiol. 104: 725-746.

Frolhich O. 1988. The "tunneling" mode of biological carrier-mediated transport. J. Membr. Biol. 101:189-198.

Gadsby DC, Rakowski RF, Weer PDe. 1993. Extracellular access to the Na, K Ppump: pathway similar to ion channel. Science 260:100-103.

Handel TM, Williams SA, DeGrado WF. 1993. Metal ion-dependent modulation of the dynamics of a designed protein. *Science 261* 879-885.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA 92*:8408–8412.

Harvey WR, Nelson N. 1994. *Transporters*. J. Exp. Biol. Volume 196.

Hasegawa H, Skach W, Baker O, Calayag MC, Lingappa V, and Verkman AS. 1992. A multifunctional aqueous channel formed by CFTR. Science 258: 1477-1479.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science 249* 884-891.

Hellinga HW, Caradonna JP, Richards FM. 1991. Construction of new ligand-binding sites in proteins of know structure 2. Grafting of buried transition-metal binding site into *Escherichia coli* thioredoxin. *J Mol Biol 222*:787–803.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.

Hernandez JA, Fischbarg J. 1994. Transport properties of single-file pores with two conformational states. Biophys. J. 67:996-1006.

Hilgemann DW, Nicoll DA, Philipson KD. 1991. Charge movement during Na^+ translocation by native and cloned cardiac Na^+/CA^{2+} exchanger. Nature. 352:715-718.

Hill TL. 1977. *Free Energy Transduction in Biology*. Academic Press, New York.

Hille B. 1992. *Ionic Channels of Excitable Membranes*. Sinauer Associates Inc., Sunderland MA.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*:1142–1154.

Ippolito JA, Alexander RS, Christianson DW. 1990. Hydrogen bond stereochemistry in protein structure and function. *J Mol Biol 215* 457-471.

Kanner BI, Schuldiner S. 1987. Mechanism of transport and storage of neurotransmitters. CRC Crit. Rev. 22,Biochem. 22: 1-38.

Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. 1995. Novel metal-binding proteins by design. *Nature Structure Biol 2*:368–373.

Krupka RM, Deves R. 1998. The choline carrier of erythrocytes; l. Location of the NEM-reactive thiol group in the inner gated channel. KJ. Membr. Biol. 101:43-47.

Krupka RM. 1989. Role of substrate binding forces in exchange-only transport systems: II Implications for the mechanism of the anion-exchanger of red cells. J. Membr. Biol. 109:159-171.

Lagnado L, Cervetto K, McNaughton PA. 1988. Ion transport by the Na-Ca exchange in isolated rod outer segments. Proc. Natl. Acad. Sci. USA. 85:4548-4552.

Lauger P, Jauch P. 1986. Microscopic description of voltage effects on ion-driven cotransport systems. J. Membrane Biol. 91:275-284.

Lauger P. 1991. *Electrongenic Ion Pumps*. Sinauer Associates Inc., Sunderland, M.A.

Lester HA, Mager S, Quick MW, Corey JL. 1994. Permeation properties of neurotransmitter ion-coupled transporters. Ann. Rev. Pharmacol. Toxicol. 34:219-249.

Lim WA, Hodel A, Sauer RT, Richards FM. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423–427.

Mager S, Naeve J, Quick M, Guastella J, Davidson N, Lester HA. 1993. Steady states, charge movements, and rates for a cloned GABA ion-coupled transporter expressed in Xenopus oocytes. Neuron. 10:177-188.

Mager S, Min C, Henry D, Davidson N, Chavkin C, Hoffman B, Lester HA. 1994. Conducting states of a mammalian serotonin ion-coupled transporter. Neuron. 12:845-859.

McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol 238* 777-793.

Nakamoto RK, Rao R, Slayman CW. 1989. Transmembrane segments of the P-type cation-transporting ATPases. A comparative study. Ann. N. Y. Acad. Sci. 574:165-179.

Nautiyal S, Woolfson DN, King DS, Alber T. 1995. A designed heterotrimeric coiled coil. *Biochemistry 34*:11645–11651.

Osterhout JJ, Handel T, Na G, Toumadge A, Long RC, Connolly PJ, Hoch JC, Johnson WC, Live D, DeGrado WF. 1992. Characterization of the structural properties of $\alpha 1\beta$, a peptide designed to form a four-helix bundle. *J Am Chem Soc 114* 331-337.

Pabo CA. 1983. Designing proteins and peptides.*Nature 301* 200.

Parent L, Supplosson S, Loo DDF, Wright EM. 1992a. Electrogenic properties of the cloned Na^+/Glucose coion-coupled /glucose cotransporter: I. Voltage clamp studies. J. Membr. Biol. 125:49-62.

Parent L, Supploisson S, Loo DDF, Wright EM. 1992b. Electrogenic properties of the cloned Na^+/Glucose coion-coupled transporter: II. A ion-coupled transporter /glucose cotransporter: II. A transport model under nonrapid equilibrium conditions. J. Membr. Biol. 125:63-79.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Rakowski RF. 1993. Charge movement by the Na/K pump in *Xenopus* oocytes. J. Gen Physiol. 101:117-114.

Regan L, DeGrado WF. 1988. Characterization of a helical protein designed from first principles. *Science 241* 976-978.

Rudnick G, Clark J. 1993. From synapse to vesicle: the reuptake and storage of biogenic amine neurotransmitters. Biochim. Biophys. Acta. 1144: 249-263.

Salemme FR. 1983. Structural properties of protein b-sheets. *Prog Biophys molec Biol 42* 95-133.

Schwartz EA, Tachibana M. 1990. Electrophysiology of glutamate and sodium co-transport in a glial cell of the salamander retina. J. Physiol. 426:32-80.

Schultz SG. 1980. *Basic Principles of Membrane Transport.* Cambridge. University Press, Cambridge. 33-41, 85, 95, 96.

Schultz SG. 1986. Ion-coupled transport of organic solutes across biological membranes. In: *Membrane Physiology*. T.E. Andreoli, J.F. Hollfman, D. D. Fanestil, and S. G. Schultz, editors, pp. 283-294. Plenum. New York.

Stein WD. 1986. *Transport and Diffusion across Cell Membranes.* Academic Press, Orlando, FL. 337-361, 613-616.

Stickle DF, Presta LG, Dill KA, Rose GD. 1992. Hydrogen bonding in globular proteins. *J Mol Biol 226* 1143-1159.

Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6* 1701-1707.

Umbach JA, Coady MJ, Wright EM. 1990. Intestinal Na^+/glucose cotransporters /glucose cotransporter expressed in *Xenopus* oocytes is electrogenic. Biophys. J. 57, 1217-1224.

Wadiche JI, Amara SG, Kavanaugh MP. 1995a . Ion fluxes associated with excittatory amino acid transport. Neuron, in press.

Wadiche JI, Arriza JL, Amara SG, Kavanaugh MP. 1995b. Kinetics of a human glutamate transporter. Neuron. 14:1019-1027.

Wright EM. 1993. The intestinal Na^+ /glucose ion-coupled cotransporter. Annu. Rev. Physiol. 55:575-89.

Yue K, Dill KA. 1992. Inverse protein folding problem: designing polymer sequences. *Proc Natl Acad Sci USA 89* 4163-4167.

Chapter 2

Coupling Backbone Flexibility and Amino Acid Sequence Selection in Protein Design

Protein Science (1997), *6*:1701–1707. Cambridge University Press. Printed in the USA.

Coupling backbone flexibility and amino acid sequence selection in protein design

ALYCE SU[1] AND STEPHEN L. MAYO[2]

[1]Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, California 91125
[2]Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California 91125

(RECEIVED January 27, 1997; ACCEPTED March 21, 1997)

Abstract

Using a protein design algorithm that considers side-chain packing quantitatively, the effect of explicit backbone motion on the selection of amino acids in protein design was assessed in the core of the streptococcal protein G β1 domain (Gβ1). Concerted backbone motion was introduced by varying Gβ1's supersecondary structure parameter values. The stability and structural flexibility of seven of the redesigned proteins were determined experimentally and showed that core variants containing as many as 6 of 10 possible mutations retain native-like properties. This result demonstrates that backbone flexibility can be combined explicitly with amino acid side-chain selection and that the selection algorithm is sufficiently robust to tolerate perturbations as large as 15% of Gβ1's native supersecondary structure parameter values.

Keywords: backbone degrees of freedom; protein design; protein G; supersecondary structure parameters

Several groups have proposed and tested systematic, quantitative methods for protein design that screen possible sequences for compatibility with a desired backbone fold (Ponder & Richards, 1987; Hellinga et al., 1991; Hurley et al., 1992; Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Harbury et al., 1995; Klemba et al., 1995; Nautiyal et al., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996). In these methods, the backbone is held fixed and a search is performed to find side chains, whose conformations are often discretized as rotamers, that lead to sterically acceptable packing arrangements. Such algorithms correctly predict highly homologous core sequences to be acceptable. A significant problem, however, is that there are cases in which certain residue combinations are predicted to be sterically incompatible with a given fold, but these combinations can yield proteins that are as stable as the wild-type protein. Crystal structures of such mutants have revealed concerted backbone movements adopted by the protein to accommodate potentially disruptive residues, which indicates that the protein possesses a significant degree of backbone flexibility that must be accounted for in order to predict the full spectrum of compatible sequences (Baldwin et al., 1993; Lim et al., 1994).

Explicit backbone flexibility can be introduced into the design process by considering supersecondary structure parameterization. Supersecondary structure parameterization has been described for fold classes that include α/α (Crick 1953a, 1953b; Chothia et al., 1981; Chou et al., 1988; Murzin & Finkelstein, 1988; Presnell & Cohen, 1989; Harris et al., 1994), α/β (Chothia et al., 1977; Janin & Chothia, 1980; Cohen et al., 1982; Chou et al., 1985), and β/β (Cohen et al., 1980, 1981; Chothia & Janin, 1981, 1982; Chou et al., 1986; Lasters et al., 1988; Murzin et al., 1994a, 1994b). Within this framework, a protein is first parsed into a collection of secondary structural elements that are then abstracted into geometrical objects. For example, an α-helix is represented by its helical axis and geometric center. The relative orientation and distance between these objects are summarized as supersecondary structure parameters. Concerted backbone motion can be introduced by simply modulating a protein's supersecondary structure parameter values.

Recent studies using coiled coils have demonstrated that core side-chain packing can be combined with explicit backbone flexibility (Harbury et al., 1995; Offer & Sessions, 1995). In these cases, the goal was to search for backbone coordinates that satisfied a fixed amino acid sequence. Our goal is to develop methodology that allows both backbone flexibility and amino acid sequence selection.

This study is concerned primarily with coupling backbone flexibility and the selection of amino acids for protein cores and an assessment of the tolerance of our side-chain selection algorithm to perturbations in protein backbone geometry. An ideal model system for these purposes is the β1 immunoglobulin-binding domain of streptococcal protein G (Gβ1) (Gronenborn et al., 1991) (Fig. 1). Its small size, 56 residues, renders computations more tractable and simplifies production of the protein by either synthetic or recombinant methods. A solution structure (Gronenborn et al., 1991) and several crystal structures (Gallagher et al., 1994) are available to provide backbone templates for the side-chain selection algorithm. In addition, the energetics and structural dynamics of Gβ1 have been

Reprint requests to: Stephen L. Mayo, mail code 147-75, Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California 91125; e-mail: steve@mayo.caltech.edu.

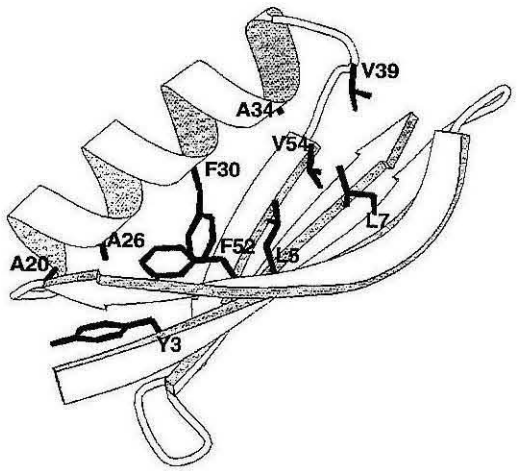


Fig. 1. Ribbon diagram of Gβ1 showing the positions of the 10 core residues examined in this study. Figure prepared with MOLSCRIPT (Kraulis, 1991).

characterized extensively (Alexander et al., 1992; Barchi et al., 1994; Kuszewski et al., 1994; Orban et al., 1995). Gβ1 contains no disulfide bonds and does not require a cofactor or metal ion to fold, but relies upon the burial of its hydrophobic core for stability. Further, Gβ1 contains sheet, helix, and turn structures and is without the repetitive side-chain packing patterns found in coiled coils and some helical bundles. This lack of periodicity reduces the bias from a particular secondary or tertiary structure and necessitates the use of an objective algorithm for side-chain selection. Perhaps most important for this study, the Gβ1 backbone can be classified as an α/β fold, a class for which extensive supersecondary structure analysis has been performed (Chothia et al., 1977; Janin & Chothia, 1980; Cohen et al., 1982; Chou et al., 1985).

Results and discussion

Sequence positions that constitute the core were chosen by examining the side-chain solvent-accessible surface area of Gβ1. We selected the 10 most buried positions, which include residues 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54 (Fig. 1). The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations at the 10 core positions.

Four sets of perturbed backbones were generated by varying Gβ1's supersecondary structure parameter values (Fig. 2). All possible core sequences consisting of alanine, valine, leucine, isoleucine, phenylalanine, tyrosine, and tryptophan (A, V, L, I, F, Y, and W) were considered for each perturbed backbone. The rotamer library used in this work has been described previously (Dahiyat & Mayo, 1996). Optimizing the sequences of the cores of Gβ1 and its structural homologues with 217 possible hydrophobic rotamers considered at each of the 10 core positions results in 217^{10} ($\sim 10^{23}$) rotamer sequences. Our scoring function consisted of two components: a van der Waals energy term and an atomic solvation term favoring burial of hydrophobic surface area. The van der Waals radii of the atoms in the simulation were scaled by either 1.0 or 0.9 in order to reduce the effects of using discrete rotamers (Mayo et al., 1990; B.I. Dahiyat & S.L. Mayo, 1997). Global optimum sequences for each of the backbone variants were found using the Dead-End Elimination (DEE) theorem (Desmet et al., 1992, 1994; Goldstein, 1994). Optimal sequences, and their corresponding proteins, are named by the backbone perturbation type, the size of the perturbation, and the radius scale factor used in their design. For example, the sequence designed using a template whose helix was translated by +1.50 Å along the sheet axis and a radius scale factor of 0.9 is called $\Delta h_{0.9}$[+1.50 Å]. Backbone perturbations that result in the same calculated core sequence are named by the perturbation with the greatest magnitude. For example, $\Delta h_{0.9}$ backbone perturbations of +1.25 and +1.50 Å result in the same sequence, which is called $\Delta h_{0.9}$[+1.50 Å]. The calculated core sequences corresponding to various backbone perturbations are listed in Tables 1, 2, 3, 4, and 5.

The optimal sequence for the 10 core positions of Gβ1 that is calculated using the native backbone (i.e., no perturbation) contains three conservative mutations relative to the wild-type sequence (Table 1) (B.I. Dahiyat & S.L. Mayo, 1997). Y3F and V39I are likely the result of the hydrophobic surface area burial term in the scoring function. L7I reflects a bias in the rotamer library used for these calculations. The crystal structure of Gβ1 has the leucine at position 7 with a nearly eclipsed χ_2 of 111°. This strained χ_2 is

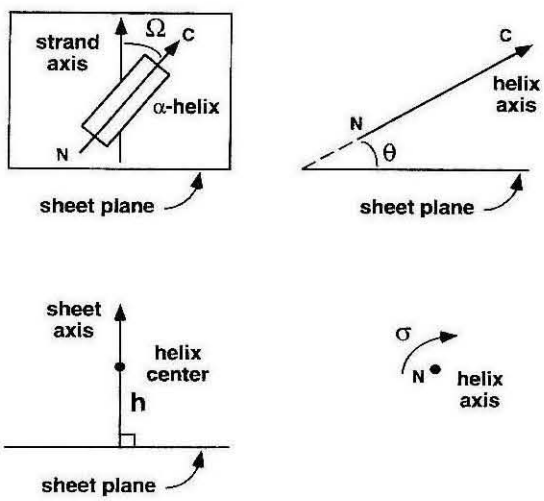


Fig. 2. Definitions of the supersecondary structure parameters used for Gβ1. The definitions are similar to those developed previously for α/β proteins (Janin & Chothia, 1980; Cohen et al., 1982). The helix center is defined as the average C_α position of residues 23–36. The helix axis is defined as the principal moment of the C_α atoms of residues 23–36 (Chothia et al., 1981). The strand axis is defined as the average of the least-squares lines fit through the midpoints of sequential C_α positions of the two central β-strands, residues 4–7 and 51–54. The sheet plane is defined as the least-squares plane fit through the C_α positions of residues 4–7 and 51–54. The sheet axis is defined as the vector perpendicular to the sheet plane that passes through the helix center. Ω is the angle between the strand axis and the helix axis after projection onto the sheet plane; θ is the angle between the helix axis and the sheet plane; h is the distance between the helix center and the sheet plane; σ is the rotation angle about the helix axis. The supersecondary structure parameter values for native Gβ1 are $\Omega = -26.49°$, $\theta = 3.20°$, $h = 10.04$ Å, and $\sigma = 0°$.

Table 1. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta h_{0.9}$[a]

$\Delta h_{0.9}$ (Å)	Vol	Gβ1 sequence										T_m (°C)	NMR
		Tyr 3	Leu 5	Leu 7	Ala 20	Ala 26	Phe 30	Ala 34	Val 39	Phe 52	Val 54		
−1.50	1.04	Phe	Ile	Val	Val	|	|	|	Ile	|	|	69	+
−1.25	1.04	Phe	Ile	Val	Val	|	|	|	Ile	|	|	69	+
−1.00	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−0.75	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−0.50	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−0.25	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
0.00	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
−0.25	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+0.50	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+0.75	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+1.00	1.13	Phe	|	Ile	|	|	|	Ile	Ile	Trp	|	85	+
+1.25	1.20	Phe	|	Ile	|	Leu	|	Ile	Ile	Trp	|	53	−
+1.50	1.20	Phe	|	Ile	|	Leu	|	Ile	Ile	Trp	|	53	−

[a]The Gβ1 wild-type sequence and position numbers are shown at the top of the table. A vertical bar indicates identity with the Gβ1 sequence. Δh is the change in the supersecondary structure parameter, h; Vol is the fraction of core side-chain volume relative to the Gβ1 sequence; T_m is the melting temperature measured by CD; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra. The T_m's for $\Delta h_{0.9}$[−1.50 Å] and $\Delta h_{0.9}$[+1.50 Å] were determined for 56-residue proteins (compared to 57-residue proteins for Gβ1 and all other mutants), which overstates the melting temperature by about 2 °C, the melting temperature difference between the 56- and 57-residue versions of Gβ1.

unlikely to be an artifact of the structure determination because it is present in two crystal forms and a solution structure (Gronenborn et al., 1991; Gallagher et al., 1994). Our rotamer library does not contain eclipsed rotamers and no staggered leucine rotamers pack well at this position. Instead, the side-chain selection algorithm chose an isoleucine rotamer that conserves the χ_1 dihedral and is able to pack well. We expect the removal of the strained leucine rotamer to stabilize the protein, a prediction that is tested in the experimental section of this work. The sequences that result from varying individual supersecondary structure parameter values show two notable trends. Small variations in the parameter values tend to have little or no effect on the calculated sequences. For example, varying $\Delta h_{0.9}$ from −0.25 to −1.00 Å (Table 1) and $\Delta h_{1.0}$ from +0.25 to +1.25 Å (Table 2) has no effect on the calculated sequences, which demonstrates the side-chain selection algorithm's tolerance to small variations in the initial backbone geometry. Large variations in the parameter values tend to result in greater sequence diversity. For example, $\Delta h_{1.0}$[+1.50 Å] contains 6 of 10 possible mutations relative to Gβ1 (Table 2). The apparently anomalous result that occurs for $\Delta h_{0.9}$ at −1.25 and −1.50 Å, an increase in core volume, is explained by the observation that translating the helix toward the sheet plane results in creating a pocket of space in the vicinity of position 20 that ultimately leads to the observed A20V mutation.

Table 2. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta h_{1.0}$[a]

$\Delta h_{1.0}$ (Å)	Vol	Gβ1 sequence										T_m (°C)	NMR
		Tyr 3	Leu 5	Leu 7	Ala 20	Ala 26	Phe 30	Ala 34	Val 39	Phe 52	Val 54		
−1.50	0.52	Ala	Ala	Ala	|	|	Ala	|	Leu	Ala	Ala	ND	ND
−1.25	0.62	Phe	Ala	Ala	|	|	Ala	|	Leu	Ala	Ala	ND	ND
−1.00	0.62	Phe	Ala	Ala	|	|	Ala	|	Leu	Ala	Ala	ND	ND
−0.75	0.91	Phe	Ala	Val	|	|	|	|	Ile	|	|	ND	ND
−0.50	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−0.25	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
0.00	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+0.25	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+0.50	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+0.75	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+1.00	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+1.25	1.05	Phe	|	Ile	|	|	|	|	Ile	Trp	|	89	+
+1.50	1.11	Phe	|	Ile	|	|	Leu	Ile	Ile	Trp	|	73	+

[a]The Gβ1 wild-type sequence and position numbers are shown at the top of the table. A vertical bar indicates identity with the Gβ1 sequence. Δh is the change in the supersecondary structure parameter, h; Vol is the fraction of core side-chain volume relative to the Gβ1 sequence; T_m is the melting temperature measured by CD; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined.

Table 3. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta\Omega_{0.9}$[a]

ΔΩ		Gβ1 sequence										T_m	
(°)	Vol	Tyr 3	Leu 5	Leu 7	Ala 20	Ala 26	Phe 30	Ala 34	Val 39	Phe 52	Val 54	(°C)	NMR
−10.0	1.00	Val	|	Val	Val	|	|	|	Ile	|	|	ND	ND
−7.5	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−5.0	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
−2.5	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
0.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+2.5	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+5.0	1.06	Phe	|	Ile	Val	|	|	|	Ile	|	|	ND	ND
+7.5	1.06	Phe	|	Ile	Val	|	|	|	Ile	|	|	ND	ND
+10.0	1.06	Phe	|	Ile	Val	|	|	|	Ile	|	|	ND	ND

[a]The Gβ1 wild-type sequence and position numbers are shown at the top of the table. A vertical bar indicates identity with the Gβ1 sequence. ΔΩ is the change in the supersecondary structure parameter, Ω; Vol is the fraction of core side-chain volume relative to the Gβ1 sequence; T_m is the melting temperature measured by CD; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined.

Table 4. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta\theta_{0.9}$[a]

Δθ		Gβ1 sequence										T_m	
(°)	Vol	Tyr 3	Leu 5	Leu 7	Ala 20	Ala 26	Phe 30	Ala 34	Val 39	Phe 52	Va; 54	(°C)	NMR
−10.0	0.98	Phe	|	Ala	|	|	|	|	Leu	Trp	|	ND	ND
−7.5	1.00	Phe	|	Leu	|	|	|	|	Leu	Trp	Ala	ND	ND
−5.0	1.03	Phe	|	Val	|	|	|	|	Ile	Trp	|	ND$^+$	ND$^+$
−2.5	1.03	Phe	|	Val	|	|	|	|	Ile	Trp	|	ND$^+$	ND$^+$
0.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+2.5	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+5.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+7.5	1.08	Phe	|	Ile	Val	|	Trp	|	Ile	Leu	|	ND	ND
+10.0	1.08	Phe	|	Ile	Val	|	Trp	|	Ile	Leu	|	ND	ND

[a]The Gβ1 wild-type sequence and position numbers are shown at the top of the table. A vertical bar indicates identity with the Gβ1 sequence. Δθ is the change in the supersecondary structure parameter, θ; Vol is the fraction of core side-chain volume relative to the Gβ1 sequence; T_m is the melting temperature measured by CD; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined; ND$^+$ indicates a property that was not determined, but that is expected to be "positive" based on sequence similarity to other DEE solutions (see $\Delta h_{0.9}$[+0.75 Å]).

Table 5. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta\sigma_{0.9}$[a]

Δσ		Gβ1 sequence										T_m	
(°)	Vol	Tyr 3	Leu 5	Leu 7	Ala 20	Ala 26	Phe 30	Ala 34	Val 39	Phe 52	Val 54	(°C)	NMR
−10.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
−7.5	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
−5.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
−2.5	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
0.0	1.01	Phe	|	Ile	|	|	|	|	Ile	|	|	91	+
+2.5	0.99	Phe	|	Val	|	|	|	|	Ile	|	|	89	+
+5.0	1.03	Phe	|	Val	|	|	|	|	Ile	Trp	|	ND$^+$	ND$^+$
+7.5	1.04	Phe	|	Val	|	|	Tyr	|	Ile	Trp	|	ND	ND
+10.0	1.04	Phe	|	Val	|	|	Tyr	|	Ile	Trp	|	ND	ND

[a]The Gβ1 wild-type sequence and position numbers are shown at the top of the table. A vertical bar indicates identity with the Gβ1 sequence. Δσ is the change in the supersecondary structure parameter, σ; Vol is the fraction of core side-chain volume relative to the Gβ1 sequence; T_m is the melting temperature measured by CD; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined; ND$^+$ indicates a property that was not determined, but that is expected to be "positive" based on sequence similarity to other DEE solutions (see $\Delta h_{0.9}$[+0.75 Å]).

Experimental validation of the designed cores focused on seven of the Δh-series mutants that contain between three and six sequence changes relative to Gβ1. The designed sequences resulting from $\Delta\Omega$, $\Delta\theta$, and $\Delta\sigma$ perturbations are, however, in many cases identical to various Δh-series sequences. Typical far-UV CD spectra are shown in Figure 3. $\Delta h_{0.9}[-1.00$ Å], $\Delta h_{0.9}[0.00$ Å], $\Delta h_{0.9}[+0.75$ Å], and $\Delta h_{0.9}[+1.00$ Å] have CD spectra that are indistinguishable from that of Gβ1, whereas $\Delta h_{0.9}[+1.50$ Å], $\Delta h_{1.0}[+1.50$ Å], and $\Delta h_{0.9}[-1.50$ Å] have CD spectra similar to that of Gβ1, suggesting that all of the mutants have a secondary structure content similar to the wild-type protein. Thermal melts monitored by CD are shown in Figure 4. All of the mutants have cooperative transitions with melting temperatures (T_m's) ranging from 53 °C for $\Delta h_{0.9}[+1.50$ Å] to 91 °C for $\Delta h_{0.9}[0.00$ Å] (Table 1). The T_m for Gβ1 is 85 °C. The measured T_m's for $\Delta h_{0.9}[-1.50$ Å] and $\Delta h_{0.9}[+1.50$ Å] are for 56-residue proteins compared to 57-residue proteins in all other cases (see Materials and methods), which results in T_m's that are estimated to be about 2 °C higher than what would be expected for the corresponding 57-residue proteins based on the T_m difference between the 56- and 57-residue versions of Gβ1. As suggested previously (B.I. Dahiyat & S.L. Mayo, 1997), the removal of the strained leucine at position seven (L7I), along with the increased hydrophobic burial generated by the Y3F and V39I mutations in $\Delta h_{0.9}[0.00$ Å], result in a protein that is measurably more stable than wild-type Gβ1. The extent of chemical shift dispersion in the 1D ^{1}H NMR spectrum of each mutant was assessed to gauge each protein's degree of native-like character (Fig. 5). All of the mutants, except $\Delta h_{0.9}[+1.50$ Å], have NMR spectra with chemical shift dispersion similar to that of Gβ1, suggesting that the proteins form well-ordered structures. $\Delta h_{0.9}[+1.50$ Å] has a spectrum with broad peaks and no dispersion, which is indicative of a collapsed, but disordered and fluctuating structure or nonspecific association. All seven mutant proteins retain their ability to bind IgG as measured by binding to an IgG-Sepharose affinity column. The stability and native-like character of $\Delta h_{0.9}[-1.50$ Å] and $\Delta h_{1.0}[+1.50$ Å] indicate that the sequence selection algorithm is sufficiently robust to tolerate Δh perturbations that are as large as 15% of Gβ1's native height supersecondary structure parameter value of 10 Å.

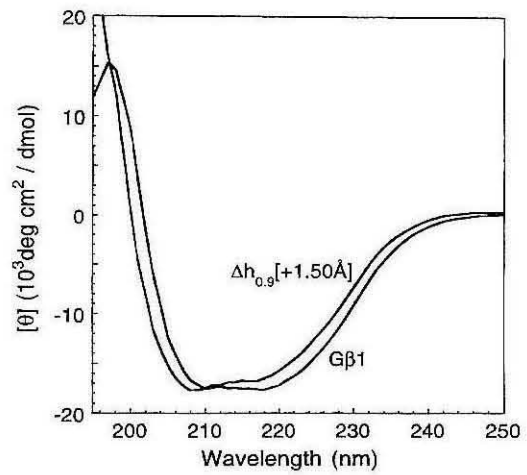


Fig. 3. Far-UV CD spectra of Gβ1 and the most perturbed of the Δh-series mutants, $\Delta h_{0.9}[+1.50$ Å]. $\Delta h_{0.9}[-1.50$ Å] and $\Delta h_{1.0}[+1.50$ Å] have CD spectra similar to $\Delta h_{0.9}[+1.50$ Å], whereas the remaining mutants have CD spectra similar to Gβ1.

