Dynamic and Stochastic Protein Simulations:

From Peptides to Viruses

Thesis by

Alan Martin Mathiowetz

In partial fulfillment of the requirements
for the degree of Doctor of Philosophy

California Institute of Technology

Pasadena, California

1993

(Submitted October 30, 1992)

To Laura.

Acknowledgements

My years at Caltech have had an intensity which I may miss some day, many years from now. The people here, however, I will miss as soon as I'm gone. I am thankful that almost every person I have met at Caltech has made the stay more enjoyable. As an advisor, Bill Goddard has been an unforgettable example of what happens when you mix hard work with inspiration, insight, and intelligence. I can only hope that a small part of his chemical intuition, breadth of knowledge, and depth of understanding has rubbed off on me. And I wouldn't mind gaining a bit of his patience and humor, as well.

I cannot possibly acknowledge all the people who have had an impact on me over the years, both as coworkers and as friends. The Goddard group alone has had more than fifty members during my stay. (Hey! I'm not the only one who has been here this long!) However, I would like to acknowledge several people who have contributed to my work. Adel Naylor was my guide to the entire field of computational biology during the early years and has been an invaluable colleague and good friend ever since. Kevin Plaxco has been a seemingly limitless source of information about "real" biology, while Naoki Karasawa has helped me to understand just about every computational method I've ever needed to employ. I would also like to thank Terry Coley, Kian-Tat Lim, and Jon Hurley for answering thousands of questions about computers in general and programming in particular. Abhinandan Jain of JPL has been great to work with over the past ten months, while I've worked on making a BIOGRAF verion of his NEIMO dynamics. And Jerry Solomon has provided great insights into protein folding through his work on lattice simulations. I also

acknowledge the great work Siddharth Dasgupta has done managing the MMSC and thank him for years of being a mentor to me and the rest of the group. Finally, I want to thank Adria McMillan, Carol Scrivner, and Debbie Chester for doing all the really hard work that has to be done, and for making the Goddard Group a nicer place to work.

It would take several more pages to list all the people who have been good friends over the years, and who have made the time here more enjoyable. However, I have to single out Bob Driscoll and Mike Kosinski, as well as Terry Coley, for being dedicated weightlifting partners for many years. But Bob wasn't supposed to get *better* than I am. I also want to thank Erik Bierwagen, Charles Musgrave, Jason Perry, and the rest of the people in the Goddard Group, for being such entertaining coworkers, friends, and, sometimes, neighbors. And I especially thank Guanhua Chen for being as great a roommate as he is a scientist, if that's possible. Finally, I want to acknowledge my great appreciation for Paul Smith, who has been both a friend and an inspiration. Science is a noble pursuit, and those who pursue it have given their lives to a worthy cause. But there is a higher calling and a deeper truth to pursue. And Paul has had the courage to follow God wherever He leads.

I feel fortunate to have had a home away from Caltech at Lake Avenue Congregational Church. So I have to thank the men and women of the 7:45 Network Class for keeping me sane and giving me perspective, as well as providing friendship, prayer support, and a lot of fun. Most of all, I thank Dave Plumb, who has been listening to me talk about protein folding for years, without telling me to shut up and talk about something interesting. Dave stood by me through some very rough times as well as some of the best days of my life.

But the people who have been there the longest are my family and I cannot possibly say how much they mean to me. I thank my Dad for always challenging me to go as far as I can, but also for being a friend no matter where I am. And I thank my Mom for her unfailing support and endless optimism. She is the perfect mother for a graduate student and I am eternally thankful for her. My sister, Denise, is also a great source of encouragement and her sense of humor is often like a cool drink on a hot, dry Southern California day. Finally, my brother Brad has always been my best friend, even when he lives 3000 miles away. I hope he always will be. Now that I'm getting out of here, maybe I can see him and Shelly more often – and, of course, Joshua and Jessica, too!

I've saved the best for last and the best is Laura, my wife. I feel like God has blessed me in a thousand ways and has given me more than I could ever deserve. But His greatest gift, besides His own Son, has been the gift of a wife like Laura. Laura has made my life a million times better than it used to be. She will always be my greatest love, my truest friend, and my closest companion. And she helped a lot with this Thesis – proof-reading and aligning figures, certainly, but also by making the rest of my life so enjoyable during the process.

So, I close by thanking God for these people and for getting me through this phase of my life. Whatever good may come of this work, the credit belongs to Him. After all, He already knows how proteins fold.

Thesis Abstract

In order to increase the efficiency of protein simulations, both deterministic and stochastic methods can be formulated in terms of the most important degrees of freedom in polypeptide and protein systems: the torsions. Two such methods are presented here. The first is Newton-Euler Inverse Mass Operator (NEIMO) Dynamics, an internal-coordinate molecular dynamics method originally designed to study the dynamics of general multibody systems. The second is the Probability Grid Monte Carlo (PGMC) method, developed for searching the conformational space of polypeptides using a weighted sampling of the most favorable dihedral angles.

The first use of the NEIMO Dynamics method for studying molecular systems is reported here. The method is used to study the dynamics of a wide range of peptide and protein systems. These range from the pentapeptide Met-enkephalin to the crystallographic asymmetric unit of the tomato bushy stunt virus (TBSV), an assembly of three chains totaling 893 residues. Bond lengths and angles do not vary during the dynamics simulations; this enables timesteps larger than 10 femtoseconds to be used for small peptides, a substantial improvement over Cartesian coordinate molecular dynamics. Timesteps of 10 fs do not work well for NEIMO simulations of large proteins because of unacceptably large energy fluctuations. However, timesteps of 2-5 fs give acceptable results, even for very large systems. The NEIMO method is applied to TBSV coat proteins, in an investigation of the effect of Ca^{2+} ions on the coat stability.

The PGMC method provides efficient conformational searches for polypeptide systems by assigning probabilities to different discrete values of the ϕ, ψ, and χ

dihedral angles. These probabilities were derived by investigation of the protein structures in the Brookhaven Protein Database. The PGMC method is applied successfully to several important problems in protein modeling: studies of the low-energy conformations of a peptide, prediction of the all-atom conformation of a protein from its C_α coordinates alone, and the prediction of antibody loop conformations. The success of the C_α modeling is further extended by its application to structures with coordinates constrained to a lattice, through the use of a simple C_α Forcefield.

Contents

Acknowledgements iv

Thesis Abstract vii

List of Figures xii

List of Tables xix

1 Simulation Techniques for Proteins 1

I. Protein Simulations 1

II. Molecular Mechanics 3

II.A. Forcefields 4

II.B. Energy Minimization 8

II.C. Molecular Dynamics 10

III. Monte Carlo 12

References 15

2 NEIMO Dynamics of Polypeptides 17

I. Introduction 18

II. Methodology 20

III. Results 26

III.A. Timing 28

III.B. Energy Fluctuations 28

III.C. Dihedral Distributions 42

IV. Conclusions 50

A. Leapfrog Verlet for Internal Coordinates 56

References 58

3 Simulations of the Tomato Bushy Stunt Virus Capsid 60

I. Introduction 61

II. Methodology 63

III. Results 70

IV. Conclusions 78

References 79

4 Probability Grid Monte Carlo 80

I. Introduction 81

II. Methodology 82

II.A. The Protein Database Subset 85

II.B. Backbone (ϕ, ψ) Probability Grids 86

II.C. Sidechain (χ) Probability Grids 97

II.D. Note on Computations 105

III. Conformations of Met-Enkephalin 105

IV. Conclusions 125

References 128

5 Protein modeling from C_α coordinates 130

I. Introduction 131

II. Methodology 134

II.A. General Methodology 134

II.B. Variables . . . 138

III. Results . . . 146

III.A. Crambin . . . 146

III.B. Larger Proteins . . . 159

IV. The C_α Forcefield . . . 163

V. Conclusions . . . 173

References . . . 174

6 Prediction of Loop Conformations in Antibodies 176

I. Introduction . . . 177

II. Methodology . . . 178

III. Antibody Hypervariable Loops . . . 183

IV. Conclusions . . . 192

References . . . 199

List of Figures

2.1 The peptide Met-enkephalin is shown, with the bonds/hinges numbered. Clusters are the chemical units connected by the hinges, such as the phenyl ring in tyrosine 1. The hinges are numbered to allow for analysis of the dihedral angles. 22

2.2 Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of Met-enkephalin. Simulations were run for 1 ps using timesteps ranging from 1 to 20 fs. N* and C* are the scaled fluctuations, $\mathcal{E}$*, where $\mathcal{E}$ is divided by the number of degrees of freedom: $\mathcal{N}$ for NEIMO simulations and 3n-6 for Cartesian coordinates. 31

2.3 Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of (Ala)$_9$. Simulations were run for 1 ps using timesteps ranging from 1 to 35 fs. N* and C* are the scaled energy fluctuations, $\mathcal{E}$*. 33

2.4 δ_{E} for 100 timesteps of NEIMO dynamics on (Ala)$_9$. 35

2.5 Avian pancreatic polypeptide (aPP), with the sidechain atoms removed for clarity. From the crystal structure 1PPT [17]). 35

2.6 Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of avian pancreatic polypeptide (aPP). Simulations were run for 1 ps using timesteps ranging from 1 to 15 fs, but all those above 11 fs caused the energy to blow up. N* and C* are the scaled energy fluctuations, $\mathcal{E}$*. 37

2.7 Scaled energy fluctuations, $\mathcal{E}^*$, for 1 ps NEIMO simulations of aPP. "Rigid H" differs from "Normal" NEIMO in that hinges which rotate only hydrogen atoms are held fixed. The "Counterions" simulation used the standard NEIMO method for the protein, but concurrently solved the Cartesian equations of motion for counterions (5 Na^+ and 3 Cl^-) added to neutralize unpaired charges. 39

2.8 The average energy fluctuations, $\langle\mathcal{E}\rangle$, during 5 ps simulations of avian pancreatic polypeptide. Fluctuations in NEIMO(N) and Cartesian(C) dynamics were determined at 0.1 ps intervals during the course of the simulation, after which velocities could be rescaled and the CMM nonbond farfield calculation was updated. 41

2.9 $\langle\mathcal{E}\rangle$ vs. protein size for Cartesian dynamics at 1fs and NEIMO dynamics at various timesteps. 42

2.10 During 5 ps molecular dynamics simulations of Met-enkephalin, the 22 dihedral angles were written out at 0.1 ps intervals. The fifty values for each dihedral are plotted here for Cartesian and NEIMO dynamics simulations using 1 fs timesteps. 44

2.11 The average dihedrals from the distributions in Figure 2.10 are shown here with error bars indicating $\pm\sigma$, the standard deviations. 47

2.12 The average dihedrals from NEIMO simulations using timesteps ranging from 1 to 10 fs. 48

2.13 The overlaps S_{12} between 1 fs and 2-10 fs NEIMO simulations of Met-enkephalin. 51

2.14 The overlaps S_{12} between 1 fs and 2, 5, and 10 fs NEIMO simulations of Met-enkephalin shown at higher resolution than Figure 2.13. 52

2.15 The overlaps S_{12} between dihedral distributions from Cartesian dynamics vs. those from NEIMO dynamics simulations with timesteps ranging from 1 fs. 53

3.1 The tomato bushy stunt virus asymmetric unit, showing the A, B, and C conformations as well as the associated Ca^{2+} ions. From the crystal structure by Olson, *et al.*[3]. 62

3.2 Detail of the aspartic acid/Ca^{2+} interactions at the contact site between the S domains of two coat proteins. 64

3.3 The TBSV protein coat, with each P and S domain represented by a sphere. 65

3.4 A second view of the TBSV coat, with the outer P domains removed to emphasize the symmetry of the virus. 66

3.5 CPU times for CMM calculations, NEIMO acceleration calculations, and overhead, including coordinate updating for NEIMO. 71

3.6 $\langle\mathcal{E}\rangle^*$ during 1.0 ps NEIMO simulations. 73

3.7 The radius of gyration during energy minimization. 74

3.8 The potential energy of the TBSV triad, including CMM nonbonded interactions with the full viral coat, during energy minimization. . . . 75

3.9 The TBSV radius during 2.0 ps Cartesian and NEIMO dynamics simulations. 76

3.10 The radius of the PH7 and NOCA models of TBSV during 4.0 ps of Cartesian (NVT) and NEIMO dynamics. 77

3.11 Potential energy during the 3.0 ps Cartesian canonical dynamics simulations. 77

4.1 The backbone (ϕ, ψ, and ω) and sidechain (χ) dihedrals of arginine. The two outermost sidechain dihedrals, $\chi^{6,1}$ and $\chi^{6,2}$ rotate hydrogens, only, so they are not varied in the Probability Grid Monte Carlo method. 84

4.2 30° ϕ, ψ grids for the three standard residue types. 89

4.3 ϕ, ψ grids for standard (non-Proline, non-Glycine) residues at grid spacings of 10, 15, 30, and 60°. 90

4.4 ϕ, ψ grids of different structural types for standard (non-Proline, non-Glycine) residues in the SS58 dataset. 98

4.5 χ grids for sidechains with one PGMC dihedral. 103

4.6 χ grids for sidechains with two significant dihedrals. 104

4.7 Results from several Monte Carlo simulations of Met-enkephalin. The simulations were identical except for the temperature used. The starting structure of each simulation was the 15° peak conformation and 15° ϕ, ψ and χ grids were used for conformational sampling. At 100 step intervals, the energy of the current conformation and the best overall conformation were recorded. The two graphs plot the current(a) and best energy(b) vs. Monte Carlo step. 115

4.8 The best energy from 40 simulations of Met-enkephalin using 15° grid spacing at 1000 K. 20 simulations began with conformations generated at random (R) and 20 began with the peak 15° conformation, 15°_i, (I). 116

4.9 PGMC simulations were run at temperatures of 0 K, 300 K, 600 K, 1000 K and 5000 K, for each grid spacing (5°, 10°, 15°, 30°, and 60°). For each spacing/temperature pair, ten simulations of 10,000 steps each were run. Here, the overall acceptance rate for the ten runs is shown for each spacing/temperature combination. 118

4.10 For each of 25 temperature/grid spacing combinations, ten 10,000-step simulations were run and the minimum energy from each run was recorded. The plots show both the overall lowest energy for the ten runs (a) and the average minimum energy (b) for the ten runs. . . 120

4.11 20 simulations of 50,000 steps each were run at 300 K using grid spacings of 5° and 10°. The best conformation from each run was energy minimized and the energy before and after minimization was recorded. The top two lines plot the unminimized energies while the bottom two, labeled "(min)", show the minimized energies. 122

4.12 The peptide Met-enkephalin, with the dihedrals numbered as in Table 4.15. 123

4.13 The dihedral angles from each of the 20 minima produced by 50,000-step simulations at 300 K. Only the dihedrals unsampled by the PGMC method, χ^1 (dihedral 4) and the ω's (6, 9, 12, and 17) are the same in every conformation. Even these degeneracies are broken during minimization. 124

4.14 The top plot shows the distribution of dihedral angles for the eight conformations with energies within 1 kcal/mol of LS_i, the conformation with the lowest known energy . The bottom plot shows the same distribution minus the one conformation which differs significantly from the others at dihedrals 16, 18, and 19. The other seven conformations have virtually identical conformations for residues Phe 4 and Met 5. 126

5.1 Three views of crambin: the C_α's (46 atoms), the peptide backbone (185 atoms), and the all-atom structure (402 atoms). From the crystal structure by Hendrickson and Teeter[1]. 132

5.2 Definition of virtual angle, ζ, and dihedral, γ, for residue i. 136

5.3 The average rms deviation from the crystal structure for models of the crambin backbone built using various temperatures and pulse sizes.141

5.4 The average rms deviation in backbone atoms (RMSB) and C_α coordinates (RMSC) for crambin backbone models built using different C_α constraint force constants. 142

5.5 Average backbone rms vs. the number of Monte Carlo steps. Also shown is the average time to build each backbone conformation. . . . 144

5.6 Average all-atom rms deviations for sidechain Monte Carlo simulations of crambin at different temperatures. 146

5.7 Backbone rms per residue for crambin model. 151

5.8 The deviation in χ^1 for each residue of the crambin model. 153

5.9 The peptide backbone of the model and crystal structures of crambin. The rms deviation is 0.538 Å. 155

5.10 A comparison of helix 2 (residues 23 to 30) in the model and crystal structures. The rms deviation is 1.026 Å for all atoms and 0.394 Å for the backbone atoms. 156

5.11 Helix 1 (residues 7 to 19) in the model and crystal structures. The rms deviation is 1.658 Å for all atoms and 0.209 Å for the backbone atoms. 157

5.12 The Hierarchical Protein Folding Strategy, which converts lattice C_α coordinates into all-atom protein conformations. 166

5.13 The probability distribution (0.01 Å resolution) for C_α–C_α bonds. The actual distribution is compared to that derived from the bond-stretch term of the C_αFF. 167

5.14 2000 virtual angle, dihedral (ζ, γ) values. 168

5.15 The ζ, γ probability grids for the entire H64 dataset, using 15° bins. . 169

5.16 The ζ, γ probability grids for HELIX and SHEET residues in the H64 dataset. 170

6.1 A schematic diagram of an IgG molecule. 184

6.2 The six hypervariable loops of the immunoglobulins HyHEL-5[14] and McPC603[15], shown as C_α coordinates only. 185

6.3 C_α traces of the predicted and actual loop conformations of HyHEL-5. 193

6.4 C_α traces of the predicted and actual loop conformations of McPC603. 194

6.5 Backbone conformations of the predicted (yellow) and actual (red) loop conformations in HyHEL-5. 195

6.6 Backbone conformations of the predicted (yellow) and actual (red) loop conformations in McPC603. 196

List of Tables

2.1 The clusters, hinges, and related dihedrals of Met-enkephalin, shown in Figure 2.1. 24

2.2 Proteins and peptides used in NEIMO simulations. The structures listed are the initial Protein Database files, except for the peptides "MEnk" and "Ala9", which were created using the BIOGRAF peptide builder. 27

2.3 Timing results for the ten protein/peptides systems studied here. The average times per timestep of the NEIMO calculation and the nonbond calculation are given, along with the NEIMO time divided by $\mathcal{N}$, and the nonbond time divided by the number of atoms, n. 29

2.4 The average values of the Met-enkephalin dihedrals from 5 ps NEIMO ($\langle\theta\rangle_N$) and Cartesian ($\langle\theta\rangle_C$) dynamics simulations, compared to the initial values θ_0 and compared to each other. 46

4.1 Crystal structures used in the H64 dataset. 87

4.2 The number of ϕ, ψ samples of the three residue types in the protein datasets. 91

4.3 This table lists the number of (and percentage of the maximum possible) ϕ, ψ gridpoints which have non-zero values for each of the three residue types at different grid spacings, S. 92

4.4 This table lists the number of high-probability gridpoints: the number which have probabilities above $\langle P\rangle$. 93

4.5 This highest-probability gridpoint of each residue type for each grid spacing. 95

4.6 The percentage of sample ϕ, ψ's falling within each quadrant of ϕ, ψ conformational space. 96

4.7 The distribution of secondary structure designations in the crystal structures of the SS58 dataset. 99

4.8 The number of PGMC χ dihedrals (N_χ) for each amino acid, the number of occurrences of each amino in the H64 crystal structures, and the number of populated (non-zero) gridpoints at each grid spacing. $N_\chi = 0$ for alanine and glycine. The numbers in italics are more than 95% of the sample size, indicating that nearly every conformation occupies a different gridpoint. 100

4.9 The total number of gridpoints for N_χ dihedrals at different grid spacings. 101

4.10 The number of populated gridpoints in the χ^1,χ^2 distribution of amino acids with $N_\chi \geq 2$. Lysine appears to be the most flexible while tryptophan is the least flexible. 102

4.11 The highest-probability gridpoint for each amino acid for $S = 30°$. P_{max} is the probability of this particular conformation. 106