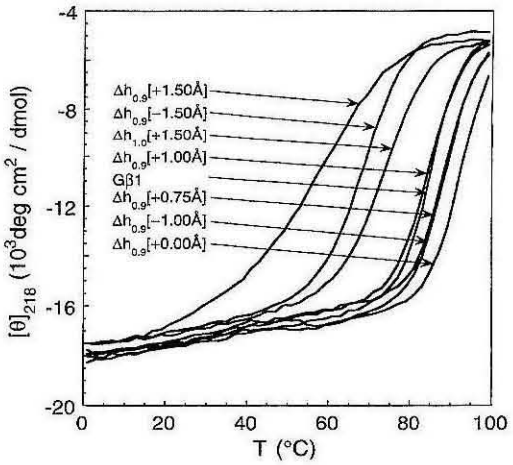


Fig. 4. Thermal denaturation of Gβ1 and the Δh-series mutants monitored by CD at 218 nm.

Although structures have not yet been determined for the six mutants that show good chemical shift dispersion in their NMR spectra, the magnitude of the backbone perturbations used to calculate these sequences are similar to the backbone perturbations observed for core mutations in other proteins (Baldwin et al., 1993; Lim et al., 1994). Elucidation of the structures of several of the mutants should contribute to our general understanding of the deformation modes available to protein backbones of the α/β class and should help define ranges of supersecondary structure parameter value perturbations that will be useful in future sequence calculations.

Conclusions

The objectives of this study were an assessment of the prospect of coupling backbone flexibility with amino acid sequence selection and an assessment of the robustness of our sequence selection algorithm to perturbations in protein backbone geometry. The results show clearly that backbone flexibility, introduced by varying supersecondary structure parameter values, can be coupled to amino acid sequence selection. This observation supports the notion that de novo protein design can be accomplished by following a two-step procedure that includes generation of backbone structures based on consensus supersecondary structure parameter values followed by automated amino acid sequence selection. The results also show that our sequence selection algorithm is tolerant to variations in starting conditions, which has the benefit of reducing the dependence on the exact details of the initial conditions. This robustness allows exploration of large backbone perturbations, which can led to greater sequence diversity.

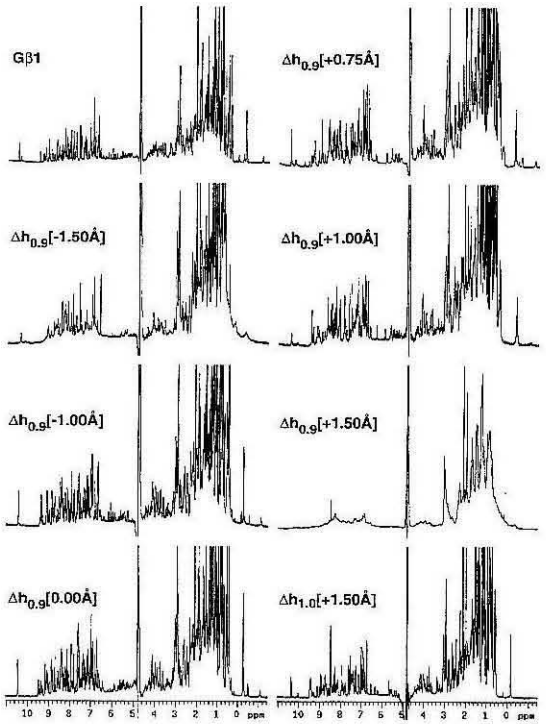


Fig. 5. 1D ^{1}H NMR spectra of Gβ1 and the Δ*h*-series mutants.

Methods and materials

Backbone perturbation, scoring functions, and DEE

The initial Gβ1 structure was taken from PDB entry 1pga (Bernstein et al., 1977; Gallagher et al., 1994). The program BIOGRAF (Molecular Simulations Incorporated, San Diego, California) was used to generate explicit hydrogens on the structure, which was then conjugate gradient minimized for 50 steps using the DREIDING forcefield (Mayo et al., 1990). The coordinate positions of atoms not involved in core sequence selection or backbone perturbations were kept fixed. Concerted backbone movements were performed by repositioning the α-helix (residues 23–36) to reflect the desired change in the indicated supersecondary structure parameter value. The coordinate positions of atoms belonging to residues 23, 24, 25, 27, 28, 29, 31, 32, 33, 35, and 36 were kept fixed after repositioning the helix. The distorted peptide bonds that result from backbone perturbations were left unchanged. The Δ*h*-, ΔΩ-, and Δσ-series perturbations were performed by translating the helix along the sheet axis, rotating the helix about the sheet axis, and rotating the helix about the vector parallel to the helix axis that passes through the helix center, respectively (see Fig. 2). The Δθ-series perturbations were performed by rotating the helix about the vector resulting from the cross product of the sheet axis and the vector parallel to the helix axis that passes through the helix center. A Lennard–Jones 12-6 potential was used for van der Waals interactions with atomic radii scaled by either 1.0 or 0.9 as indicated (B.I. Dahiyat & S.L. Mayo, 1997). The Richards definition of solvent-accessible surface area (Lee & Richards, 1971) was used and areas were calculated with the Connolly algorithm (Connolly, 1983). An atomic solvation parameter of 23.2 cal/mol/Å^2 was used to favor hydrophobic burial (Dahiyat & Mayo, 1996). The rotamer library and DEE optimization followed the methods of our previous work (Dahiyat & Mayo, 1996). Calculations were performed on either a 12 processor, R10000-based Silicon Graphics Power Challenge or a 512 node Intel Delta.

Mutagenesis and protein purification

A synthetic Gβ1 gene (Minor & Kim, 1994) was cloned into pET11a (Novagen) and used as the template for inverse PCR mutagenesis (Hemsley et al., 1989). 5′ phosphorylated oligos (Genosys) were used without further purification. Mutant sequences were confirmed by DNA sequencing. The expression and purification of the mutant proteins followed published procedures (Minor & Kim, 1994). Incomplete N-terminal processing resulted in a mixture of 56- and 57-residue proteins, which were separated by HPLC (Minor & Kim, 1994). The 57-residue proteins were used in all cases except for mutants $\Delta h_{0.9}[-1.50$ Å] and $\Delta h_{0.9}[+1.50$ Å], where the 56-residue proteins were used. Molecular weights were confirmed by mass spectrometry.

CD and NMR

CD spectra were measured on an Aviv 62DS spectrometer at pH 6.0, 50 mM sodium phosphate buffer, 25 °C, and 50 μM protein. A 1-mm pathlength cell was used and the temperature was controlled by a thermoelectric unit. Thermal melts were performed in the same buffer using 2° temperature increments with an averaging time of 10 s and an equilibration time of 90 s. T_m values were derived from the ellipticity at 218 nm ($[\theta]_{218}$) by evaluating the maximum of a $d[\theta]_{218}/dT$ versus T plot. The T_m's were reproducible to within 2°. Protein concentrations were determined by UV spectrophotometry. NMR samples were prepared in 90/10 H_2O/D_2O and 50 mM phosphate buffer at pH 6.0. Spectra were acquired on a Varian Unity Plus 600 MHz spectrometer at 25 °C. One-thousand twenty-four transients were acquired with 1.5 s of solvent presaturation used for water suppression. Samples were approximately 0.5 mM.

Acknowledgments

We thank M. Ary, S. Malakauskas, and S. Ross for technical assistance, G. Hathaway of the Caltech Protein and Peptide Microanalytical Facility for mass spectra, the Caltech DNA Sequencing Core Facility for DNA sequencing, and the Caltech Center for Advanced Computing Research for use of the Intel Delta. This work was supported by the Rita Allen Foundation, the David and Lucile Packard Foundation, and the Searle Scholars Program/The Chicago Community Trust.

References

Alexander P, Fahnestock S, Lee T, Orban J, Bryan P. 1992. Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains β1 and β2—Why small proteins tend to have high denaturation temperatures. *Biochemistry 31*:3597–3603.

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science 262*:1715–1718.

Barchi JJ, Grasberger B, Gronenborn AM, Clore GM. 1994. Investigation of the backbone dynamics of the IgG-binding domain of streptococcal protein G

by heteronuclear two-dimensional ^{1}H-^{15}N nuclear magnetic resonance spectroscopy. *Protein Sci 3*:15–21.
Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.
Betz SF, DeGrado WF. 1996. Controlling topology and native-like behavior of de novo-designed peptides—Design and characterization of antiparallel 4-stranded coiled coils. *Biochemistry 35*:6955–6962.
Chothia C, Janin J. 1981. Relative orientation of close-packed β-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:4146–4150.
Chothia C, Janin J. 1982. Orthogonal packing of β-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:3955–3965.
Chothia C, Levitt M, Richardson D. 1977. Structure of proteins: packing of α-helices and pleated sheets. *Proc Natl Acad Sci USA 74*:4130–4134.
Chothia C, Levitt M, Richardson D. 1981. Helix to helix packing in proteins. *J Mol Biol 145*:215–250.
Chou KC, Maggiora GM, Nemethy G, Scheraga HA. 1988. Energetics of the structure of the four-α-helix bundle in proteins. *Proc Natl Acad Sci USA 85*:4295–4299.
Chou KC, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. 1985. Interactions between an α-helix and β-sheet energetics of α/β packing in proteins. *J Mol Biol 186*:591–609.
Chou KC, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. 1986. Interactions between two β-sheets energetics of β/β packing in proteins. *J Mol Biol 188*:641–649.
Cohen FE, Sternberg MJE, Taylor WR. 1980. Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature 285*:378–382.
Cohen FE, Sternberg MJE, Taylor WR. 1981. Analysis of the tertiary structure of protein β-sheet sandwiches. *J Mol Biol 148*:253–272.
Cohen FE, Sternberg MJE, Taylor WR. 1982. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J Mol Biol 156*:821–862.
Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*:709–713.
Crick FHC. 1953a. The Fourier transform of a coiled-coil. *Acta Crystallogr 6*:685–689.
Crick FHC. 1953b. The packing of α-helices. *Acta Crystallogr 6*:689–697.
Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.
Dahiyat BI, Mayo SL. 1997. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA*, in press.
Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.
Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*:539–542.
Desmet J, De Maeyer M, Lasters I. 1994. The dead-end elimination theorem: A new approach to the side-chain packing problem. In: Merz KM Jr, Le Grand SM, eds. *The protein folding problem and tertiary structure prediction*. Boston: Birkhäuser. pp 307–337.
Gallagher T, Alexander P, Bryan P, Gilliland GL. 1994. Two crystal structures of the β1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry 33*:4721–4729.
Goldstein RF. 1994. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys J 66*:1335–1340.
Gronenborn AM, Filpula DR, Essign NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science 253*:657–661.
Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: Structure prediction for coiled coils. *Proc Natl Acad Sci USA 92*:8408–8412.
Harris NL, Presnell SR, Cohen FE. 1994. Four helix bundle diversity in globular proteins. *J Mol Biol 236*:1356–1368.
Hellinga HW, Caradonna JP, Richards FM. 1991. Construction of new ligand-binding sites in proteins of know structure 2. Grafting of buried transition-metal binding site into *Escherichia coli* thioredoxin. *J Mol Biol 222*:787–803.
Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.
Hemsley A, Arnheim N, Toney MD, Cortopassi G, Galas DJ. 1989. A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Res 17*:6545–6551.
Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*:1142–1154.
Janin J, Chothia C. 1980. Packing of α-helices onto β-pleated sheets and the anatomy of α/β proteins. *J Mol Biol 143*:95–128.
Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. 1995. Novel metal-binding proteins by design. *Nature Struct Biol 2*:368–373.
Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.
Kuszewski J, Clore GM, Gronenborn AM. 1994. Fast folding of a prototypic polypeptide—The immunoglobulin binding domain of streptococcal protein G. *Protein Sci 3*:1945–1952.
Laster I, Wodak SJ, Alard P, Custem EV. 1988. Structure principles of parallel β-barrels in proteins. *Proc Natl Acad Sci USA 85*:3338–3342.
Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.
Lim WA, Hodel A, Sauer RT, Richards FM. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423–427.
Mayo SL, Olafson BD, Goddard WA III. 1990. Dreiding—A generic force-field for molecular simulations. *J Phys Chem 94*:8897–8909.
Minor DL, Kim PS. 1994. Measurements of the β-sheet-forming propensities of amino acids. *Nature 367*:660–663.
Murzin AG, Finkelstein AV. 1988. General architecture of the α-helical globule. *J Mol Biol 204*:749–769.
Murzin AG, Lesk AM, Chothia C. 1994a. Principles determining the structure of β-sheet barrels. I. A theoretical analysis. *J Mol Biol 236*:1369–1381.
Murzin AG, Lesk AM, Chothia C. 1994b. Principles determining the structure of β-sheet barrels. II. The observed structures. *J Mol Biol 236*:1382–1400.
Nautiyal S, Woolfson DN, King DS, Alber T. 1995. A designed heterotrimeric coiled coil. *Biochemistry 34*:11645–11651.
Offer G, Sessions R. 1995. Computer modelling of the α-helical coiled coil: Packing of side-chains in the inner core. *J Mol Biol 249*:967–987.
Orban J, Alexander P, Bryan P, Khare D. 1995. Assessment of stability differences in the protein G β1 and β2 domains from hydrogen deuterium exchange—Comparison with calorimetric data. *Biochemistry 34*:15291–15300.
Ponder JW, Richards FM. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.
Presnell SR, Cohen FE. 1989. Topological distribution of four-α-helix bundles. *Proc Natl Acad Sci USA 86*:6592–6596.

Chapter 3

Assembling De Novo Backbone Templates for Amino Acid Sequence Selection in Protein Design - (I) Theory

Abstract

Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms. However, all of these methods rely on the use of a fixed backbone template chosen from the Protein Data Bank (Bernstein et al., 1977). We sought to expand the range of computational protein design by developing a general, quantitative design method for computing de novo backbone templates. The method had to compute atomic resolution backbones compatible with the atomistic sequence selection algorithm we were using (Dahiyat & Mayo, 1997) and it had to be applicable to all protein motifs. The algorithm we developed uses supersecondary structure parameters to determine the orientation among secondary structural elements, given a target protein fold. All possible backbone arrangements are screened using a cost function which evaluates core packing, hydrogen bonding, loop closure, and backbone torsional geometry. Given a specified number of residues for each secondary structural element, a single optimal configuration is found. We chose three motifs to test our method ($\beta\beta\alpha$, $\beta\alpha\beta$, and $\alpha\alpha$) since their combination can be used to approximate most possible backbone fold. The best structure found for the $\beta\beta\alpha$ motif is similar to a zinc finger, and the best structure for the $\beta\alpha\beta$ motif is similar to a segment of a β-barrel (PDB entry 8adh). The backbone obtained for the $\alpha\alpha$ motif resembles minimized protein A. These results suggest that de novo backbones assembled using this method may serve as adequate input templates for atomistic sequence selection algorithms.

Introduction

De novo protein design has received considerable attention recently and significant advances have been made towards the goal of producing stable, well-folded proteins with novel sequences. Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the aim of developing general design algorithms (Ponder & Richards, 1987; Hellinga et al., 1991; Hurley et al., 1992; Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Harbury et al., 1995; Klemba et al., 1995; Nautiyat et al., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996). These techniques screen possible sequences for compatibility with a desired protein motif. To date, they all have required a fixed backbone template as input. We sought to expand the range of computational protein design by developing a general, quantitative design method for computing de novo backbone templates. Two critical components for a backbone design algorithm are its compatibility with the atomistic sequence selection algorithm we have been using (Dahiyat & Mayo, 1996) and its applicability to all protein motifs. Representations of designed backbones have ranged from ribbon-diagrams (Chirgadze, 1987; Efimov, 1993a, b; Efimov, 1994; Efimov, 1997) and lattice models (Park & Levitt, 1995; Dill & Stigter, 1995) to the more precise atomistic representations of poly-Ala or poly-Val peptides (Chou & Scheraga, 1982; Chou et al., 1983). The ribbon-diagrams and lattice models provide insufficient atomic information for sequence selection algorithms, while the atomistic polypeptides tend to bias the final optimized sequence by including side chain data in determining the backbone structure (Su & Mayo, unpublished results). In addition, previously designed backbones mainly focused on two specific protein motifs — the coiled-coil (Harbury, 1995) and the β-barrel (Murzin et al., 1994a, b). Our goal is a general backbone design algorithm that computes

atomic resolution backbones for an arbitrary motif, but provides no extra information to bias the final optimized sequence. This work reports the initial computational results of our backbone design algorithm for the ββα, βαβ, and αα motifs.

In selecting the motif set to demonstrate the generality of the algorithm, we chose the ββα, βαβ, and αα motifs, since their combination can be used to approximate most possible backbone fold. The combination of these three motifs includes all the possible connecting loop classes (Donate et al., 1996) and all the possible secondary structural elements (α-helix, parallel β-sheet, and anti-parallel β-sheet).

Our design methodology consists of an atomistic backbone design algorithm that quantitatively considers specific interactions among secondary structural elements. The backbone design algorithm screens all possible geometric orientations of secondary structural elements and finds the optimal placement of these elements given a target protein fold. In order to accurately specify the atomic positioning of these secondary structural elements, we consider a discrete set of geometric orientation parameters for each element, called supersecondary structure parameters (Chothia et al., 1997; Janin & Chothia, 1980; Su & Mayo, 1997). Within each element, we further consider the mapping of backbone atoms N, H, C, O, C_α, and C_β, according to the allowed backbone torsional distribution. The immense search problem presented by supersecondary structure parameter and backbone torsional angle optimization (10^{16} to 10^{24} possible configurations) is surmounted by applying the Adaptive Simulated Annealing Algorithm (Ingber, 1993). This is a standard optimization algorithm that searches in continuous space. The cost function used by the algorithm scores a particular backbone arrangement by evaluating its core packing (Ponder & Richards, 1987), hydrogen bonding

(Ippolito et al., 1990; Stickle et al., 1992; McDonald & Thornton, 1994), loop closure (Bruccoleri, 1993; Donate et al., 1996), and backbone torsional geometry (Salemme, 1983). Given a specified number of residues for each secondary structural element, this procedure yields a family of optimal configurations for each motif. No amino acid sequence information is required.

Backbone Design Algorithm

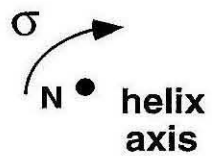
The backbone design algorithm computes all backbone atoms (N, HN, C, O, C_α, and C_β) for a target protein motif by iteratively generating secondary structural elements (α-helix, β-strand), connecting them with proper loop fragments (Donate et al., 1996), and evaluating their cost function. All loop fragments from the appropriate loop class were allowed: ee for $\beta\beta$, eh for $\beta\alpha$, he for $\alpha\beta$, and hh for $\alpha\alpha$. Seven supersecondary structure parameters were used to describe the $\beta\alpha\beta$ and $\beta\beta\alpha$ motifs (σ, λ, d, τ, Ω, θ, and h). For the $\alpha\alpha$ motif, only four were used (σ, λ, d, and τ). The parameters are illustrated in Figure 1. Parameter σ describes the rotation angle about the helix axis and parameter λ describes the translation distance along the helix axis. Parameter d describes the distance between the center of the two strands or helices and parameter τ describes the angle between them. Parameter Ω describes the angle between the strand axis and the helix axis after projection onto the sheet. Parameter θ describes the angle between the helix axis and the sheet plane. Parameter h describes the distance between the center of the α-helix and the plane of the β-sheet. The range of values used for each of the parameters is shown in Table 1. Range values were chosen to include all values reported in the literature (Janin & Chothia, 1980; Chothia et al., 1981; Cohen et al., 1981).

Figure 1 Definition of generic supersecondary structure parameters used in assembling motifs $\beta\beta\alpha$, $\beta\alpha\beta$, and $\alpha\alpha$. (a) σ is the rotation angle about a single axis; (b) λ is the translation distance along a single axis; (c) d is the distance between the centers of two parallel axes; (d) τ is the dihedral angle between the two parallel axes; (e) Ω is the angle between an axis projected onto a plane and a selected direction contained in the same plane; (f) θ is the angle between an axis and a plane; (g) h is the distance between the center of an axis and a plane.

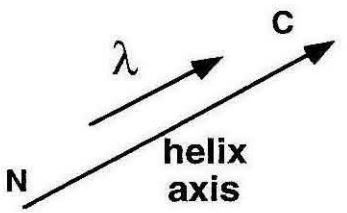
In general, σ and λ characterize a single-axis interaction, e.g. rotation about or translation along a helix or strand axis; d and τ characterize a two-axis interaction, e.g. two β-strands or two α-helices; h, Ω, and θ characterize a three (or more)-axis interaction, e.g. an α-helix with a β-sheet containing two (or more) β-strands.

In the $\beta\beta\alpha$ and $\beta\alpha\beta$ motifs, parameters d and τ determine the relative orientation of the two strands, parameters h, Ω, and θ determine the relative orientation of the helix to the sheet; in the $\alpha\alpha$ motif, parameters d and τ determine the relative orientation of the two helices. In all cases, parameters σ and λ associated with each axis allow individual rotation and translation of each axis.

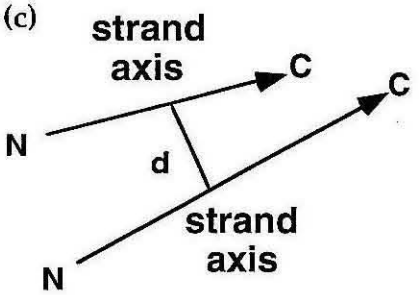
(a)


σ
N
helix
axis

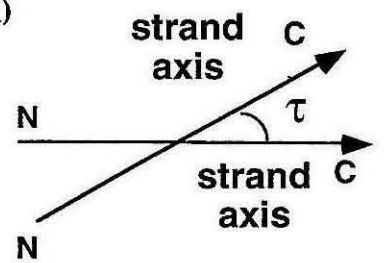
(b)


λ
C
N
helix
axis

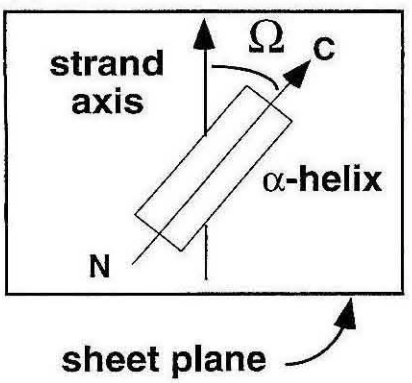
(c)


strand
axis
C
C
N
d
strand
axis
N

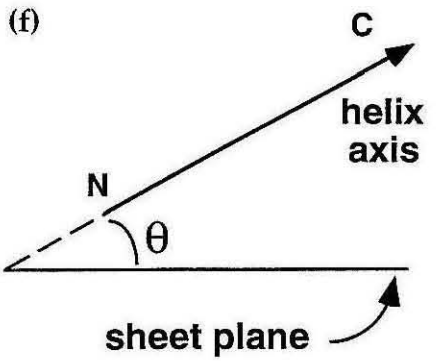
(d)


strand
axis
C
N
τ
strand
axis
C
N

(e)


strand
axis
Ω
C
α-helix
N
sheet plane

(f)


C
helix
axis
N
θ
sheet plane

(g)

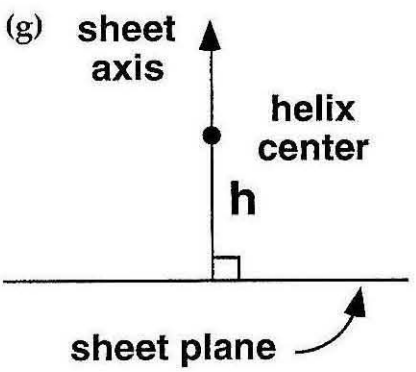

sheet
axis
helix
center
h
sheet plane

Table 1 Optimized parameter values for de novo backbones of the ββα, βαβ, and αα motifs.

variables	ββα ranges	ββα optimized values	βαβ ranges	βαβ optimized values	αα ranges	αα optimized values
local						
n1 (residues)	[5, 7]	7	[5, 7]	5	[10, 20]	18
φ1 (degrees)	[-144.0, -70.0]	-101.0	[-144.0, -70.0]	-115.0	[-103.0, -41.0]	-67.0
ψ1 (degrees)	[94.0, 156.0]	104.0	[94.0, 156.0]	124.0	[-60.0, -41.0]	-47.0
n2 (residues)	[5, 7]	7	[10, 20]	15	[10, 20]	18
φ2 (degrees)	[-144.0, -70.0]	-101.0	[-103.0, -41.0]	-60.0	[-103.0, -41.0]	-58.0
ψ2 (degrees)	[94.0, 156.0]	104.0	[-60.0, -41.0]	-42.0	[-60.0, -41.0]	-42.0
n3 (residues)	[10, 20]	15	[5, 7]	5		
φ3 (degrees)	[-103.0, -41.0]	-58.0	[-144.0, -70.0]	-115.0		
ψ3 (degrees)	[-60.0, -41.0]	-54.0	[94.0, 156.0]	124.0		
non-local						
σ1 (degrees)	[0.0, 90.0]	55.0	[0.0, 90.0]	47.0	[-100.0, 100.0]	-26.0
λ1 (Å)	[-1.5, 1.5]	-0.5	[-1.5, 1.5]	-1.1	[-1.8, 1.8]	0.9
σ2 (degrees)	[0.0, 90.0]	60.0	[-100.0, 100.0]	99.0	[-100.0, 100.0]	87.0
λ2 (Å)	[-1.5, 1.5]	0.5	[-1.8, 1.8]	-0.8	[-1.8, 1.8]	-0.9
σ3 (degrees)	[-100.0, 100.0]	61.0	[0.0, 90.0]	44.0		
λ3 (Å)	[-1.8, 1.8]	-0.6	[-1.5, 1.5]	-0.6		
d (Å)	[4.0, 5.0]	4.8	[4.0, 5.0]	4.8	[7.0, 11.0]	9.0
τ (degrees)	[-25.0, 25.0]	12.0	[13.0, 25.0]	15.0	[-35.0, 35.0]	-3.0
Ω (degrees)	[-20.0, 0.0]	-6.0	[-22.0, 10.0]	-20.0		
θ (degrees)	[0.0, 20.0]	9.0	[0.0, 20.0]	5.0		
h (Å)	[8.0, 11.5]	9.8	[7.0, 13.0]	8.8		
loop						
axis1 to axis2	EE loops	Eaa-2-E	EH loops	Eet_abH	HH loops	HbabaHs
axis 2 to axis 3	EH loops	EabH	HE loops	HagbaE		

Local variables ni, φi, and ψi, are the number of residues and backbone φ/ψ dihedral angles for the i-th axis, respectively. i = {1, 2, 3} for both the ββα and βαβ motifs, i = {1, 2} for the αα motif. In this notation, the i-th axis represents the i-th secondary structural element. For exmaple, in the ββα motif axis 1 is the first β-strand, axis 2 is the second β-strand, and axis 3 is the first α-helix. Non-local variables σi, λi, d, t, Ω, θ, h are the supersecondary structure parameters described in Figure 1. The loop terminologies follow that of Donate (Donate et al., 1996), where E denotes a β-strand and H denotes an α-helix. Therefore, a ββα fold will need an EE loop connecting the first two β-strands and an EH loop connecting the second β-strand and the following α-helix. The optimized values yield structures shown in Figure 3.

The atoms of the secondary structures were first generated around the z-axis (Shimanouchi & Mizushima SI, 1955; Miyazawa, 1961; Sugeta, 1967) using the ϕ and ψ angles shown in Table 1 (Flory, 1969). These atoms were then mapped to the target locations via standard Euler matrices where the matrix elements consisted of combinations of the supersecondary structure parameters (see Table 2a, b, c).

The mapped structures were then evaluated as shown in Figure 2. Clashing structures (with positive van der Waals energies) and structures where none of the appropriate loops provided closure were rejected. The shortest loops were chosen which provided closure. The cost function value was then determined for each of the remaining structures based on core-forming C_β-C_β interactions and the total number of hydrogen bonds. C_β-C_β distances were measured between the helix and each of the β-strands for the ββα and βαβ motifs, and between the two helices for the αα motif. Distances within 4.1 to 6.6 Å were counted as −1, while distances less than 4.1 Å were counted as +1. The number of hydrogen bonds between the β-strands was determined as described in Figure 2; each hydrogen bond was counted as −1. Counts were totaled to obtain the cost function value for each structure. These values were then fed into the Adaptive Simulated Annealing Algorithm (Ingber, 1993) and the structure with the lowest cost function value was determined.

The best backbone structures obtained for each of the three motifs are shown in Figure 3. The structure for the ββα motif is similar to that of a zinc finger and the structure for the βαβ motif is similar to a segment of a β-barrel (PDB entry 8adh). The backbone obtained for the αα motif resembles minimized protein A. The parameter values used to generate these backbones are listed in Table 1.

Table 2a Euler matrices used in ββα backbone construction.

$\vec{x} \rightarrow \tilde{R}\vec{x}+\vec{T}+\lambda\vec{n}$

$$\tilde{R}=\begin{Bmatrix} cos\beta cos\gamma - cos\alpha sin\beta sin\gamma, & -cos\beta sin\gamma - cos\alpha sin\beta cos\gamma, & sin\alpha sin\beta \\ sin\beta cos\gamma + cos\alpha cos\beta sin\gamma, & -sin\beta sin\gamma + cos\alpha cos\beta cos\gamma, & -sin\alpha cos\beta \\ sin\alpha sin\gamma, & sin\alpha cos\gamma, & cos\alpha \end{Bmatrix}$$

$$\vec{T}=\begin{Bmatrix} t_{mx} \\ t_{my} \\ t_{mz} \end{Bmatrix} - \tilde{R}\begin{Bmatrix} x_{mx} \\ x_{my} \\ x_{mz} \end{Bmatrix}$$

$$\vec{n}=\begin{Bmatrix} n_x \\ n_y \\ n_z \end{Bmatrix}$$

where $\alpha = tan^{-1}\frac{\sqrt{nx^2+ny^2}}{n_z}$, $\beta = \frac{\pi}{2}+cos^{-1}\frac{n_x}{sin\alpha}$, $\gamma = \sigma$.

	n_x	n_y	n_z	tm_x	tm_y	tm_z
1st β-strand	$\frac{1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	d/2	0
2nd β-strand	$\frac{-1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	-d/2	0
1st α-helix	$cos\theta cos(-\Omega)$	$cos\theta sin(-\Omega)$	$sin\theta$	0	0	h

The β-strands and α-helix are first built along the z-axis then mapped according to $\vec{x} \rightarrow \tilde{R}\vec{x}+\vec{T}+\lambda\vec{n}$, where $\vec{x}$ runs through all atoms in the strand or helix. $\vec{x}_m = \{x_{mx},\ x_{my},\ x_{mz}\}$ is the midpoint of the strand or helix, $\vec{n} = \{n_x,\ n_y,\ n_z\}$ is the unit vector along the target axis orientation, $\vec{t}_m = \{t_{mx},\ t_{my},\ t_{mz}\}$ is the desired position for $\vec{x}_m = \{x_{mx},\ x_{my},\ x_{mz}\}$ after the mapping $\vec{x} \rightarrow \tilde{R}\vec{x}+\vec{T}$. The mapping elements are computed from supersecondary structure parameters shown in Figure 1.

Table 2b (continued) Euler matrices used in βαβ backbone construction.

$\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T} + \lambda\bar{n}$

$$\tilde{R} = \begin{Bmatrix} cos\beta cos\gamma - cos\alpha sin\beta sin\gamma, & -cos\beta sin\gamma - cos\alpha sin\beta cos\gamma, & sin\alpha sin\beta \\ sin\beta cos\gamma + cos\alpha cos\beta sin\gamma, & -sin\beta sin\gamma + cos\alpha cos\beta cos\gamma, & -sin\alpha cos\beta \\ sin\alpha sin\gamma, & sin\alpha cos\gamma, & cos\alpha \end{Bmatrix}$$

$$\bar{T} = \begin{Bmatrix} t_{mx} \\ t_{my} \\ t_{mz} \end{Bmatrix} - \tilde{R} \begin{Bmatrix} x_{mx} \\ x_{my} \\ x_{mz} \end{Bmatrix}$$

$$\bar{n} = \begin{Bmatrix} n_x \\ n_y \\ n_z \end{Bmatrix}$$

where $\alpha = tan^{-1} \frac{\sqrt{nx^2 + ny^2}}{n_z}$, $\beta = \frac{\pi}{2} + cos^{-1} \frac{n_x}{sin\alpha}$, $\gamma = \sigma$.

	n_x	n_y	n_z	tm_x	tm_y	tm_z
1st β-strand	$\frac{1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$d/2$	0
1st α-helix	$cos\theta cos(-\Omega)$	$cos\theta sin(-\Omega)$	$sin\theta$	0	0	h
2nd β-strand	$\frac{1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{-tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$-d/2$	0

The β-strands and α-helix are first built along the z-axis then mapped according to $\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T} + \lambda\bar{n}$, where $\bar{x}$ runs through all atoms in the strand or helix. $\bar{x}_m = \{x_{mx},\ x_{my},\ x_{mz}\}$ is the midpoint of the strand or helix, $\bar{n} = \{n_x,\ n_y,\ n_z\}$ is the unit vector along the target axis orientation, $\bar{t}_m = \{t_{mx},\ t_{my},\ t_{mz}\}$ is the desired position for $\bar{x}_m = \{x_{mx},\ x_{my},\ x_{mz}\}$ after the mapping $\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T}$. The mapping elements are computed from supersecondary structure parameters shown in Figure 1.

Table 2c (continued) Euler matrices used in αα backbone construction.

$\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T} + \lambda\bar{n}$

$$\tilde{R} = \begin{Bmatrix} cos\beta cos\gamma - cos\alpha sin\beta sin\gamma, & -cos\beta sin\gamma - cos\alpha sin\beta cos\gamma, & sin\alpha sin\beta \\ sin\beta cos\gamma + cos\alpha cos\beta sin\gamma, & -sin\beta sin\gamma + cos\alpha cos\beta cos\gamma, & -sin\alpha cos\beta \\ sin\alpha sin\gamma, & sin\alpha cos\gamma, & cos\alpha \end{Bmatrix}$$

$$\bar{T} = \begin{Bmatrix} t_{mx} \\ t_{my} \\ t_{mz} \end{Bmatrix} - \tilde{R} \begin{Bmatrix} x_{mx} \\ x_{my} \\ x_{mz} \end{Bmatrix}$$

$$\bar{n} = \begin{Bmatrix} n_x \\ n_y \\ n_z \end{Bmatrix}$$

where $\alpha = tan^{-1} \frac{\sqrt{nx^2 + ny^2}}{n_z}$, $\beta = \frac{\pi}{2} + cos^{-1} \frac{n_x}{sin\alpha}$, $\gamma = \sigma$.

	n_x	n_y	n_z	tm_x	tm_y	tm_z
1st α-helix	$\frac{1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	d/2	0
2nd α-helix	$\frac{-1}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	$\frac{tan^2(\frac{\tau}{2})}{\sqrt{1+tan^2(\frac{\tau}{2})}}$	0	-d/2	0

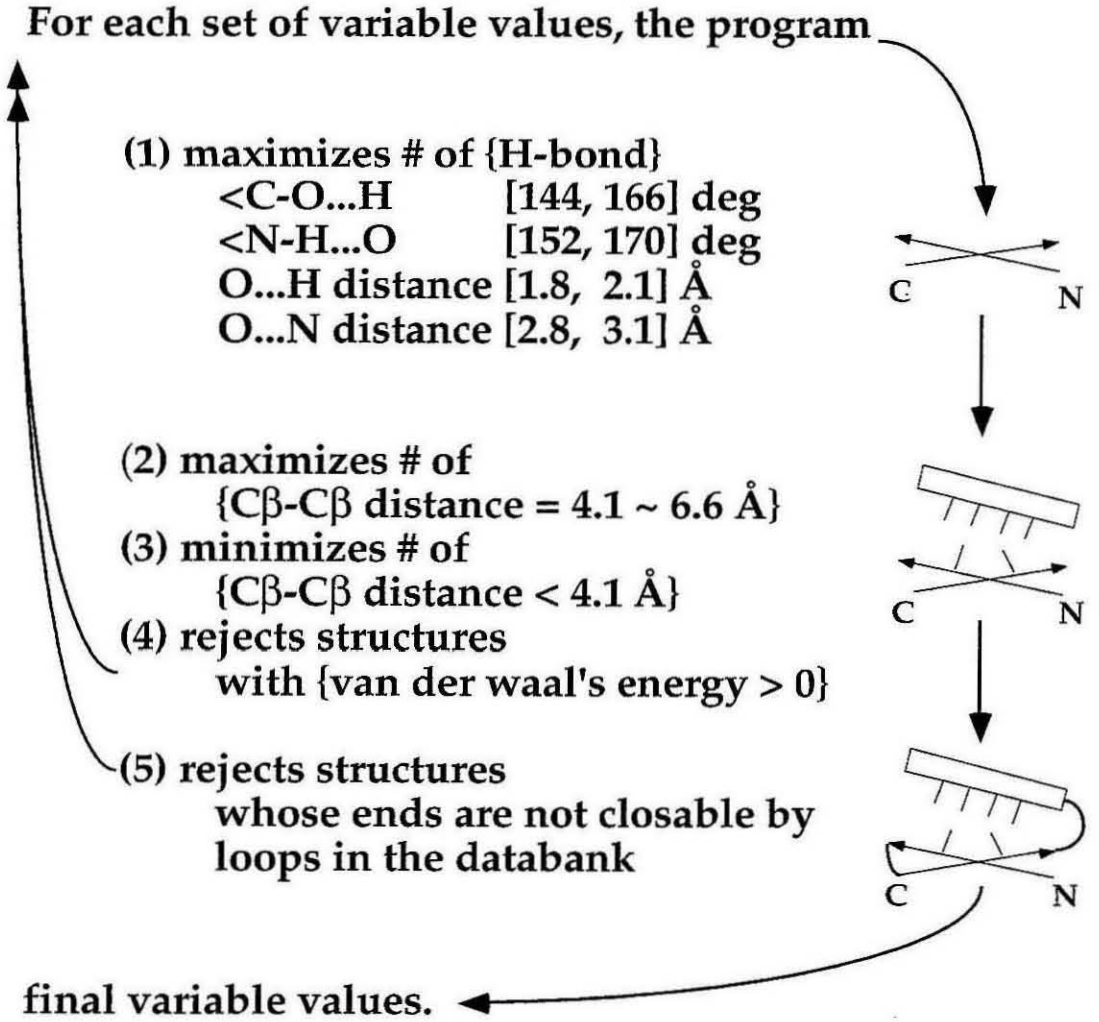
The α-helices are first built along the z-axis then mapped according to $\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T} + \lambda\bar{n}$, where $\bar{x}$ runs through all atoms in the helix. $\bar{x}_m = \{x_{mx}, \ x_{my}, \ x_{mz}\}$ is the midpoint of the helix, $\bar{n} = \{n_x, \ n_y, \ n_z\}$ is the unit vector along the target axis orientation, $\bar{t}_m = \{t_{mx}, \ t_{my}, \ t_{mz}\}$ is the desired position for $\bar{x}_m = \{x_{mx}, \ x_{my}, \ x_{mz}\}$ after the mapping $\bar{x} \rightarrow \tilde{R}\bar{x} + \bar{T}$.

The mapping elements are computed from supersecondary structure parameters shown in Figure 1.

Figure 2 Schematics of the backbone design algorithm using the $\beta\beta\alpha$ motif as an example. A set of randomly selected parameter values is used to start a simulation. The algorithm first computes atomic positions for HN, N, C, O, Cα, Cβ of a strand or helix using a single pair of backbone ϕ/ψ values. The algorithm then maps these backbone atoms into space by standard Euler matrices whose matrix elements are calculated from values of supersecondary structure parameters (see Table 2a,b,c).

Once positioned, the backbone configuration consisted of these atoms is evaluated by four terms: (1) the number of hydrogen bonds between paired strands; (2) the number of inter-segment Cβ-Cβ pairs that are within 4.1 Å to 6.6 Å of each other; (3) the number of inter-segment Cβ-Cβ pairs that are closer than 4.1 Å to each other; (4) the van der Waals interaction; (5) the existence of database loops (Donate et al., 1996) to connect the ends between two sequencely-adjacent secondary structural elements.

The cost is calculated by combining these four terms. The cost is set to a big positive number, if either the van der Waals interaction is positive or no loops from the database can successfully close the ends, otherwise it is set to the negative sum of the number of core-forming Cβ-Cβ pairs plus the number of hydrogen-bonds. The adaptive simulated annealing procedure (Ingber, 1993) was employed to optimize parameter values by minimizing the cost. The entire procedure was repeated over different lengths of secondary structural elements.



Note our upperbounds of <C-O...H (166 deg) and <N-H...O (170 deg) did not include 180 deg, which is a good geometry as evaluated by physico-chemical formulae. This is because our hydrogen bond definition is derived from the observed hydrogen bond geometries in proteins (Stickle et al., 1992) rather than from ab-initio physico-chemical calculations.

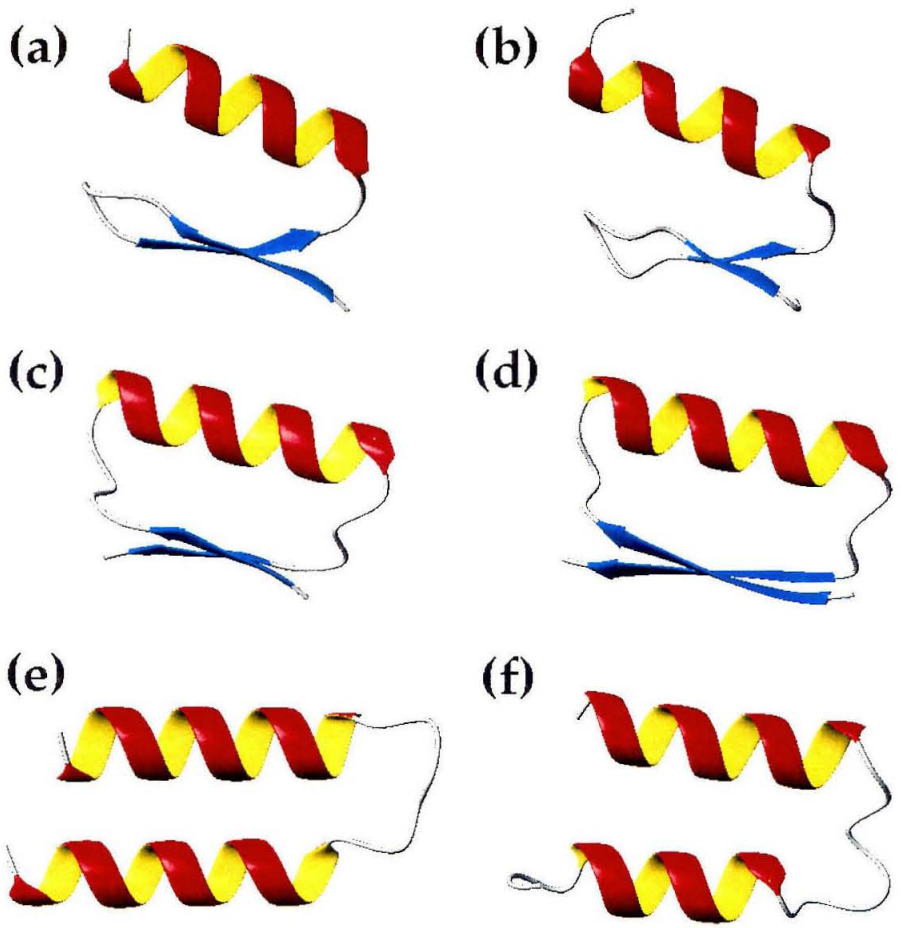


Figure 3 Optimized de novo backbones for the ββα, βαβ, and αα motifs. The structure for the ββα motif (Figure 3a) is similar to that of a zinc finger (Figure 3b) and the structure for the βαβ motif (Figure 3c) is similar to a segment of a β-barrel (PDB entry 8adh, Figure 3d). The backbone obtained for the αα motif (Figure 3e) resembles minimized protein A (Figure 3f). The parameter values used to generate these backbones are listed in Table 1.

Our backbone design algorithm is not restricted to any particular motif and allows for the possibility of novel folds. Given a ribbon diagram for an arbitrary protein, a set of coordinates can be generated describing its optimal backbone configuration. This is done as follows. N segments are defined, consisting α-helices and/or β-strands. This relationship to each other is then described. They can be connected directly by loops or associated with each other through hydrogen bonding or packing. The location of the loops is determined from the ribbon diagram and their appropriate class is selected from the protein loop database compiled by Donate (Donate et al., 1996). The ribbon diagram defines which segments need to be considered as partners for hydrogen bonding or packing and this information is input into the cost function. For each added helix, eight new variables are included (ϕ/ψ angles, number of residues, σ, λ, Ω, θ, h) and packing is considered when evaluating the cost function. For each added β-strand, seven new variables are included (ϕ/ψ angles, number of residues, σ, λ, d, τ), and both packing and hydrogen bonding are considered when evaluating the cost function. For larger systems, it may be advisable to build the backbone structure incrementally. For example, for a $\beta\alpha\beta\alpha$ motif, it would probably be best to first build the $\beta\alpha\beta$ structure and then add on the last α.

Sequence Optimization

We used the backbone coordinates computed from our backbone design algorithm as input to our automated sequence selection algorithm (Dahiyat & Mayo, 1997). Our sequence selection algorithm uses a different scoring function for each of the residue positions depending on whether they are located in the buried core, on the solvent exposed surface, or on the boundary between the core and the surface. Residue positions were classified as core,

surface, or boundary based on the extent of side-chain burial and the direction of their C_α-C_β vectors (see Figure 4 and Table 3a,b). As in our previous studies, only hydrophobic amino acids were considered at the core positions (A, V, L, I, F, Y and W), only hydrophilic amino acids were considered at the surface positions (A, S, T, H, D, N, E, Q, K, and R), and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions. For each amino acid, a discrete set of conformers (rotamers) from an expanded backbone-dependent rotamer library was used (Dahiyat & Mayo, 1997).

All residues were split into two sets and optimized separately to speed up the calculation. In all calculations, one set contained the core and boundary positions, and the other contained the remaining surface positions. For the $\beta\beta\alpha$ motif, the core and boundary set contained $3.37*10^{12}$ possible amino acid sequences corresponding to $2.91*10^{26}$ rotamer sequences; the surface set contained 10^{17} possible amino acid sequences corresponding to $2.34*10^{52}$ rotamer sequences. For the $\alpha\alpha$ motif, the core and boundary set contained $1.23*10^{14}$ possible amino acid sequences corresponding to $1.31*10^{30}$ rotamer sequences; the surface set contained 10^{23} possible amino acid sequences corresponding to $5.36*10^{67}$ rotamer sequences. For the $\beta\alpha\beta$ motif, the core and boundary set contained $5.75*10^{9}$ possible amino acid sequences corresponding to $2.10*10^{21}$ rotamer sequences; the surface set contained 10^{20} possible amino acid sequences corresponding to $1.28*10^{42}$ rotamer sequences. The two groups do not interact strongly with each other making their sequence optimization mutually independent, though there are strong interactions within each group. Each optimization was carried out with the non-optimized positions in the template set to Alanine.

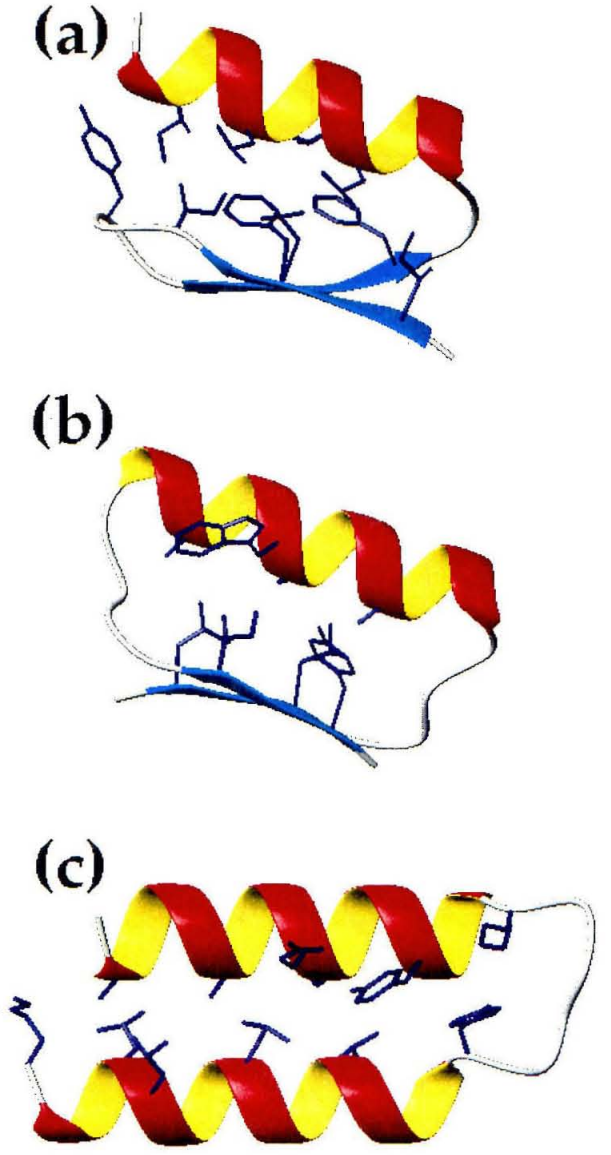


Figure 4 Optimized structures for ββα, βαβ, and αα backbone folds. (a) ββα; (b) βαβ; (c) αα. These structures have well-packed cores (core & boundary residues); this is mainly attributable to the Cβ-Cβ interaction constraint (between 4.1 Å to 6.6 Å).

Table 3a Optimized amino acid sequence for the ββα motif.

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
residue class	S	B	S	B	S	C	S	B	S	S	B	S	B	S	S	S	S	S	C	S
sequence	K	I	T	E	T	I	R	Y	K	T	F	T	F	E	R	K	E	E	L	K
position	21	22	23	24	25	26	27	28												
residue class	S	B	B	S	S	B	S	S												
sequence	T	L	L	E	R	I	E	K												

Table 3b Optimized amino acid sequence for the βαβ motif.

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
residue class	S	B	S	B	S	S	S	S	S	S	B	S	S	B	B	S	S	C	S	S
sequence	R	Y	T	I	E	G	H	D	K	K	A	D	K	W	A	K	K	A	D	K
position	21	22	23	24	25	26	27	28	29											
residue class	S	S	C	S	S	B	S	B	S											
sequence	K	G	A	D	E	L	T	L	K											

Table 3c Optimized amino acid sequence for the αα motif.

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
residue class	S	B	S	S	S	C	S	S	B	B	S	S	B	S	S	B	S	S	S	S
sequence	R	A	N	E	E	A	E	R	E	Q	K	R	Y	N	D	K	E	N	K	E
position	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35					
residue class	B	S	S	S	B	S	S	B	S	S	B	B	S	S	B					
sequence	Y	R	D	R	I	R	E	L	N	R	L	L	N	E	K					

Table 3 Residue positions were classified as core (C), surface (S), or boundary (B) based on the extent of side-chain burial and the direction of their C_α-C_β vectors. As in our previous studies, only hydrophobic amino acids were considered at the core positions (A, V, L, I, F, Y and W), only hydrophilic amino acids were considered at the surface positions (A, S, T, H, D, N, E, Q, K, and R), and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions.

The optimal sequences obtained from the two calculations were combined to yield the structures shown in Figure 4. These structure have well-packed cores; this is mainly attributable to the C_β-C_β clustering criteria used in the cost function (4.1 to 6.6 Å).

Applications

Our backbone design algorithm and our sequence selection algorithm are similar in that they both use discrete representation, scan a specified search space, and employ a scoring function (Ponders & Richards, 1987). The backbone design algorithm treats secondary structural elements as basic units and searches in the supersecondary structure parameter space, while the sequence selection algorithm treats side-chain rotamers as basic units and searches in the side-chain torsional space. Within the search space, each backbone configuration is scored by examining its core packing, hydrogen bonding, loop closure, and local backbone torsional geometries, whereas each sequence configuration is scored by evaluating its core packing, hydrogen bonding, surface area burial, and local secondary structure propensity.

Compared with previous backbone design methods, our atomistic backbone design algorithm is advanced. It provides the correct amount of information for sequence selection and applies to all classes of protein folds. These factors make it a valid backbone design tool. However, there is still room for improvement. The algorithm cannot pre-determine the optimal number of residues in each secondary structural element. A method that may provide some insight into this issue is presented in our experimental paper on this topic (Su & Mayo, 1997). Also, since the backbone torsion angles are fixed for each secondary structural element, curved helices and strands cannot be simulated. However, this problem can be eliminated by

introducing systematic variation in the backbone torsion angles as part of the design (Salemme & Weatherford 1981a, b, c; Salemme, 1983).

Our backbone design method is potentially applicable to several areas including protein minimization, protein folding and threading, and the design of novel protein folds. A major challenge in these areas is the construction of atomistic backbone templates upon which different sequences or side-chain rotamers can be impartially screened. In this paper, we describe a computational procedure capable of accomplishing this task. The experimental validation of the sequences selected using these de novo backbone templates is described in the following chapter.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol Biol 112*:535–542.

Betz SF, DeGrado WF. 1996. Controlling topology and native-like behavior of de novo-designed peptides — Design and characterization of antiparallel 4-stranded coiled coils. *Biochemistry 35*:6955–6962.

Bruccoleri RE. 1993. Application of systematic conformational search to protein modeling. *Mol Sim 10* 151-174.

Chirgadze, YN. Deduction and systematic classification of spatial motifs of the antiparallel β structure in globular proteins. *Acta Cryst A43* 405-417.

Chothia C, Levitt M, Richardson D. 1977. Structure of proteins: packing of *a*-helices and pleated sheets. *Proc Natl Acad Sci USA 74*:4130–4134.

Chothia C, Levitt M, Richardson D. 1981. Helix to helix packing in proteins. *J Mol Biol 145*:215–250.

Chou KC, Scheraga HA. 1982. Origin of the right-handed twist of β-sheets of poly(L Val) chains. *Proc Natl Acad Sci USA 79*. 7047-7051.

Chou KC, Nemethy G, Scheraga HA. 1983. Effect of amino acid composition on the twist and the relative stability of parallel and antiparallel β-sheets. *Biochemistry 22* 6213-6221.

Cohen FE, Sternberg MJE, Taylor WR. 1981. Analysis of the tertiary structure of protein *b*-sheet sandwiches. *J Mol Biol 148*:253–272.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.

Dahiyat BI, Mayo SL. 1997. De novo protein design - fully automated sequence selection. *Science 278* 82-87.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci* 4:2006–2018.

Dill KA, Stigter D. 1995. Modeling protein stability as heteropolymer collapse. *Adv Protein 46* 59-104.

Donate LE, Rufino SD, Canard LHJ, Blundell TL. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database of modeling and prediction. *Protein Sci 5* 2600-2616.

Efimov, AV. 1993a. Standard structures in proteins. *Prog Biophys molec Biol 60* 201-239.

Efimov, AV. 1993b. Patterns of loop regions in proteins. *Curr Op Str 3* 379-384.

Efimov, AV. 1994. Favoured structural motifs in globular proteins. *Structure 15* 999-1002.

Efimov, AV. 1997. Structure trees in protein superfamilies. *Proteins 28* 241-260.

Flory PJ. 1969. Statistical mechanics of chain molecules. *John Wiley & Sons, Inc.*

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA* 92:8408–8412.

Hellinga HW, Caradonna JP, Richards FM. 1991. Construction of new ligand-binding sites in proteins of know structure 2. Grafting of buried transition-metal binding site into *Escherichia coli* thioredoxin. *J Mol Biol* 222:787–803.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA* 91:5803–5807.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224:*1142–1154.

Ingber L. 1993. Simulated annealing: Practice versus theory. *J. Mathl. Comput. Modelling. 18(11):*29-57.

Ippolito JA, Alexander RS, Christianson DW. 1990. Hydrogen bond stereochemistry in protein structure and function. *J Mol Biol 215* 457-471.

Janin J, Chothia C. 1980. Packing of *a*-helices onto *b*-pleated sheets and the anatomy of *a/b* proteins. *J Mol Biol 143:*95–128.

Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. 1995. Novel metal-binding proteins by design. *Nature Structure Biol 2:*368–373.

McDonald IK, Thornton JM. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol 238* 777-793.

Miyazawa T. 1961. Molecular vibrations and structure of high polymers. II. helical parameters of infinite polymer chains as functions of bond lengths, bond angles, and internal rotation angles. *J Pol Sci 55* 215-231.

Murzin AG, Lesk AM, Chothia C. 1994a. Principles determining the structure of *b*-sheet barrels. I. A theoretical analysis. *J Mol Biol 236:*1369–1381.

Murzin AG, Lesk AM, Chothia C. 1994b. Principles determining the structure of *b*-sheet barrels. II. The observed structures. *J Mol Biol 236:*1382–1400.

Nautiyal S, Woolfson DN, King DS, Alber T. 1995. A designed heterotrimeric coiled coil. *Biochemistry 34:*11645–11651.

Park BH, Levitt M. 1995. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol 249* 493-507.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193:*775–791.

Salemme FR, Weatherford DW. 1981a. Conformational and geometrical properties of β-sheets in proteins I. Parallel β-sheets. *Mol Biol 146* 101-117.

Salemme FR, Weatherford DW. 1981b. Conformational and geometrical properties of β-sheets in proteins II. Antiparallel and Mixed β-sheets. *Mol Biol 146* 119-141.

Salemme FR, Weatherford DW. 1981c. Conformational and geometrical properties of β-sheets in proteins III. Isotropically Stressed Configuration. *Mol Biol 146* 143-156.

Salemme FR. 1983. Structural properties of protein β-sheets. *Prog Biophys molec Biol 42* 95-133.

Shimanouchi T, Mizushima SI. 1955. On the helical configuration of a polymer chain. J Chem Phys 23 707-711.

Stickle DF, Presta LG, Dill KA, Rose GD. 1992. Hydrogen bonding in globular proteins. *J Mol Biol 226* 1143-1159.

Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6* 1701-1707.

Sugeta H, Miyazawa T. 1967. General method for calculating helical parameters of polymer chains from bond lengths, bond angles, and internal rotation angles. *Biopolymers 5* 673-679.