4.12 Conformations generated for Met-enkephalin from peak ϕ, ψ, and χ gridpoints at different grid spacings. Because of steric overlap, the energy of the 5° conformation is greater than 1×10^6 kcal/mol. 109

4.13 Conformations generated from peak ϕ, ψ, and χ gridpoints at different grid spacings, after minimization with the DREIDING forcefield. The energy of each conformation is given, along with the RMS deviation from its original (un-minimized) structure. 111

4.14 LS_i is the structure produced by rotating the dihedrals of our starting structure to the values reported for the ECEPP/2 global minimum [17]. LS_m was obtained by a Cartesian-coordinate conjugate-gradients minimization of LS_i using the DREIDING forcefield. 113

4.15 Dihedral numbering used in Figure 4.13 and Figure 4.14. 125

5.1 ζ, γ regions indicating residue i is likely to be in an α helix or β sheet conformation; i.e., its ϕ and ψ fall within the corresponding ϕ, ψ region listed in the lower table. ζ_i and γ_i are defined in Figure 5.2. 137

5.2 Variables used in PGMC C_α Builder. For Phase 1, "steps" refers to the number of conformations sampled as each residue is added. For Phase 2, it refers to the total number of conformations sampled. . . . 139

5.3 Values used for production runs of the C_α Builder. 147

5.4 The energy, rms deviation in backbone atoms (RMSB), and rms deviation in ϕ, ψ dihedrals (RMSD) for each of the 20 backbone conformations generated for crambin. 148

5.5 The energy and rms deviation in atomic coordinates for each of the crambin models produced by the PGMC C_α Builder. 149

5.6 RMS deviations for different regions of the crambin model. 150

5.7 The rms deviations in various types of dihedrals for the crambin model and the percentage of each type of dihedral with deviations less than 30° or more than 90°. 154

5.8 The proteins modeled by the PGMC C_α Builder. The reference crystal structure is given along with the number of residues in the protein and the percent of these which are in α helices and β sheets. 160

5.9 The results from Phase 1 constructions of the backbone conformations of several proteins. 161

5.10 The results from Phase 2 constructions of the sidechains of crambin, plastocyanin, and flavodoxin. 162

5.11 A comparison of the results for flavodoxin vs. other methods. "Correct" refers to dihedrals predicted to within 20° of their crystal structure values. 163

5.12 Equilibrium geometries and force constants in the C_αFF. 171

5.13 Best-fit fcc lattice conformations of several proteins, before and after minimization with the C_αFF. 171

5.14 Results from building all-atom conformations from various C_α conformations of crambin using the PGMC C_α Builder. 172

6.1 The loop residues of HyHEL-5 and McPC603. 187

6.2 The rms deviations for all atoms, backbone atom (BB), and C_α coordinates are given for the predicted conformations of the McPC603 hypervariable loops. 188

6.3 The lowest-energy loops created in phases 1 and 2 of simulations of HyHEL-5. 189

6.4 The best rms deviation of C_α coordinates from the crystal structure from among the 1000 loops generated in Phase 1. 191

6.5 The results from this work (PGMC) compared to results from three different methods: a conformational-search algorithm[18], a method which uses the canonical structures of loops other than H3 (L3 of HyHEL-5 also does not fit one of the canonical structures)[16], and a method which combines conformational searching with comparisons to database conformations[8]. 197

Chapter 1

Simulation Techniques for Proteins: Molecular Dynamics and Monte Carlo

I. Protein Simulations

Advances in many fields have led to an accelerated demand for computational tools which can be used to study proteins. Biophysical techniques such as X-ray crystallography and multidimensional NMR are heavily dependent on computers for data retrieval and analysis, as well as for model-building of protein structures from those data. These techniques have provided a wealth of information about protein structures, much of which can serve as a starting point for computational studies of the protein's dynamics, thermodynamics, and substrate interactions. Spectacular advances in gene cloning and sequencing have provided tremendous amounts of protein sequence data, which demand computational analysis. New sequences are often compared against a gigantic library of other sequences in hope of discovering any structural similarities and evolutionary relationships. Sequence data is accumulating even faster than structural data, so there is great demand for computational techniques which can provide structural informational about the protein using both sequence and homology data. The eventual goal is a technique for predicting the three-dimensional structure of a protein from the sequence alone. Advances in computer power and computational techniques and the promise of continual advances in the future, have encouraged scientists to believe that computational solutions to these and other problems are possible, today or in the near future.

A wide variety of computational techniques have been applied to the study of proteins. The area of protein structure prediction is particularly wide-open, because the task seems too daunting for more traditional techniques. Among the many theoretical approaches applied to this goal are neural networks[1], lattice simulations[2, 3], structural profiles[4], and analysis of patterns in sequence alignments[5]. The most popular techniques for studying proteins, however, remain molecular modeling and molecular mechanics. The two terms are often used interchangeably, but the former refers primarily to the graphical display of molecules and the manipulation of these structures to obtain structural insights, while the latter refers to underlying computational techniques for analyzing molecular structure, dynamics, thermodynamics, and other properties. Molecular modeling techniques are widely used for predicting protein structures from structural homology[6] and understanding enzyme-substrate interactions[7]. Molecular mechanics techniques, especially molecular dynamics, are used in a wide variety of applications. It is not now feasible to simulate the entire folding process of a protein using molecular dynamics (folding can take several seconds or even minutes, but molecular dynamics usually uses timesteps of 10^{-15} seconds), but the unfolding process can be studied to gain insight into intermediates in the folding process[8]. Other applications of molecular dynamics to proteins include the calculation of relative free energies of substrate binding to enzymes[9] and analysis of protein-solvent interactions[10].

II. Molecular Mechanics

"Molecular Mechanics" refers to computational techniques which use classical mechanics to analyze the structure and dynamics of molecular systems including biological macromolecules, organic compounds, polymers, and materials. These systems are composed of atoms which are treated as classical particles, whose interactions are described by simple two-, three-, and four-body potential energy functions. This classical forcefield-based approach is a great simplification over quantum chemistry, which describes systems in terms of nuclei, electrons, and orbitals. This simplicity allows molecular mechanics to be applied to much larger systems than can be studied by *ab initio* methods. Tremendous improvements in computer power and computational methodology have accelerated the pace towards simulation of larger and larger systems, so that today simulations of a million particles is possible. Such advances have also enabled researchers to obtain much greater information from their simulations: more accurate calculation of physical and chemical properties or simulations of much longer dynamical processes.

The simplest calculation in molecular mechanics is a calculation of the potential energy of the system, which is performed by summing the numerous energy terms for the given conformation of the system using the given set of potential energy functions and parameters. Optimizing the structure of a system can be done by "energy minimization" which improves the conformation by reducing the gradients and the energy of the system. A great deal more information about a system can be obtained from molecular dynamics simulations. In these calculations, the motions of the particles are followed by calculating the forces from the forcefield and, from this, the accelerations and velocities. Careful control of the energy and temperature of the system ensures that the conformations produced form a statistical ensemble, from which thermodynamic and other properties can be calculated.

II.A. Forcefields

Forcefields enable the potential energy of a molecular system to be calculated rapidly and fairly accurately. A typical forcefield represents each atom in the sytem as a single point and energies as a sum of two-, three-, and four-particle interactions. The potential energy of a particular interaction is described by an equation which involves the positions of the particles and a small number of parameters which have been determined experimentally or by quantum mechanical calculations. For large systems with many particles, the equations are usually quite simple in order to allow for a rapid calculation of the total energy. There is often a trade-off between simplicity and accuracy. For instance, the bond between two particles i and j can be described by a harmonic potential,

$$V_b(i,j) = \frac{1}{2}K_b(R_{ij} - R_{eq})^2, \tag{1.1}$$

or a Morse potential:

$$V_b(i,j) = \frac{1}{2}D_0[e^{-\alpha(R_{ij}-R_{eq})} - 1]^2. \tag{1.2}$$

Here, R_{ij} is the distance between the two particles; K_b, R_{eq}, and D_0 are force constant, equilibrium geometry, and bond energy parameters, respectively; and $\alpha = \sqrt{K_b/2D_0}$. The Morse potential is more accurate, especially when R_{ij} is significantly larger than the equilibrium bond distance. The harmonic potential, however, is calculated very fast and gives reasonable answers for bonds near their equilibrium geometries. Forcefields designed for proteins and nucleic acids almost always use the simpler harmonic form.

During the past ten years, several forcefields have been developed for protein simulations. Those having the most widespread usage are AMBER [11] and CHARMm [12]. Recently, the DREIDING forcefield was enhanced and published[13]. Although this forcefield is much more general than either AMBER or CHARMm, it is equally effective for protein simulations. All three are "united atom" forcefields: hydrogens bonded to carbons are not treated specially but are treated as a unit with the carbon atom. The three forcefields also share many of the same potential functions. However, specific parameters, such as force constants and equilibrium geometries, and atom type assignments, i.e., which parameters should be used for which atoms in a simulation, are different. Most of the calculations presented here used the DREIDING forcefield. However, in some instances results are reported for simulations using AMBER for the sake of comparison.

Valence interactions. The overall potential energy of a molecular system is typically described by a forcefield like

$$V(\mathbf{r}) = V_b + V_\theta + V_\phi + V_i + V_{el} + V_{vdw} + V_{hb}, \tag{1.3}$$

where the energy potential terms are either between bonded atoms ("valence interactions") or through-space ("nonbonded interactions"). Valence interactions include bonds(V_b), angles(V_θ), torsions(V_ϕ), and inversions(V_i). The nonbonded interactions are electrostatic(V_{el}), van der Waals(V_{vdw}), and hydrogen bonds(V_{hb}). The valence interactions are generally quite simple, like the bond energy terms in Equation (1.1). The total bond energy is a sum over all bonds in the system, which is typically very close to the number of atoms:

$$V_b = \sum_{bonds} \frac{1}{2} K_b (R_{ij} - R_{eq})^2. \tag{1.4}$$

The angle term is very similar and is also the same in both DREIDING and AMBER:

$$V_\theta = \sum_{angles} \frac{1}{2} K_\theta (\theta_{ijk} - \theta_{eq})^2. \tag{1.5}$$

Here, θ_{ijk} is the angle formed by the atoms i, j, and k. Torsion terms are also treated identically in DREIDING and AMBER, but the equation is very different from the equations for bonds and angles:

$$V_\phi = \sum_{torsions} \sum_{n=1}^{6} \frac{1}{2} K_{\phi,n}[1 - d\cos(n\phi)]. \tag{1.6}$$

Here, each four-body torsion is itself a sum of up to six terms, each of which can have its own periodicity. The periodicity is determined by n, while $d(= \pm 1)$ determines whether the term has a maximum at $\phi = 0°$ or at $\phi = 180°/n$.

The most complex term in a typical protein forcefield is the inversion term, which is added to ensure that a particular atom i, which is bonded to three other atoms j,k, and l, remains planar or non-planar. AMBER and DREIDING treat this term differently. AMBER uses the angle ψ between the lij and kil planes and the equation:

$$V_i = \sum_{inv} \frac{1}{2} K_\psi \cos[n(\psi - \psi_{eq})]. \tag{1.7}$$

Planarity is enforced by $n = 2$ and a tetrahedral geometry is enforced by $n = 3$. The DREIDING forcefield uses a different angle ϕ between the ijk and ljk planes and a simpler harmonic term:

$$V_i = \sum_{inv} \frac{1}{2} K_\phi (\phi - \phi_{eq})^2. \tag{1.8}$$

Note that ϕ and ψ are unrelated to the important ϕ and ψ backbone dihedrals of proteins. For more details, see the AMBER[11] and DREIDING[13] papers.

Nonbonded interactions. The number of valence interactions that must be calculated for a molecule is usually proportional to the number of atoms, n. The number of nonbonded terms, however is roughly proportional to n^2, because they involve almost all possible pairs of atoms. It is slightly less than n^2 because two atoms involved in a particular bond or angle are not considered to have a through-space interaction. Also, it is very common to ignore interactions between atoms too far apart in space (typically, more than 9 Å), even though this technique can be quite inaccurate (N. Karasawa and H.Q. Ding, unpublished data). Nevertheless, for large systems, the bulk of computational time is spent calculating the nonbonded interactions, so a great deal of work has been done to optimize these calculations for vector and parallel processors. There has also been considerable work in developing new methods of accurate nonbond calculations which are proportional to n (see Reference [4]). Nevertheless, most calculations are done using the n^2 techniques.

Both van der Waals and electrostatic interactions are calculated over pairs of atoms, so they are usually done concurrently:

$$V_{el} + V_{vdw} = \sum_i \sum_{j>i} \left\{ \frac{q_i q_j}{\epsilon R_{ij}} + D_0 \left[\left(\frac{R_{eq}}{R_{ij}} \right)^{12} - 2 \left(\frac{R_{eq}}{R_{ij}} \right)^{6} \right] \right\}. \qquad (1.9)$$

If all atoms in a system are explicitly included in a calculation, the vacuum dielectric constant ($\epsilon = 1$) should be used. However, ϵ is often set proportional to R_{ij}, ostensibly to represent the electrostatic screening effect of solvent atoms when they not present, but more practically to make the electrostatic term proportional to $1/R_{ij}^2$ rather than $1/R_{ij}$. This speeds calculations substantially because R_{ij}^2 can be directly calculated from the Cartesian coordinates of i and j without requiring a lengthy square-root calculation. Each forcefield includes van der Waals radii and well depths for each atom type. The equilibrium bond strength D_0 in Equation (1.9) is the geometric mean of the van der Waal's well depths of the individual atoms i and j. The equilibrium bond length R_{eq} is the arithmetic mean of the two van der Waals radii.

Hydrogen atoms are treated specially. The AMBER forcefield assigns charges to hydrogens but does not give them van der Waals parameters. Instead, it uses "off-diagonal" van der Waals terms. In other words, these are special terms for i, j interactions when $i \neq j$, rather than simply using the averages of the individual atomic terms. These interactions do not use the Lennard-Jones 12-6 potential, but rather a Lennard-Jones 12-10 potential, which goes to zero much more quickly:

$$V_{vdw,H} = \sum_i \sum_{j>i} D_0 \left[5 \left(\frac{R_{eq}}{R_{ij}} \right)^{12} - 6 \left(\frac{R_{eq}}{R_{ij}} \right)^{10} \right]. \tag{1.10}$$

DREIDING treats hydrogens even more unusually. Hydrogens are not given charges or van der Waals parameters, so Equation (1.9) does not apply at all. Rather, the DREIDING forcefield has a special hydrogen bond term for D–H–A interactions, where D is the hydrogen bond donor, H is the hydrogen bonded to it covalently, and A is the hydrogen bond acceptor, non-covalently attached. The DREIDING hydrogen bond uses both a radial R_{DA} and an angular θ_{DHA} part:

$$V_{hb} = \sum_i \sum_{j>i} D_0 \left[5 \left(\frac{R_{eq}}{R_{\mathrm{DA}}} \right)^{12} - 6 \left(\frac{R_{eq}}{R_{\mathrm{DA}}} \right)^{10} \right] \cos^2 \theta_{\mathrm{DHA}}. \tag{1.11}$$

Both the radial and angular parts are set to zero beyond certain cutoff values and switching functions are used to make the transition to $V_{hb} = 0$ smooth. See Reference [15] for details.

II.B. Energy Minimization

The potential energy calculated by summing the energies of various interactions is a numerical value for a single conformation. This number can be used to evaluate a particular conformation, but it may not be a useful measure of a conformation because it can be dominated by a few bad interactions. For instance, a large molecule with an excellent conformation for nearly all atoms can have a large overall energy because of a single bad interaction, for instance two atoms too near each other in space and having a huge van der Waals repulsion energy. It is often preferable to carry out energy minimization on a conformation to find the best nearby conformation. Energy minimization is usually performed by gradient optimization: atoms are moved so as to reduce the net forces on them. The minimized structure has small forces on each atom and therefore serves as an excellent starting point for molecular dynamics simulations.

Energy minimization is usually performed in Cartesian coordinates, by optimizing along pathways in $3n$-dimensional space, where n is the number of particles. This pathway can be the gradient, ∇, where

$$\nabla_x = \frac{\partial V}{\partial x}. \tag{1.12}$$

In other words, each Cartesian component, x, of the gradient equals the derivative of the potential energy with respect to that component. Only those interactions involving particle i contribute to the gradients of the Cartesian coordinates of i (x_i, y_i, z_i). The $3n$ components of ∇ constitute a path, $\mathbf{P}$, in $3n$-dimensional space. Finding the minimum along this pathway typically involves an interpolation of two points in $3n$-space to find a new point where $\nabla \cdot \mathbf{P} = 0$. Usually, however, $|\nabla| \neq 0$ at the new point, so a new path is chosen and minimization proceeds. It is possible to set $\mathbf{P} = \nabla$ at each new point, but it is more efficient to choose the new pathway to be orthogonal to all previous paths. This method of "conjugate gradients" is perhaps the most popular method of energy minimization. Details of this method can be found in Reference [16].

It is also possible to minimize the energy of a conformation by optimizing the dihedral angle degrees of freedom, rather than the Cartesian coordinates. The minimization occurs in M-dimensional space, where M is the number of dihedral angles. Torques, or derivatives of the forcefield with respect to dihedral angles, take the place of the gradient. We have found that "torque minimization," when followed by Cartesian minimization, produces an overall lower-energy conformation than Cartesian minimization alone. Neither method, however, can guarantee that the lowest possible conformation (the global minimum) will be reached. The process of moving along pathways in conformational space usually ends at a "local minimum" – a well in the potential energy surface, where the energy is lower than for all other nearby conformations, but not necessarily lower than other local minima.

II.C. Molecular Dynamics

The most important application of forcefields has been molecular dynamics. Molecular dynamics simulates the motion of particles in a system as they react to forces caused by interactions with other particles. In itself, this dynamical view of molecular systems can be important for studying time-dependent processes. However, two aspects of these simulations give them importance which goes far beyond their fundamental use. First, these calculations allow a system to sample conformational space. While an energy minimization calculation will find a local minimum in the potential energy surface, molecular dynamics calculations can cover a far broader sample of conformations. By giving each particle a velocity, molecular dynamics imparts kinetic energy to the system. This energy can be sufficient to enable the system to progress over barriers in the potential surface which could not be crossed in a gradient minimization procedure. A second very important factor in molecular dynamics is that the conformations produced during a simulation can form a thermodynamic ensemble. For instance, maintaining constant total energy, volume, and particles produces a microcanonical ensemble of conformations. Microcanonical dynamics is the easiest and most common form, but new methods have been developed to form other types of ensembles. This property allows one to calculate thermodynamics properties, such as relative free energies, from molecular dynamics simulations. This has been exploited recently with great success by free-energy perturbation calculations[17].

Molecular dynamics calculations evaluate the forces acting on each particle and use these to determine the accelerations these particles undergo. Particle velocities are initially determined by a random distribution calibrated to give a Maxwell-Boltzmann distribution at a given simulation temperature, but the velocities are updated according to the calculated accelerations. Most molecular dynamics methods work in Cartesian coordinates, allowing the maximum $3n$ degrees of freedom for n particles. Each particle i has three Cartesian degrees of freedom ($\mathrm{r}_i = x_i, y_i, z_i$). These degrees of freedom are uncoupled: forces, velocities, and accelerations are determined for each degree of freedom independently of the other degrees of freedom, with the exception that the overall translation and rotation of the system are subtracted out. The forces acting on particle i are the opposite of the gradient:

$$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i}. \tag{1.13}$$

Since the Cartesian degrees of freedom are uncoupled, each force component, F_x is calculated separately:

$$F_x = -\frac{\partial V}{\partial x}. \tag{1.14}$$

The accelerations, $\ddot{x}$, are calculated from Newton's equation of motion:

$$\ddot{x} = \frac{F_x}{m_i}, \tag{1.15}$$

where m_i is the mass of particle i. Ideally, velocities would be updated from accelerations by analytical integration of the equations of motion as in Equation (1.16), where $v_x^{t_1} = \dot{x}^{t_1}$ is the x-component of the velocity vector at time t_1:

$$v_x^{t_2} = v_x^{t_1} + \int_{t_1}^{t_2} \ddot{x} dt. \tag{1.16}$$

Unfortunately, an analytical equation for $\ddot{x}$ would be extraordinarily unwieldy, except for very simple systems, so the integration in Equation (1.16) must be done numerically.

There are numerous methods for doing numerical integrations[16] and many of these have been used in molecular dynamics. The simulations reported here use the most popular of the methods, the Verlet algorithm[18]. The verlet algorithm, itself, has many formulations[19], of which we use the "leapfrog formulation," so named because velocities and coordinates are updated at half-timestep intervals after one another. Methods for numerically integrating the equations of motion generally divide the simulation into timesteps, h, which are shorter than the periodicity of the fastest motions in the system. Typically, a timestep of one femtosecond (1×10^{-15} s) is used, to enable accurate integration of O–H and N–H bond stretches. In the leapfrog Verlet algorithm, the velocities at timestep $n + \frac{1}{2}$ are obtained from the previous velocities and the new accelerations:

$$v_x^{n+\frac{1}{2}} = v_x^{n-\frac{1}{2}} + h\ddot{x}^n. \tag{1.17}$$

The new velocities $v_x^{n+\frac{1}{2}}$ are then used to update the coordinates to timestep $n + 1$:

$$x^{n+1} = x^n + hv_x^{n+\frac{1}{2}}. \tag{1.18}$$

These new coordinates are then used to calculate the forces as in Equation (1.14) and the process is repeated.

III. Monte Carlo

Monte Carlo calculations represent an entirely different type of simulation from molecular dynamics. The name "Monte Carlo" comes from the random-chance nature of the simulations, akin to the games of chance at Monaco's gambling resort. Rather than being a deterministic method like molecular dynamics, where the physical properties of a system (e.g., coordinates, interatomic forces) determine its time evolution, Monte Carlo simulations are stochastic and use random numbers to generate a sample population of the system from which properties can be determined. Monte Carlo simulations are by no means limited to molecular systems, but are used in such diverse areas as integrated-circuit design and solving differential equations. But Monte Carlo calculations are very widespread in chemical simulations, primarily in studies of gases and fluids, where the random nature of the technique is readily employed.

An extremely important Monte Carlo algorithm for molecular systems was developed by Metropolis *et al.*[20]. One can calculate a molecular property F from a canonical ensemble, using the equation

$$F = \frac{\int F e^{-E/k_BT} dqdp}{\int e^{-E/k_BT} dqdp}, \tag{1.19}$$

where k_B is the Boltzmann constant, T is the system temperature, and $dqdp$ is a volume element in phase space. The integral is generally too complex to solve analytically, but it can be estimated by a computer simulation using a sufficiently large sample. A simulation with N_c sample configurations has properties calculated from:

$$F = \frac{\sum_{c=1}^{N_c} F_c e^{-E_c/k_BT}}{\sum_{c=1}^{N_c} e^{-E_c/k_BT}}. \tag{1.20}$$

The straightforward approach to calculating Equation (1.20) by generating numerous configurations and weighting them by $\exp(-E_c/k_BT)$, has problems in the (common) case where most generated configurations have high energies. The great majority of the conformations will be those which are least important, i.e., they have the lowest weighting factors. The Metropolis algorithm avoids this problem by generating conformations according to the probability $\exp(-E_c/k_BT)$ and weighting them all equally. This ideal distribution is established by giving each conformation a conditional probability of being accepted into the average. Each conformation c is perturbed in some way to produce conformation $c+1$. If the energy of the new conformation, E_{c+1} is smaller than that of E_c, the new conformation is accepted. If its energy is higher, the probability of it being accepted is $P = \exp(-\Delta E/k_BT)$ where

$\Delta E = E_{c+1} - Ec$. The standard method for enforcing this probability is to generate a random number n_r and to accept the new conformation $c+1$ if $n_r < exp(-\Delta E/k_B T)$. Otherwise, the new conformation is rejected and the previous one is restored and it is included again in the summation. Although many enhancements have been made to Monte Carlo theory since the Metropolis algorithm was derived, it still has very wide popularity. In some fields, the Metropolis algorithm has practically become the definition of Monte Carlo simulations.

The random nature of Monte Carlo simulations makes them useful for sampling conformational space. Although they are generally not as efficient as molecular dynamics simulations for sampling conformational space[21], Monte Carlo simulations can incorporate large conformational changes which cannot be simulated by molecular dynamics. For instance, the Dihedral Probability Grid Monte Carlo method of Chapter 3 can rotate a dihedral angle in a single step without regard to an energy barrier which might prevent the same rotation in molecular dynamics. In general, we have found that Monte Carlo simulations are excellent for coarse-grained sampling of conformational space while molecular dynamics and minimization techniques are excellent for performing the complementary role of local conformational optimization.

References

[1] M.S. Friedrichs, R.A. Goldstein, and P.G. Wolynes, *J. Mol. Biol.*, 222, 1013-1034 (1991).

[2] A. Godzik, J. Skolnick, and A. Kolinski, *Proc. Natl. Acad. Sci., USA*, 89, 2629-2633 (1992).

[3] D.G. Covell and R.L. Jernigan, *Biochemistry*, 29, 3287-3294 (1990).

[4] J.B. Bowie, R. Lüthy, and D. Eisenberg, *Science*, 253, 407-414 (1991).

[5] S.A. Benner and D. Gerloff, *Adv. Enz. Regul.*, 31, 121-181 (1991).

[6] I.T. Weber *et al.*, *Science*, 243, 928-931 (1989); R.S. Struthers, D.H. Kitson, and A.T. Hagler, *Proteins: Structure, Function, Genet.*, 9, 1-11 (1991).

[7] E.C. Meng, B.K. Shoichet, and I.D. Kuntz, *J. Comp. Chem.*, 13, 505-524 (1992).

[8] J. Tirado-Rives and W.L. Jorgensen, *Biochemistry*, 30, 3864-3871 (1991).

[9] T.P. Straatsma and J.A. McCammon, *Meth. Enzym.*, 202, 497-511 (1991).

[10] K. Sharp, *J. Comp. Chem.*, 12, 454-468 (1991).

[11] S.J. Weiner *et al.*, *J. Am. Chem. Soc.*, 106, 765-784 (1984).

[12] B.R. Brooks *et al.*, *J. Comp. Chem*, 4, 187-217 (1983).

[13] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[14] H.Q. Ding, N. Karasawa, and W.A. Goddard III, *J. Chem. Phys.*, 97, 4309-4315 (1992).

[15] *BIOGRAF Reference Manual*, Molecular Simulations, Inc. (1992).

[16] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge (1989).

[17] D.L. Beveridge and F.M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.*, 18, 431-492 (1989).

[18] A. Rahman, *Phys. Rev.*, 136, A405 (1964); L. Verlet, *Phys. Rev.*, 159, 98 (1967).

[19] D.W. Heermann, *Computer Simulation Methods in Theoretical Physics*, Springer-Verlag, Berlin (1986).

[20] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller, *J. Chem. Phys.*, 21, 1087-1092 (1953).

[21] G. Jacucci and A. Rahman, *Il Nuovo Cimento*, 4, 341-355 (1984).

Chapter 2

Newton Euler Inverse Mass Operator (NEIMO) Dynamics of Polypeptides

Abstract

Newton-Euler Inverse Mass Operator (NEIMO) Dynamics is a fast method for calculating internal-coordinate molecular dynamics. Unlike other exact methods for solving these equations of motion, computational time for the NEIMO method is proportional to $\mathcal{N}$, rather than $\mathcal{N}^3$, where $\mathcal{N}$ is the number of degrees of freedom. This allows internal-coordinate dynamics to be solved for very large systems. The first use of the NEIMO method for molecular dynamics is presented here. Results are given for simulations of a wide range of peptide and protein systems. The computational time is shown to be rigorously proportional to $\mathcal{N}$. Additionally, the dynamics are shown to be accurate for timesteps much larger than those used in Cartesian-coordinate dynamics. For small peptides, timesteps as large as 20 fs are achievable.

I. Introduction

Molecular dynamics simulations have become a very important tool in computational chemistry and biology. Once primarily useful for understanding the dynamic properties of molecular systems, molecular dynamics has become invaluable for such diverse tasks as building protein models from crystallographic data[1] and determining the relative free energy of binding of different drug molecules to the same receptor[2]. As these calculations become more accurate and applicable to a wider variety of problems, researchers seek to apply them to larger and more complex systems as well as to use them to study processes occurring over increasingly longer timescales. This requires continual improvements in both computer hardware and software performance. The former challenge is being addressed by such developments as vector processing supercomputers, RISC workstations and now massively parallel supercomputers. The software problem is being tackled on several overlapping fronts: optimization of programs for new computer architectures[3], improved efficiency in calculating interatomic forces[4, 5], and development of techniques for allowing larger timesteps in molecular dynamics simulations[6, 12].

Molecular dynamics simulations typically involve numerical integration of Newton's equations of motion. Timesteps for the integration must be small enough for the fastest modes to be handled accurately. Typically, a timestep of 1 femtosecond (1×10^{-15} s) is necessary to handle bondstretches involving hydrogen atoms. Although success has been achieved lately by separating short and long-range forces and using different timesteps for the different forces[6], the most popular approach for increasing timesteps is to fix the fastest degrees of freedom (bond stretches and angles) and to solve the equations of motion for the slower (dihedral) degrees of freedom. Such an approach is especially justified for studies of large biological molecules, where bond lengths and angles vary little from one structure to another and nearly all important conformational transitions are due to dihedral angle motions.

The SHAKE algorithm[7] has become the standard approach for doing molecular dynamics with fixed bond lengths. It can also be used to hold angles fixed, but this is less common. SHAKE is a modification of the Verlet algorithm for integrating the equations of motion for the $3n-6$ Cartesian coordinates degrees of freedom in an n-particle system. Particle velocities are first calculated for the unconstrained system, then modified to meet each constraint. An iterative process is required to meet all the constraints concurrently. The SHAKE algorithm works well for timesteps up to 5 fs[9, 10], thereby enabling a five-fold speedup in computational time as long as the process of iteratively solving the constraint equations does not consume too much time[9].

An alternative to the SHAKE methodology of solving Cartesian coordinate dynamics plus constraints, is to solve the equations of motion directly for the internal degrees of freedom. Solutions to these equations automatically fulfill the desired bond length and/or angle constraints, so their efficiency is not limited by a secondary constraint-solving step. Indeed, Mazur *et al.*[14] were able to simulate accurately a small polypeptide, $(\mathrm{Ala})_9$, with timesteps as large as 13 fs. This is a significant improvement over the SHAKE algorithm. Unfortunately, their method requires the direct solution of matrix equations; the computational time for solving these equations is usually proportional to $\mathcal{N}^3$, where $\mathcal{N}$ is the number of degrees of freedom. This can be prohibitive for large systems.

Recently, Jain *et al.*[11, 12] have developed an alternative method for solving the equations of motion for internal coordinates. This new method, the Newton-Euler Inverse Mass Operator (NEIMO) method, does not require direct manipulation of matrices and is proportional to $\mathcal{N}$, rather than $\mathcal{N}^3$. This method was developed for spacecraft dynamics, but in a separate report, Jain *et al*[12] described how the method could be applied to molecular dynamics. This report presents the first implementation of the NEIMO method for molecular systems. We have studied the dynamics of polypeptide systems and have found that we are able to calculate the dynamics of some systems accurately with timesteps as large as 20 fs. Because the NEIMO method is computationally proportional to $\mathcal{N}$, it can be applied to very large systems. The method is shown to be rigorously proportional to $\mathcal{N}$ for systems as large as the tomato bushy stunt virus crystal structure[24], which has three chains of nearly 300 residues each.

II. Methodology

In Cartesian coordinate molecular dynamics calculations, the $3n$ degrees of freedom are uncoupled, and Newton's equations of motion can be solved independently for each degree of freedom, x:

$$m_i \ddot{x} = F_x. \tag{2.1}$$

Here, the mass of particle i, m_i, is used for that particle's three degrees of freedom, $\ddot{x}$ is the acceleration, and F_x is the force. In internal coordinates, similar equations of motion can be written for systems with a tree-topology:

$$\mathcal{M}(\theta)(\ddot{\theta}) + \mathcal{C}(\theta, \ddot{\theta}) = T(\theta). \tag{2.2}$$

Here, θ is the $\mathcal{N}$-length vector of internal coordinates, $\mathcal{C}$ is the $\mathcal{N}$-length vector of nonlinear forces and T is the $\mathcal{N}$-length vector of generalize forces, or torques in the case of dihedral-angle degrees of freedom. In internal coordinates, the $\mathcal{N}$ degrees of freedom are not uncoupled. The $\mathcal{N} \times \mathcal{N}$ mass matrix, $\mathcal{M}$, has off-diagonal elements and has a nonlinear dependence on θ, so the equations cannot be solved independently for each degree of freedom, θ_i as is done for the Cartesian degrees of freedom in Equation (2.1). The accelerations $\ddot{\theta}$ can be determined by solving the matrix equation in Equation (2.2), as was done for the Lagrange equations of motion by Mazur *et al.*[14]. The computational cost of these matrix equations is proportional to $\mathcal{N}^3$, so they are prohibitive for large molecules. Recently, however, Jain *et al.*[12], have developed a recursive algorithm for solving the equations of motion which does not require explicit calculation of the mass matrix, $\mathcal{M}$, or direct solution of the matrix equation 2.2. The method uses *spatial operator algebra* in a recursive approach which is computationally proportional to $\mathcal{N}$. This linear dependence on $\mathcal{N}$ opens up a new class of molecular systems to study by internal-coordinate molecular dynamics.

The details of the recursive spatial operator equations for solving the equations of motion can be found in Reference [12]. The methodology has been developed for general multibody systems configured as serial chains, topological trees, or closed-loop systems[11]. The implementation for molecular systems reported here has not yet been extended to closed-loop topologies, so only the serial chain and tree topology will be discussed here. The NEIMO method uses the concepts of "clusters" and "hinges" to describe a mechanical system. A cluster is a body which moves as a unit; in a molecule, this could be a single atom, a multiple-atom group such as a methylene group, a phenyl ring, or even an entire domain of a protein. A "hinge" describes the relative orientation between two connected clusters; in a molecular system, the hinges correspond to the bonds connecting adjacent clusters. There are six possible degrees of freedom for each hinge, including such modes as bond stretching, ball and socket rotation, and torsional rotational. Our interest is primarily in the torsional degrees of freedom, so each hinge in our molecular implementation is limited to a single torsional degree of freedom. The one exception per molecule is the "base" cluster, which is connected to the reference coordinate system by a hinge with the full six degrees of freedom, thereby enabling each molecule to be oriented correctly

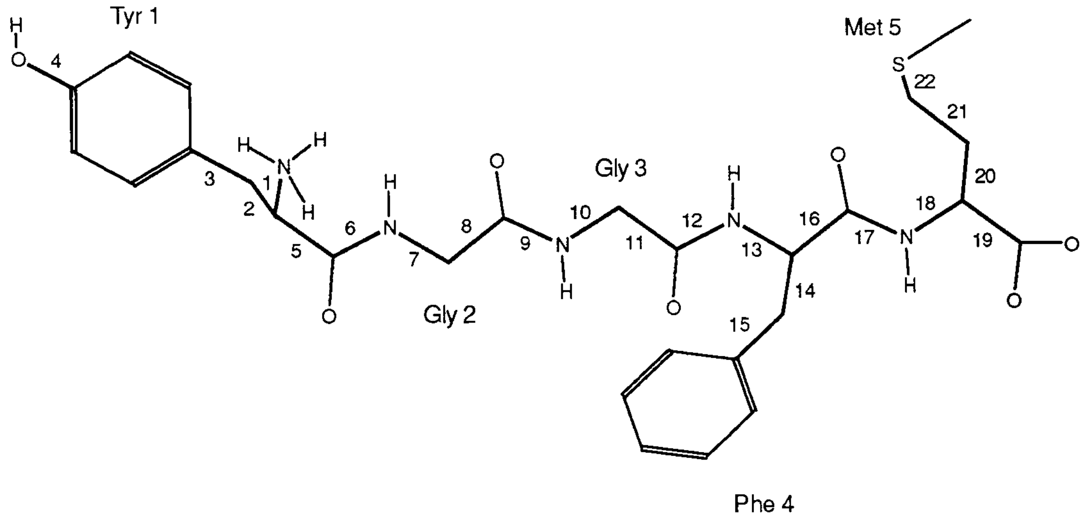


Figure 2.1. The peptide Met-enkephalin is shown, with the bonds/hinges numbered. Clusters are the chemical units connected by the hinges, such as the phenyl ring in tyrosine 1. The hinges are numbered to allow for analysis of the dihedral angles.

in space.

The relationship between adjacent clusters is described in terms of "parents" and "children." The base cluster can have one or more child clusters; it is the parent cluster of each of these children. Each of these clusters, in turn, may have one or more children, branching outward from the base. In a topological tree, outward branching can continue with each cluster having zero, one, or more children, but each child having only one parent cluster. Clusters at the far extent of each branch are termed "tips" and have no children. In a serial chain, such as a linear polymer, there is a single tip cluster; between the base and tip clusters, each cluster has a unique parent and unique child. In a protein system, most of the C_α atoms are branch points with two children, and the outermost cluster of every sidechain (other than alanine and glycine, which have no sidechain clusters) is a tip. These concepts are visualized in Figure 2.1, where the pentapeptide Met-enkephalin is shown with the hinges numbered. This numbering system is not equivalent to the numbering used in Reference [12], but is used here to facilitate analysis of the dihedral angles. Hinge 0, which connects the base cluster to the reference frame, is not shown. The clusters can be assigned the same number as the hinge connecting it to its parent cluster and are listed that way in Table II. In addition, the torsional degree of freedom for each hinge, other than hinge 0, can be defined to correspond to a specific dihedral angle. These dihedrals are also listed in Table II, using the standardize nomenclature for protein dihedrals.

The NEIMO method uses spatial operator algebra to formulate the equations of motion in terms of relationships between parent and child clusters, and to solve the equations during several recursions from the base cluster to the tips, and the tips to the base. These Newton-Euler recursive equations of motion are described in detail in Reference [12], but can be summarized as:

1. A base to tips recursion, during which each cluster's spatial velocity is determined from the torsional motion of its associated hinge as well as the motion of its parent cluster.

2. A tips to base recursion, during which the effective forces acting on each cluster are derived from the explicit Cartesian forces, hinge torques, and Coriolis and related forces acting on the cluster, as well as the forces imparted to it from its children. Several other quantities related to the inverse of the mass matrix are also calculated during this recursion.

3. A final base to tips recursion, during which the acceleration of each cluster and its related hinge acceleration, are determined from the cluster's acceleration within its own frame of reference plus acceleration due to the acceleration of its parent cluster.

The use of spatial operator algebra has another important consequence in that an

Cluster	Hinge	Dihedral	Residue
-NH_3^+	0	(none)	Tyr 1
-CH-	1	ϕ	
-CH_2-	2	χ^1	
-C_6H_4-	3	χ^2	
-OH	4	χ^6	
-(CO)-	5	ψ	
-(NH)-	6	ω	Gly 2
-CH_2-	7	ϕ	
-(CO)-	8	ψ	
-(NH)-	9	ω	Gly 3
-CH_2-	10	ϕ	
-(CO)-	11	ψ	
-(NH)-	12	ω	Phe 4
-CH-	13	ϕ	
-CH_2-	14	χ^1	
-C_6H_5	15	χ^2	
-(CO)-	16	ψ	
-(NH)-	17	ω	Met 5
-CH-	18	ϕ	
-COO^-	19	ψ	
-CH_2-	20	χ^1	
-CH_2-	21	χ^2	
-SH	22	χ^3	

Table 2.1. The clusters, hinges, and related dihedrals of Met-enkephalin, shown in Figure 2.1.

operator expression can be derived for the inverse of the mass matrix, $\mathcal{M}$, without actually having to invert the matrix. This allows the accelerations, $\ddot{\theta}$, to be determined by recursive equations, computationally proportional to $\mathcal{N}$, the number of clusters, rather than by matrix equations of the order of $\mathcal{N}^3$. The use of an operator expression for $\mathcal{M}^{-1}$ has given the Newton-Euler Inverse Mass Operator (NEIMO) Dynamics method its name.