Chapter 4

Assembling De Novo Backbone Templates for Amino Acid Sequence Selection in Protein Design - (II) Experiment

Abstract

The automation of de novo protein design can be reduced to two steps: computing a desired tertiary structure and finding the amino acid sequence that will stablize this fold. Regarding the second step, several groups (Ponder & Richards, 1987; Hellinga et al., 1991; Hurley et al., 1992; Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Harbury et al., 1995; Klemba et al., 1995; Nautiyat et al., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996) have successfully developed computer algorithms capable of selecting the best sequence for backbone templates chosen from the Protein Data Bank (Bernstein et al., 1977). However, less has been done about the first step (Su & Mayo, 1997). In the previous theory paper, we proposed a computational method to tackle the first step. The method computes atomic resolution backbones compatible with the atomistic sequence selection algorithm we have been using (Dahiyat & Mayo, 1997). In this experimental paper, we evaluate our backbone design method by testing the thermal stability and structural properties of the designed peptides. We also explored relevant issues for integrating computer-generated backbones with the sequence-selection algorithm. Starting with a computer-generated ββα motif, we optimized five sequences for this backbone and characterized them using CD and 1D NMR. It was found that small differences in the number and location of the hydrophobic residues can significantly change the thermodynamic behaviour of the designed peptides. This supports the previously acknowledged importance of binary patterning (Hecht et al., 1990). We compared our five sequences with FSD-1, a thermally stable ββα protein with a well-defined structure developed in our lab using the same sequence-selection algorithm (Dahiyat & Mayo, 1997). We found that the sequence that had the binary pattern most similar to FSD-1 also had the best

thermodynamic properties. Based on the results of these five peptides, a set of heuristic rules was derived which could be used to improve the computer-generated backbones. In addition to the number and location of hydrophobic residues, we found that it is critical to determine the optimal number of residues for each secondary structural element. Validation of these rules will be the focus of future experimental efforts.

Introduction

The problem of de novo protein design can be reduced to two steps: selecting a desired tertiary structure and finding a sequence that will stablize this fold. The automation of the second step has been successfully accomplished by Dahiyat and Mayo (Dahiyat & Mayo, 1997). In their recent work, a sequence selection algorithm was developed to find the optimal sequence for a given backbone. However, all of these protein design algorithms use backbone coordinates selected from the Protein Data Bank as input (Bernstein et al, 1977). We sought to broaden the scope of these methods by using computationally-generated backbones (Chapter 4) as input. If successful, this would allow novel backbone folds to be included as part of the protein design process.

We developed an algorithm to compute backbones which uses supersecondary structure parameters and selects a family of optimal configurations based on a cost function. This algorithm is described in the Chapter 4. A family of optimal configurations was determined for each of the three motifs. In order to validate our backbone design algorithm we decided to measure the thermal stability and structural properties of one representative of the family of optimal configurations using CD and 1D NMR. Initially we tested the $\beta\beta\alpha$ motif since this motif had been used to validate the sequence selection algorithm and we knew that it produced a thermally stable and well-ordered structure (Dahiyat & Mayo, 1997).

If the measurements on our representative configuration do not indicate a sufficiently stable or well-defined structure, it does not necessarily imply that our backbone design algorithm is invalid. It may merely mean that additional factors must be taken into consideration. For example, binary patterning (Hecht et al., 1990) and the relative length of the secondary

structural elements may also play a role. If we could refine our representative configuration by modifying its binary patterning, new sequences could be generated which have a different number or location of hydrophobic residues. Ultimately the results of these modifications could be used as feedback to improve our backbone design algorithms (Chapter 4).

Results and Discussion

Initial configuration results. The thermal stability and structural properties of the representative configuration for the ββα motif (bba1, see Figure 1 & Table 1a) were measured using CD and 1D NMR. In order to characterize it at its most stable solution state, we first performed a pH scan. As the pH was increased from 3.9 to 7.1, the CD spectrum changed from a random coil to a more structured state (see Figure 2a). pH 7.1 was therefore used to measure its thermal reversibility and perform 1D NMR. As shown in Figure 2b, the peptide was thermally irreversible. Its NMR spectrum showed a broad baseline and ill-dispersed peaks (Figure 2c). These characteristics suggested that our initial configuration was unstable and prone to aggregation. This might be improved by reducing the overall hydrophobicity of the peptide. We also noticed that the residues of the first strand did not follow the typical alternating hydrophobic/hydrophilic pattern. Position 4 was hydrophilic but should have been hydrophobic. We therefore decided to generate new sequences which incorporated these modifications to see if we could produce a more stable and well-ordered protein.

New sequence design. The binary patterning of the first β-strand of our initial configuration was altered by replacing the hydrophilic residue in position 4 with a hydrophobic one. That is, the sequence selection algorithm was required to select one of the hydrophobic residues for this position. The

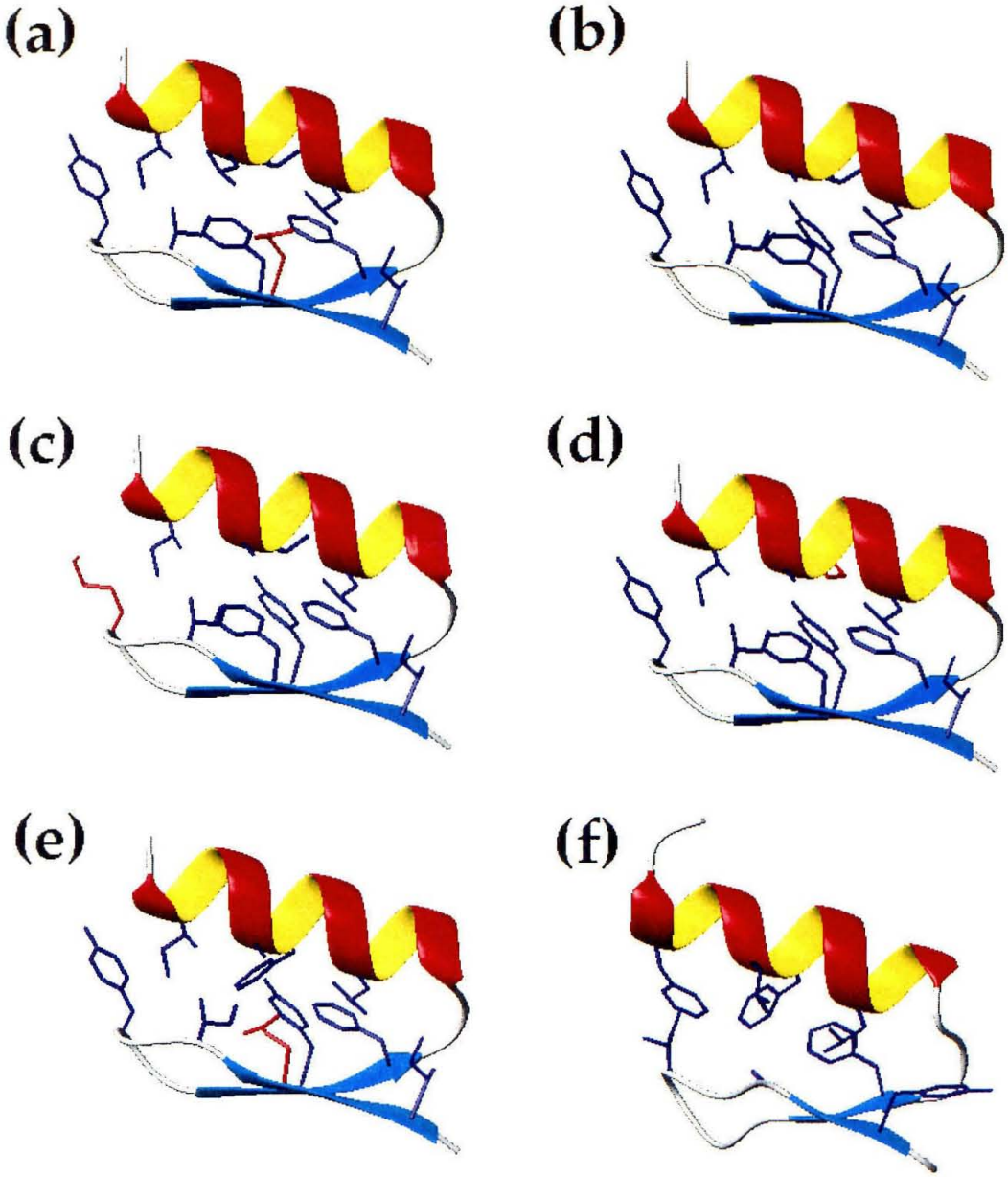


Figure 1 Structures of (a) bba1; (b) bba2; (c) bba3; (d) bba4; (e) bba5; (f) fsd1. Internal (core & boundary) side-chains were shown. Red side-chains indicated positions converted from hydrophobic to hydrophilic with respect to bba2's internal side-chain composition.

Table 1a Optimized amino acid sequences.

	1									1									2							
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8
bba1	K	I	T	E	T	I	R	Y	K	T	F	T	F	E	R	K	E	E	L	K	T	L	L	E	R	I	E	K
bba2	K	I	T	Y	T	I	E	Y	K	T	F	T	F	E	R	K	E	E	L	K	D	L	A	E	K	I	K	R
bba3	K	I	T	Y	T	I	E	K	K	T	F	T	F	K	N	K	R	E	L	K	E	L	A	E	R	I	E	R
bba4	K	I	T	Y	T	I	E	Y	K	T	F	T	F	K	N	K	R	E	L	K	E	K	A	E	E	I	K	R
bba5	K	I	T	Y	T	I	E	Y	K	T	E	T	F	D	N	K	R	N	L	K	E	F	A	E	R	I	K	R
fsd1	Q	Q	Y	T	A	K	I	K	G	R	T	F	R	N	E	K	E	L	R	D	F	I	E	K	F	K	G	R

The amino acid sequences were optimized according to methods described in Chapter 3.

Table 1b Residue classification for each position.

	1									1									2							
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8
bba1	S	B	S	B	S	C	S	B	S	S	B	S	B	S	S	S	S	S	C	S	S	B	B	S	S	B	S	S
bba2	S	C	S	C	S	C	S	C	S	S	C	S	C	S	S	S	S	S	C	S	S	B	B	S	S	B	S	S
bba3	S	C	S	C	S	C	S	S	S	S	C	S	C	S	S	S	S	S	C	S	S	B	B	S	S	B	S	S
bba4	S	C	S	C	S	C	S	S	C	S	C	S	C	S	S	S	S	S	C	S	S	S	B	S	S	B	S	S
bba5	S	C	S	C	S	C	S	C	S	S	S	S	C	S	S	S	S	S	C	S	S	B	B	S	S	B	S	S
fsd1	S	S	B	S	C	S	B	S	S	S	S	B	S	S	S	S	S	B	S	S	B	B	S	S	B	S	S	S

S = surface, B = boundary, C = core residues. The residues were classified according to methods described in Chapter 3.

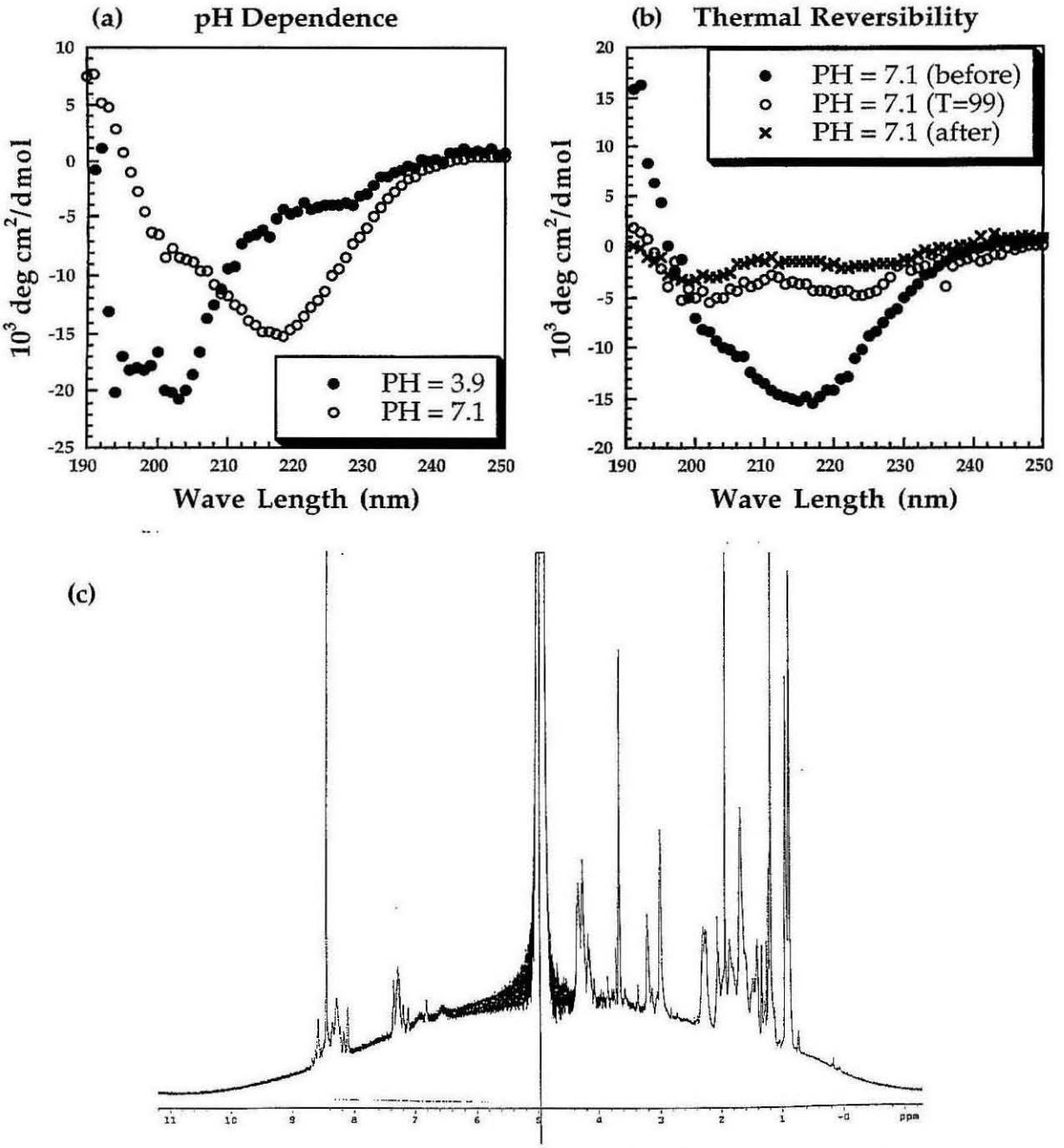


Figure 2 bba1's CD and 1D NMR spectra. (a) pH scan from pH 3.9 to 7.1; (b) thermal reversibility measured at pH 7.1; (c) 1D NMR spectrum showed a broad baseline and ill-dispersed peaks.

sequence generated (bba2) is shown in Figure 1. A pH scan, thermal reversibility test and 1D NMR were again performed. At the optimal pH, 6.8, the structure was again found to be thermally irreversible and its 1D NMR spectrum still exhibited broad baseline and ill-dispersed peaks (Figure 3a,b,c). These results suggested that designed sequences were too hydrophobic. This led us to design two additinal sequences, bba3 and bba4, where the hydrophobicity was reduced.

We kept the modification made in bba2 and made two additional modifications in each new sequence. We first attempted to improve the stability of the α-helix by restricting position 15 to residues that have high N-cap propensity (D, N, S, T) (Aurora & Rose, 1998). This was done given both bba3 and bba4. We then increased the overall hydrophilicity of the structures. In the first one, bba3, we replaced the hydrophobic residue in position 8 of the β-turn with a hydrophilic one. This position was chosen since its Cα-Cβ vector points slightly outward and the residue does not strongly interact with the hydrophobic core. The same thing was done at a different location to generate bba4. The hydrophobic residue in position 22 of the α-helix was changed to a hydrophilic one. This location was again chosen since its Cα-Cβ vector points outward, away from the hydrophobic one.

A pH scan, thermal reversibility test, and 1D NMR spectrum were again performed on bba3. As the pH increased from 4.0 to 10.0, the CD signal changed from a random coil (pH 4.0 and 5.6), to a more structured state (pH 6.7 and 8.1), and finally to an aggregated form (pH 10.0) (Figure 4a). At the optimal pH, 6.7, TFE was added to further stablize the peptide. As the concentration of TFE was increased from 0% to 20%, bba3 became more structured but was still thermally irreversible (Figure 4b). 1D NMR spectra were acquired both with and without TFE. The sample with TFE showed an

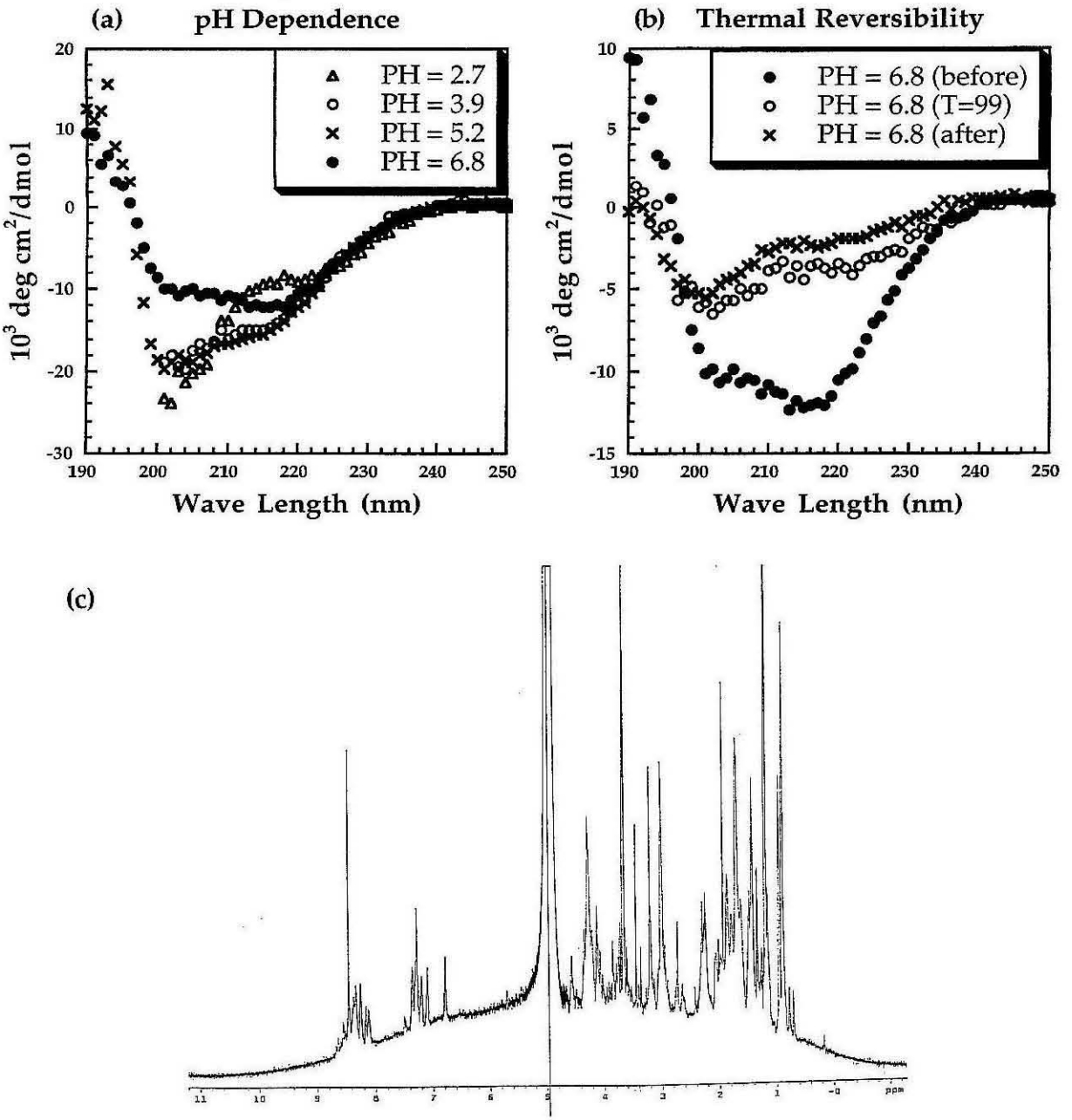


Figure 3 bba2's CD and 1D NMR spectra. (a) pH scan from pH 2.7 to 6.8; (b) thermal reversibility measured at pH 6.8; (c) 1D NMR spectrum showed a broad baseline and ill-dispersed peaks.

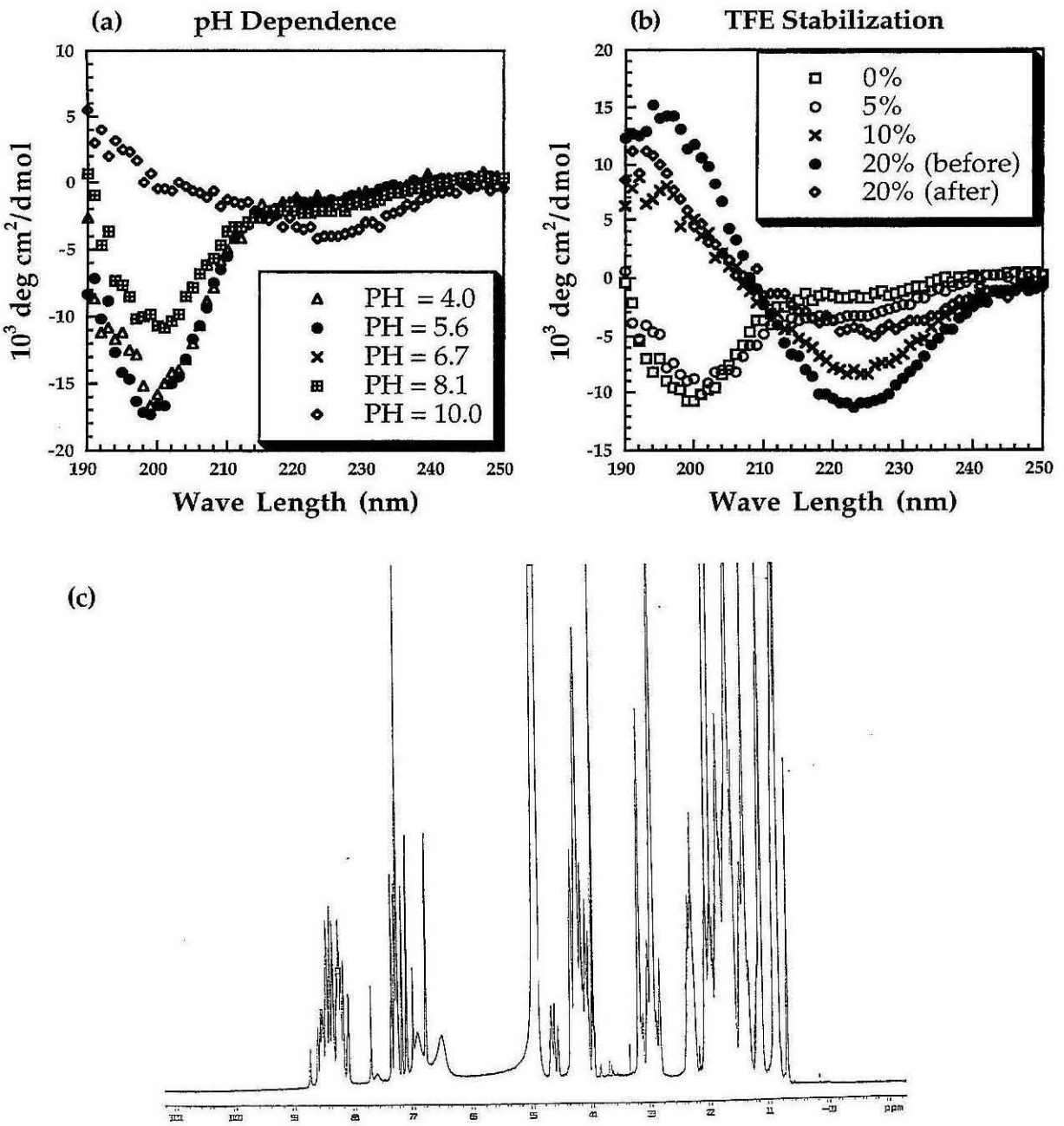


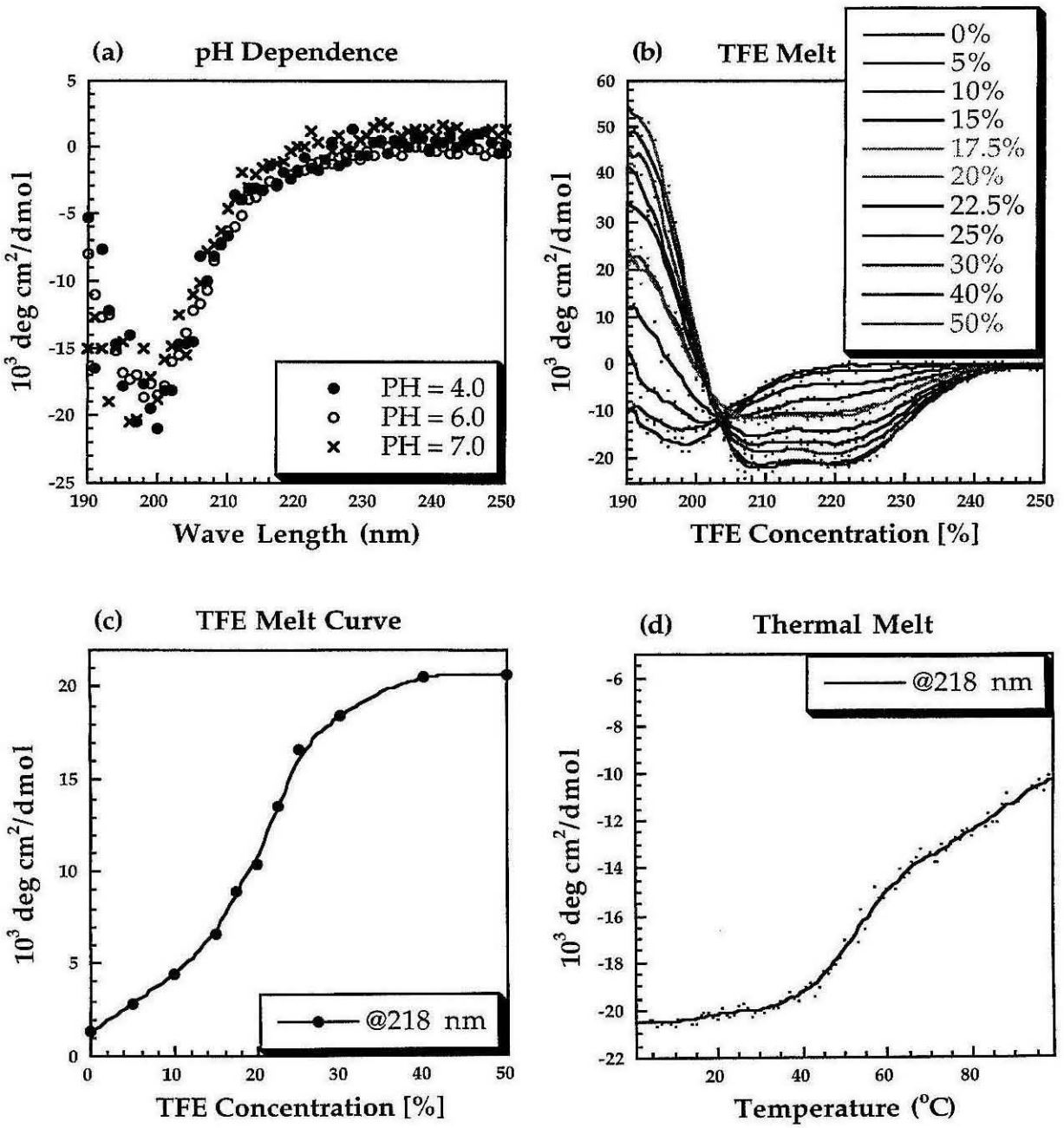
Figure 4 bba3's CD and 1D NMR spectra. (a) pH scan from pH 4.0 to 10.0; (b) TFE stablization at pH 6.7, thermal melt is still irreversible at 20% TFE; (c) 1D NMR spectrum showed more dispersion than bba1 and bba2 and did not have broad baseline, suggesting an improved overall hydrophobicity.

aggregated spectrum (not shown). The sample without TFE showed more dispersion than bba1 and bba2 and did not have a broad baseline, suggesting an improved overall hydrophobicity (Figure 4c).

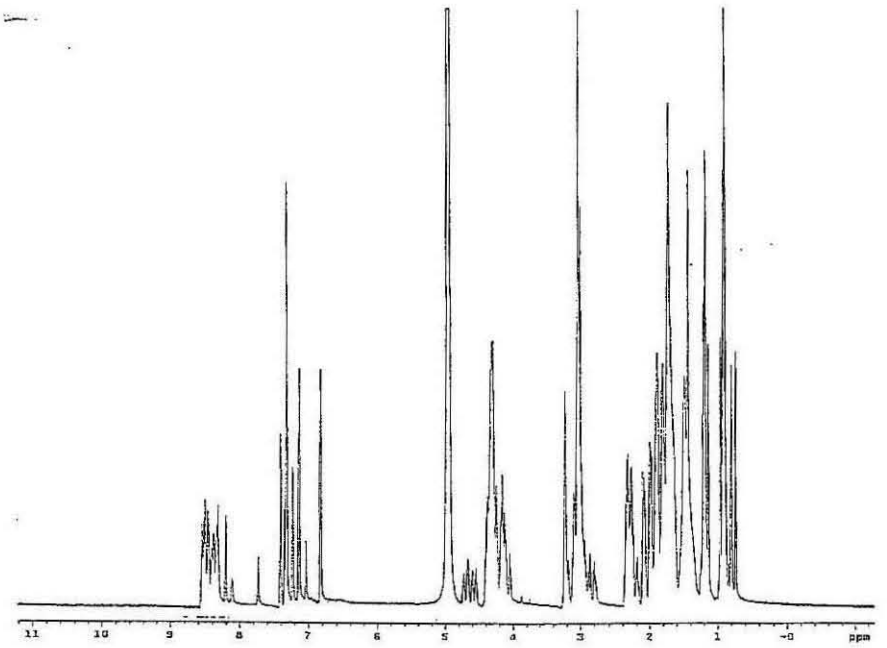
As for bba4, the CD spectrum remained random-coil like for pH values ranging from 4.0 to 7.0 (Figure 5a). At pH 6.0, TFE was added to further stablize the peptide. To determine the minimum TFE concentration that would enable the peptide to exhibit the most structure, a TFE-melt was conducted. We found that 40% TFE was the minimum concentration where the CD spectrum exhibited the most structure and was still thermally reversible (Figure 5b). We extracted ΔG from the TFE melt curve and obtained a value of 4.588 kcal (Figure 5c). This suggested that an extra 4.588 kcal would be necessary to stabilize the peptide in pure buffer (Luo & Baldwin, 1997). A thermal melt was done on bba4 at 40% TFE. It showed that bba4 was only weakly cooperative (Figure 5d). 1D NMR spectra were again acquired for bba4 both with and without TFE. In pure buffer the spectra showed a random-coil pattern (Figure 5e). However in 40% TFE, the more dispersion was obtained (Figure 5f).