Currently, we use a "leapfrog" Verlet algorithm[8] to integrate the equations of motion, using the accelerations determined by the recursive spatial operator equations. The Verlet algorithm is a very successful and popular method in Cartesian-space dynamics, but may be less suitable for NEIMO dynamics because it calculates accelerations and velocities at alternating half-timesteps. In NEIMO dynamics, accelerations are not independent of velocities, so the half-timestep separation of accelerations and velocities must be modified. This can be done by iteratively solving for the velocities at integer timesteps. Fortunately, this iteration is very fast in practice. A major advantage of the Verlet algorithm is that it requires only a single calculation of the forces at each timestep. In simulations of large systems, the force calculation consumes the vast majority of computational time, so methods which require only a single force calculation are preferable to methods which require two or more force calculations per timestep, such as the Gear predictor-corrector algorithm[15]. Currently, other integration schemes are being investigated for use in NEIMO dynamics, but all results presented here use the leapfrog Verlet algorithm. Details of this integration method are given in Appendix A.

The NEIMO calculations presented here were performed using a version of the program written to work with the BIOGRAF/POLYGRAF program from Molecular Simulations, Inc[16]. All calculations were performed on Iris PowerSeries and Iris Indigo workstations from Silicon Graphics, Inc.

III. Results

NEIMO calculations were carried out on a wide variety of peptide and protein systems, ranging from the five-residue peptide Met-enkephalin to the tomato bushy stunt virus (TBSV) protomer, which contains three proteins totaling 893 residues. Table 2.2 contains a list of the ten systems studied. The two peptides were built using the Peptide Builder of BIOGRAF[16], which uses standard amino acid geometries. They were initially configured as alpha helices, but were minimized to a local potential energy minimum using conjugate gradients minimization. As in all calculations reported here, the DREIDING forcefield was used for these minimizations. No solvent or counterions were used, but the dielectric constant for each pair of atoms i and j was set proportional to r_{ij}, the distance between them. This crudely represents the electrostatic shielding of aqueous solvent. For these peptides, no nonbond cutoff was used; i.e., all possible pairs were included in the van der Waals and electrostatic calculations. The initial conformations of the proteins were derived from the X-ray crystal structures listed in Table 2.2. All metal ions, solvent molecules, and disulfide bridges were removed, leaving only protein chains which conformed to a tree topology. (As mentioned above, sidechain aromatic rings and proline rings are treated as single clusters). Hydrogen atoms were then added to non-carbon atoms. As was done for the peptides, the DREIDING forcefield was used to energy-minimize these conformations. Nonbonds, however, were treated differently. The large size of the proteins precluded the inclusion of all possible nonbond pairs, a number close to n^2 for an n-atom protein. Therefore, the cell-multipole method (CMM) of Ding *et al.*[4] was used to calculate the van der Waals and electrostatic interactions. This method is roughly proportional to n, but provides far greater accuracy than the standard approach of excluding all nonbond interactions greater than 9 Å.

Protein	Structure	Ref.	Residues	Atoms	$\mathcal{N}$
Met-Enkephalin	MEnk	-	5	48	28
$(\mathrm{Ala})_9$	Ala9	-	9	57	32
Avian Pancreatic Polypeptide	1ppt	[17]	36	368	192
Crambin	1crn	[18]	46	402	216
Plastocyanin	7pcy	[19]	98	857	460
Troponin-C	5tnc	[20]	161	1514	857
Alpha-Lytic Protease	2alp	[21]	198	1748	959
Carbonic Anhydrase	2ca2	[22]	256	2482	1305
Carboxypeptidase A_α	5cpa	[23]	307	2986	1581
Tomato Bushy Stunt Virus	2tbv	[24]	893	8083	4335

Table 2.2. Proteins and peptides used in NEIMO simulations. The structures listed are the initial Protein Database files, except for the peptides “MEnk” and “Ala9,” which were created using the BIOGRAF peptide builder.

III.A. Timing

Timing results for the ten systems are shown in Table 2.3. The times represent the average of 100 dynamics steps run on an Iris Indigo workstation. Times are given for both the NEIMO calculations and the nonbond calculations, the latter of which consumes the vast majority of cpu time, even when a very fast method such as CMM is used. The NEIMO timing is shown to be rigorously proportional to $\mathcal{N}$ for Crambin and the large proteins. The times for the small peptides apparently are shorter, but the resolution of the timing routine used is 0.01 s, so the results may be off by as much as ± 50%. The nonbond calculations are less consistent, even for CMM, which is proportional to n, the number of atoms in the system. The lack of perfect proportionality is due to the asymmetry of the protein conformations. The Cell-Multipole algorithm[4] creates a cubic cell around the system being simulated and divides this cell into a hierarchy of smaller cubic cells, for which dipole and quadrupole terms are calculated. An oblong protein molecule would require a particularly large outer cell and would have many of its smaller cells empty. These are the least efficient conditions for CMM. Considering this limitation, the method is very nearly proportional to n, and is much faster than the n^2 method of calculating all possible nonbond pairs.

III.B. Energy Fluctuations

The primary advantage of internal-coordinate methods of molecular dynamics is the ability to use larger timesteps than the 1 femtosecond step size typically used in Cartesian molecular dynamics. An important measure of the accuracy of the dynamics calculation is the energy fluctuations. In microcanonical dynamics, the total energy of the system, E, should be constant, even though its two components, the potential energy, V, and kinetic energy, K, fluctuate. The energy fluctuation $\mathcal{E}$

Dynamics Timing

	NEIMO		Nonbonds		
Protein	Time (s)	Time/$\mathcal{N}$ (ms)	Method	Time (s)	Time/n (ms)
MEnk	0.011	0.393	All NB	0.044	0.92
Ala9	0.012	0.375	All NB	0.061	1.07
1ppt	0.084	0.438	All NB	1.933	5.25
1crn	0.102	0.472	All NB	2.322	5.78
7pcy	0.220	0.478	All NB	10.121	11.81
1ppt	0.084	0.438	CMM	1.408	3.83
1crn	0.102	0.472	CMM	1.950	4.85
7pcy	0.220	0.478	CMM	3.541	4.13
5tnc	0.411	0.480	CMM	9.180	6.06
2alp	0.460	0.480	CMM	11.153	5.26
2ca2	0.629	0.482	CMM	15.612	6.29
5cpa	0.762	0.482	CMM	22.733	7.61
2tbv	2.094	0.483	CMM	55.439	6.86

Table 2.3. Timing results for the ten protein/peptides systems studied here. The average times per timestep of the NEIMO calculation and the nonbond calculation are given, along with the NEIMO time divided by $\mathcal{N}$, and the nonbond time divided by the number of atoms, n.

is defined by

$$\mathcal{E} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T}, \tag{2.3}$$

where k_B is the Boltzmann constant and T is the temperature of the simulation. Figure 2.2 shows the value of E during 1 picosecond (1×10^{-12} s) simulations of the pentapeptide Met-enkephalin (NH_3^+-Tyr-Gly-Gly-Phe-Met-COO^-) for NEIMO(N) and Cartesian(C) dynamics simulations at timesteps ranging from 1 fs to 20 fs. The Cartesian dynamics simulations had an initial 1 fs equilibration phase, where the fluctuations were significantly higher. These fluctuations were not included in these results. The NEIMO simulations did not require an equilibration phase. Cartesian dynamics simulations were not possible for timesteps greater than 3 fs. When timesteps are too large for the motions being simulated, particle motions from one timestep to the next are exaggerated and the energy quickly "blows up" (the energy goes to infinity). For Cartesian dynamics of large systems, timesteps greater than 1 fs are usually unstable. Even for the small peptide Met-enkephalin, a 2 fs timestep gives rise to far larger energy fluctuations than a 1 fs simulation. The NEIMO dynamics simulation is far more stable, with timesteps as large as 18 fs giving small fluctuations, smaller even than the Cartesian dynamics simulation with a 1 fs timestep. A fairer comparison is to divide the energy fluctuations by the number of degrees of freedom. For Met-enkephalin, $\mathcal{N} = 28$ (22 dihedral angles plus the six degrees of freedom for the base body), while the number of degrees of freedom in Cartesian dynamics is $3n - 6$, or 138. The scaled fluctuations, $\mathcal{E}^*$, are also shown in Figure 2.2 and are labeled with an asterisk (N* and C*). NEIMO timesteps as large as 12 fs gave smaller scaled fluctuations than the 1 fs Cartesian simulation.

Similar results were obtained for nine-residue polyalanine, $(\mathrm{Ala})_9$. Cartesian dynamics were only possible at 1 fs and 2 fs timesteps. The 3 fs simulation did not blow up, but the fluctuations were extremely large. The scaled fluctuations, $\mathcal{E}^*$, were very

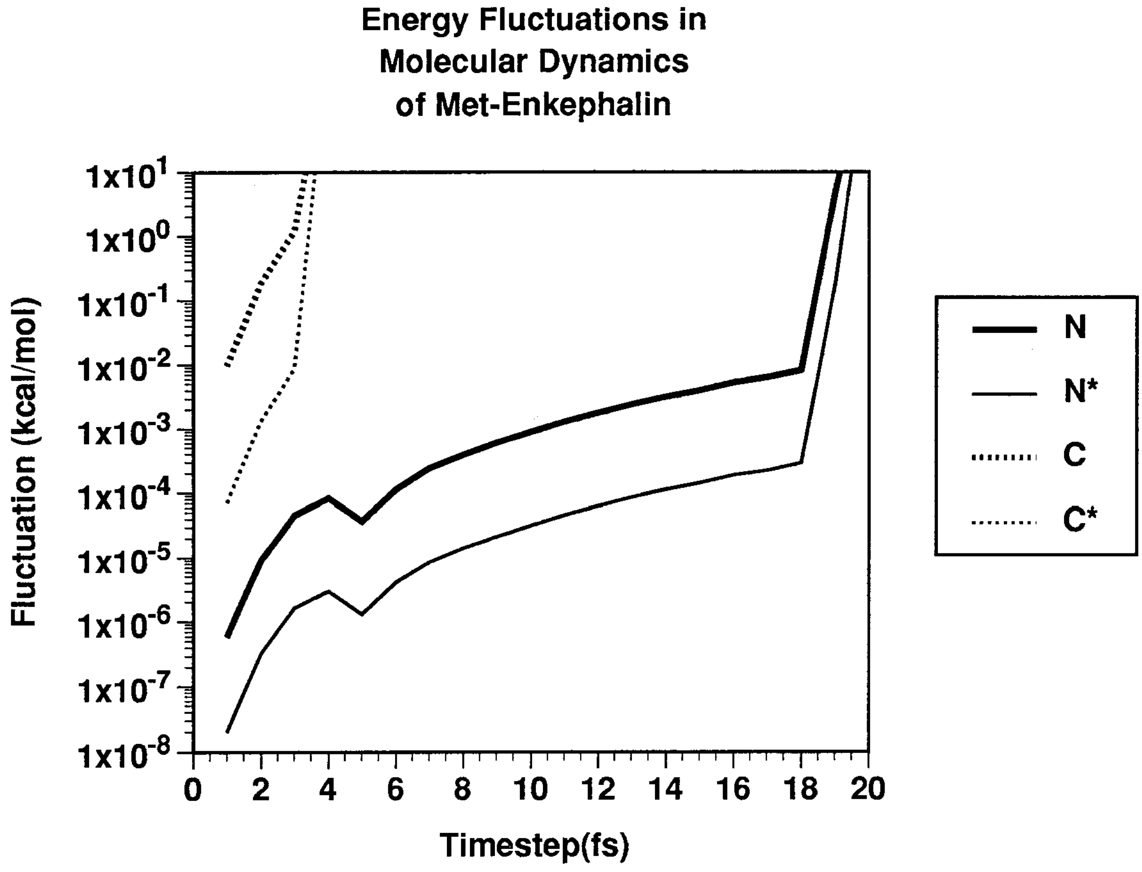


Figure 2.2. Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of Met-enkephalin. Simulations were run for 1 ps using timesteps ranging from 1 to 20 fs. N* and C* are the scaled fluctuations, $\mathcal{E}^*$, where $\mathcal{E}$ is divided by the number of degrees of freedom: $\mathcal{N}$ for NEIMO simulations and 3n-6 for Cartesian coordinates.

similar for 1 fs and 2 fs Cartesian dynamics of both peptides. The NEIMO simulations of $(\mathrm{Ala})_9$ gave larger values of $\mathcal{E}$ and $\mathcal{E}^*$ than for Met-enkephalin at almost every timestep, but the fluctuations did not blow up until timesteps larger than 30 fs were used. It is likely that $(\mathrm{Ala})_9$ is able to tolerate such large timesteps because it has no light sidechain clusters which would be expected to have higher rotational velocities. Since the DREIDING forcefield uses a united-atom approach, the CH_3 units of the Alanine sidechains are treated as a single atom and, therefore, do not form clusters which move independently in the NEIMO model. In contrast, the tyrosine, phenylalanine, and methionine sidechains of Met-enkephalin all contain individual clusters with low moments of inertia. As indicated below in the analysis of Met-enkephalin dihedral angle fluctuations, the long, unbranched methionine sidechain is particularly flexible.

As the simulations are carried out for longer periods of time, the fluctuations $\mathcal{E}$ gradually increased. For instance, a 1 ps NEIMO simulation of Met-enkephalin using a 5 fs timesteps had a value of $\mathcal{E}$ less than 0.0001 kcal/mol. The same simulation run for 5 ps had $\mathcal{E} = 0.0042$ kcal/mol, even though no 0.1 ps stretch of the simulation had $\mathcal{E} > 0.0004$ kcal/mol, and the average fluctuation over the 50 0.1 ps stretches was only 0.0001 kcal/mol. Run for 25 ps, the simulation has an overall $\mathcal{E}$ of 0.0360 kcal/mol, even though the average 0.1 ps stretch had $\mathcal{E} = 0.0005$ kcal/mol. This discrepancy is caused by very slow fluctuations in the total energy which cause $\langle E^2 \rangle$ to slowly diverge from $\langle E \rangle^2$. The cause of this long-term fluctuation is unknown.

In order to compare NEIMO directly to the matrix method of Mazur *et al.*[14], the quantity δ_E was calculated from simulations of $(\mathrm{Ala})_9$ at timesteps ranging from 1 fs to 20 fs. For each timestep, the simulation was run for 4.0 ps during which the velocities were rescaled, when necessary, to equilibrate the system. At the end of the 4.0 ps run, 110 additional steps were run. The first ten of these were discarded,

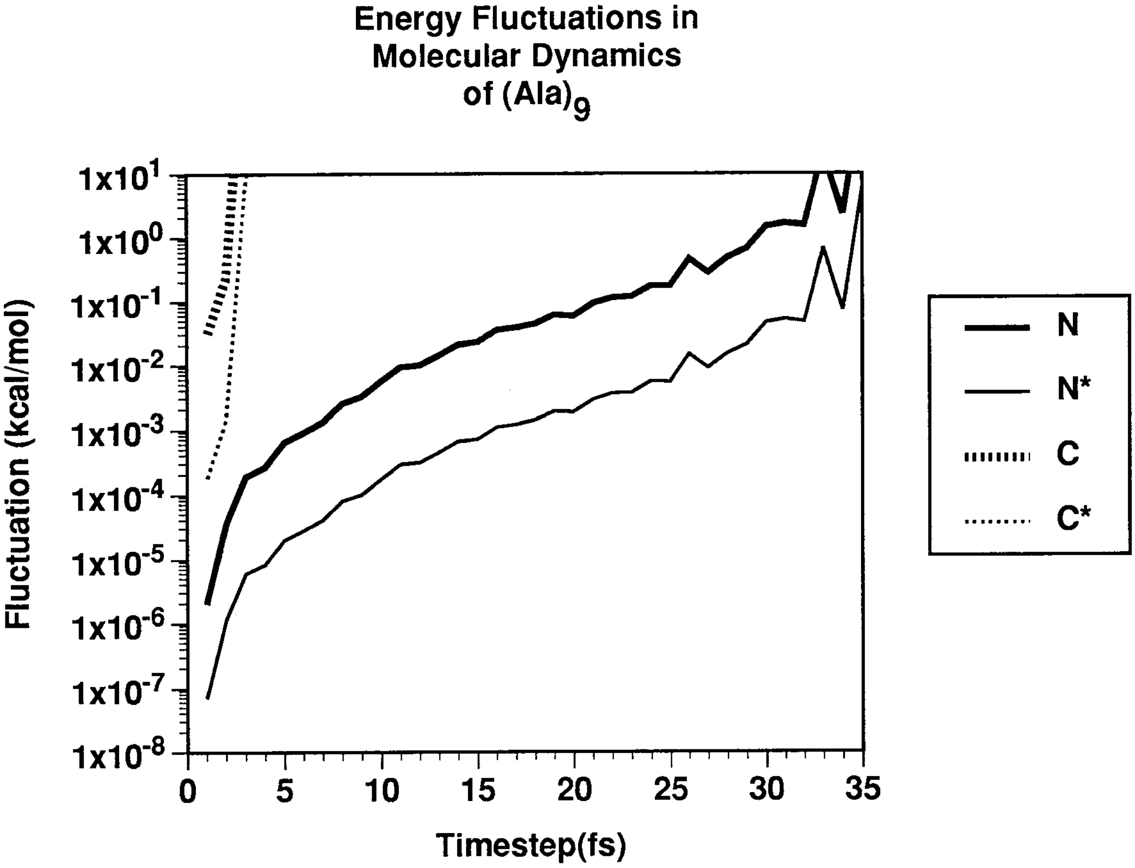


Figure 2.3. Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of $(\mathrm{Ala})_9$. Simulations were run for 1 ps using timesteps ranging from 1 to 35 fs. N* and C* are the scaled energy fluctuations, $\mathcal{E}^*$.

but the final 100 steps were used to determine δ_{E}, which is defined by

$$\delta_{\mathrm{E}} = \frac{\sqrt{\langle \Delta E \rangle}}{\langle E \rangle}. \tag{2.4}$$

$\langle E \rangle$ is the average energy during the 100 steps, and $\langle \Delta E \rangle$ is the root-mean-square deviation in the energy. Mazur *et al.* reported simulations on $(\mathrm{Ala})_9$ using a variety of models including some containing explicit hydrogens. The DREIDING/NEIMO calculation corresponds to their third model: united atoms are used rather than explicit hydrogens, and all bond lengths and angles are fixed. Only dihedral degrees of freedom are allowed plus the six degrees of freedom of the base body, for a total of 32 degrees of freedom. Mazur *et al.* obtained a value of $\delta_{\mathrm{E}} = 0.8 \times 10^{-6}$ using timesteps of 0.5 fs. Values of δ_{E} increased linearly with increasing timesteps, but they were able to achieve their desire level of accuracy, $\delta_{\mathrm{E}} \approx 10^{-2}$, using timesteps as large as 13 fs. NEIMO simulations using a 0.5 fs timestep had a larger value of $\delta_{\mathrm{E}} = 4.0 \times 10^{-6}$, but timesteps as large as 15 fs gave $\delta_{\mathrm{E}} \approx 10^{-2}$, as can be seen in Figure 2.4. These results are very consistent with the results of Mazur *et al.*, even though they used a different forcefield (a combination of CHARMm[26] and ECEPP[27]) and a different integration scheme. It should be noted that the choice of forcefield should not affect the results dramatically, since our average energy (25.402 kcal/mol) is nearly equal, though opposite in sign, to the average energy they report (-26.8 kcal/mol) for their 0.5 fs timestep simulations.