Developing guidelines for binary patterning. We decided to analyze previous results to see if guidelines could be developed to further improve the designed peptides' stability. These guidelines focused optimizing the peptide's binary patterning. For the 28-residue $\beta\beta\alpha$ peptide, we optimized (a) the number of residues in each secondary structural element, (b) the balance between hydrophobic and hydrophilic residues, (c) the number of hydrophobic residues, and (d) the location of hydrophobic residues. We applied these above four guidelines to generate a new peptide, bba5. We describe individual guidelines and the stability of bba5 as follows.

Figure 5 bba4's CD and 1D NMR spectra. (a) bba4's CD spectrum remained random-coil like for pH values ranging from 4.0 to 7.0; (b) TFE melt conducted at pH 6.0 to determine the minimum TFE concentration that would enable the peptide to exhibit the most structure, 40% TFE was found; (c) ΔG extracted from the TFE melt curve was 4.588 kcal, suggesting an extra 4.588 kcal would be necessary to stabilize the peptide in pure buffer (Luo & Baldwin, 1997); (d) a thermal melt was done on bba4 at 40% TFE. It showed that bba4 was only weakly cooperative. Furthermore, the shape of the melting curve suggested the possibility of intermediates.


(a) pH Dependence
10^3 deg cm^2/dmol
Wave Length (nm)
PH = 4.0
PH = 6.0
PH = 7.0
(b) TFE Melt
10^3 deg cm^2/dmol
TFE Concentration [%]
0%
5%
10%
15%
17.5%
20%
22.5%
25%
30%
40%
50%
(c) TFE Melt Curve
10^3 deg cm^2/dmol
TFE Concentration [%]
@218 nm
(d) Thermal Melt
10^3 deg cm^2/dmol
Temperature (°C)
@218 nm

(e)



(f)

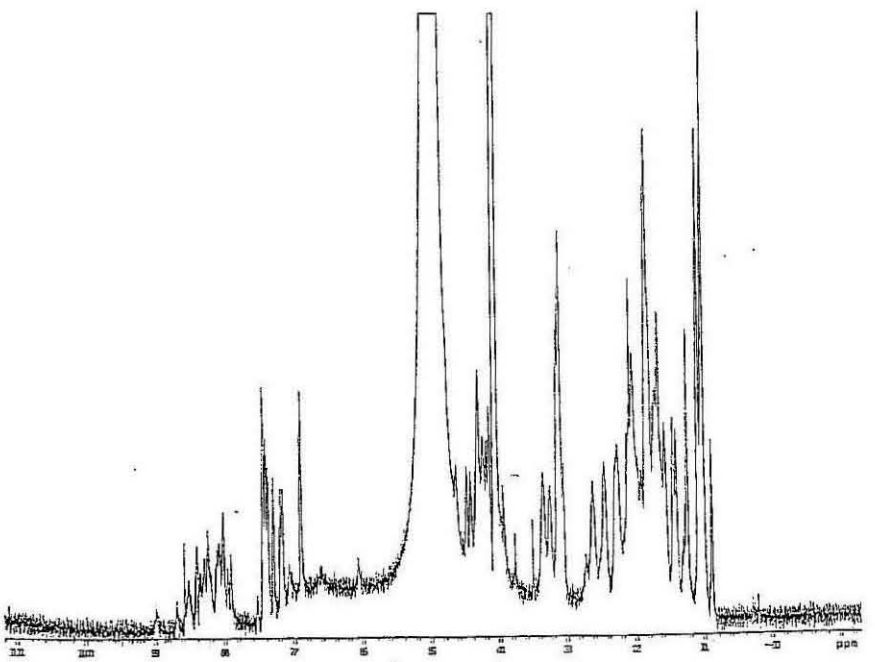


Figure 5 (continued) bba4's CD and 1D NMR spectra. 1D NMR spectra were acquired for bba4 both with and without TFE. (e) in pure buffer the spectra showed a random-coil pattern; (f) however in 40% TFE, more dispersion was obtained.

(a) Determine the number of residues in each secondary structural element. The number of residues in each secondary structural element can be determined by considering the geometry of the ββα fold. That is, the relative orientation of the β-strands and the α-helix determines the length of the secondary structural element. As shown in Figure 6, the 28-residue sequence is first partitioned into 6 segments: the N-terminal residues, the first β-strand, the second β-strand, the β–α loop, the α-helix, and the C-terminal residues. First we assumed that the number of residues at the N and C termini match the periodicity of the adjacent secondary structural elements. For the ββα motif, this leads to two residues at the N terminus and four residues at the C terminus. Secondly, we chose the shortest β–α loop from the databank (Donate et al., 1996) that could properly align the α-helix on top of the β-sheet. In this case, the shortest loop has two residues. Next we optimized the number of residues for each of the β-strands and the α-helix. Since the β-strands were in an anti-parallel β-sheet, they were assumed to be the same length. The length of the β-strand and α-helix segments can be determined from the following equations:

$$2B + A = L - \text{N terminal residues} - \text{C terminal residues} - \beta\text{–}\alpha \text{ loop}$$
$$= 28 - 2 - 4 - 2 = 20; \quad (1)$$

$$L(A) = 1.5A; \quad (2)$$

$$L(B) = 3.3B; \quad (3)$$

$$L(A) * \cos(\theta) = L(B) * \cos(\tau/2.0) = L(A) * \cos(20°) = L(B) * \cos(10°) \quad (4)$$

where B is the number of residues in each β-strand, A is the number of residues in the α-helix, L is the total length of the peptide (= 28), L(A) is the length of the α-helix, L(B) is the length of the β-strands, θ is the angle between the α-helix and the averaged direction of the β-strands, τ is the twist angle between the two β-strands, 3.3 Å and 1.5 Å were used as the average Cα-Cα

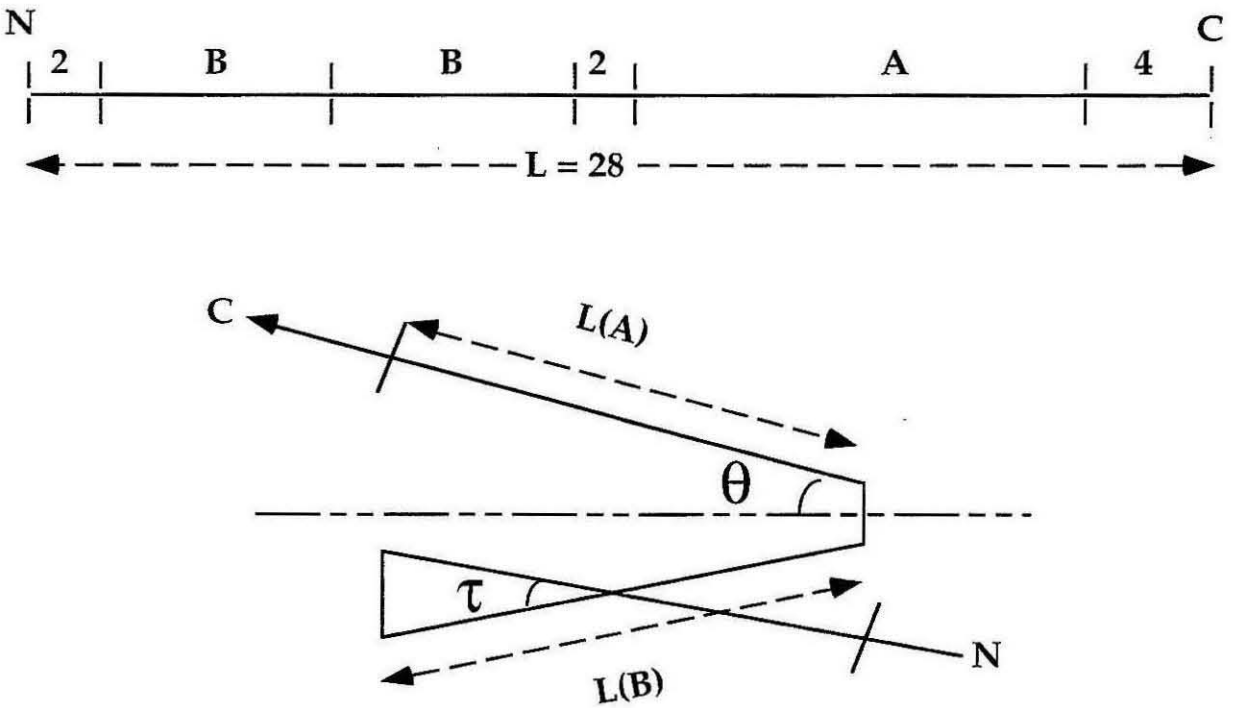


Figure 6 The 28-residue sequence is first partitioned into 6 segments: the N-terminal residues (2 residues long), the first β-strand (B residues long), the second β-strand (B residues long), the β–α loop (2 residues long), the α-helix (A residues long), and the C-terminal residues (4 residues long). The length of the β-strand and α-helix segments can be determined from the following equations (see the text for the descriptions of the following variables):

$$2B + A = L - \text{N terminal residues} - \text{C terminal residues} - \beta\text{–}\alpha \text{ loop}$$
$$= 28 - 2 - 4 - 2 = 20; \quad (1)$$

$$L(A) = 1.5A; \quad (2)$$

$$L(B) = 3.3B; \quad (3)$$

$$L(A) * \cos(\theta) = L(B) * \cos(\tau/2.0) = L(A) * \cos(20^\circ) = L(B) * \cos(10^\circ) \quad (4)$$

Solving these equations simultaneously, we get B = 5 and A = 10.

distances for an α-helix and a β-strand, and θ and τ were both approximated to be 20 degrees. Solving these equations simultaneously, we get B = 5 and A = 10. Therefore, out of the 28 residues in the ββα fold, the first 12 residues will belong to the β-strands (2B + N terminal residues), the next 2 residues will belong to the β–α loop, and the last 14 residues will belong to the α-helix (A + C terminal residues).

(b) Determine the balance of hydrophobic and hydrophilic residues. The overall hydrophobicity had strongly affected the thermal stabilities of bba1 to bba4. We quantitate the hydrophobicity of the four designed peptides. We partitioned the 20 amino acids into five hydrophobicity classes based on Radzicka and Wolfenden (Radzicka & Wolfenden, 1988), where the scores ranged from +2 to -2 with +2 being the most hydrophilic. Class one (++), with its hydrophobicity greater than 10 kcal/mol, consisted of only one residue, Arginine. Class 2 (+), with its hydrophobicity betwen 5 to 10 kcal/mol, consisted of residues Aspartic Acid, Glutamic Acid, Asparagine, Lysine, Glutamine, and Histidine. Class 3 (==), with its hydrophobicity betwen 2 to 5 kcal/mol, consisted of residues Serine and Threonine. Class 4 (-), with its hydrophobicity betwen -3 to 2 kcal/mol, consisted of residues Tyrosine, Cystidine, Alanine, Trptophan, Methionine, Phenylalanine, and Glycine. Class 5 (--), with its hydrophobicity betwen less than -3 kcal/mol, consisted of residues Valine, Isoleucine, and Leucine. The final hydrophobicity score for each peptide was calculated by summing over the five classes (the number of residues in each class was multiplied by the integer value associated with each class, then these values were totalled). For example, bba1 has 3 residues in class 1, 11 residues in class 2, 5 residues in class 3, 3 residues in class 4, and 6 residues in class 5. This leads to a hydrophobicity score of 3 * 2 + 11 * 1 + 5 * 0 + 5 * (-1) + 3 * (-2) = 2 for bba1. Table 2 shows the Hydrophobicity scores

Table 2 Hydrophobicity scores calculated for all peptides.

class	class members	score	fsd1	bba1	bba2	bba3	bba4	bba5
++ (hydrophiblic)	R	2	4	3	2	3	2	r
+ (hydrophilic)	DENKQH	1	12	11	12	12	13	[28-x-r-(s+t)]; t = 4
== (neutral)	ST	0	2	5	4	4	4	(s+t); t = 4
- (hydrophobic)	YCAWMFG	-1	7	3	5	4	5	(1-α)x
-- (hydrophobic)	VIL	-2	3	6	5	5	4	αx
hydrophobic score			7	2	-1	4	4	must be >=4

The final hydrophobicity score for each peptide was calculated by summing over the five classes (the number of residues in each class was multiplied by the integer value associated with each class, then these values were totalled). For example, bba1 has 3 residues in class 1, 11 residues in class 2, 5 residues in class 3, 3 residues in class 4, and 6 residues in class 5. This leads to a hydrophobicity score of 3 * 2 + 11 * 1 + 5 * 0 + 5 * (-1) + 3 * (-2) = 2 for bba1. Sequences with net hydrophobicity scores less than 4 were too hydrophobic and hence thermally irreversible.

calculated for the four peptides. Sequences with net hydrophobicity scores less than 4 were too hydrophobic and hence thermally irreversible.

(c) Determine the number of hydrophobic residues. We estimated the number of hydrophobic residues by combining guidelines (a) and (b). Guideline (a) suggested each β-strand consisted of 5 residues. In order to maximize the interactions between the β-sheet and the α-helix interface, we assumed positions 1, 3, 5 along the strands to have their Cα-Cβ vectors pointing towards the helix, positions 2 and 4 to have their Cα-Cβ vectors pointing away from the helix. These two positions were therefore classified as surface positions. Because of the use of intrinsic β-sheet forming propensities in this calculation, four Threonines were specified.

Since guideline (b), which requires a total hydrophobic score >= 4, we can estimate the number of hydrophobic residues, x. As shown in Table 2, the number of residues in class 1 through 5 can be described as follows: r, 28 - x - r - (s + t), s + t, (1 - α) * x, α * x, where r is the number of Arginines, s is the number of Serines, t = 4 is the number of Threonines, and α is a number between 0 and 1 designating the fraction of x belonging to Class 5. The number of residues in each class added together gives us the total length for the peptide:

$$r + [28 - x - r - (s + 4)] + [s + 4] + [(1 - \alpha) * x] + [\alpha * x] = 28 \qquad (5)$$

Applying guideline (b), multiplying the appropriate score for each class we obtain:

$$[r] * 2 + [(28 - x - r - s - t)] * 1 + [(s + t)] * 0 + [(1 - \alpha) * x] * (-1) + [\alpha * x] * (-2) >= 4 \qquad (6)$$

This sets an upperbound (20 + r - s) / (2 + α) of hydrophobic residues for the sequences compatible with guidelines (a) and (b). This upperbound depends on the amino acid composition in the sequence. More Arginines allow more

hydrophobic residues, more Serines decrease the room for hydrophobic residues, and more hydrophobic residues like Leucine, Isolucine, or Valine reduce the overall number for hydrophobic residues. Since bba3 and bba4 possessed appropriate hydrophobic scores (= 4), we can use their averaged values for r, s, and α to etimate x ($r = 2.5$, $s = 0$, $\alpha = 0.5$). Substituting these values in the equation, we therefore obtain 9 residues as the upperbound.

(d) Determine the location of hydrophobic residues. We estimated the location of the hydrophobic residues by combining guidelines (a) and (c). Guideline (a) suggested 5 residues for each β-strand, and 10 residues for the α-helix, exclusive of N and C terminal residues. As described in guideline (c), positions 1, 3, 5, of the β-strands were assumed to have Cα-Cβ vectors pointing towards the α-helix, thus forming the hydrophobic core. This implies a total of 6 hydrophobic residues for the anti-parallel β-sheet. In addition, we can expect a total of 4 hydrophobic residues in the α-helix based on its periodicity, as shown in Figure 7. Together, this amounts to a total of 10 residues, exceeding the upperbound of 9 hydrophobic residues as derived in guideline (c). Therefore, at least one hydrophobic position had to be converted to hydrophilic. The question was which one.

It wouldn't be wise to further reduce the number of hydrophobic residues in the α-helix, as it already has a much lower density (4/10) than the β-strand (6/10). Furthermore, when only 3 hydrophobic residues in the α-helix were used, the resulting peptide bba4 needed co-solvent TFE to induce structure, suggesting that three hydrophobic residues were inadequate for helix stabilization. We therefore considered hydrophobic positions in the β-sheet. Taking into account the right-handed twist of a β-strand, namely, the progressively inward-pointedness of Cα-Cβ vectors when viewed from the N

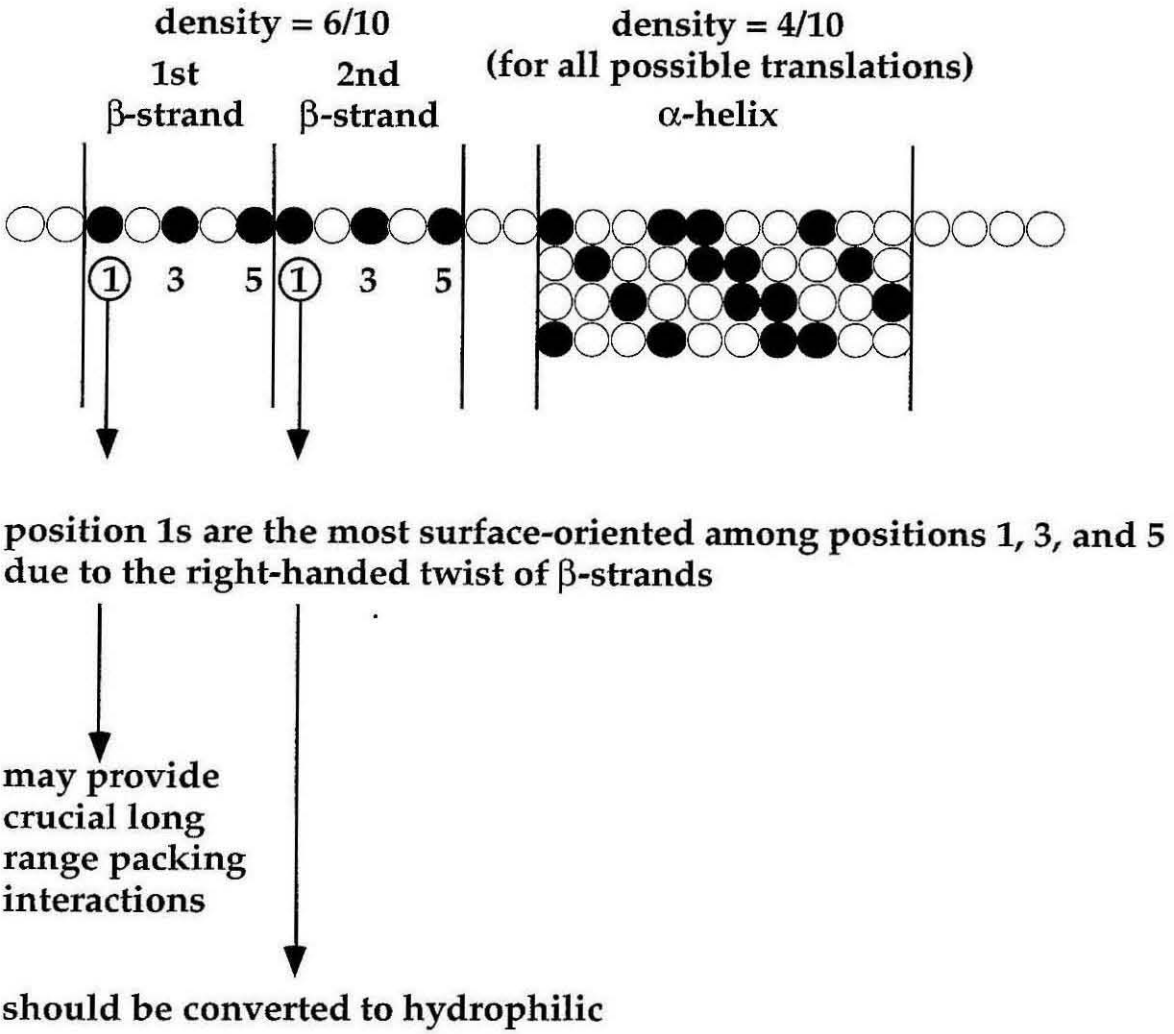
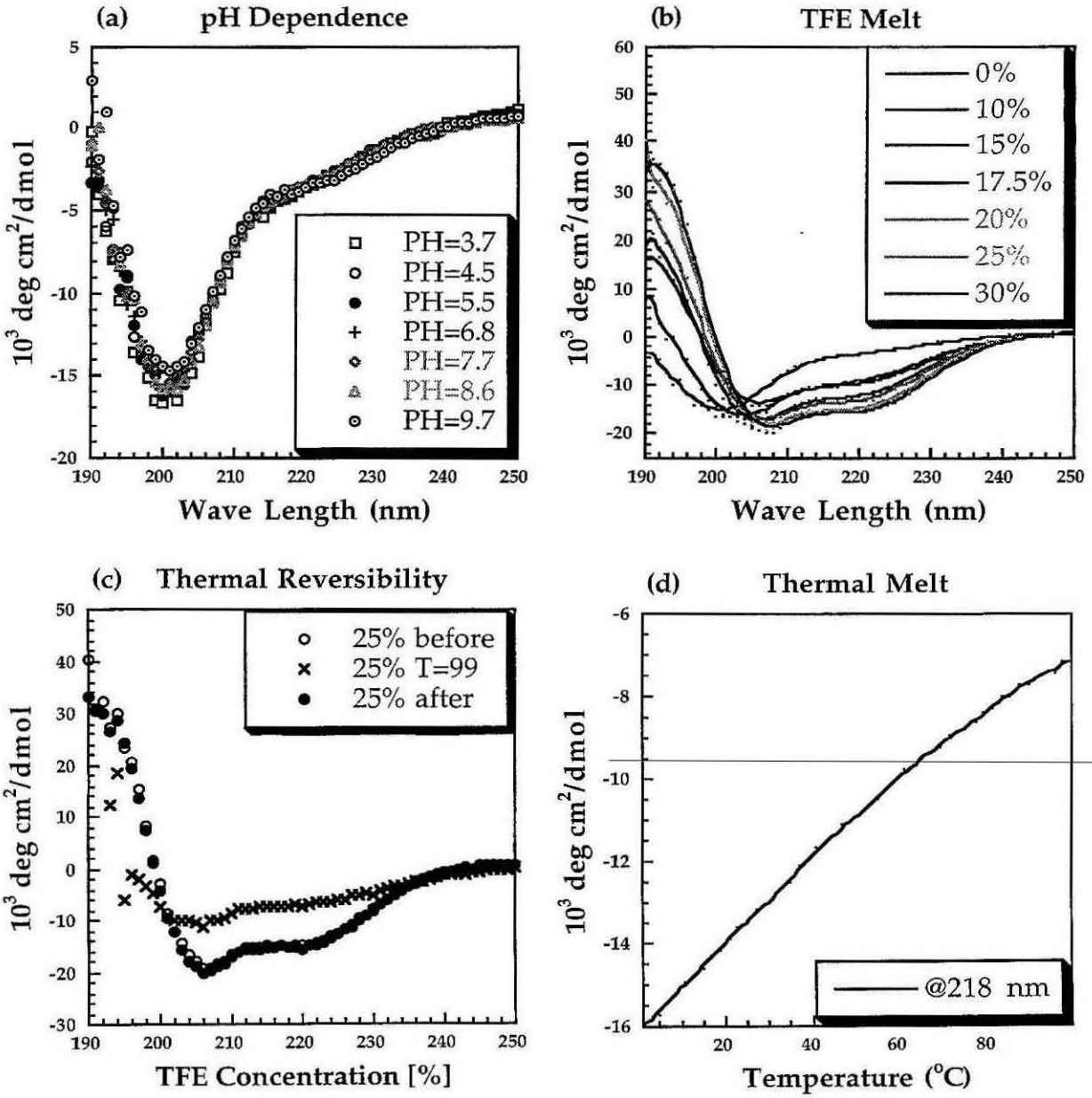


Figure 7 Determine the location of hydrophobic residues. Arguments were made based on (i) the density of hydrophobic residues in an α-helix versus a β-strand, and (ii) the right-handed twist of β-strands. This binary pattern guided us to design bba5. Starting with the original backbone, we changed position 11 from hydrophobic to hydrophilic. Position 11 corresponded to the first hydrophobic residue on the second β-strand.

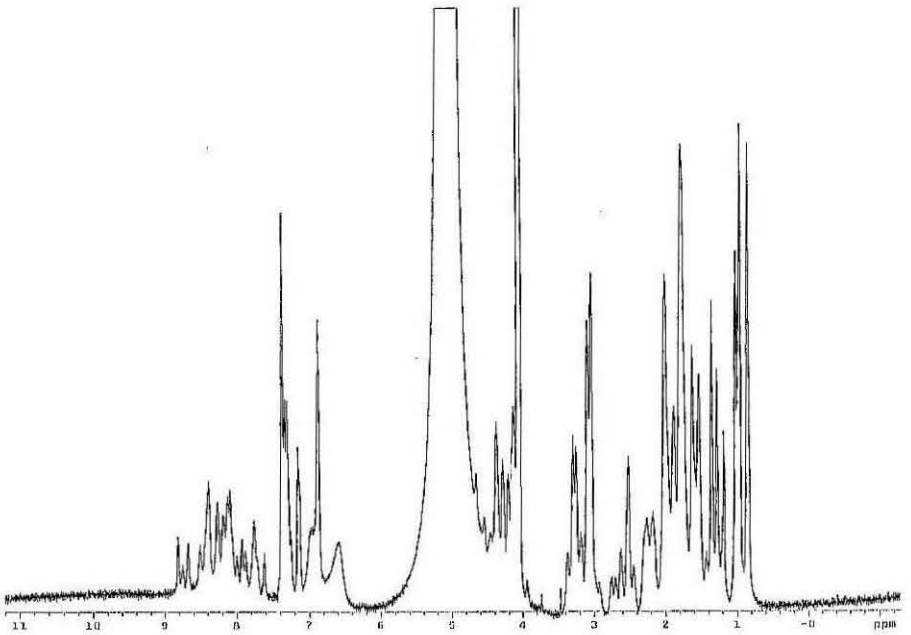
to C terminus, the first positions in each of the β-strands are the most surface-oriented and would perhaps be the best to convert to a hydrophilic residue. Given a choice of position 1 in the first or second β-strand, we chose position 1 in the second strand because we wanted to retain position 1 in the first strand since it is closer to the N terminus and may provide crucial long range packing interactions. This binary pattern guided us to design bba5. Starting with the original backbone, we changed position 11 from hydrophobic to hydrophilic. Among bba1 to bba5, bba5 has binary patterning most similar to FSD-1 (Figure 1 & Table 1a), a thermally stable ββα protein with a well-defined structure developed in our lab using the same sequence-selection algorithm (Dahiyat & Mayo, 1997).

Final configuration results. The above guidelines were used to design our final peptide, bba5, see Figure 1. The CD spectrum for bba5 remained random-coil like for pH values ranging from 3.7 to 9.7 (Figure 8a). At pH 5.5, TFE was added to further stabilize the peptide. To determine the minimium TFE concentration that would enable the peptide to exhibit the most structure, a TFE-melt was conducted. We found that 25% TFE was the minimum concentration where the CD spectrum exhibited the most structure (Figure 8b) and was still thermally reversible (Figure 8c). This was an improvement from bba4, where 40% of TFE was required. A thermal melt was done on bba5 at 25% TFE. It showed that bba5 was weakly cooperative (Figure 8d). 1D NMR spectra were acquired for bba5 in 25% TFE. bba5 showed good dispersion, significantly better than bba1 to bba4, suggesting a more structured peptide (Figure 8e). FSD-1's 1D NMR spectrum was also included in Figure 8f for comparison.

Figure 8 bba5's CD and 1D NMR spectra. (a) pH scan from pH 3.7 to 9.7; (b) TFE-melt conducted at pH 5.5 to determine the minimum TFE concentration that would enable the peptide to exhibit the most structure, a TFE concentration of 25% was found; (c) bba5 was thermally reversible at pH 5.5 in 25% TFE; (d) bba5's thermal melt at pH 5.5 in 25% TFE. It showed that bba5 was weakly cooperative. However, unlike the melting curve for bba4, bba5's melting curve did not suggest the possibility of intermediates


(a) pH Dependence
10^3 deg cm^2/dmol
Wave Length (nm)
PH=3.7
PH=4.5
PH=5.5
PH=6.8
PH=7.7
PH=8.6
PH=9.7
(b) TFE Melt
10^3 deg cm^2/dmol
Wave Length (nm)
0%
10%
15%
17.5%
20%
25%
30%
(c) Thermal Reversibility
10^3 deg cm^2/dmol
TFE Concentration [%]
25% before
25% T=99
25% after
(d) Thermal Melt
10^3 deg cm^2/dmol
Temperature (°C)
@218 nm

(e)



(f)

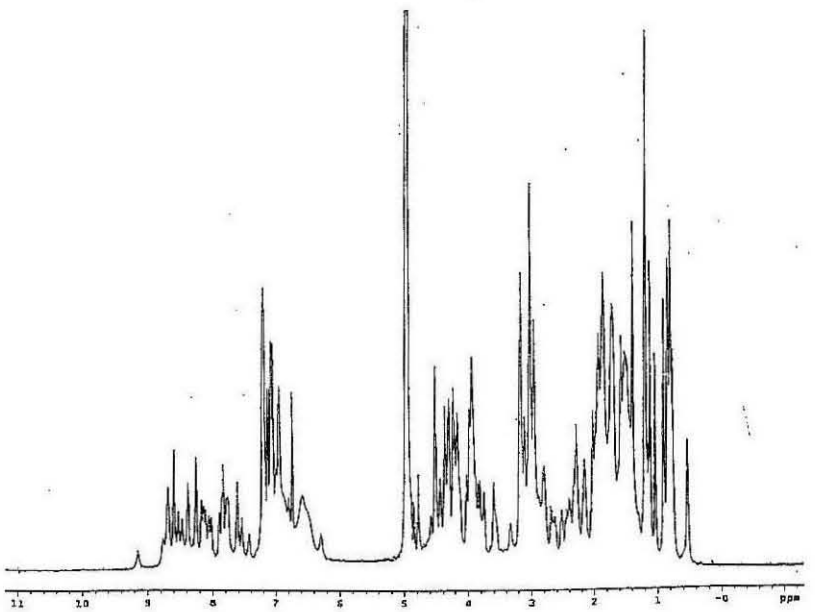


Figure 8 (continued) bba5's CD and 1D NMR spectra. (e) 1D NMR spectra for bba5 acquired at pH 5.5 in 25% TFE. bba5 showed good dispersion, significantly better than bba1 to bba4, suggesting a more structured peptide; (f) FSD-1's 1D NMR spectrum obtained in TFE-free but otherwise similar experimental conditions.

Conclusions

In this experiment paper, we evaluate our backbone design method by testing the thermal stability and structural properties of the designed peptides. Starting with a computer generated ββα motif, we optimized five sequences for this backbone and characterized them using CD and 1D NMR. It was found that small differences in the number and location of the hydrophobic residues can significantly change the thermodynamic behaviour of the designed peptides. This supports the previously acknowledged importance of binary patterning in protein design (Hecht et al., 1990). We compared our five sequences with FSD-1, a thermally stable ββα protein with a well-defined structure developed in our lab using the same sequence-selection algorithm (Dahiyat & Mayo, 1997). We found that the sequence that had binary patterning most similar to FSD-1 also had the best thermodynamic properties. Based on the results of these five peptides, a set of heuristic rules were derived which could be used to improve the computer-generated backbones. In addition to the number and location of hydrophobic residues, we found that it is critical to determine the optimal number of residues for each secondary structural element. Validation of these rules will be the focus on future experimental efforts.

Materials and Methods

Sequence optimization. The optimal configuration for ββα motif was detrmined using the sequence selection and the backbone design algorithm as described in Chapter 4. The binary pattern of the optimal configuration (bba1) was altered to generate four additional sequences. With bba2, we replaced hydrophilic residues on position 4 of the first strand of the β-sheet with a hydrophobic one. With bba3, we made the same change as in bba2, and we

also replaced the hydrophobic residue in position 8 in the β-turn with a hydrophilic one. bba4 was generated in the same way as bba3. The same change was made as in bba2 and a hydrophobic residue was replaced with a hydrophilic one. However, this time, a different location was chosen -- position 22 on the α-helix. For bba5 we did the same thing again, this time choosing position 11 on the 2nd strand of the β-sheet as our location for the new hydrophilic residue. The residue were reclassified in accordance with these changes. Core residues (hydrophobic) were reclassified as surface (hydrophilic) and vice versa (see Table 1b).

Peptide synthesis and purification. Peptides were synthesized by using standard fluorenylmethoxycarbonyl chemistry, and were purified by reverse -phase HPLC.

CD and NMR spectroscopy. CD spectra were measured on an Aviv 62DS spectrometer at 7 °C, 50 mM sodium phosphate buffer, and 50 mM protein, for pH values range from 2.7 to 10.0. For each pH value, a wavelength scan from 190 nm to 250 nm was recorded. The pH value which produced CD signal with the most secondary structure was used for a thermal reversibility test. The test was performed by recording the wavelength scan at 7 °C, 99 °C and 7 °C again. If the CD traces remained constant across all pH values, the thermal reversibility test was performed at pH 6. If the peptide was thermally reversible it was retested using different concentrations of TFE.

In all these experiments, a 1 mm pathlength cell was used and the temperature was controlled by a thermoelectric unit. Thermal melts were then performed on the thermally reversible peptides at 218 nm using the best solvent and pH conditions. Temperature increments of 2 °C were used with an averaging time of 10 s and an equilibration time of 90 s. Peptide concentrations were determined by UV spectrophotometry.

NMR samples for the thermally irreversible peptides (bba1, bba2 and bba4) were prepared in 90/10 H_2O/D_2O and 50 mM phosphate buffer at the optimal pH as determined from the CD experiments. NMR samples for bba3 and bba5 were prepared in 10% D_2O with 40% and 25% deuterated TFE, respectively. Spectra were acquired on a Varian Unity Plus 600 MHZ spectrometer at 7 °C. 1024 transients were acquired with 1.5 s of solvent presaturation used for water suppresion. NMR Samples were prepared at the highest peptide concentration which did not result in obvious aggregation.

References

Aurora R, Rose GD. 1998. Helix capping. *Prot Sci 7* 21-38.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol Biol 112*:535–542.

Betz SF, DeGrado WF. 1996. Controlling topology and native-like behavior of de novo-designed peptides — Design and characterization of antiparallel 4-stranded coiled coils. *Biochemistry 35*:6955–6962.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*:895–903.

Dahiyat BI, Mayo SL. 1997. De novo protein design - fully automated sequence selection. *Science 278* 82-87.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

Donate LE, Rufino SD, Canard LHJ, Blundell TL. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database of modeling and prediction. *Protein Sci 5* 2600-2616.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA 92*:8408–8412.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science 249* 884-891.

Hellinga HW, Caradonna JP, Richards FM. 1991. Construction of new ligand-binding sites in proteins of know structure 2. Grafting of buried

transition-metal binding site into *Escherichia coli* thioredoxin. *J Mol Biol 222*:787–803.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*:1142–1154.

Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. 1995. Novel metal-binding proteins by design. *Nature Structure Biol 2*:368–373.

Luo PZ, Baldwin RL. 1997. Mechanism of helix induction by trifluoroethanol - a framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. *Biochemistry 36* 8413-8421.

Nautiyal S, Woolfson DN, King DS, Alber T. 1995. A designed heterotrimeric coiled coil. *Biochemistry 34*:11645–11651.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Radzicka A, Wolfenden R. 1988. Comparing the polarities of the amino acids - sidechain distribution coefficients between the vapor-phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry 27* 1664-1670.

Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6* 1701-1707.

Chapter 5

A Multi-Substrate Single-File Model for Ion-Coupled Transporters

762 Biophysical Journal Volume 70 February 1996 762–777

A Multi-Substrate Single-File Model for Ion-Coupled Transporters

Alyce Su,* Sela Mager,‡ Stephen L. Mayo,‡§ and Henry A. Lester‡
*Division of Physics, Mathematics and Astronomy, and ‡Division of Biology, §Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125 USA

ABSTRACT Ion-coupled transporters are simulated by a model that differs from contemporary alternating-access schemes. Beginning with concepts derived from multi-ion pores, the model assumes that substrates (both inorganic ions and small organic molecules) hop a) between the solutions and binding sites and b) between binding sites within a single-file pore. No two substrates can simultaneously occupy the same site. Rate constants for hopping can be increased both a) when substrates in two sites attract each other into a vacant site between them and b) when substrates in adjacent sites repel each other. Hopping rate constants for charged substrates are also modified by the membrane field. For a three-site model, simulated annealing yields parameters to fit steady-state measurements of flux coupling, transport-associated currents, and charge movements for the GABA transporter GAT1. The model then accounts for some GAT1 kinetic data as well. The model also yields parameters that describe the available data for the rat 5-HT transporter and for the rabbit Na^+-glucose transporter. The simulations show that coupled fluxes and other aspects of ion transport can be explained by a model that includes local substrate-substrate interactions but no explicit global conformational changes.

INTRODUCTION

Several classes of membrane transport proteins use electrochemical gradients for ions (usually Na^+ or H^+) to accumulate organic molecules (neurotransmitters, sugars, amino acids, osmolytes) in plant and animal cells (Schultz, 1986; Harvey and Slayman, 1994). The tight flux coupling between these inorganic and organic substrates constitutes a hallmark of ion-coupled transporters and contrasts with properties of ion channels, another major class of membrane transport proteins (Hille, 1992). To explain the mechanism of flux coupling, two major classes of model have been proposed (Hill, 1977; Schultz, 1980; Stein, 1986; Kanner and Schuldiner, 1987; Rudnick and Clark, 1993). Early models envisioned a recirculating carrier whose motions were largely governed by the binding and dissociation of the substrates. More recently, sequence analysis of cloned transporters suggests 6–12 putative transmembrane domains, rendering a recirculating carrier less plausible.

Most contemporary mechanistic concepts of ion-coupled transport employ the alternating-access scheme first enunciated by Jardetzky (1966) and developed in many papers by Laüger (see Laüger, 1979, 1991; Wright, 1993; Lester et al., 1994). In this scheme, ion-coupled transporters are viewed as pores or channels that have two gates. Whereas the pore has sites that bind, or perhaps merely accept, all of the permeant substrates, the gates have most of the (poorly understood) properties that ensure coupled transport. When all of the substrates are bound appropriately, the gates undergo conformational changes; and these conformational changes account for the differences in compartmentalization of the substrates during the transport cycle. Some alternating-access schemes incorporate ordered binding and dissociation of substrates (see, for instance, Rudnick and Clark, 1993). Now that cloned transporters can be expressed at high densities and studied with good temporal resolution in heterologous expression systems, additional measurements are available on pre-steady-state kinetics and charge movements associated with one or a few steps in the transport cycle (Parent et al., 1992a,b; Mager et al., 1993, 1994; Cammack et al., 1994; Wadiche et al., 1995a). Several studies build on these time-resolved data in the context of the alternating-access model (Parent et al., 1992b; Mager et al., 1993; Wadiche et al., 1995a).

However, the newer measurements have also revealed several additional classes of complexities that cannot be explained by straightforward alternating-access models: 1) There are leakage currents—Na^+ fluxes in the absence of organic substrate (Schwartz and Tachibana, 1990; Umbach et al., 1990; Cammack et al., 1994). 2) There are major departures from accepted stoichiometry, so that transport-associated currents are several times larger than the flux of organic substrate (Mager et al., 1994; Wadiche et al., 1995b; Galli et al., 1995; Picaud et al., 1995; Risso et al., 1995). 3) There are actual or inferred quantized current events that exceed by several orders of magnitude the single-charge events expected from the model (Mager et al., 1994; Wadiche et al., 1995b; DeFelice et al., 1995; Cammack and Schwartz, 1995).

Although more complex alternating-access models can be developed to account for some of these new phenomena, the time seemed ripe for an alternative class of models. Our formulation is termed the multi-substrate single-file transport model. We borrow heavily from ion channel models that incorporate a pore with several simultaneously bound ions (Hille, 1992). In particular, we do not explicitly allow conformational changes that change the compartmentalization of the

Received for publication 25 July 1995 and in final form 21 October 1995.
Address reprint requests to Dr. Henry A. Lester, Division of Biology 156-29, California Institute of Technology, Pasadena, CA 91125. Tel.: 818-395-4946; Fax: 818-564-8709; E-mail: lester@caltech.edu.

0006-3495/96/02/762/16 $2.00

substrates. The gates of the alternating-access model have been de-emphasized. Instead, functional compartmentalization arises because the pore (or lumen or channel) of the transporter mediates multiple substrate bindings and substrate-substrate interactions that favor, albeit only statistically, permeation in fixed ratios of inorganic ions to organic substrate.

In this first paper on the topic, we test the hypothesis of "multi-substrate single-file transport" in a quantitative, physically realistic fashion. We consider that the following three steady-state properties are most appropriate for detailed simulation: the ratio between the fluxes of organic and inorganic substrate, the electrical currents associated with the full transport cycle, and charge movements associated with partial steps of transport. We therefore simulate these properties of three ion-coupled transporters for which high-resolution functional studies have been reported: the GABA transporter GAT1 (Guastella et al., 1990; Mager et al., 1993), the serotonin transporter 5-HTT (Mager et al., 1994), and the Na^+-glucose transporter SGLT1 (Parent et al., 1992a,b). In each case, the multi-substrate single-file transport model has been found to reproduce available experimental data within experimental error (although charge movements have not been measured for 5-HTT). The model also accounts for newer phenomena such as leakage currents of all these transporters and variable stoichiometry of 5-HT transport, among other permeation properties.

Our approach has certainly been foreshadowed by many previous suggestions that transporters have channel-like mechanisms, for instance, in mediated ionic transport (Frohlich, 1988; Krupka, 1989; Dani and Levitt, 1990; Hasegawa et al., 1992), in electrogenic membrane systems (Andersen et al., 1985; Lagnado et al., 1988; Nakamoto et al., 1989; Hilgemann et al., 1991; Lauger, 1991; Wang et al., 1992; Gadsby et al., 1993; Rakowski, 1993), in neurotransmitter transporters (Krupka and Deves, 1988; Schwartz and Tachibana, 1990), and in facilitative sugar transporters (Barnett et al., 1975; Lowe and Walmsley, 1986; Kimmich and Randles, 1988; Walmsley, 1988; Baldwin, 1993; Hernandez and Fischbarg, 1994). Detailed theories have been based on electrodiffusion (Chen and Eisenberg, 1993; Eisenberg, 1994) and have treated channels that can simultaneously contain two ionic species at once (Franciolini and Nonner, 1994). Although molecular cloning has given us knowledge about the amino acid sequence of many ion-coupled transporters (Harvey and Nelson, 1994), there is still little relevant structural information at the atomic scale, or even at the level of tertiary structure or membrane topology or tertiary structure. Therefore the model is cast in purely formal terms at present.

THEORY AND METHODOLOGY

Structure of the model

The multi-substrate single-file transport model treats a transporter as a single file of binding sites, with ends open to the external and internal solutions (Fig. 1 *A*). The solutions are infinite sources and sinks: substrate concentrations within the solutions are not changed by the events of the simulation. Substrates are loaded from the solutions into the ion-coupled transporter lumen at rates governed partially by mass action. Once a substrate is within the ion-coupled transporter, it is allowed to move between sites and to the solutions. The kinetics of the ion-coupled transporter are reconstructed by combining all possible transitions.

Each of the N sites can be either empty or occupied by one of the m substrates. A total of N^{m+1} arrangements are possible. A state of the transporter is defined to be one such arrangement.

A hopping event is an allowed hop taken by a single substrate. Fig. 1 *B* shows the set of allowed hops. An allowed hop brings one substrate to an adjacent, empty site. Special rules apply to sites 1 and N (the external and internal solutions are considered to be adjacent to sites 1 and N, respectively). A substrate from the solution can hop onto site 1 or N only when the site is empty; however, a substrate on site 1 or N can always hop into the solution.

Hopping frequencies are rate constants that describe the probability of individual hopping events. Hopping frequencies $k_{a\to b}$ are described as the product $fg_{vo}g_{ar}$. The first term f is called the "intrinsic hopping frequency" of a substrate; it is defined from mass action if the hop is from the solution and by a unimolecular event if the hop is from a binding site to another site within the membrane or to the solution. For

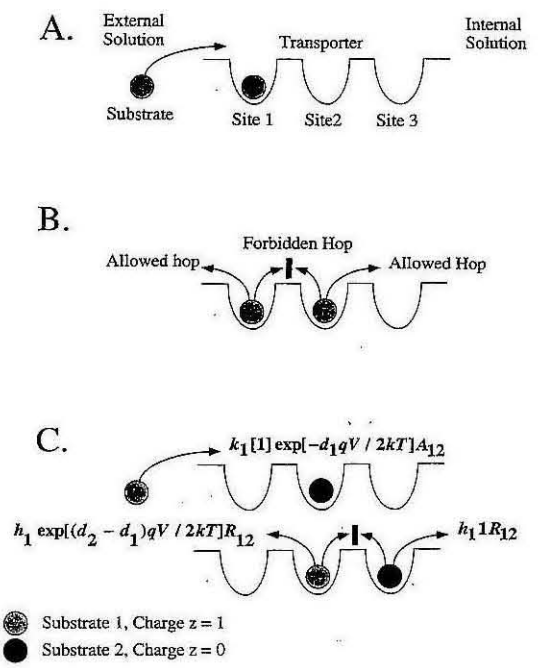


FIGURE 1 Multi-substrate single-file transport model. (*A*) Structure of the transporter, including three binding sites. The state, with a single bound substrate, corresponds to state 2 in Fig. 2. (*B*) Allowed hopping events. (*C*) Rules for obtaining the hopping frequencies, including the effects of voltage and of attraction/repulsion.

the present version of the model, the intrinsic hopping frequency varies with the substrate but is constant across all sites within the pore.

The second term, a dimensionless factor g_v, describes the effect of membrane potential on the intrinsic hopping frequency. The simple equations given below for the effect of potential are consistent with a model in which a) the membrane potential produces a linear field, b) there is a rate-limiting energetic barrier midway between each pair of sites, and c) the energy of the barrier with respect to each site is modified by the interaction between the charge on the substrate and the field.

The third term, a dimensionless factor g_{ar}, describes the effect of interactions with other substrates that occupy adjacent sites or, if an adjacent site is empty, the next site. g_{ar} is the product of an attraction term A and a repulsion term R. Examples are shown in Fig. 1 *C*. These attraction and repulsion terms also allow for the possibility that the field changes with the ionic contents of the channel.

For hops by a substrate i with charge zq in the presence of another substrate j (we include the possibility that $j = 0$ denotes an empty site)

$f = k_i[i]$ If it hops from the external solution to site 1, or from the internal solution to site N

$= h_i$ If it hops from a site within the transporter

$g_v = \exp[-(d_{s+1} - d_s)zq_iV/2kT]$ If it hops from site s to $s + 1$

$= \exp[(d_{s+1} - d_s)zq_iV/2kT]$ If it hops from site $s + 1$ to s

$g_{ar} = A_{ij}R_{ij}$;

A_{ij} Describes a hop that brings substrate i adjacent to substrate j

R_{ij} Describes a hop that separates substrate i from substrate j,

where

$[i]$ = concentration of substrate i in the solution

k_i = forward binding constant of substrate i in $M^{-1}\ s^{-1}$

h_i = intrinsic hopping frequency of substrate i in s^{-1}

d_s = fraction of electrical field sensed by a substrate with charge zq ($s = 0$ to $N + 1$, $d_0 = 0$, $d_{N+1} = 1$, $d_0 \le d_s \le d_{N+1}$)

V = membrane potential in mV

z = valence of a charged substrate

kT/q_i = ±25 mV, where
k = Boltzmann constant
T = Absolute temperature
q_i = Elementary charge on substrate i

A_{ij} = attractive coupling between two substrates separated by an empty site

R_{ij} = repulsive coupling between two substrates in adjacent sites

A_{ij} = A_{ji}, and $R_{ij} = R_{j,i}$, for i, j > 0; that is, the substrate-substrate couplings are symmetric

For each transporter configuration a, all transporter configurations b accessible from a by a single hopping event are linked. The corresponding hopping frequencies $k_{a\to b}$ are derived as described above. If configurations a and b cannot be linked by a single hopping event, then $k_{a\to b} = 0$. The final diagram is called a hopping diagram (Fig. 2), whose mathematical representation is a matrix $\boldsymbol{H}$ constructed as

$$H_{ab} = k_{b\to a} \tag{1}$$

$$H_{aa} = -\sum_{b=1,b\neq a}^{b=N^{m+1}} k_{a\to b}. \tag{2}$$

Given the hopping matrix $\boldsymbol{H}$, the time dependence of the ion-coupled transporter configuration distribution is

$$\frac{d\bar{x}}{dt} = \boldsymbol{H}\bar{x} \tag{3}$$

or equivalently,

$$\frac{dx_a}{dt} = \sum_{b=1}^{b=N^{m+1}} H_{ab}x_b = \sum_{b=1,b\neq a}^{b=N^{m+1}} H_{ab}x_b + H_{aa}x_a = \sum_{b=1,b\neq a}^{b=N^{m+1}} k_{b\to a}x_b - \sum_{b=1,b\neq a}^{b=N^{m+1}} k_{a\to b}x_a, \tag{4}$$

where $\bar{x} = (x_1, x_2, \ldots x_N{}^{m+1})$, $(0 \le x_a \le 1)$ represents the distribution of ion-coupled transporter configurations. For example, $x_1 = 0.25$ denotes that 25% of the transporters are in configuration 1.

At steady state, Eq. 4 reduces to

$$\frac{dx_a}{dt} = 0, \tag{5}$$

resulting in N^{m+1} linearly dependent homogeneous equations, of which $N^{m+1} - 1$ are linearly independent. The solution to Eq. 5 is a null space of dimension 1, which can be solved numerically by the Gauss-Jordan elimination method (Press et al., 1992). This solution yields predictions for several functional measurements, described in detail below.

A three-site two-substrate special case

For most of the simulations in this paper, we assume the following special case:

1. The transporter has $N = 3$ binding sites.

2. There are only $m = 2$ substrates, Na^+ (substrate 1) and the organic substrate (substrate 2). In particular, we disregard permeation by Cl^-. Our justification for this choice is a) that the role of Cl^- is rather less clear than that of the other substrates for most transporters (Mager et al., 1993; Lester et al., 1994) and b) that introducing another ion would complicate the simulations without adding much insight at this early stage.

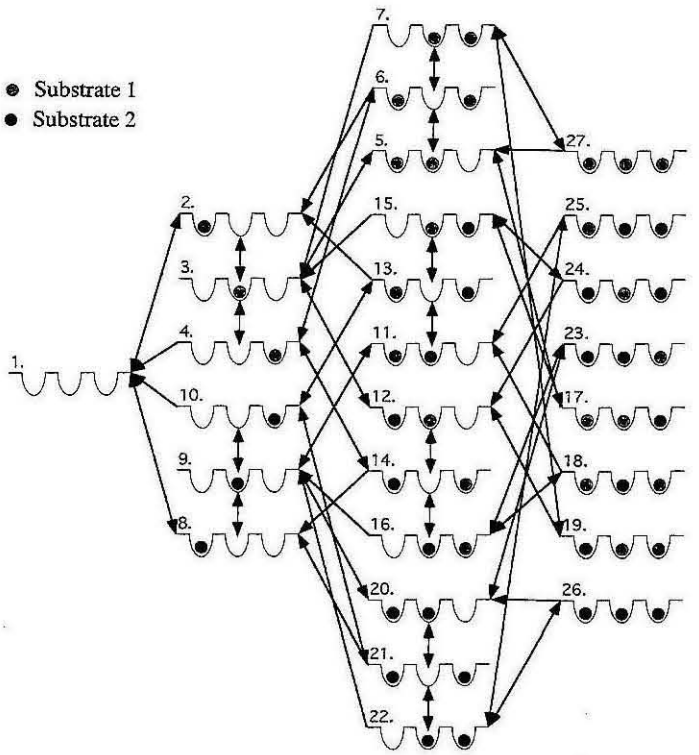


FIGURE 2 Hopping diagram for GAT1, 5-HT transporter, and SGLT1. This special case is used in most of our simulations. The ion-coupled transporter is modeled with three sites and two substrates, which generate a total of $3^{2+1} = 27$ transporter configurations. All transitions are reversible, except for those from the site 3 (rightmost) to the internal solution (an infinite sink).

3. The organic substrate has $q_2 = 0$ charge. This assumption holds for GABA and for the sugars but not for 5-HT$^+$ and is modified in some of the simulations described later.

4. The internal solution has zero concentration of all substrates. This assumption chiefly reflects the facts a) that oocyte experiments performed in our laboratory do not control intracellular concentrations and b) that these concentrations are thought to be small in most modern experiments. However, the assumption is modified in some simulations described below (for instance, data presented in Fig. 6).

The matrix elements corresponding to this special case are listed in Table 1 and the corresponding hopping diagram is shown in Fig. 2.

Steady-state predictions of the special case

Under these special conditions, Eq. 4 becomes

$$\frac{dx_a}{dt} = \sum_{b=1}^{27} H_{ab}x_b = \sum_{b=1,b\neq a}^{27} H_{ab}x_b + H_{aa}x_a = \sum_{b=1,b\neq a}^{27} k_{b\to a}x_b - \sum_{b=1,b\neq a}^{27} k_{a\to b}x_a, \tag{4a}$$

where $\bar{x} = (x_1, x_2, \ldots x_{27})$, $(0 \leq x_a \leq 1)$ represents the distribution of ion-coupled transporter configurations. Eq. 5 then becomes 27 linearly dependent homogeneous equations, of which 26 are linearly independent. The model makes predictions about three types of steady-state data.

Charge distribution

The simulated steady-state charge occupancy, $Q_\infty(V)$, is defined as follows:

$$Q_\infty(V) \equiv T\sum_{a=1}^{27} x_a Q_a = T\sum_{a=1}^{27} x_a \sum_{i=1}^{3} d_i q_i, \tag{6}$$

where T is the total number of transporters in the membrane and Q_a is the equivalent charge of the ath configuration. For instance, $Q_{26} = (d_1 + d_2 + d_3)q$ and $Q_{27} = 0$. We also define the maximum charge movement, from hyperpolarizing to depolarizing potentials, by $Q_{max} = Q(+\infty) - Q(-\infty)$.

Transport-associated current

The steady-state transport-associated current, I_∞, results from charged substrates moving fully across the membrane. This current is simulated by summing the charge movements that result from charged substrates hopping into the

TABLE 1 Hopping matrix elements for GAT1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1		R01 h1 exp[d1 (v/50)]		R01 h1 exp[-(1-d3)(v/50)]				R02 h2		A02 h2																	
2	A01 P1		R01 h1 exp[(d2-d1)(v/50)]			R01 h1 exp[-(1-d3)(v/50)]							R02 h2														
3		A01 h1 exp[-(d2-d1)(v/50)]		A01 h1 exp[d2-d1)(v/50)]	R11 h1 exp[d1 (v/50)]		R11 h1 exp[-(1-d3)(v/50)]								R12 h2												
4			R01 h1 exp[-(d3-d2)(v/50)]			R01 h1 exp[d1 (v/50)]								R02 h2			R12 h2										
5			A11 P1			R11 h1 exp[(d1-d2)(v/50)]																					R11 h1 exp[-(1-d3)(v/50)]
6				A01 P1	R11 h1 exp[-(d3-d2)(v/50)]		R11 h1 exp[-(d2-d1)(v/50)]																				
7						A11 h1 exp[(d2-d1)(v/50)]												R12 h2								R11 h1 exp[d1(v/50)]
8	A01 P2								R02 h2					R01 h1 exp[-(1-d3)(v/50)]							R02 h2						
9								A01 h2		A02 h1	R12 h1 exp[d1(v/50)]					R12 h2 exp[-(1-d3)(v/50)]				R22 h2		R22 h2					
10									R02 h2				R01 h1 exp[d1(v/50)]								R02 h2						
11									A12 P1				A12 h2												R22 h2		
12			A12 P2											A12 h1										R12 h2			
13										A01 P1	R12 h2	R12 h2															
14				A01 P2								R12 h1 exp[-(d2-d1)(v/50)]															
15													A01 h1 exp[-(d2-d1)(v/50)]											R12 h2			
16		R12 h2												A12 h2									R22 h2				
17	A11 P1																										
18		A12 P1																									
19							A12 P2																			R22 h2	
20									A22 P2												A22 h2		R12 h2 exp[-(1-d3)(v/50)]				
21										A02 P2												R22 h2					
22																				A22 h2					R12 h2		
23		A22 P2																									
24	A12 P2																										
25																						A12 P1					
26																						A22 P2					
27							A11 P1																				

The matrix elements corresponding to the special case given in the text; the corresponding hopping diagram is shown in Fig. 2. H_{ba} corresponds to a hopping event from configuration a to configuration b. For example, the matrix element H_{21} is the same as $k_{1\to2}$. Each diagonal element is the negative of the sum of all the entries in the same column. For GAT1, substrate 1 is Na^+ and substrate 2 is GABA. P1 = $k_1[Na^+]\exp(d_1V/50)$. P2 = $k_{21}[GABA^+]$.

internal solution, e.g., $k_{27\to5}$. For our special case in which Na^+ is the only charged substrate, the steady-state transport-associated current is

$$I_\infty \equiv Tq[(x_4k_{4\to1}) + (x_6k_{6\to2}) + (x_7k_{7\to3}) + (x_{14}k_{14\to8}) + (x_{16}k_{16\to9}) + (x_{18}k_{18\to11}) + (x_{19}k_{19\to12}) + (x_{23}k_{23\to20}) + (x_{27}k_{27\to5})]. \quad (7)$$

Flux ratios

Steady-state flux coupling between substrates i and j is summarized by the ratio between the fluxes of i and j ($flux_i/flux_j$). The flux ratio between Na^+ and the organic substrate is

$$\frac{flux_1}{flux_2} \equiv \frac{[(x_4k_{4\to1}) + (x_6k_{6\to2}) + (x_7k_{7\to3}) + (x_{14}k_{14\to8}) + (x_{16}k_{16\to9}) + (x_{18}k_{18\to11}) + (x_{19}k_{19\to12}) + (x_{23}k_{23\to20}) + (x_{27}k_{27\to5})]}{[(x_{10}k_{10\to1}) + (x_{13}k_{13\to2}) + (x_{15}k_{15\to3}) + (x_{17}k_{17\to5}) + (x_{21}k_{21\to8}) + (x_{22}k_{22\to9}) + (x_{24}k_{24\to12}) + (x_{25}k_{25\to11}) + (x_{26}k_{26\to20})]}. \quad (8)$$

Available experimental data

The predictions described above were compared with the appropriate experimental data using algorithms described in the next section. Our most complete experimental data are available for GAT1. For instance, the experimental data on steady-state charge distribution was measured from voltage jumps in the absence of GABA (Mager et al., 1993). The integration of the transient response gives the total number of charges within the membrane. For purposes of these calculations, we abstracted the data of Mager et al. (1993) into the following functional form, which describes the data:

$$Q_\infty(V) = \frac{Nq}{1 + e^{-(V-V_h)/28.1}}. \quad (9)$$

A time interval of 0.2 ms was adequate for the simulations because the steady-state $\bar{x}$ obtained by integration and the Gauss-Jordan elimination method deviated by less than 0.001%. The steady-state transport-associated currents I_∞ were also taken from the data of Mager et al. (1993). The flux ratios $flux_1/flux_2$ for Na^+/GABA were assumed to lie between 1 and 2 (Kanner and Schuldiner, 1987).