Although timesteps of 15 fs and longer are clearly possible for NEIMO simulations of small peptides such as Met-enkephalin and $(\mathrm{Ala})_9$, such timesteps are not yet possible for large polypeptides and proteins. Simulations were run on the 36 residue hormone peptide avian pancreatic polypeptide (aPP), a very interesting case because it is one of the smallest known polypeptides to fold into a stable globular form. Figure 2.5 shows the polypeptide backbone of aPP, which has two helices, an α helix and a collagen-like polyproline helix. Hydrophobic sidechains line the cleft between

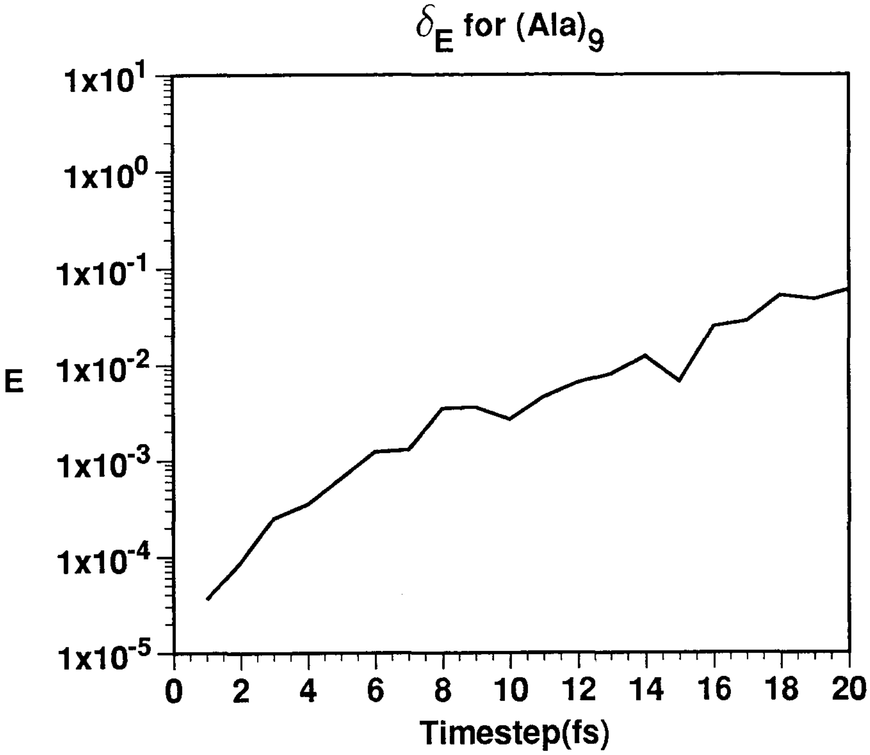


Figure 2.4. δ_E for 100 timesteps of NEIMO dynamics on (Ala)$_9$.

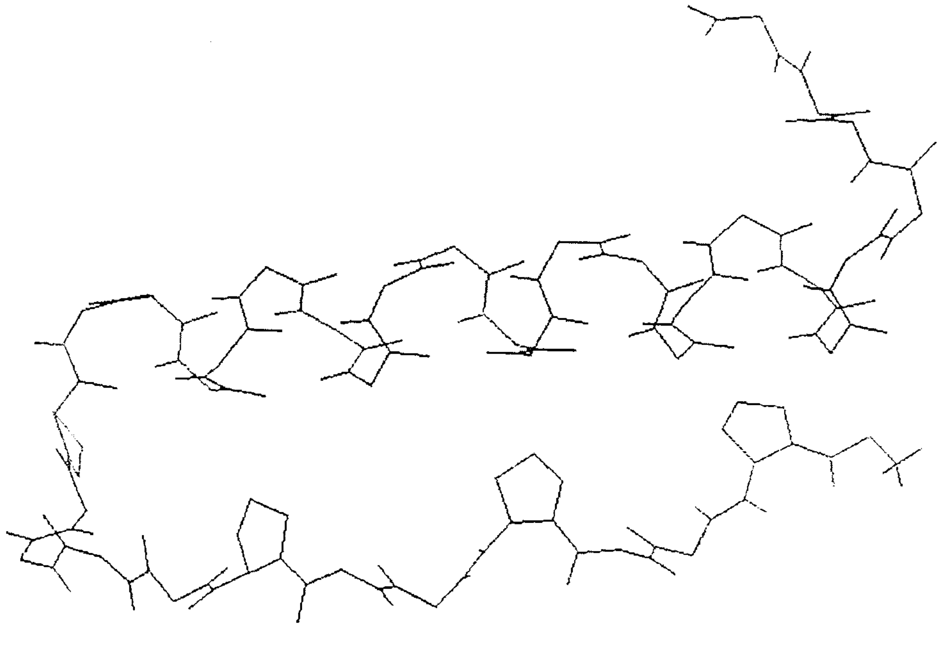


Figure 2.5. Avian pancreatic polypeptide (aPP), with the sidechain atoms removed for clarity. From the crystal structure 1PPT [16]).

the two helices, allowing for unusual stability in a peptide this size.

Figure 2.6 shows $\mathcal{E}$ and $\mathcal{E}^*$ using different timesteps for 1 ps simulations of aPP. NEIMO simulations of aPP break down when timesteps above 10 fs are used. Although timesteps as large as 9 fs give values of $\mathcal{E}$ as good or better than the 1 fs Cartesian simulation, the scaled fluctuations, $\mathcal{E}^*$, are approximately equal for the 6 fs NEIMO and 1 fs Cartesian cases. Several factors may cause folded polypeptides and proteins to have substantially larger fluctuations than small peptides at large timesteps. Complex secondary structure elements such as helices, turns, and beta sheets, are held together by hydrogen bonds, which are very short-range interactions. Large timesteps may cause rapid destabilization of these hydrogen bond networks. Another potential problem, due to the nature of the dynamics algorithm rather than the chemical nature of the system being studied, is the possible buildup of errors as the recursive equations are solved from the base cluster to the tips and the tips to the base. This is a more fundamental problem that may be substantially improved by higher precision in the calculations and a more central choice for the base cluster. Currently, the amino-terminal -NH_3^+ group is usually chosen as the base cluster, and the carboxy terminal residue tip clusters are maximally distant.

The fastest dynamical modes in the NEIMO model are those with the smallest spatial inertia. In protein systems, these are clusters containing explicit hydrogens, where rotation of the hinge moves only the hydrogen atoms. For instance, the hydroxyl group of Tyrosine forms a two-atom cluster. Rotation of the hinge between the aromatic ring C_ζ and the hydroxyl O_η only modifies the hydroxyl hydrogen. These are the fastest degrees of freedom in the system. These dihedrals can be fixed by including the moiety in its parent cluster. In tyrosine, the hydroxyl and aromatic ring can be treated as a single cluster. This “Rigid H” model removes the fastest degrees of freedom of the system and enables even longer timesteps to be used than

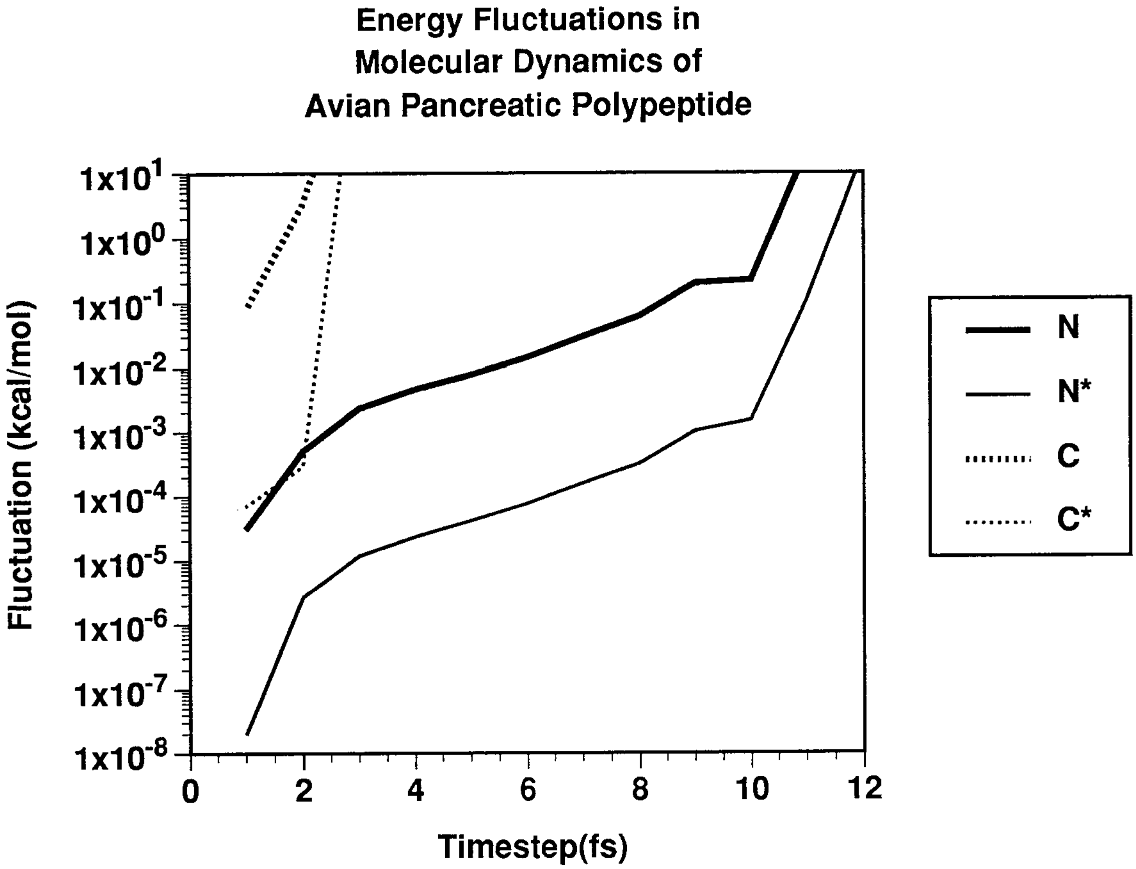


Figure 2.6. Energy fluctuations, $\mathcal{E}$, for NEIMO(N) and Cartesian(C) dynamics simulations of avian pancreatic polypeptide (aPP). Simulations were run for 1 ps using timesteps ranging from 1 to 15 fs, but all those above 11 fs caused the energy to blow up. N* and C* are the scaled energy fluctuations, $\mathcal{E}^*$.

the standard NEIMO model. This is seen clearly in Figure 2.7, where the 18 -OH and -NH_2 groups have been incorporated with their parent clusters. Although the scaled fluctuations, $\mathcal{E}^*$, are very similar for small timesteps, the standard model blows up when timesteps above 10 fs are used, while the "Rigid H" fluctuations increase only slowly above this point. Simulations using even longer timesteps displayed the same gradual increase in fluctuations, without any sharp jump in $\mathcal{E}^*$, as is seen in the standard model when timesteps above 10 fs are used. The "Rigid H" model would be useful for studies primarily interested in large-scale motions, where the hydrogen-bonding interactions of these sidechain groups are less important, and the advantage of longer timesteps is pre-eminent.

Detailed studies of protein systems require the inclusion of solvent, which plays an important role in stabilizing the native conformation of most proteins. Solvent includes both water (and/or lipids in the case of membrane-bound proteins) and ionic charges, which may be present to stabilize charged groups on the protein. In order to test the ability of NEIMO simulations to include such factors, we ran calculations where NEIMO dynamics were used to solve the equations of motion for the protein, while standard Cartesian dynamics were solved simultaneously for counterions. Avian pancreatic polypeptide was used as a test system. Oppositely-charged groups within 10 Å of each other were considered paired. This left eight unpaired charges, which were then neutralized by adding counterions (five Na^+ and three Cl^-). After the counterion locations were optimized by minimizing their energies, simulations were run for 2 ps using various timesteps. The first picosecond was used for equilibrating the counterion motions, and $\mathcal{E}$ was determined from 1 ps to 2 ps. The results are shown in Figure 2.7, along with the results from standard and "Rigid H" simulations of the protein alone. The addition of counterions increases the energy fluctuation substantially, but timesteps as large as 8-10 fs are still practical. This

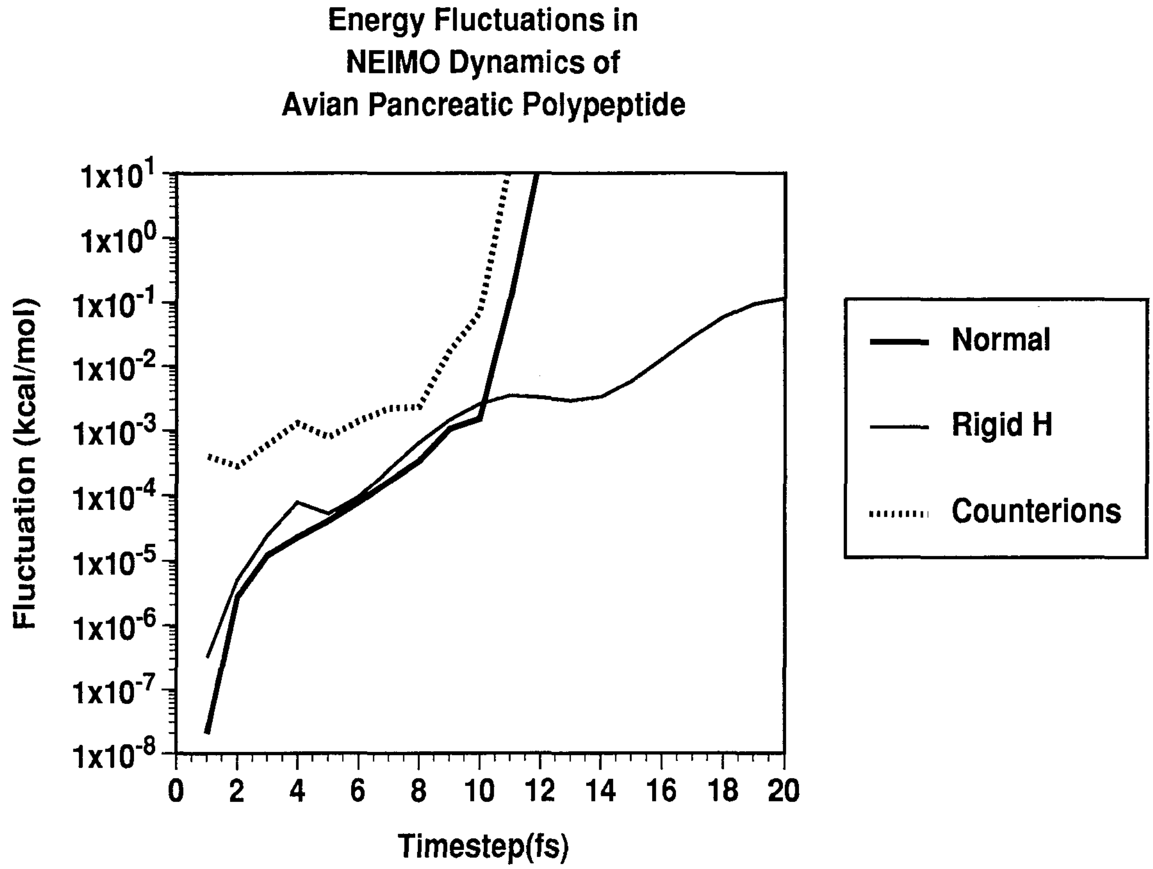


Figure 2.7. Scaled energy fluctuations, $\mathcal{E}^*$, for 1 ps NEIMO simulations of aPP. "Rigid H" differs from "Normal" NEIMO in that hinges which rotate only hydrogen atoms are held fixed. The "Counterions" simulation used the standard NEIMO method for the protein, but concurrently solved the Cartesian equations of motion for counterions (5 Na$^+$ and 3 Cl$^-$) added to neutralize unpaired charges.

is, nevertheless, a great improvement over simulations where all atoms are treated with Cartesian-space molecular dynamics.

It is common practice to keep the temperature of a microcanonical dynamics simulation roughly constant by periodically scaling the velocities. Other calculations which must be done periodically, such as updating a list of nonbond pairs within a given cutoff distance, can be done at the same time the velocities are rescaled. This is particularly important for large systems, where calculation of all nonbonded interactions for every timestep is prohibitive. Under such conditions, where nonbonds and velocities are updated periodically, the total energy, E, does not remain constant throughout the entire time of the simulation. The energy fluctuation $\mathcal{E}$ is, therefore, no longer an accurate measure of the dynamics because $\langle E^2 \rangle$ diverges from $\langle E \rangle^2$. Instead, we have used the average fluctuation, $\langle \mathcal{E} \rangle$, determined by calculating $\mathcal{E}$ during each 0.100 ps interval, and averaging. If the total calculation has N_i 0.100 ps intervals, $\langle \mathcal{E} \rangle$ is defined by

$$\langle \mathcal{E} \rangle = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathcal{E}_i, \tag{2.5}$$

where $\mathcal{E}_i$ is the energy fluctuation calculated during the i-th interval. In such calculations, timesteps should be chosen so that they give an integral number of dynamics steps per 0.100 ps – for instance, a timestep of 3.0303 fs is used, rather than 3.0 fs.

Figure 2.8 shows the variation in $\langle \mathcal{E} \rangle$ during 5 ps simulations of aPP. In these calculations, the Cell-Multipole Method was used for the nonbond calculations. The $\mathcal{E}$ was calculated during 0.1 ps intervals, during which the average kinetic energy was calculated and the farfield contribution to the CMM energy was held constant[4]. At the end of each 0.1 ps interval, the velocities were rescaled if necessary, the CMM farfield was recalculated, and the $\mathcal{E}$ was recorded. At the end of the 5 ps simulations, the $\mathcal{E}$ values were averaged to give $\langle \mathcal{E} \rangle$. These values are plotted in Figure 2.8. For very short timesteps (1 and 2 fs), the $\langle \mathcal{E} \rangle$ values are much larger than the $\mathcal{E}$ values

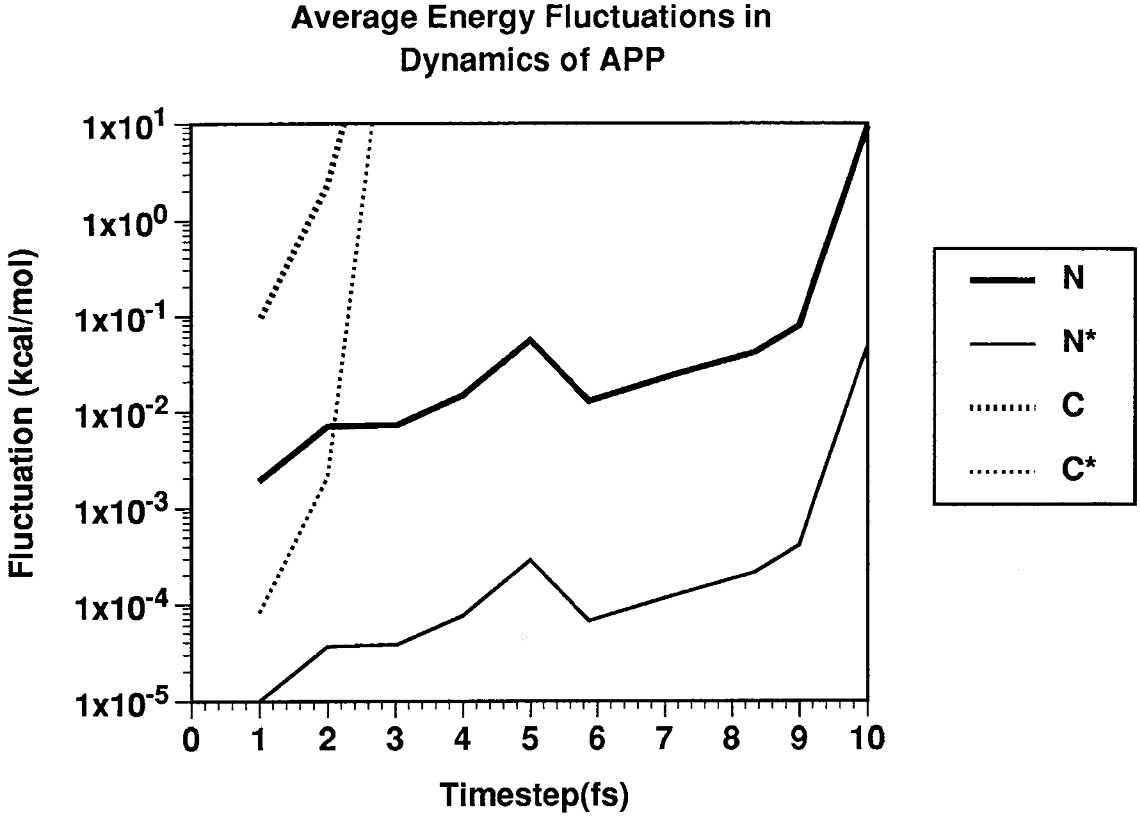


Figure 2.8. The average energy fluctuations, $\langle \mathcal{E} \rangle$, during 5 ps simulations of avian pancreatic polypeptide. Fluctuations in NEIMO(N) and Cartesian(C) dynamics were determined at 0.1 ps intervals during the course of the simulation, after which velocities could be rescaled and the CMM nonbond farfield calculation was updated.

from the 1 ps simulations in Figure 2.6. At large timesteps, however, the results are very consistent with the shorter simulations.

Figure 2.9 shows the average value of E* during 5 ps simulations of several of the proteins in Table 2.2. Clearly the energy fluctuations in NEIMO dynamics simulations, even when scaled by the number of degrees of freedom, increase linearly with protein size. This is in contrast to the fluctuations during Cartesian dynamics simulations, which are roughly constant when divided by the number of degrees of freedom. Some of the possible causes of this phenomenon have been mentioned above. It should be noted that for NEIMO simulations using a 1 fs timestep, the

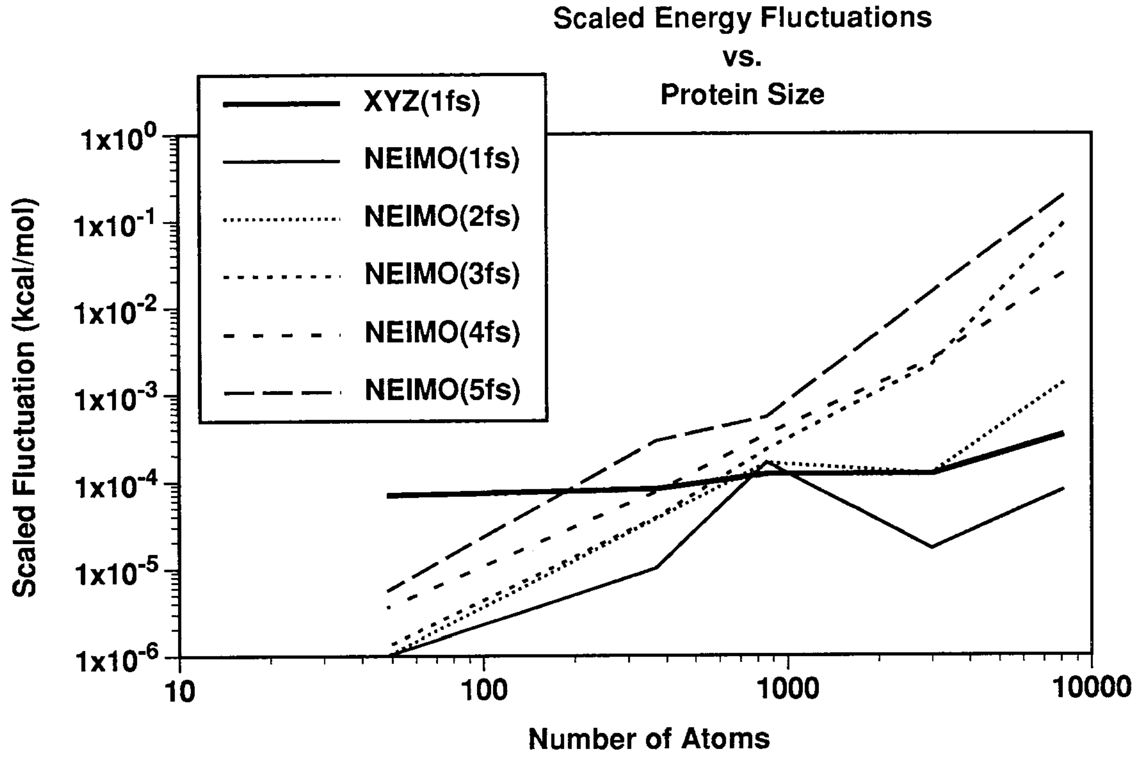


Figure 2.9. $\langle\mathcal{E}\rangle$ vs. protein size for Cartesian dynamics at 1fs and NEIMO dynamics at various timesteps.

rise in $\langle\mathcal{E}\rangle^*$ is much flatter than at higher timesteps. The very large proteins studied here may have problems with large timesteps simply because the actual number of nonbond interactions is extremely large, being proportional to n^2, and large dynamics steps can cause rearrangement of a substantial proportion of these interactions. Currently, work is being done to improve the results for large timesteps.

III.C. Dihedral Distributions

Analysis of energy fluctuations indicates that the NEIMO method accurately solves its equations of motion for molecular systems, but does not verify that NEIMO produces molecular motions similar to Cartesian dynamics, which solve the equations of motion for the full $3n$ - 6 degrees of freedom. In order to determine how well NEIMO simulations represent the dynamics of the Met-enkephalin peptides, the distribution of dihedral angles during these simulations was determined. Cartesian dynamics using a 1fs timestep and NEIMO dynamics using a variety of timesteps were run for 5.0 ps at a simulation temperature of 300 K, during which the dihedral angles were output every 0.1 ps. Figure 2.10 shows the resulting distributions from simulations of Met-enkephalin. The numbering of the dihedral angles is shown in Figure 2.1 and further identified in Table 2.1. The top graph in Figure 2.10 shows the distribution from Cartesian dynamics and the bottom shows the distribution from a NEIMO dynamics simulation; both simulations used a 1 fs timestep. The distributions from the two simulations are very similar, with the backbone ω dihedrals (6, 9, 12, and 17) showing the least flexibility, as would be expected, and the methionine sidechain dihedrals showing the greatest variation during the simulation. The average values for each dihedral, θ, can be calculated from such distributions. Because dihedral angles have a periodicity of 2π (360°), the average cannot be calculated directly, but is derived from the average cosine and sine[28]:

$$\langle\theta\rangle = \arctan\left(\langle\sin\theta\rangle/\langle\cos\theta\rangle\right). \tag{2.6}$$

Once $\langle\theta\rangle$ is known, the standard deviations can be calculated easily from N_i simulations:

$$\sigma = \left[\frac{\sum_{i=1}^{N_i}(\delta\theta)^2}{N_i - 1}\right]^{1/2}, \tag{2.7}$$

where

$$\begin{gathered}\delta\theta_i = (\theta_i - \langle\theta\rangle)\\ -\pi < \delta\theta_i < \pi.\end{gathered} \tag{2.8}$$

Because of the periodicity of dihedral angles, Equation (2.8) can always be enforced by appropriate additions or subtractions of 2π.

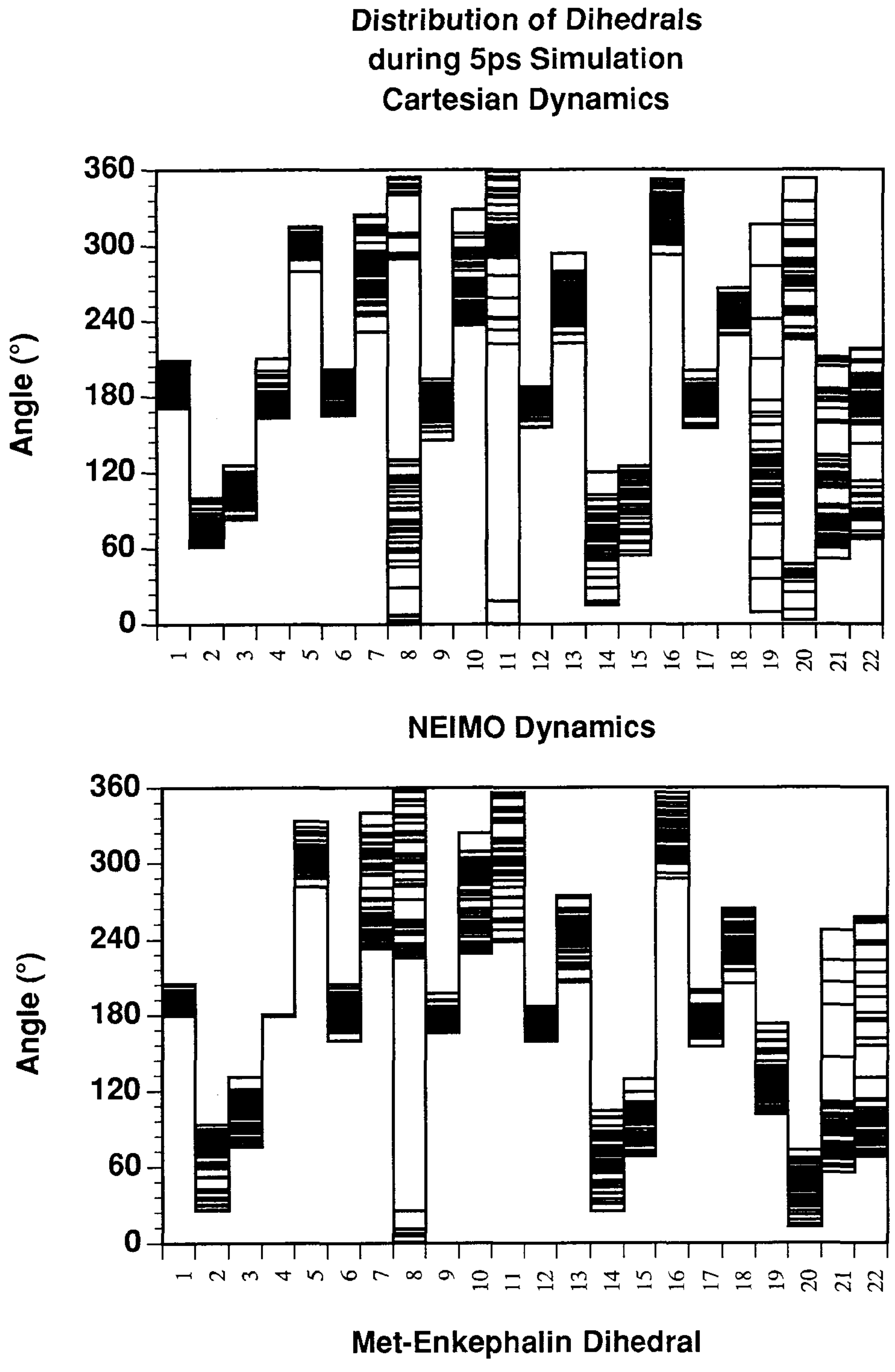


Figure 2.10. During 5 ps molecular dynamics simulations of Met-enkephalin, the 22 dihedral angles were written out at 0.1 ps intervals. The fifty values for each dihedral are plotted here for Cartesian and NEIMO dynamics simulations using 1 fs timesteps.

The average values, $\langle\theta\rangle$, and standard deviations, σ, for the distributions in Figure 2.10 are shown in Figure 2.11. The average values are also shown in Table 2.4, and are compared to the initial conformation. The NEIMO results are very similar to the results from the Cartesian simulations, indicating that the reduction in the number of degrees of freedom does not, in general, affect the torsional flexibility of the molecules. There are two exceptions to this here: χ^1 of Met 5 undergoes a transition from roughly 30°to -60°(300°) in the Cartesian simulation, but remains near 45°in the NEIMO simulation. Secondly, the ψ angle of Gly 2 is rotated from -60°to 60°in the Cartesian simulation, but remains near -60°during the NEIMO calculation. It is possible that fixing the angle terms increases the barriers to rotation enough to prevent these transitions during 5 ps NEIMO simulation at 300 K. The rotational transition of Met 5 χ^1 did occur after approximately 40 ps of a 50 ps NEIMO simulation using 5 fs timesteps. A 600 K NEIMO simulation using 5 fs timesteps saw both transitions occur by 20 ps, but the temperature was high enough that further transitions continued in both directions over these barriers. It is important to note that the NEIMO formalism explicitly includes the capacity for bond stretches and angle bends between clusters, but the current implementation uses only the dihedral degrees of freedom.

A slightly different type of plot shows the average dihedrals from 10 different NEIMO simulations in Figure 2.12. The simulations were identical except for the timestep, which ranged from 1 fs to 10 fs. As explained above, timesteps were chosen to give an integer number of dynamics steps per 0.100 ps. It is clear that the results are extremely consistent for timesteps up to 10 fs. Only the two outer sidechain dihedrals χ^1 and χ^2 of Met 5 have significantly different distributions at different timesteps. χ^1 has $\langle\theta\rangle \approx 145°$ for 8, 9, and 10 fs timestep simulations, but $\langle\theta\rangle \approx 90°$ for the smaller timesteps. It is possible that the larger timesteps occasionally enable

Dihedral	θ_0	$\langle\theta\rangle_N$	$\delta\theta_{N0}$	$\langle\theta\rangle_C$	$\delta\theta_{C0}$	$\delta\theta_{CN}$
1	186.3	191.1	4.8	189.6	3.3	-1.5
2	70.6	66.6	-4.0	76.1	5.5	9.5
3	107.4	99.9	-7.5	104.6	-2.8	4.7
4	178.0	178.0	0.0	180.2	2.2	2.2
5	300.1	306.8	6.7	301.4	1.3	-5.4
6	185.1	183.4	-1.7	186.0	2.6	2.6
7	311.5	272.0	-39.5	279.4	-32.1	7.4
8	306.8	300.8	-6.0	55.8	109.0	115.0
9	182.1	177.7	-4.4	173.0	-9.1	-4.7
10	294.8	271.7	-23.1	263.4	-31.4	-8.3
11	353.2	303.8	-49.4	309.8	-43.4	6.0
12	173.4	173.9	0.5	172.3	-1.1	-1.6
13	246.1	241.6	-4.5	254.7	-8.6	13.1
14	61.3	65.5	4.2	71.2	9.9	5.7
15	72.0	92.8	20.8	99.8	27.0	7.6
16	351.7	320.0	-31.7	320.4	-31.3	0.4
17	184.0	177.2	-6.8	176.3	-7.7	-0.9
18	238.8	241.0	2.2	245.6	6.8	4.6
19	116.3	128.8	12.5	116.0	-0.3	-12.8
20	33.3	46.0	12.7	289.8	-103.5	-116.4
21	70.1	89.2	19.1	113.5	43.4	24.3
22	82.0	117.2	35.2	159.4	77.4	-39.8

Table 2.4. The average values of the Met-enkephalin dihedrals from 5 ps NEIMO ($\langle\theta\rangle_N$) and Cartesian ($\langle\theta\rangle_C$) dynamics simulations, compared to the initial values θ_0 and compared to each other.

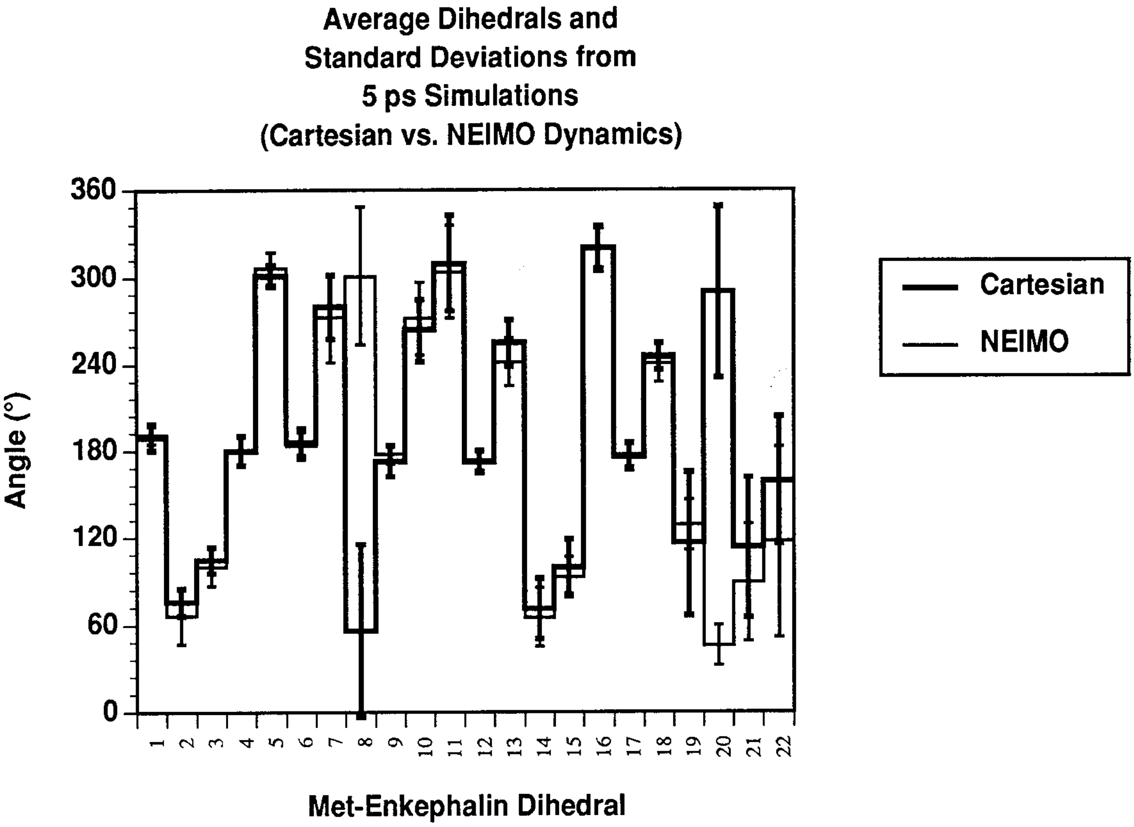


Figure 2.11. The average dihedrals from the distributions in Figure 2.10 are shown here with error bars indicating $\pm\sigma$, the standard deviations.