Parameter searches

A set of randomly selected parameter values was used to start a simulation (Fig. 3). Gauss-Jordan elimination (Press et al., 1992) was used to solve Eq. 5 for the 27 x_a values. These values in turn yielded the three steady-state

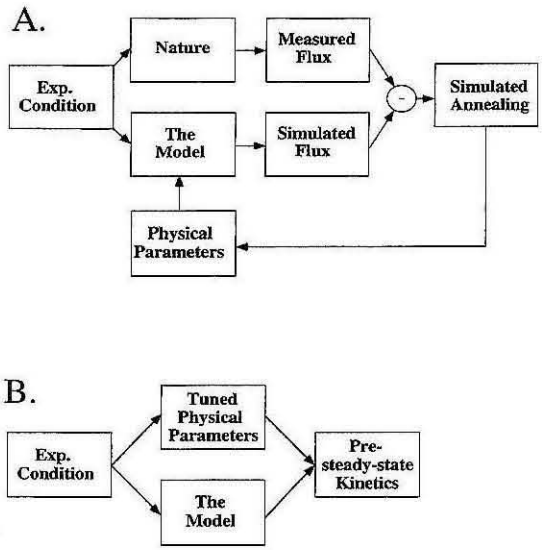


FIGURE 3 Simulation scheme. (*A*) Fitting mode. Simulated annealing is employed to optimize the parameter values by minimizing the error between the experimental and the simulated steady-state data. The resulting values are then used to predict the pre-steady-state kinetics. (*B*) Predicting mode. The resulting values from *A* are then used to predict transient currents.

predictions above: charge distribution, steady-state current, and flux ratio. Simulated annealing (Ingber, 1993) was employed to optimize the parameter values by minimizing the error between the experimental and the simulated steady-state data. The simulated annealing algorithm written in c was obtained by FTP from http://www.alumni.caltech.edu/~ingber/. Convergence was specified by an annealing temperature of 10^{-7} (Ingber, 1993). Fitting sessions required ~8 h on a Silicon Graphics Challenge workstation with 200-MHz R4400 processors.

Finally, we note that the exemplar physiological experiments in the paper by Mager et al. (1993) were performed on distinct oocytes, each expressing an unknown number of transporters. To calculate the total number of transporters T expressed in these exemplar oocytes from the Q_{max} data for each oocyte, we would have to assume values for d_1, d_2, and d_3. Therefore the flux per transporter was not accessible from the procedures presented thus far; and the hopping rate constants were uncertain by a common multiplicative constant. We could, however, make use of the experimentally determined value, $I_\infty/Q_{max} = 5.3\ s^{-1}$, where I_∞ is measured at 96 mM Na^+, 0.1 mM GABA, and −80 mV (Mager et al., 1993). This additional constraint in the simulations allowed us to determine the absolute values of the rate constants. The calculations are presented for $T = 1.18 \times 10^{11}$ transporters per cell (presumably a *Xenopus* oocyte expressing GAT1).

Calculation of transient currents

Simulated transient data for GAT1 were not used in the parameter search algorithm; instead, they were used as independent tests of our theories and parameters after the simulated annealing yielded several sets of parameters with nearly equal cost function values for the steady-state data. Transients were simulated by numerical integration, using Mathematica. The transient currents are given by the sum of all hopping events involving the movement of charge into or within the membrane weighted by the electrical distance traveled. Voltage jump relaxations for GAT1 were simulated using the experimental protocol; that is, for fixed $[Na^+]$ and zero GABA, the membrane potential was stepped between values as shown in Fig. 5. The voltage jump transients for GAT1 are defined as

$$\frac{I(t)}{Tq} \equiv d_1\left\{\begin{array}{l}[(x_1k_{1\to2}) - (x_2k_{2\to1})] + [(x_3k_{3\to5}) - (x_5k_{5\to3})] \\ + [(x_4k_{4\to6}) - (x_6k_{6\to4})] + [(x_9k_{9\to11}) - (x_{11}k_{11\to9})] \\ + [(x_{10}k_{10\to13}) - (x_{13}k_{13\to10})] + [(x_{15}k_{15\to17}) - (x_{17}k_{17\to15})] \\ + [(x_{16}k_{16\to18}) - (x_{18}k_{18\to16})] + [(x_7k_{7\to27}) - (x_{27}k_{27\to7})] \\ + [(x_{22}k_{22\to25}) - (x_{25}k_{25\to22})]\end{array}\right\}$$

$$+ (d_2 - d_1)\left\{\begin{array}{l}[(x_2k_{2\to3}) - (x_3k_{3\to2})] + [(x_6k_{6\to7}) - (x_7k_{7\to6})] \\ + [(x_{13}k_{13\to15}) - (x_{15}k_{15\to13})]\end{array}\right\} \quad (10)$$

$$+ (d_3 - d_2)\left\{\begin{array}{l}[(x_3k_{3\to4}) - (x_4k_{4\to3})] + [(x_5k_{5\to6}) - (x_6k_{6\to5})] \\ + [(x_{12}k_{12\to14}) - (x_{14}k_{14\to12})]\end{array}\right\}$$

$$+ (1 - d_3)$$

$$\times\left\{\begin{array}{l}[(x_4k_{4\to1}) + (x_6k_{6\to2})] + [(x_7k_{7\to3}) + (x_{14}k_{14\to8})] \\ + [(x_{16}k_{16\to9}) + (x_{18}k_{18\to11})] + [(x_{19}k_{19\to12}) + (x_{23}k_{23\to20})] \\ + [(x_{27}k_{27\to5})]\end{array}\right\},$$

where all x values are time dependent. Concentration-jump relaxations were simulated by stepping one or two substrate concentrations while holding other parameters fixed. For simplicity, relaxation kinetics were characterized by fitting to a single exponential component.

RESULTS

Figs. 4, 5, 6, 7, and 8 show the simulation results for the ion-coupled transporters GAT1, 5-HTT, and SGLT1. Optimized parameter values are listed in Table 2.

Simulation of GAT1 function

We devoted most effort to simulating GAT1 function, because we have access to extensive data for this ion-coupled transporter. GAT1 is a neurotransmitter transporter belonging to the plasma membrane Na^+/Cl^- coupled subfamily (Amara and Arriza, 1993). It drives the neurotransmitter GABA up its concentration gradient by cotransporting Na^+ down its electrochemical potential gradient (Keynan and Kanner, 1988). Electrophysiological data, tracer flux measurements, and thermodynamic measurements all suggest that 2 Na^+ ions are transported along with a single GABA molecule (Lester et al., 1994); and we have therefore used a special case of our model that includes three binding sites, so that these three molecules may bind simultaneously. Cl^- is another possible substrate for GAT1, but its role is not clear at the moment (Lester et al., 1994); therefore we felt it inappropriate at present to complicate the simulations by including Cl^-.

The experimentally determined permeation properties of GAT1 include steady-state currents as a function of Na^+ and GABA, leakage currents (Na^+ flux in the absence of GABA), substrate flux ratio, charge distribution, and gating currents induced by voltage jumps (Mager et al., 1993; Cammack et al., 1994). We note that leakage currents were not reported in our study of GAT1 function (Mager et al., 1993) but were reported by Cammack et al. (1994). In unpublished recent work, we have indeed found small inward currents that are suppressed by GABA uptake inhibitors in the absence of GABA itself; although we lack systematic data, these leakage currents are <5% of the maximal GABA-induced currents.

The model accounts well for the data that we used to fit the parameters, as summarized in Fig. 4. Five sets of parameters, each resulting from a simulated annealing run, had nearly equal cost function values and provided similar simulations of these data; the simulations of Fig. 4 were calculated with the parameters that also provided good simulations of transient currents (these parameters are given in Table 2). GAT1 steady-state currents are shown as a function of $[Na^+]$ in Fig. 4 *A*; note that the simulation dose-response relations show the observed sigmoidal dependence on $[Na^+]$. The simulations also show the expected hyperbolic dependence on [GABA] in Fig. 4 *B*. The simulations showed leakage currents of ~1% of the saturation current in the absence of external GABA (Fig. 4 *C*). A Na^+/GABA flux ratio between 1 and 2 was obtained (Fig. 4 *D*); the simulated flux ratio depends on membrane potential, a point that has not been tested experimentally for GAT1. The steady-state charge distribution exhibited a sigmoidal dependence on membrane voltage, and the midpoint was shifted to more negative voltages with decreasing $[Na^+]$; these features agree quantitatively with the experimental data (Fig. 4 *E*).

Transient currents

As a further qualitative test of the model, parameter sets from the simulated annealing were tested for their ability to simulate various transient currents. One parameter set survived this test; Fig. 5 displays some results. The simulated voltage-jump relaxations in the absence of GABA, shown in Fig. 5 *A*, display the general features described by Mager et al. (1993) and Cammack et al. (1994): for hyperpolarizing jumps, the transients are negative and consist of a rapid

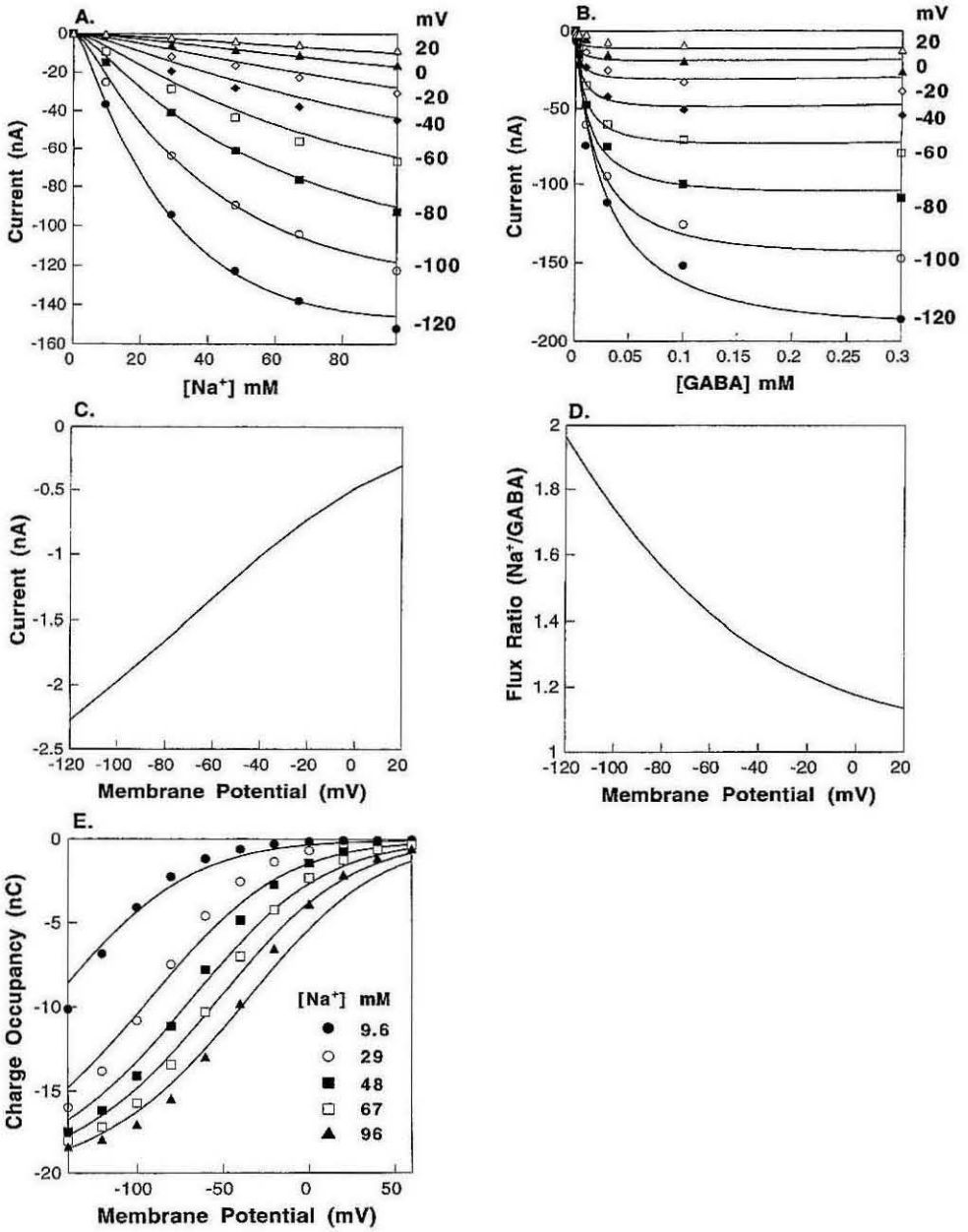


FIGURE 4 Simulated versus experimental GAT1 steady-state properties. Smooth lines are the simulations, using the parameters in Table 2. (*A*) Steady-state current as a function of $[Na^+]$. Experimental data from Mager et al. (1993). [GABA] = 0.1 mM, membrane potentials as indicated. (*B*) Steady-state current as a function of [GABA]. Experimental data from Mager et al. (1993). $[Na^+]$ = 96 mM; membrane potentials as in *A*. (*C*) Simulated leakage current at zero [GABA] versus voltage. (*D*) Flux ratio of Na^+ to GABA simulated at $[Na^+]$ = 96 mM, [GABA] = 0.3 μM, for the voltage range from −120 to 20 mV. (*E*) Steady-state charge distribution as a function of membrane potential in the absence of GABA. Experimental data from Mager et al. (1993), represented by Eq. 9. The curves are parametric in $[Na^+]$ as indicated.

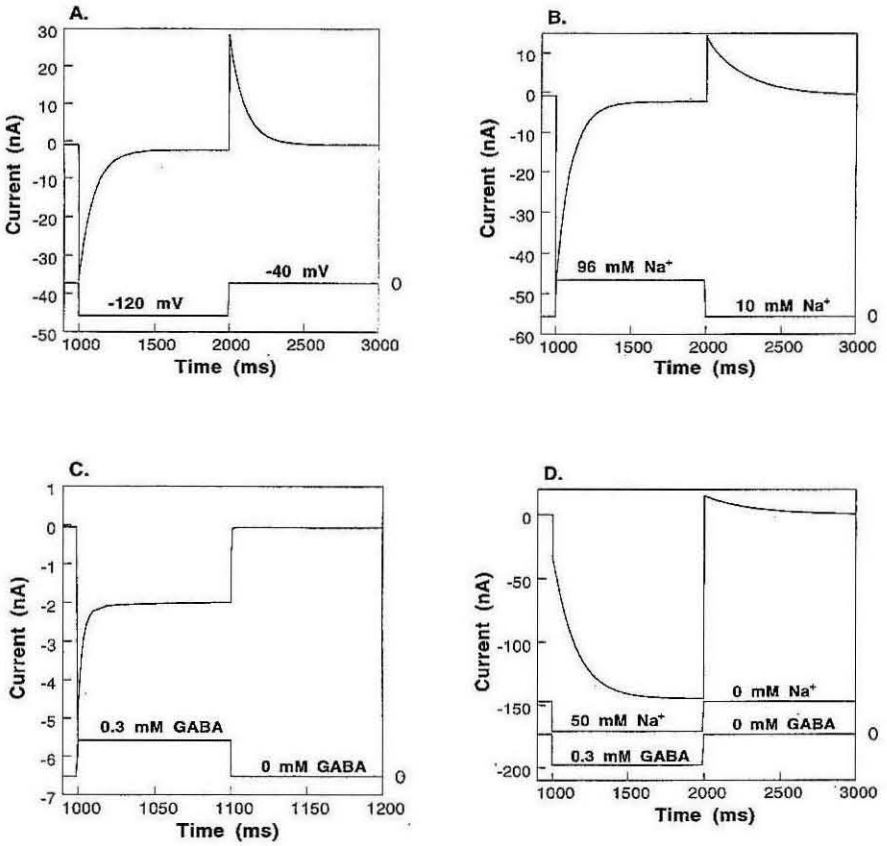


FIGURE 5 Simulated GAT1 transient currents. (*A*) Transient current for voltage jumps in the absence of GABA, at $[Na^+]$ = 96 mM. The membrane potential was held at −40 mV, stepped to −120 mV at time = 1 s, and stepped back to −40 mV at time = 2 s. (*B*) Transient current during $[Na^+]$ jumps in the absence of GABA, with membrane voltage clamped at −120 mV. $[Na^+]$ was held at 10 mM, stepped to 96 mM at time = 1 s, and stepped back to 10 mM at time = 2 s. (*C*) Transient current during [GABA] jumps, with $[Na^+]$ = 10 mM and membrane voltage clamped at −10 mV. [GABA] was held at 0 mM, stepped to 0.3 mM at time = 1 s, and stepped back to 0 mM at time = 1.1 s. (*D*) Transient current during simultaneous $[Na^+]$ and [GABA] jumps, with voltage clamped at −120 mV. $[Na^+]$ was held at 0 mM, stepped to 50 mM at time = 1 s, and stepped back to 0 mM at time = 2 s. [GABA] was held at 0 mM, stepped to 0.3 mM at time = 1 s, and stepped back to 0 mM at time = 2 s.

rising phase followed by a decay over several hundred milliseconds. For depolarizing jumps, the direction of the transient is reversed. The integral of these relaxations is represented by the steady-state charge distributions (Fig. 4 *E*); and we have verified that these simulated charge movements are equal and opposite for the jumps in both directions between a pair of potentials (Mager et al., 1993). The simulated relaxation time constant lies in the range of 70–150 ms and displays a maximum at negative membrane potentials (not shown), in agreement with experimental data (Mager et al., 1993; Cammack et al., 1994; S. Mager, preliminary unpublished results).

Simulated $[Na^+]$ and [GABA] jump-induced transients are shown in Fig. 5, *B* and *C*, respectively; and Fig. 5 *D* shows simulated simultaneous jumps of both substrate concentrations. There are no published data available for comparison with Fig. 5 *B*. The waveforms in Fig. 5, *C* and *D*, however, reproduce the general features observed by Cammack et al. (1994). Thus, the simulated jump from 0 to 0.3 mM GABA produces a transient peak on the millisecond time scale, followed by a sustained transport-associated current (Fig. 5 *C*) and compares well with figure 9 of Cammack et al. (1994). Furthermore, the simulated jump to 50 mM Na^+ and 0.3 mM GABA produces a rapid rise followed by a slower rise on the time scale of several hundred milliseconds, in agreement with figure 9 of Cammack et al. (1994). The simulated relaxations are more rapid than the published records for the jump back to 0 [GABA] and 0 $[Na^+]$, and there is a small simulated outward charge movement in Fig. 5 *D*; but these discrepancies could arise if the experimental solution changes are slightly slower than expected.

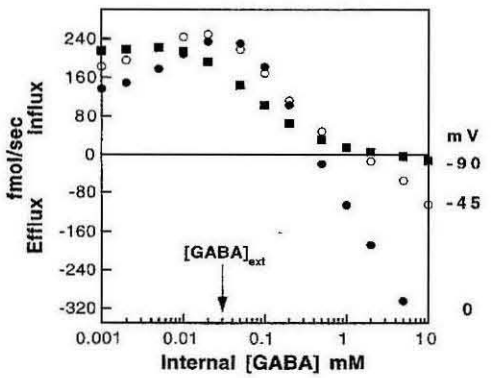


FIGURE 6 GAT1 transport up a concentration gradient. Simulation of Na^+-coupled GABA transport against a GABA concentration gradient. The external and internal [Na^+] were set to 96 and 5 mM, respectively. The external [GABA] was set to 0.03 mM, and the internal [GABA] was varied. GABA fluxes were simulated for membrane potentials of 0 mV, −45 mV, and −90 mV. Thus the electrochemical gradient for Na^+ alone favors Na^+ influx. Net GABA *influx* still occurs despite the fact that the [GABA] gradient alone favors GABA *efflux*.

Transport against a [GABA] gradient

In all of the simulations above, the internal concentrations of substrates are assumed to be zero. We relaxed this assumption to demonstrate the most important functional role of ion-coupled cotransporters: they drive the organic substrate up its concentration gradient by using the electrochemical potential gradient of an ionic substrate. For simplicity—and in the absence of real data on this point—we assumed that the forward binding constants k_i are the same for the internal and external sides of the membrane. The external and internal [Na^+], and membrane potential are adjusted to drive the Na^+ flow inward. The internal [GABA] was set higher than the external [GABA]; therefore net GABA efflux would occur if GABA cannot couple to the electrochemical potential gradient of Na^+. However, the simulation results in a net GABA influx (Fig. 6), suggesting that the multi-substrate single-file transport model does support transport up the chemical gradient of the organic substrate.

The attraction and repulsion terms

In the multi-substrate single-file transport model, the physical meanings of k_i, h_i, and d_s are transparent and represent common concepts in membrane transport. We comment here on the physical meaning of the attraction and repulsion parameters, A_{ij} and R_{ij}, respectively. Consider, for example, A_{12} and R_{12}, and focus on the transitions between transporter configurations 14 and 16.

$$k_{14\to16} = A_{12}h_2 \tag{11}$$

$$k_{16\to14} = R_{12}h_2 \tag{12}$$

We can define an equilibrium "association constant" for the state in which substrate 2 (GABA) is adjacent to substrate 1 (Na^+):

$$K_{12} = \frac{k_{16\to14}}{k_{16\to14}} = \frac{A_{12}}{R_{12}} = K_o \exp(-\Delta G_{12}/kT) \tag{13}$$

where K_o is a constant standard state factor involving the translational partition function of the ion (McQuarrie, 1976), k is the Boltzmann constant, T is the absolute temperature, and ΔG_{12} is the relative free energy between any pair of states, such as 14 and 16, in which these two substrates move adjacent to each other. If there are no couplings between substrates 1 and 2, then we expect no difference in the relative free energy, which will lead to a ratio of $A_{12}/R_{12} = 1$. However, if A_{12}/R_{12} is not equal to 1, the two substrates are said to be coupled. For instance, if the optimized values of A_{12} and R_{12} are such that $A_{12}/R_{12} > 1$, moving the two substrates together is favored. By more classical arguments, the ratio of k_i $[i]/h_i$ controls the energetics of moving substrate i into the ion-coupled transporter. Generalizing this argument to attractive and repulsive couplings between substrates, we obtain the following rules for GAT1:

1. Moving Na^+ into the transporter is favored at high [Na^+].
2. Moving GABA into the transporter is favored at high [GABA].
3. Moving Na^+ and GABA adjacent to each other is favored.
4. Moving Na^+ and Na^+ adjacent to each other is favored. This is the most surprising result of our analysis.
5. Moving GABA and GABA adjacent to each other is not favored.

According to these microscopic rules, at high [Na^+] and [GABA], we expect transporter configurations with more than two GABA molecules to be underpopulated. This is indeed the case. Under such conditions, the simulated population of transporter configurations 20, 21, 22, 23, 25, and 26 is negligible (Fig. 2). This may be due to the favored adjacencies between Na^+ and GABA (Rule 3).

Simulation of 5-HT transporter function

5-HTT is another ion-coupled neurotransmitter transporter belonging to the plasma membrane Na^+/Cl^--coupled family. It is known to accumulate 5-HT^+ in the presence of Na^+, Cl^-, and K^+. Na^+ and K^+ are cotransported and countertransported with 5-HT^+, respectively. The argument that Cl^- is cotransported with 5-HT^+ is somewhat less direct, as it has been difficult to demonstrate 5-HT^+ accumulation with only the Cl^- gradient as a driving force. The K^+ is coupled to the 5-HT^+ accumulation but is not required, for the transport still occurs in the absence of K^+ (Rudnick and Clark, 1993). Therefore, only Na^+ and 5-HT^+ are included in our model. Our 5-HT transporter model, like our GAT1 model, has three binding sites.

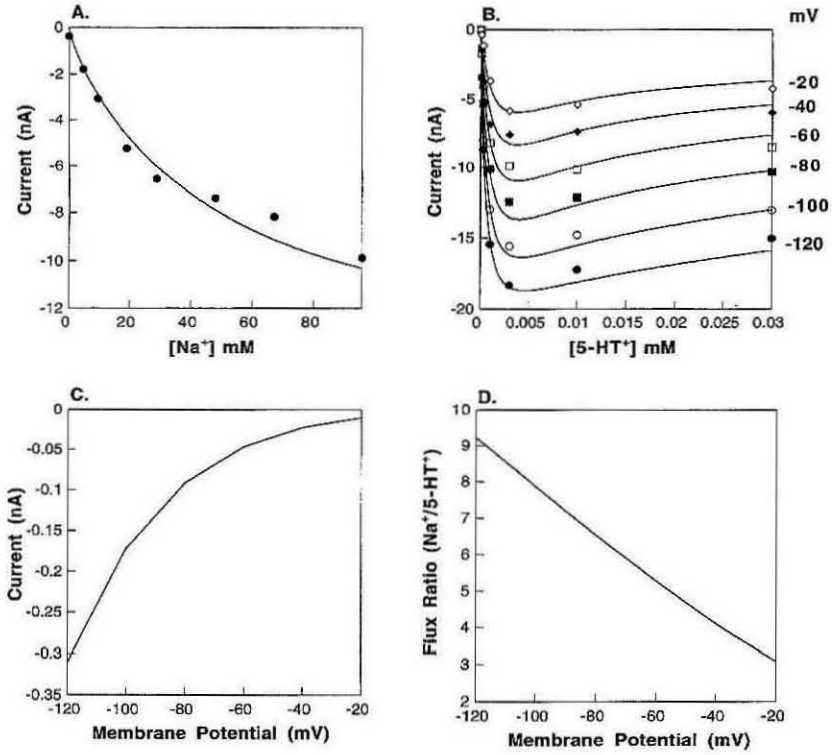


FIGURE 7 Simulated versus experimental 5-HTT steady-state properties. (*A*) Steady-state current as a function of [5-HT$^+$]. Experimental data were taken at [Na$^+$] = 96 mM, with the membrane potential varying from −120 to −20 mV (Mager et al., 1994). (*B*) Steady-state current as a function of [Na$^+$]. Data were taken at [5-HT$^+$] = 3 μM, with membrane potential set to −60 mV (Mager et al., 1994). (*C*) Simulated leakage current at zero 5-HT, under the same experimental conditions as in *A*. (*D*) Flux ratio of Na$^+$ to 5-HT$^+$ simulated at [Na$^+$] = 96 mM, [5-HT$^+$] = 1 μM, versus membrane voltage.

The experimentally determined permeation properties of 5-HTT include its steady-state currents as a function of Na$^+$ and 5-HT$^+$, leakage currents (Na$^+$ flux in the absence of 5-HT$^+$), and variable stoichiometry (5–12 Na$^+$ transported per 5-HT$^+$) (Mager et al., 1994). The phenomena of leakage currents and variable stoichiometry are not expected from the alternating access model. The simulation reproduced the above experimental results under one set of model parameters. The simulated 5-HTT steady-state current showed a nonsigmoidal dependence on [Na$^+$] (Fig. 7 *A*) and a biphasic dependence on [5-HT$^+$] (Fig. 7 *B*). Leakage currents on the order of 1 nA were observed in the absence of external 5-HT$^+$ (Fig. 7 *C*). An important feature of 5-HT transport is its variable stoichiometry (voltage-dependent flux ratio of Na$^+$/5-HT$^+$). This is simulated in Fig. 7 *D*.

We comment here on several critical parameters that underlie two unique permeation properties of the 5-HT transporter: the nonsigmoidal dependence on [Na$^+$] and the biphasic dependence on [5-HT$^+$]. The nonsigmoidal dependence can be made sigmoidal by increasing the population of transporter configurations 5, 6, 7 (occupied by two Na$^+$) relative to 2, 3, 4 (occupied by one Na$^+$). The biphasic dependence can be abolished either a) by increasing the intrinsic hopping frequency ratio between Na$^+$ and 5-HT$^+$ or b) by increasing the positive coupling between 5-HT$^+$ and its modulators. This implies that the slower movement of 5-HT$^+$ can block the movement of Na$^+$, thereby causing decreasing current at higher [5-HT$^+$].