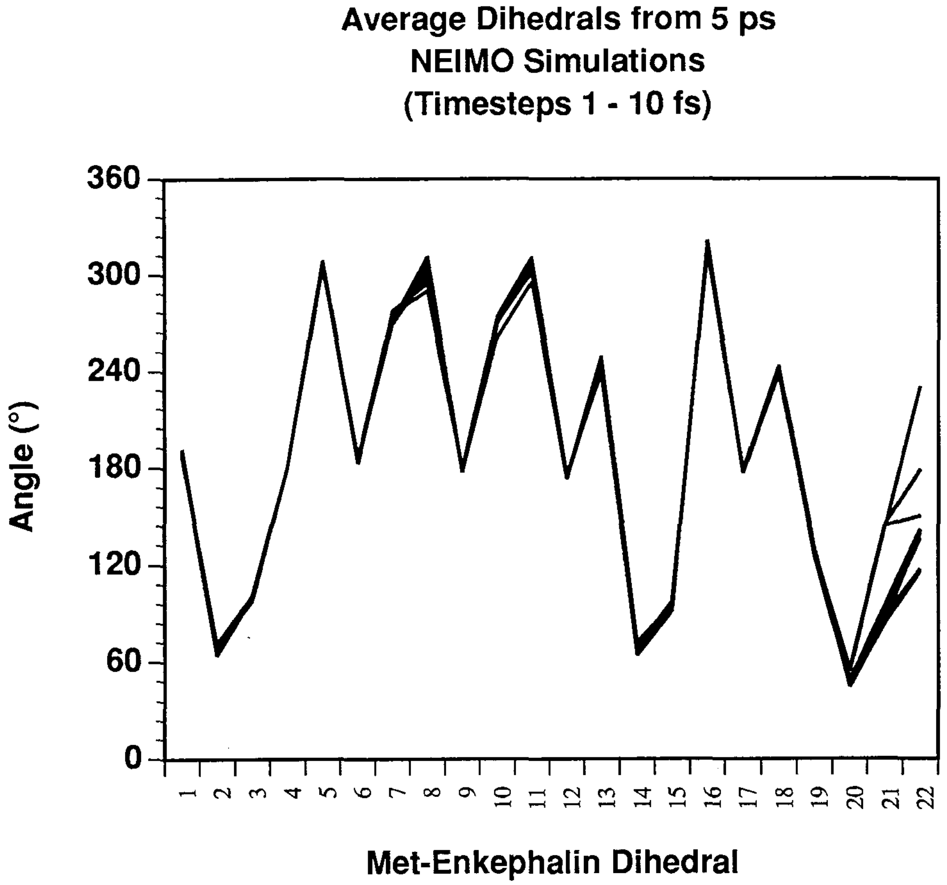


Figure 2.12. The average dihedrals from NEIMO simulations using timesteps ranging from 1 to 10 fs.

the molecule to jump over rotational energy barriers which cannot be cleared by simulations using smaller timesteps which, in effect, calculate energies and forces at more points along the trajectory.

In order to quantify the dihedral distributions, we represented each distribution by a gaussian, using the average, $\langle\theta\rangle$, and standard deviation, σ, from the 50 datapoints:

$$\Psi(\theta, \langle\theta\rangle, \sigma) = \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^{1/2} e^{-\delta\theta^2/4\sigma^2}. \tag{2.9}$$

These gaussians were normalized to give a probability function,

$$\int_{-\pi}^{\pi} P(\theta)d\theta = \int_{\pi}^{\pi} [\Psi]^2 d\theta = 1, \tag{2.10}$$

which is a normal distribution. Note that the constant in Equation (2.9) is derived for nonperiodic variables, and the total probability in Equation (2.10) does not exactly equal 1.0 if σ is so large that the probability is non-zero for every value of θ. This is not the case for any of the distributions we have analyzed.

Distributions from two different simulations can be compared by calculating the overlap, S_{12}, of the functions Ψ_1 and Ψ_2:

$$S_{12} = \int_{-\pi}^{\pi} \Psi_1 \Psi_2 d\theta. \tag{2.11}$$

If Ψ_1 and Ψ_2 are defined as

$$\Psi_1 = \left[\tfrac{2\alpha}{\pi}\right]^{1/4} e^{-\alpha(\delta\theta_1)^2}, \quad \Psi_2 = \left[\tfrac{2\beta}{\pi}\right]^{1/4} e^{-\beta(\delta\theta_2)^2}, \tag{2.12}$$

where $\alpha = 1/4\sigma_1^2$ and $\beta = 1/4\sigma_2^2$, and $\delta\theta_1$ and $\delta\theta_2$ are defined as in Equation (2.8) for $\langle\theta\rangle_1$ and $\langle\theta\rangle_2$, then the product of these functions is also a gaussian:

$$\Psi_1 \Psi_2 = \left[\frac{4\alpha\beta}{\pi^2}\right]^{1/4} K_{12} e^{-(\alpha+\beta)\delta\theta_{12}}. \tag{2.13}$$

Here, $\delta\theta_{12}$ is defined as usual from $\langle\theta\rangle_{12}$, where

$$\langle\theta\rangle_{12} = \frac{\alpha\langle\theta\rangle_1 + \beta\langle\theta\rangle_2}{\alpha+\beta}. \tag{2.14}$$

The constant in Equation (2.13) is

$$K_{12} = \exp\left[\frac{(\alpha\langle\theta\rangle_1 + \beta\langle\theta\rangle_2)^2}{\alpha+\beta} - (\alpha\langle\theta\rangle_1^2 + \beta\langle\theta\rangle_2^2)\right]. \tag{2.15}$$

Inserting Equation (2.13) into Equation (2.11) gives a formula for the overlap:

$$S_{12} = \left[\frac{4\alpha\beta}{(\alpha+\beta)^2}\right]^{1/4} K_{12}. \tag{2.16}$$

S_{12} equals 1 if the two distribution functions are identical and equals 0 if there is no overlap.

The overlaps from the 10 NEIMO simulations plotted in Figure 2.12 are shown in Figure 2.13. Each line represents the overlap between the 1 fs timestep simulation and one of the simulations with a larger timestep. A second figure, Figure 2.14, specifically shows the overlaps between the 1 fs simulation and the 2, 5, and 10 fs simulations at a higer resolution. As expected, there is almost 100% overlap among the NEIMO simulations, which indicates clearly that the molecular dynamics are very consistent across a range of timesteps of 1 fs to 10 fs. The only exceptions to these are χ dihedrals of Met 5 and the ω of Gly 2. The relatively small overlap of the latter is actually due to the very small value of σ (0.3°) for the 1 fs NEIMO simulation. The discrepancy in the methionine sidechains is also due primarily to differences in σ rather than $\langle\theta\rangle$ for the smaller timestep simulations. At larger timesteps, however, both σ and $\langle\theta\rangle$ differ.

Overlaps between the dihedral distribution from the Cartesian simulation, and those from the NEIMO simulations, are shown in Figure 2.15. Here, the overlap is quite small for χ^2 of Met 5 and the ψ backbone dihedral of Gly 2, as indicated by the large differences in $\langle\theta\rangle$ note above. A third very-low overlap is seen for the ω of Gly 2. This difference is completely hidden in Figure 2.11 since it is due entirely to the extremely low value of σ in the NEIMO simulations. The value is so low, in fact, that it does not appear in Figure 2.11 for the 1 fs NEIMO simulation. The overlaps are greater than 65% for 19 of the 22 dihedrals for every NEIMO timestep. Excluding the methionine residue, overlaps are greater than 90% for 13 of the 16 dihedrals.

IV. Conclusions

The NEIMO method has now been successfully applied to polypeptide and protein systems. The method is extremely fast compared to other internal-coordinate dynamics methods, as it scales linearly with the number of degrees of freedom.

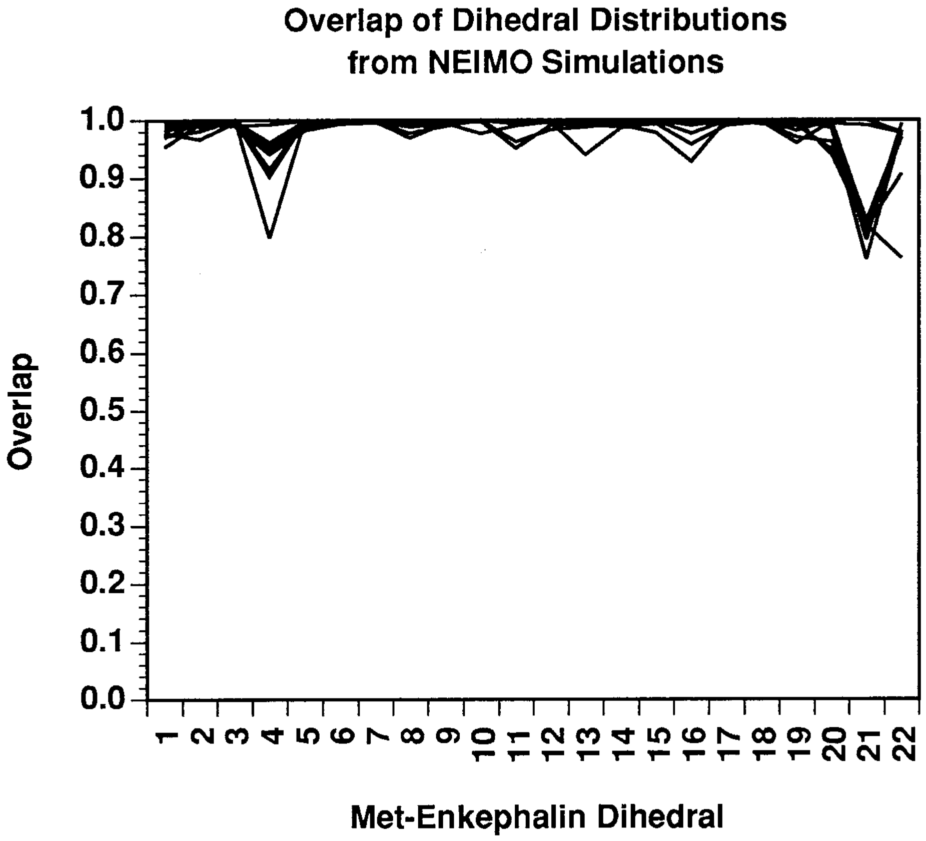


Figure 2.13. The overlaps S_{12} between 1 fs and 2-10 fs NEIMO simulations of Met-enkephalin.

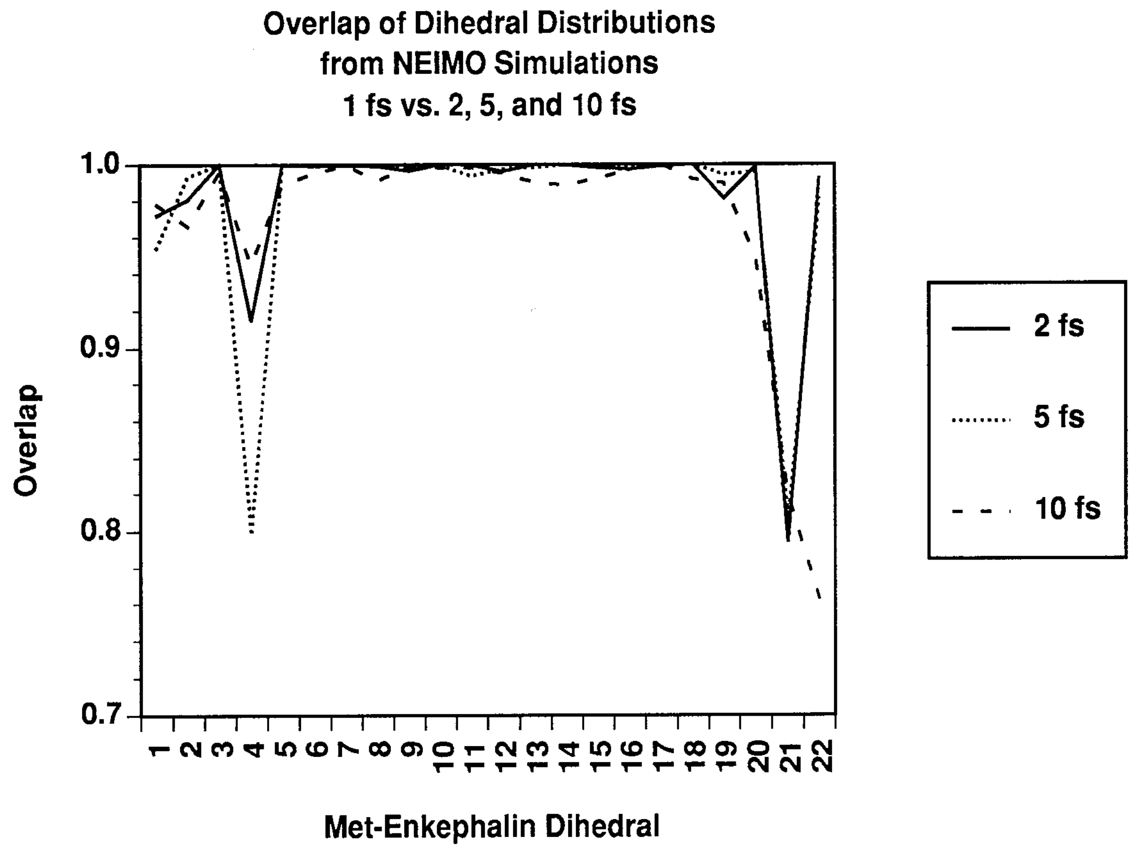


Figure 2.14. The overlaps S_{12} between 1 fs and 2, 5, and 10 fs NEIMO simulations of Met-enkephalin shown at higher resolution than Figure 2.13.

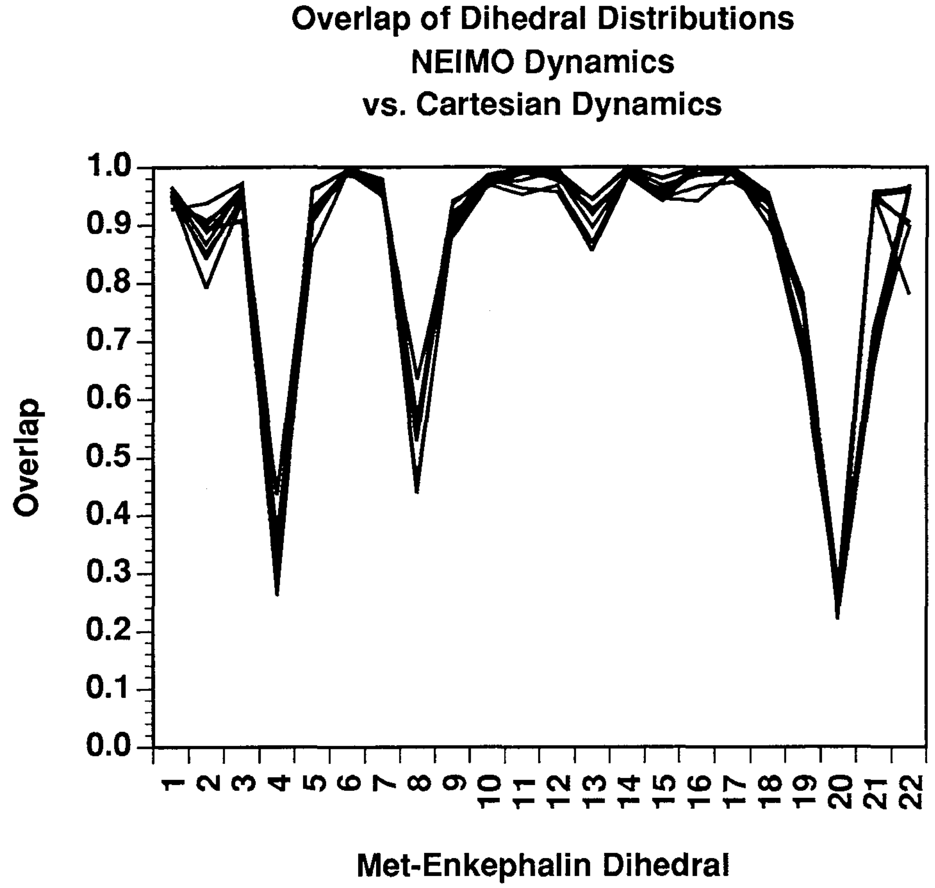


Figure 2.15. The overlaps S_{12} between dihedral distributions from Cartesian dynamics vs. those from NEIMO dynamics simulations with timesteps ranging from 1 fs.

For increasingly large systems, the NEIMO computational requirements grow more slowly than those for energy calculations, thereby allowing for its applicability to extremely large systems. Molecular dynamics of only torsional degrees of freedom can use larger timesteps than simulations allowing all possible degrees of freedom. NEIMO calculations of peptides indicate that timesteps as large as 20 femtoseconds can be used for these small systems. Timesteps of this size are not yet possible for large polypeptides and proteins, as judged by the criterion of total energy fluctuations. However, timesteps of 5 fs and longer can be used for large systems without danger of energy divergence; such calculations may be useful for conformational analyses of extremely large systems such as viruses. The use of a different integration method may improve the energy conservation for larger systems.

The dynamics of polypeptides are accurately modeled by the NEIMO method. Analyses of dihedral angle fluctuations show that NEIMO dynamics simulations produce conformational fluctuations very similar to those arising from Cartesian dynamics simulations. The few exceptions to this in simulations of Met-enkephalin appear to be cases where rotational barriers are traversed in the Cartesian dynamics simulation, but not in a NEIMO simulation at the same temperature. It is likely that fixing bond angles keeps rotational barriers higher than in a more flexible model. Future implementations of the NEIMO method for molecular systems will include the additional hinge degrees of freedom, allowing for flexibility in these angles.

Acknowledgements

We would like to acknowledge the indispensable help of Dr. Abhinandan Jain of the Jet Propulsion Laboratory in providing the spatial operator codes for the core computations in our molecular version of NEIMO dynamics. We also wish to thank Dr. Guillermo Rodriguez of JPL and Dr. N. Vaidehi of Caltech for their important contributions to the development of the methodology and the current implementation. AMM acknowledges a National Research Service Award/NIH Predoctoral Traineeship in Biotechnology.

A. Leapfrog Verlet for Internal Coordinates

The Verlet algorithm is probably the most popular algorithm for integrating the equations of motions in standard Cartesian-space molecular dynamics. Currently, NEIMO dynamics uses a modification of one formulation of the Verlet algorithm, the "leapfrog" formulation. Regardless of the formulation, the Verlet algorithm requires that accelerations be calculated at a time t, before the velocities at time t are known. However, NEIMO accelerations at time t depend upon the velocities at time t, so these velocities must first be estimated. Iteration allows for a more accurate determination of the velocities.

A "leapfrog" Verlet calculation with a timestep of h determines the accelerations at integer timesteps n ($\ddot{\theta}^n$, simulation time $= nh$) and uses these to determine the velocities at timestep $n + \frac{1}{2}$ ($\dot{\theta}^{n+\frac{1}{2}}$). These velocities are then used to determine the coordinates at timestep $n+1$ (θ^{n+1}). The next dynamics timestep then proceeds with $n = n+1$. However, in order to calculate $\ddot{\theta}^n$, NEIMO requires the unknown $\dot{\theta}^n$ as well as known θ^n. It is, therefore, first necessary to estimate the velocities $\dot{\theta}^n$ from the previously determined velocities:

$$\dot{\theta}^n = 1.5\dot{\theta}^{n-\frac{1}{2}} - 0.5\dot{\theta}^{n-\frac{3}{2}}. \tag{2.17}$$

It is then possible to calculate the NEIMO accelerations $\ddot{\theta}^n$. Accelerations are calculated by solving the spatial operator (SO) equations described in detail in References [12] and [11]. The accelerations are a function of the coordinates θ^n, the velocities $\dot{\theta}^n$, the torques $\mathbf{T}^n$, and/or the forces $\mathbf{F}^n$. The forces and torques are

$$\ddot{\theta}^n = \mathrm{SO}(\mathbf{F}^n, \mathbf{T}^n, \theta^n, \dot{\theta}^n) \tag{2.18}$$

calculated from the derivatives of the potential energy functions with respect to Cartesian and dihedral coordinates, respectively. The accelerations are used to update the velocities as they are in Cartesian dynamics (cf. Equation (1.17)).

$$\dot{\theta}^{n+\frac{1}{2}} = \dot{\theta}^{n-\frac{1}{2}} + h\ddot{\theta}^{n} \tag{2.19}$$

Because $\dot{\theta}^n$ is estimated in Equation (2.17), $\ddot{\theta}n$ is inaccurate and such errors could build up as the simulation progresses. In order to eliminate such errors, $\dot{\theta}^n$ is re-estimated from the new $\dot{\theta}^{n+\frac{1}{2}}$ and the known $\dot{\theta}n - \frac{1}{2}$.

$$\dot{\theta}^{n} = 0.5\dot{\theta}^{n+\frac{1}{2}} + 0.5\dot{\theta}^{n-\frac{1}{2}} \tag{2.20}$$

Repeating Equations 2.18 through 2.20 until $\dot{\theta}^n$ converges produces a very accurate value for $\ddot{\theta}^n$. Sufficient convergence is generally reached after a single iteration, so the effect on overall computational speed is minimal. The converged values of $\ddot{\theta}^n$ in Equation (2.19) give values for $\dot{\theta}^{n+\frac{1}{2}}$ which are then used to update the coordinates.

$$\theta^{n+1} = \theta^{n} + h\dot{\theta}^{n+\frac{1}{2}} \tag{2.21}$$

The dynamics step is completed by updating the Cartesian coordinates from the new internal coordinates θ^{n+1}.