Simulation of SGLT1 function

SGLT1 is a Na$^+$/glucose cotransporter that shares a postulated membrane topology and functional properties with GAT1 (Wright, 1993). Na$^+$ and sugar (the nonmetabolizable derivative α-methyl-D-glucose is typically used in physiological studies) appear to be the only substrates; thus the number and charge of its substrates equal those for the simplified model of GAT1 discussed above. SGLT1 has been modeled previously as a carrier (Parent et al., 1992b). Our simulations approximated the *I-V* curves (Fig. 8, *A* and *B*) and the experimental substrate flux ratio (Fig. 8 *C*) under one set of model parameters. Table 2 reveals that several of the kinetic parameters for SGLT1 are greater than those for GAT1, so that SGLT1 has an overall higher turnover rate.

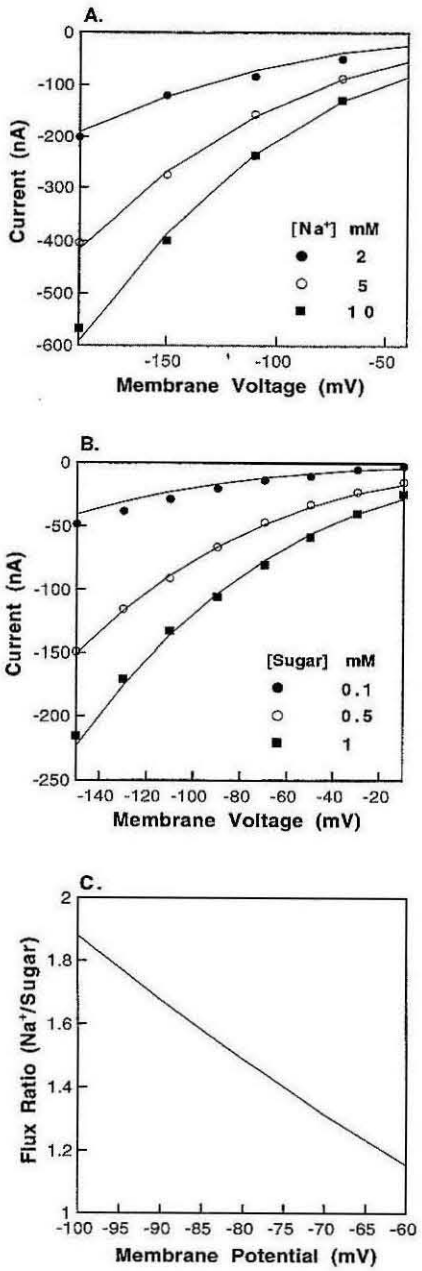


FIGURE 8 Simulated versus experimental SGLT1 steady-state properties. (*A*) Steady-state current as a function of membrane potential. Experimental data were taken at [sugar] = 1 mM, with [Na^+] varying from 2 to 10 mM and as indicated. (*B*) Steady-state current as a function of membrane potential. Data were taken at [Na^+] = 10 mM, with [sugar] varying from 0.1 to 1 mM and as indicated. Assumed number of transporters in the membrane for *A* is roughly twice that for *B*. (*C*) Flux ratio of Na^+ to [sugar] simulated at [Na^+] = 10 mM, [sugar] = 6 mM, as a function of voltage from −100 to −60 mV.

There are also differences in the electrical positions of the binding sites between these two transporters, accounting for the different voltage sensitivities of transport.

DISCUSSION

Our simulations show that it is possible to account for the major features of ion-coupled transport without invoking explicit conformational changes that control the compartmentalization of substrates. The present multi-substrate single-file transport model differs from the usual alternating access models, which suggest that the transporter possesses a "gate" at both ends of the lumen, that the gates are alternately exposed to one side of the membrane, and that the gating is controlled by substrate binding at appropriate sites and with appropriate stoichiometry. In the multi-substrate single-file transport framework, the only structural assumption is the protein's capability to provide a linear lumen and to mediate substrate-substrate interactions. This assumption seems plausible in light of present knowledge about permeation through ion channels.

Ion-coupled transporters show a wide range of permeation properties, including leakage currents, variable stoichiometry, substrate flux coupling, and voltage-dependent transients. Alternating-access models account for only some of these properties. Tight substrate flux coupling and voltage-dependent charge movements naturally result from these models, whereas variable stoichiometry and leakage currents are inherently hard to explain with these models. In contrast, the present scheme covers a broad spectrum of permeation properties. It can be applied to GAT1, which shows leakage currents, tight substrate flux coupling, and voltage-dependent gating currents. The model can also be used to interpret the variable stoichiometry of the 5-HT transporter. The flexibility of the model can be largely attributed to the minimal assumptions made about transporter structures. One of our colleagues remarked that our model resembles a diffusion pump.

Transport cycles

In an alternating-access model (see, for instance, Rudnick and Clark, 1993), there is a definite series of transitions among steps, looping back to a starting state after a complete set of substrates has been transported; and this series constitutes the transport cycle. In the multi-substrate single-file transport model, there are many more possible states than in other models; which are the major steps in the transport cycle? To address this question, we trace out major transport cycles in Fig. 2 as follows. We start from the empty transporter (configuration 1) and choose the transition with the highest net flux. If the two greatest net fluxes differ by less than 5%, we choose both. These steps are repeated until closed loops are formed. Fig. 9 *A* shows the subset of states and transitions from Fig. 2 that are included in the two major transport cycles for GAT1, under the same experimental conditions as in Fig. 4 *D*. The two transport

TABLE 2 Optimized parameter values for GAT1, 5-HT, and SGLT1

Parameter	Physical meaning	Units	GAT1	5-HT transporter	SGLT1
k_1	Forward binding constant (Na^+)	$M^{-1}s^{-1}$	170.	56.3	6.1×10^5
k_2	Forward binding constant (organic substrate)	$M^{-1}s^{-1}$	6.45×10^5	35,100	3.1×10^5
h_1	Intrinsic hopping frequency (Na^+)	s^{-1}	1.02	3.59	340
h_2	Intrinsic hopping frequency (organic substrate)	s^{-1}	12.6	1.00	1055
d_1	Electrical location of binding sites ($0 < d < 1$)		0.5	0.509	0.164
d_2			0.692	0.701	0.674
d_3			0.8	0.901	0.894
A_{10}	Attractive interaction between Na^+ and an empty site		0.133	8.62×10^{-3}	6.42×10^{-7}
A_{20}	Attractive interaction between organic substrate and an empty site		0.05	6.85	1.23
A_{11}	Attractive interaction between Na^+ and Na^+		9.83	9.79	4.04
A_{12}	Attractive interaction between Na^+ and organic substrate		40001	2.73	8.61
A_{22}	Attractive interaction between organic substrate and organic substrate		40.5	0.0526	0.297
R_{10}	Repulsive interaction between Na^+ and an empty site		9.74	1.04	0.506
R_{20}	Repulsive interaction between organic substrate and an empty site		60.5	3.35×10^{-3}	8.21
R_{11}	Repulsive interaction between Na^+ and Na^+		0.257	3.42	9.70
R_{12}	Repulsive interaction between Na^+ and organic substrate		10001	4.05	2.07
R_{22}	Repulsive interaction between organic substrate and organic substrate		701	8.38	8.73
ΔG_{10}	Repulsive interaction between Na^+ and an empty site		2.54	2.84	8.04
ΔG_{20}	Repulsive interaction between organic substrate and an empty site		4.20	−4.51	1.12
ΔG_{11}	Repulsive interaction between Na^+ and Na^+		−2.16	−0.62	0.52
ΔG_{12}	Repulsive interaction between Na^+ and organic substrate		−0.82	0.23	−0.84
ΔG_{22}	Repulsive interaction between organic substrate and organic substrate		1.69	3.00	2.00

For GAT1, 5-HTT, and SGLT1, five, eight, and two sets of parameters (respectively) yielded nearly equal cost function values for the steady-state data. For GAT1, the parameters given are those that also simulated the transient data (see Fig. 5). For the 5-HT transporter and SGLT1, parameters with the lowest cost function values are presented. Substrate 1 = Na^+ in all cases. For GAT1, 5-HTT, and SGLT1, substrate 2 = GABA, 5-HT^+, and sugar, respectively.

cycles operate simultaneously; one releases a Na^+, and the other releases both a Na^+ and a GABA. Together, they provide an explanation for the observed substrate flux ratio ($1 < Na^+/GABA < 2$).

To understand the origin of "variable stoichiometry," a similar transport cycle analysis is applied to the 5-HT transporter (Fig. 9 *B*). The transport cycle couples one 5-HT^+ release to a loop of Na^+ releases. This loop may be responsible for the observed variable stoichiometry, because various numbers of Na^+ molecules can be released per 5-HT^+ molecule. If this hypothesis is correct, it can further explain the voltage dependence of variable stoichiometry (Fig. 7 *D*). Upon hyperpolarizing, the steady-state population of transporter configurations 5, 6, 7 and the transition frequency from configuration 3 to 5 are both increased, which accelerates the loop and stimulates more Na^+ flux per 5-HT^+ molecule transported.

Computational strategy

Most previous modeling studies have used the King-Altman diagrammatic method to analyze the steady-state distribution of transporters (e.g., Hill, 1977). However, the number of King-Altman patterns increases dramatically as the number of states increases. This constrains the number of practical carrier states. One could in principle automate the King-Altman procedure to analyze more complicated kinetic schemes. However, the time required to list the King-Altman patterns grows as N^{N-1}, where N is the number of carrier states. The present model uses Gauss-Jordan elimination to solve the steady-state distribution. Gauss-Jordan elimination is computationally different from the King-Altman diagrammatic method, but gives the same solution. Because Gauss-Jordan elimination scales as N^3, where N is the number of carrier states, it allows for the consideration of a significantly larger number of states.

Limitations and future directions

The major limitation of our theory in its present form is its inability to deal with the phenomena of antiport, countertransport, exchange, or counterflow, as exemplified by Na^+/Ca^{2+} exchangers, by the glutamate trans-

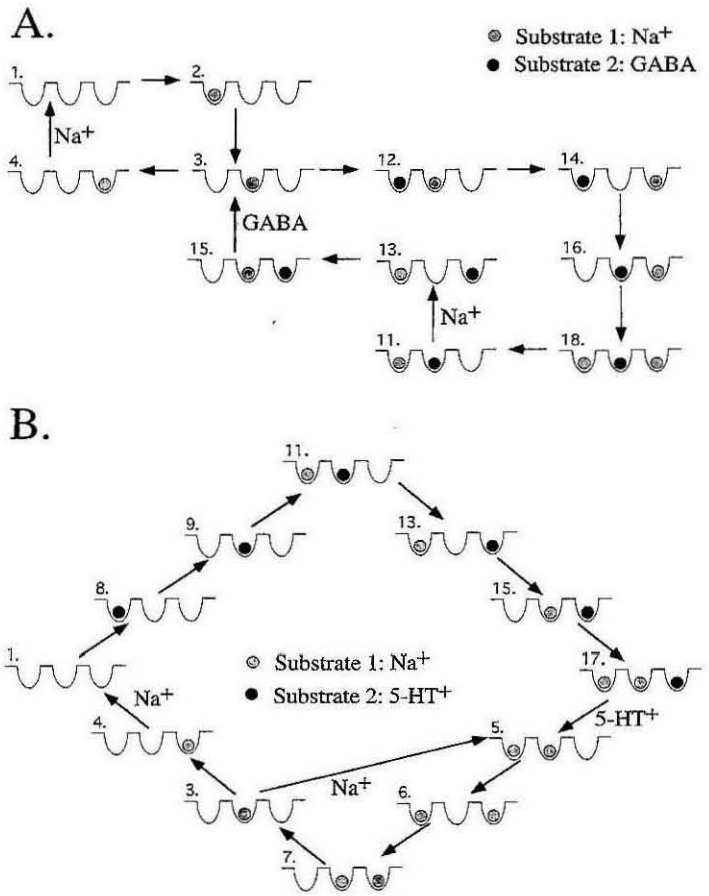


FIGURE 9 Major transport cycles for GAT1 and 5-HTT. (*A*) Major transport cycles for GAT1 traced out by the method described in text. Two transport cycles operate simultaneously. Together, they provide an explanation to the observed substrate flux ratio ($1 < Na^+/GABA < 2$). (*B*) Major transport cycles for 5-HTT. The cycle couples one $5\text{-}HT^+$ release to a loop of Na^+ releases. This loop may be responsible for the observed variable stoichiometry.

porters of the nervous system, by the 5-HT transporter (Rudnick and Clark, 1993), and under some conditions by GABA transporters (Mabjeesh and Kanner, 1989). Fig. 6 shows that our model does simulate a small (20–50%) increased influx through GAT1 as internal [GABA] is increased; but we have not systematically studied this effect. We suspect that models with alcoves or Y-shaped lumens would be required to account for robust antiport (Eisenberg, 1994).

The model is also uninformative about other aspects of transporter function. It neglects the frank channel activity, with conductances on the order of picosiemens, that has recently been observed with GAT1 (Cammack and Schwartz, 1995) and with the 5-HT transporter (F. Lin et al., unpublished observations) and inferred from noise studies (see Introduction). Finally, we have not yet considered ATPase transporters.

We do not wish to imply that our model treats the lumen of the transporter as a featureless series of chemically inert pockets. We suspect that movements of the side chains lining the lumen could provide the attractive and repulsive interactions (quantified by A_{ij} and R_{ij}) that we postulate. We emphasize, however, that we view these motions as local, probably restricted to <0.5 nm from the bound substrates in question, as opposed to the global changes in compartmentalization that characterize alternating-access models. Some of the interactions represented by the optimized parameters disagree with intuition; for instance, rule 4 suggests that two neighboring Na^+ ions constitute a relatively stable configuration. Such interactons might be produced by side chains that interpose between two substrates, leading either to charge pairs or to cation-pi interactions between substrates and side chains (Dougherty, 1995).

Structure of ion-coupled transporters

Our theory must remain at the formal level, primarily because we know so little about transporter structure. Although we do know the primary amino acid sequence

for the transporters simulated here, there is no proof for topology diagrams postulated on the basis of the sequence, no information about secondary or tertiary structure, and little data about the residues that actually line the lumen that we postulate. Until such information becomes available, there is little objective reason to choose our model over the alternating-access model and its variants. Although some may prefer the present model because of its simplicity, nature does not seem to prefer simple mechanisms for her macromolecuar machines.

CONCLUSION

The hypothesis of "multi-substrate single-file transport" has been tested and found to account quantitatively for the steady-state data of ion-coupled transporters GAT1, 5-HTT, and SGLT1. In addition, the model predicts the GAT1 transporter's transient responses to voltage and concentration jumps. The model also simulates transport up a concentration gradient, an important feature of ion-coupled transporters. Model analysis such as major transport cycles gives insights about the tight flux coupling of GAT1, leakage currents, and variable stoichiometry of 5-HT transport. Inspection of model parameter values yields a set of microscopic rules that may explain the simulated population distribution of transporter configurations. The success of the model suggests that concepts such as "alternating access" or "gates" are not necessary in interpreting existing experimental results and perhaps can be recast into the present, more general, framework.

We thank Eric Bax, Bassil Dahiyat, Norman Davidson, and Jun Li for suggestions. Scott Fraser suggested the diffusion pump analogy.

This work was supported by grants from the National Institute of Neurological Diseases and Stroke and the National Institute on Drug Abuse. SLM acknowledges support from the Rita Allen Foundation, the David and Lucille Packard Foundation, and the Searle Scholars Program.

REFERENCES

Amara, S. G., and J. L. Arriza. 1993. Neurotransmitter ion-coupled transporters: three distinct gene families. *Curr. Opin. Neurobiol.* 3:337–344.

Andersen, O. S., J. E. N. Silveira, and P. R. Steinmetz. 1985. Intrinsic characteristics of the proton pump in the luminal membrane of a tight urinary epithelium. The relation between transport rate and $\Delta\mu H$. *J. Gen. Physiol.* 86:215–234.

Baldwin, S. A. 1993. Mammalian passive glucose ion-coupled transporters: members of an ubiquitous family of active and passive transport proteins. *Biochim. Biophys. Acta.* 1165:17–49.

Barnett, J. E. G., G. D. Holman, R. A. Chalkley, and K. A. Munday. 1975. Evidence for two asymmetric conformational states in the human erythrocyte sugar-transport system. *Biochem. J.* 145:417–429.

Cammack, J. N., S. V. Rakhilin, and E. A. Schwartz. 1994. A GABA ion-coupled transporter operates asymmetrically and with variable stoichiometry. *Neuron.* 13:1–20.

Cammack, J. N., and E. A. Schwartz. 1993. Ions required for the electrogenic transport of GABA by horizontal cells of the catfish retina. *J. Physiol.* 473:81–102.

Cammack, J. N., and E. A. Schwartz. 1995. Channel behavior in a GABA transporter. *Proc. Natl. Acad. Sci. USA.* In press.

Chen, D. P., and R. S. Eisenberg. 1993. Flux, coupling, and selectivity in ionic channels of one conformation. *Biophys. J.* 65:727–746.

Dani, J. A., and D. G. Levitt. 1990. Diffusion and kinetic approaches to describe permeation in ionic channels. *J. Theor. Biol.* 146:289–301.

DeFelice, L. J., R. Galli, and D. Blakely. 1995. Current fluctuations in norepinephrine transporters. *Biophys. J.* 68:A232.

Dougherty, D. A. 1995. Cation-π interactions in chemistry and biology. *Science.* In press.

Eisenberg, R. S. 1994. Atomic biology, electrostatics, and ionic channels. *In* New Developments and Theoretical Studies of Proteins. Ron Elbor, editor. World Scientific Publishing, Philadelphia. 1–116.

Franciolini, F., and W. Nonner. 1994. A multiion permeation mechanism in neuronal background chloride channels. *J. Gen. Physiol.* 104:725–746.

Frolhich, O. 1988. The "tunneling" mode of biological carrier-mediated transport. *J. Membr. Biol.* 101:189–198.

Gadsby, D. C., R. F. Rakowski, and P. De Weer. 1993. Extracellular access to the Na, K pump: pathway similar to ion channel. *Science.* 260:100–103.

Galli, A., L. DeFelice, B.-J. Duke, K. R. Moore, and R. D. Blakely. 1995. Sodium-dependent norepinephrine-induced currents in norepinephrine-transporter-transfected HEK-293 cells blocked by cocaine and antidepressants. *J. Exp. Biol.* 198:2197–2212.

Guastella, J. H., N. Nelson, H. Nelson, L. Czyzyk, S. Keynan, M. C. Midel, N. Davidson, H. Lester, and B. Kanner. 1990. Cloning and expression of a rat brain GABA ion-coupled transporter. *Science.* 249:1303–1306.

Harvey, W. R., and C. L. Slayman. 1994. Coupling as a way of life. *J. Exp. Biol.* 196 (Transporters):1–4.

Hasegawa, H., W. Skach, O. Baker, M. C., Calayag, V. Lingappa, and A. S. Verkman. 1992. A multifunctional aqueous channel formed by CFTR. *Science.* 258:1477–1479.

Hernandez, J. A., and J. Fischbarg. 1994. Transport properties of single-file pores with two conformational states. *Biophys. J.* 67:996–1006.

Hilgemann, D. W., D. A. Nicoll, and K. D. Philipson. 1991. Charge movement during Na^+ translocation by native and cloned cardiac Na^+/Ca^{2+} exchanger. *Nature.* 352:715–718.

Hill, T. L. 1977. Free Energy Transduction in Biology. Academic Press, New York.

Hille, B. 1992. Ionic Channels of Excitable Membranes. Sinauer Associates, Sunderland, MA.

Ingber, L. 1993. Simulated annealing: practice versus theory. *J. Math. Comput. Modelling.* 18:29–57.

Jardetzky, O. 1966. Simple allosteric model for membrane pumps. *Nature.* 211:969–970.

Kanner, B. I., and S. Schuldiner. 1987. Mechanism of transport and storage of neurotransmitters. *CRC Crit. Rev. Biochem.* 22:1–38.

Keynan, S., and B. I. Kanner. 1988. γ-Aminobutyric acid transport in reconstituted preparations from rat brain: coupled sodium and chloride fluxes. *Biochemistry.* 27:12–17.

Kimmich, G. A., and J. Randles. 1988. Na^+-coupled sugar transport: membrane potential-dependent K_m and K_i for Na^+. *Am. J. Physiol.* 255:C486–C494.

Krupka, R. M. 1989. Role of substrate binding forces in exchange-only transport systems. II. Implications for the mechanism of the anion-exchanger of red cells. *J. Membr. Biol.* 109:159–171.

Krupka, R. M., and R. Deves. 1988. The choline carrier of erythrocytes. Location of the NEM-reactive thiol group in the inner gated channel. *J. Membr. Biol.* 101:43–47.

Lagnado, L., K. Cervetto, and P. A. McNaughton. 1988. Ion transport by the Na-Ca exchanger in isolated rod outer segments. *Proc. Natl. Acad. Sci. USA.* 85:4548–4552.

Lauger, P. 1979. A channel mechanism for electrogenic ion pumps. *Biochim. Biophys. Acta.* 552:143–61.

Lauger, P. 1991. Electrogenic Ion Pumps. Sinauer Associates, Sunderland, MA.

Lester, H. A., S. Mager, M. W. Quick, and J. L. Corey. 1994. Permeation properties of neurotransmitter ion-coupled transporters. *Annu. Rev. Pharmacol. Toxicol.* 34:219–249.

Lowe, A. G., and A. R. Walmsley. 1986. The kinetics of glucose transport in human red blood cells. *Biochim. Biophys. Acta.* 857:146–154.

Mabjeesh, N. J., and B. I. Kanner. 1989. Low-affinity γ-aminobutyric acid transport in rat brain. *Biochemistry.* 28:7694–7699.

Mager, S., C. Min, D. Henry, N. Davidson, C. Chavkin, B. Hoffman, and H. A. Lester. 1994. Conducting states of a mammalian serotonin transporter. *Neuron.* 12:845–859.

Mager, S., J. Naeve, M. Quick, C. Labarca, N. Davidson, and H. A. Lester. 1993. Steady states, charge movements, and rates for a cloned GABA ion-coupled transporter expressed in *Xenopus* oocytes. *Neuron.* 10:177–188.

McQuarrie, D. A. 1976. Statistical Mechanics. Harper & Row, New York.

Nakamoto, R. K., R. Rao, and C. W. Slayman. 1989. Transmembrane segments of the P-type cation-transporting ATPases. A comparative study. *Ann. N.Y. Acad. Sci.* 574:165–179.

Parent, L., S. Supplisson, D. D. F. Loo, and E. M. Wright. 1992a. Electrogenic properties of the cloned Na^+/glucose cotransporter. I. Voltage clamp studies. *J. Membr. Biol.* 125:49–62.

Parent, L., S. Supplisson, D. D. F. Loo, and E. M. Wright. 1992b. Electrogenic properties of the cloned Na^+/glucose cotransporter. II. A transport model under nonrapid equilibrium conditions. *J. Membr. Biol.* 125:63–79.

Picaud, S. A., H. P. Larsson, G. B. Grant, H. Lecar, and F. S. Werblin. 1995. A glutamate gated chloride channel with glutamate transporter-like properties in cone photoreceptors of the tiger salamander. *J. Neurophysiol.* 74:1760–1771.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical Recipes in c. 36–41. Cambridge University Press, Cambridge.

Rakowski, R. F. 1993. Charge movement by the Na/K pump in *Xenopus* oocytes. *J. Gen Physiol.* 101:117–114.

Risso, S., L. DeFelice, and R. D. Blakely. 1995. Sodium-dependent GABA-induced current in GAT1-transfected HeLa cells. *J. Physiol.* In press.

Rudnick, G., and J. Clark. 1993. From synapse to vesicle: the reuptake and storage of biogenic amine neurotransmitters. *Biochim. Biophys. Acta.* 1144:249–263.

Schultz, S. G. 1980. Basic Principles of Membrane Transport. Cambridge. University Press, Cambridge. 33–41, 85, 95, 96.

Schultz, S. G. 1986. Ion-coupled transport of organic solutes across biological membranes. *In* Membrane Physiology. T. E. Andreoli, J. F. Hollfman, D. D. Fanestil, and S. G. Schultz, editors. Plenum Press, New York. 283–294.

Schwartz, E. A., and M. Tachibana. 1990. Electrophysiology of glutamate and sodium co-transport in a glial cell of the salamander retina. *J. Physiol.* 426:32–80.

Stein, W. D. 1986. Transport and Diffusion across Cell Membranes. Academic Press, Orlando, FL. 337–361, 613–616.

Umbach, J. A., M. J. Coady, and E. M. Wright. 1990. Intestinal Na^+/glucose cotransporter expressed in *Xenopus* oocytes is electrogenic. *Biophys. J.* 57:1217–1224.

Wadiche, J. I., S. G. Amara, and M. P. Kavanaugh. 1995b. Ion fluxes associated with excitatory amino acid transport. *Neuron.* 15: 721–728.

Wadiche, J. I., J. L. Arriza, S. G. Amara, and M. P. Kavanaugh. 1995a. Kinetics of a human glutamate transporter. *Neuron.* 14:1019–1027.

Walmsley, A. R. 1988. The dynamics of the glucose ion-coupled transporter. *Trends Biochem. Sci.* 13:226–231.

Wang, J. S., J. M. Tang, and R. S. Eisenberg. 1992. A calcium conducting channel akin to a calcium pump. *J. Membr. Biol.* 130:163–181.

Wright, E. M. 1993. The intestinal Na^+/glucose ion-coupled cotransporter. *Annu. Rev. Physiol.* 55:575–89.

Chapter 6

Summary

Summary

Two independent biophysical projects were completed: the Protein Backbone Design Project and the Ion-Coupled Transporter Simulation Project. Chapter 2, 3 and 4 are dedicated to the first project, and Chapter 5 is devoted to the second project.

In chapter 2, I assessed the effect of explicit backbone motion on the selection of amino acids in designing the core of the β 1 domain of the streptococcal protein G. I used a protein design algorithm that quantitatively considers side-chain packing. Concerted backbone motion was introduced by varying Gβ1's supersecondary structure parameter values. The stability and structural flexibility of seven of the redesigned proteins were determined experimentally and showed that core variants containing as many as 6 of 10 possible mutations retain native-like properties. This result demonstrates that backbone flexibility can be combined explicitly with amino acid side-chain selection and that the selection algorithm is sufficiently robust to tolerate perturbations as large as 15% of Gβ1's native supersecondary structure parameter values.

In chapter 3, I sought to expand the range of computational protein design by developing a general, quantitative design method for computing de novo backbone templates. The method had to compute atomic resolution backbones compatible with the atomistic sequence selection algorithm I was using and it had to be applicable to all protein motifs. The algorithm I developed uses supersecondary structure paramters to determine the orientation among secondary structural elements, given a target protein fold. Possible backbone arrangements are screened using a cost function which evaluates core packing, hydrogen bonding, loop closure, and backbone torsional geometry. Given a specified number of residues for each secondary

structral element, a family of optimal configurations is found. I chose three motifs to test our method (ββα, βαβ, and αα) since their combination can be used to approximate most possible backbone fold. The best structure found for the ββα motif is similar to a zinc finger, and the best structure for the βαβ motif is similar to a segment of a β-barrel (PDB entry 8adh). The backbone obtained for the αα motif resembles minimized protein A. These results suggest that de novo backbones assembled using this method may serve as adequate input templates for atomistic sequence selection algorithms.

In chapter 4, I evaluated the backbone design method by testing the thermal stability and structural properties of the designed peptides. I also explored relevant issues for integrating computer-generated backbones with the sequence-selection algorithm. Starting with a computer-generated bba motif, I optimized five sequences for this backbone and characterized them using CD and 1D NMR. It was found that small differences in the number and location of the hydrophobic residues can significantly change the thermodynamic behaviour of the designed peptides. This supports the previously acknowledged importance of binary patterning. I compared our five sequences with FSD-1, a thermally stable ββα protein with a well-defined structure developed in our lab using the same sequence-selection algorithm (Dahiyat & Mayo, 1997). I found that the sequence that had the binary pattern most similar to FSD-1 also had the best thermodynamic properties. Based on the results of these five peptides, a set of heuristic rules was derived which could be used to improve the computer-generated backbones. In addition to the number and location of hydrophobic residues, I found that it is critical to determine the optimal number of residues for each secondary structural element. Validation of these rules will be the focus of future experimental efforts.

In chapter 5, I developed a model to simulate ion-coupled transporters. The model differs from contemporary alternating-access schemes. Beginning with concepts derived from multi-ion pores, the model assumes that substrates (both inorganic ions and small organic molecules) hop a) between the solutions and binding sites and b) between binding sites within a single-file pore. No two substrates can simultaneously occupy the same site. Rate constants for hopping can be increased both a) when substrates in two sites attract each other into a vacant site between them and b) when substrates in adjacent sites repel each other. Hopping rate constants for charged substrates are also modified by the membrane field. For a three-site model, simulated annealing yields parameters to fit steady-state measurements of flux coupling, transport-associated currents, and charge movements for the GABA transporter GAT1. The model then accounts for some GAT1 kinetic data as well. The model also yields parameters that describe the available data for the rat 5-HT transporter and for the rabbit Na^+-glucose transporter. The simulations show that coupled fluxes and other aspects of ion transport can be explained by a model that includes local substrate-substrate interactions but no explicit global conformational changes.

References

Dahiyat BI, Mayo SL. 1997. De novo protein design - fully automated sequence selection. *Science 278* 82-87.