References

[1] A.T. Brunger, *Annu. Rev. Phys. Chem.*, 42, 197-223 (1991).

[2] D.L. Beveridge and F.M. DiCapua, *Annu. Rev. Biophys. Biophys. Chem.*, 18, 431-492 (1989).

[3] B.R. Rousseau *et al.*, *Molec. Phys.*, 76, 1079-1091 (1992); N.J. Wagner *et al.*, *Phys. Rev. A*, 12, 8457-8470 (1992); C.L. Brooks, W.S. Young, and D.J. Tobias, *Int. J. Supercomp. Appl.*, 5, 98-112 (1991).

[4] H.Q. Ding, N. Karasawa, and W.A. Goddard III, *J. Chem. Phys.*, 97, 4309-4315 (1992).

[5] H.Q. Ding, N. Karasawa, and W.A. Goddard III, *Chem. Phys. Lett.*, 196, 6-10 (1992).

[6] M.E. Tuckerman and B.J. Berne, *J. Chem. Phys.*, 95, 8362-8364 (1991).

[7] J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen, *J. Comp. Phys.*, 23, 327-341 (1977).

[8] A. Rahman, *Phys. Rev.*, 136, A405-411 (1964); L. Verlet, *Phys. Rev.*, 159, 98-103 (1967).

[9] K.D. Hammonds and J.-P. Ryckaert, *Comp. Phys. Comm.*, 62, 336-351 (1991).

[10] J.A. McCammon and S.C. Harvey, *Dynamics of Proteins and Nucleic acids*, Cambridge University Press, Cambridge, UK (1987).

[11] G. Rodriguez, A. Jain, and K. Kreutzdelgado, *J. Astronaut.*, 40, 27-50 (1992).

[12] A. Jain, N. Vaidehi, and G. Rodriguez, *J. Comp. Phys.*, (to be published).

[13] H. Goldstein, *Classical Mechanics, 2nd. ed.*, Addison-Wesley, Reading, Mass., (1980).

[14] A.K. Mazur and R.A. Abagyan, *J. Comp. Phys.*, 92, 261-272 (1991).

[15] C.W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, (1971).

[16] BIOGRAF/POLYGRAF. Copyright by Molecular Simulations, Inc. (1992).

[17] T.L. Blundell *et al.*, *Proc. Natl. Acad. Sci., USA*, 78, 4175-4179 (1981).

[18] W.A. Hendrickson and M.M. Teeter, *Nature*, 290, 107-113 (1981).

[19] C.A. Collyer *et al.*, *J. Mol. Biol.*, 211, 617-632 (1990).

[20] O. Herzberg and M.N.G. James, *J. Mol. Biol.*, 203, 761-779 (1988).

[21] M. Fujinaga *et al.*, *J. Mol. Biol.*, 184, 479-502 (1985).

[22] A.E. Eriksson *et al.*, *Proteins: Structure, Function, Genet.*, 4, 274-282 (1988).

[23] D.C. Rees, M. Lewis, and W.N. Lipscomb, *J. Mol. Biol.*, 168, 367-387 (1983).

[24] S.C. Harrison *et al.*, *Nature*, 276, 368-373 (1978).

[25] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[26] B.R. Brooks *et al.*, *J. Comp. Chem*, 4, 187-217 (1983).

[27] M.J. Sippl, G. Némethy, and H.A. Scheraga, *J. Phys. Chem.*, 88, 6231-6233 (1984).

[28] M.H. Lambert and H.A. Scheraga, *J. Comp. Chem.*, 10, 817-831 (1989).

Chapter 3

Simulations of the Tomato Bushy Stunt Virus Capsid

Abstract

The spherical protein capsid of the tomato bushy stunt virus expands by 10% when calcium ions are removed from the system and the pH is raised above 7[1]. We have attempted to simulate this effect using molecular dynamics. Although we calculate velocities for only the three proteins of the asymmetric unit plus associated ions, we include the nonbonded interactions of all 180 proteins (nearly 500,000 atoms) in the virus coat. The Cell-Multipole Method (CMM)[2] enables us to calculate the nonbond interactions in this gigantic system. Also, in order to increase the timelength of the simulation, we employ internal coordinate dynamics, using Newton-Euler Inverse Mass Operator (NEIMO) Dynamics[4], which enable us to use timesteps of 2 femtoseconds for proteins of this size.

I. Introduction

The tomato bushy stunt virus is an RNA virus composed of 180 identical coat proteins arranged in $T = 3$ icosahedral symmetry. The asymmetric unit of the viral capsid, containing three copies of the coat protein having slightly different conformations, has been crystallized and is shown in Figure 3.1. The three conformations of the coat protein are designated A, B, and C, and differ primarily in the orientation of a few surface sidechains. A second major difference is that while all three conformations contain RNA-binding (R), surface (S), and projecting (P) domains, the R domain (residues 1-101) is completely unresolved in the A and B conformations while in the C conformation, residues 67-101 have an ordered structure and are resolved. The full viral coat is shown in Figure 3.3, with two spheres representing each protein. The symmetry is more apparent in Figure 3.4, where the P domains have been removed and only the S domains are shown. The picture emphasizes the five-fold symmetric axis, but three-fold and two-fold symmetries also exist for the coat. The viral RNA (molecular weight 1.5×10^6) is, of course, not icosahedral; it is disordered and does not appear in the crystal structure at all.

In addition to the three proteins, Figure 3.1 shows the location of the two Ca^{2+} ions which bind per protein. Each pair of calcium cations binds in a negatively charged pocket at the interface between adjacent S domains in the triad, the pocket being formed by five aspartic acid sidechains contributed by the two proteins. This is shown in more detail in Figure 3.2. It is postulated that the interaction between these Asp residues and the Ca^{2+} ions plays a major role in stabilizing the viral coat[1]. If the Ca^{2+} ions are removed, the virus expands as the pH is raised above 7. The hydrodynamic radius of the virus can expand by as much as 10%, but there is no loss of mass and the process is reversible. A low-resolution (8 Å) crystal structure was determined for the expanded conformation of the virus[1] and indicated that

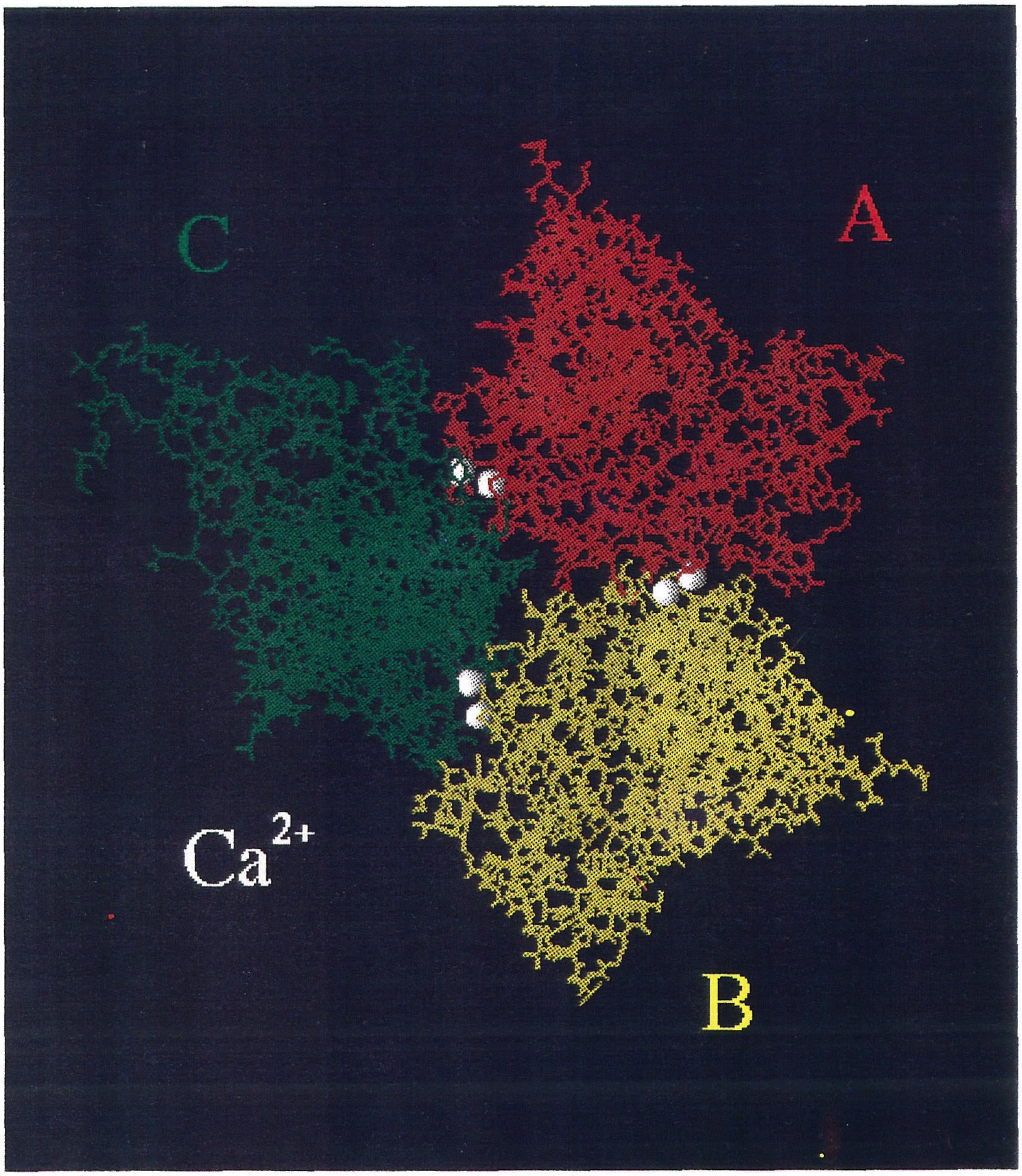


Figure 3.1. The tomato bushy stunt virus asymmetric unit, showing the A, B, and C conformations as well as the associated Ca^{2+} ions. From the crystal structure by Olson *et al.*[3].

expansion occurred by relative motions perpendicular to the interfaces where Ca^{2+} ions bind in the unexpanded conformation. However, no atomic details were available from this low resolution data.

In order to investigate the expansion phenomenon, we carried out molecular dynamics calculations on two different models of the viral coat proteins representing different possible configurations. The model systems include all resolved residues from the asymmetric unit plus counterions, Na^+ and Cl^- and, perhaps, Ca^{2+}, for a total of 8138 atoms. Through the use of the transformation matrices in the crystal structure (Brookhaven Protein Database structure 2TBV), coordinates can be generated for the entire viral coat containing 180 proteins and nearly 488,280 atoms. It is not yet practical to simulate a system this large on a standard workstation, so the dynamics were only calculated for the three proteins of the asymmetric unit. However, it is possible to include the electrostatic and van der Waals, i.e., "nonbonded" forces contributed by the rest of the viral coat. No standard method for calculating nonbonded forces could be used for a system of this size, but the Cell-Multipole Method (CMM)[2] provides a means for doing such calculations both quickly and accurately. Therefore, we are able to simulate the dynamics of the entire viral coat. The RNA is not represented in the current calculations. Newton-Euler Inverse Mass Operator (NEIMO) Dynamics (see Chapter 2) provide a means of speeding the calculations by allowing us to use internal-coordinate dynamics with timesteps of 2 fs, rather than the 1 fs timesteps required by Cartesian dynamics.

II. Methodology

In an attempt to simulate the expansion effect, two different models of the TBSV asymmetric unit were developed. The first contains the protein atoms and calcium

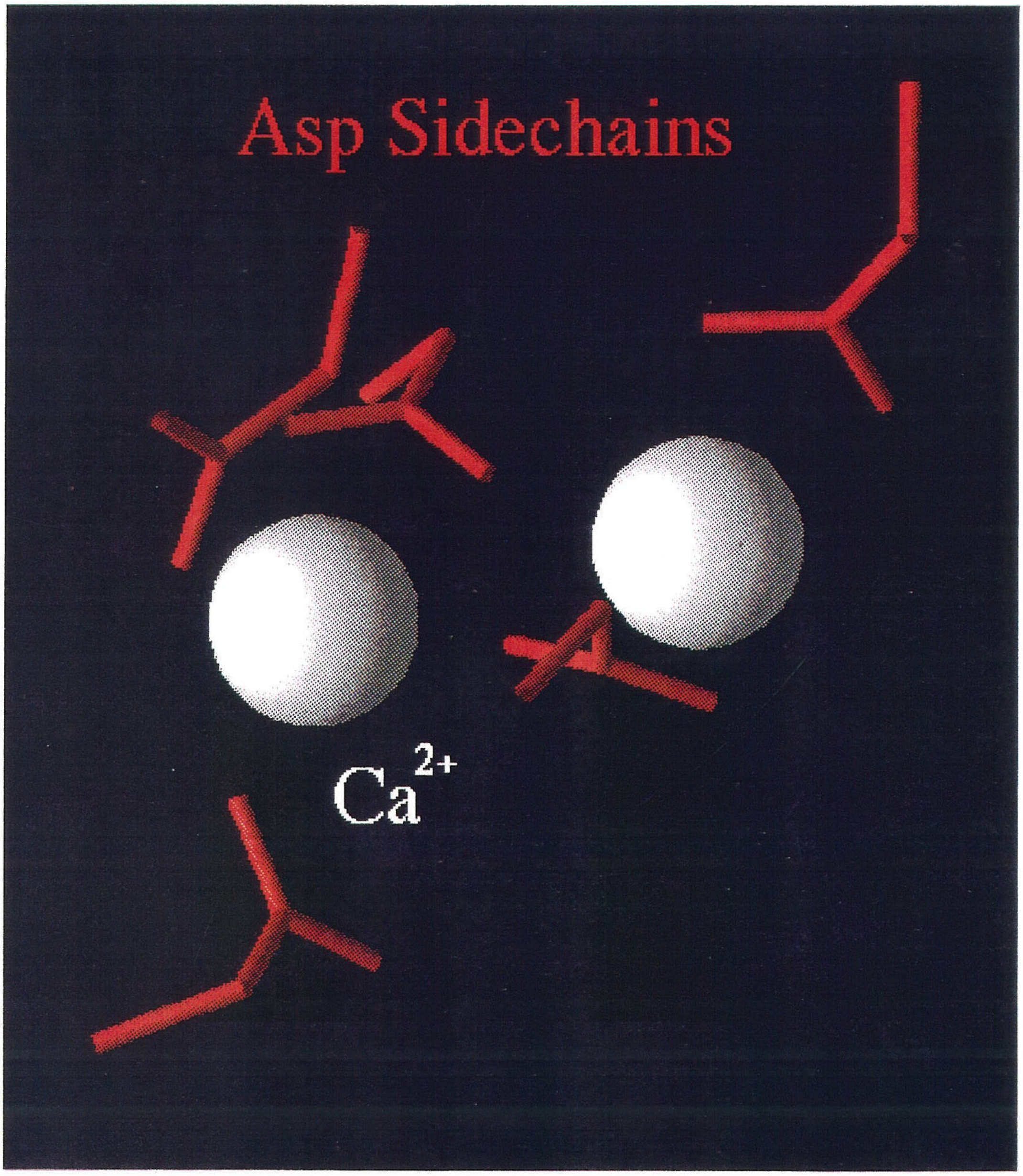


Figure 3.2. Detail of the aspartic acid/Ca^{2+} interactions at the contact site between the S domains of two coat proteins.

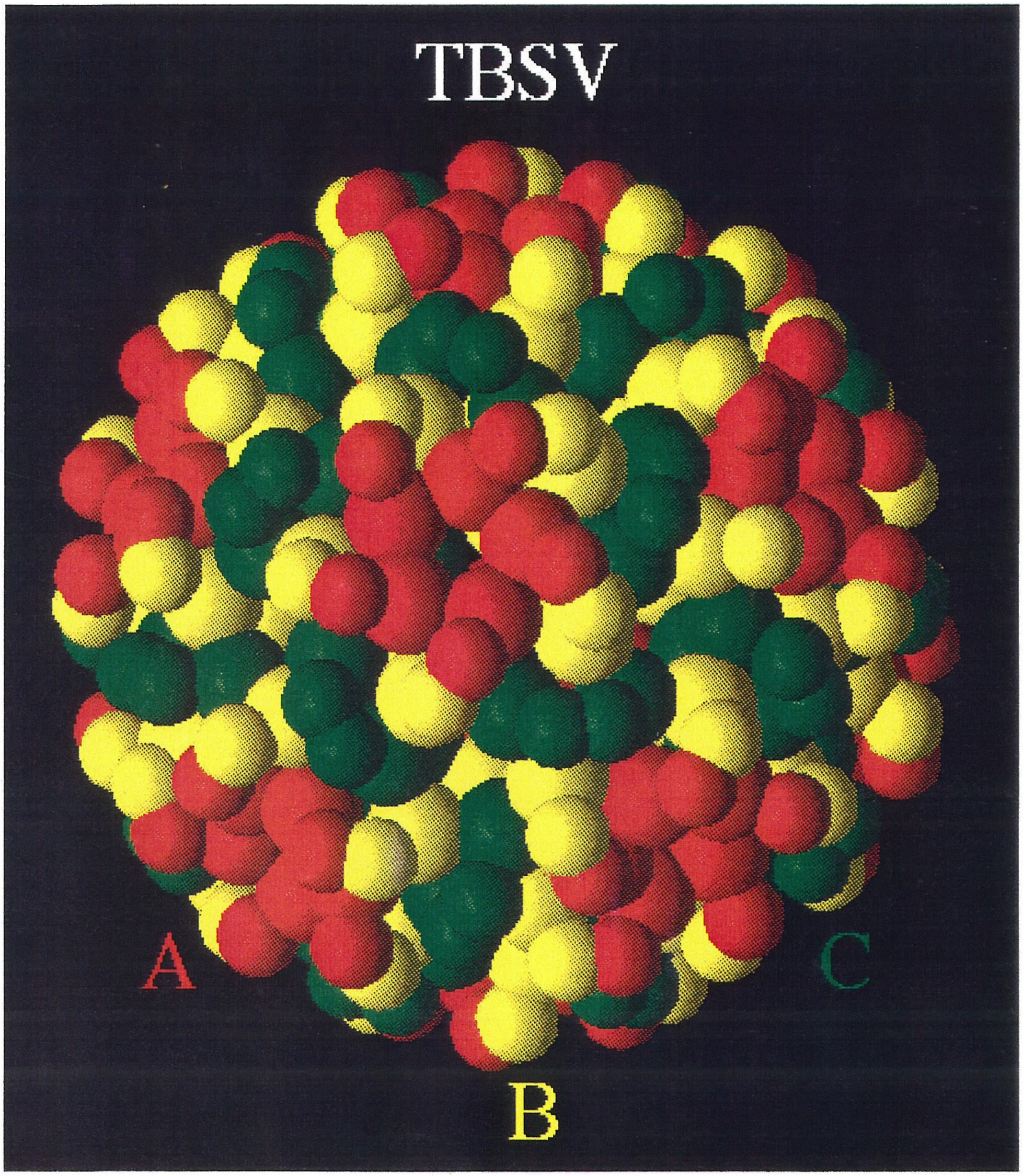


Figure 3.3. The TBSV protein coat, with each P and S domain represented by a sphere.

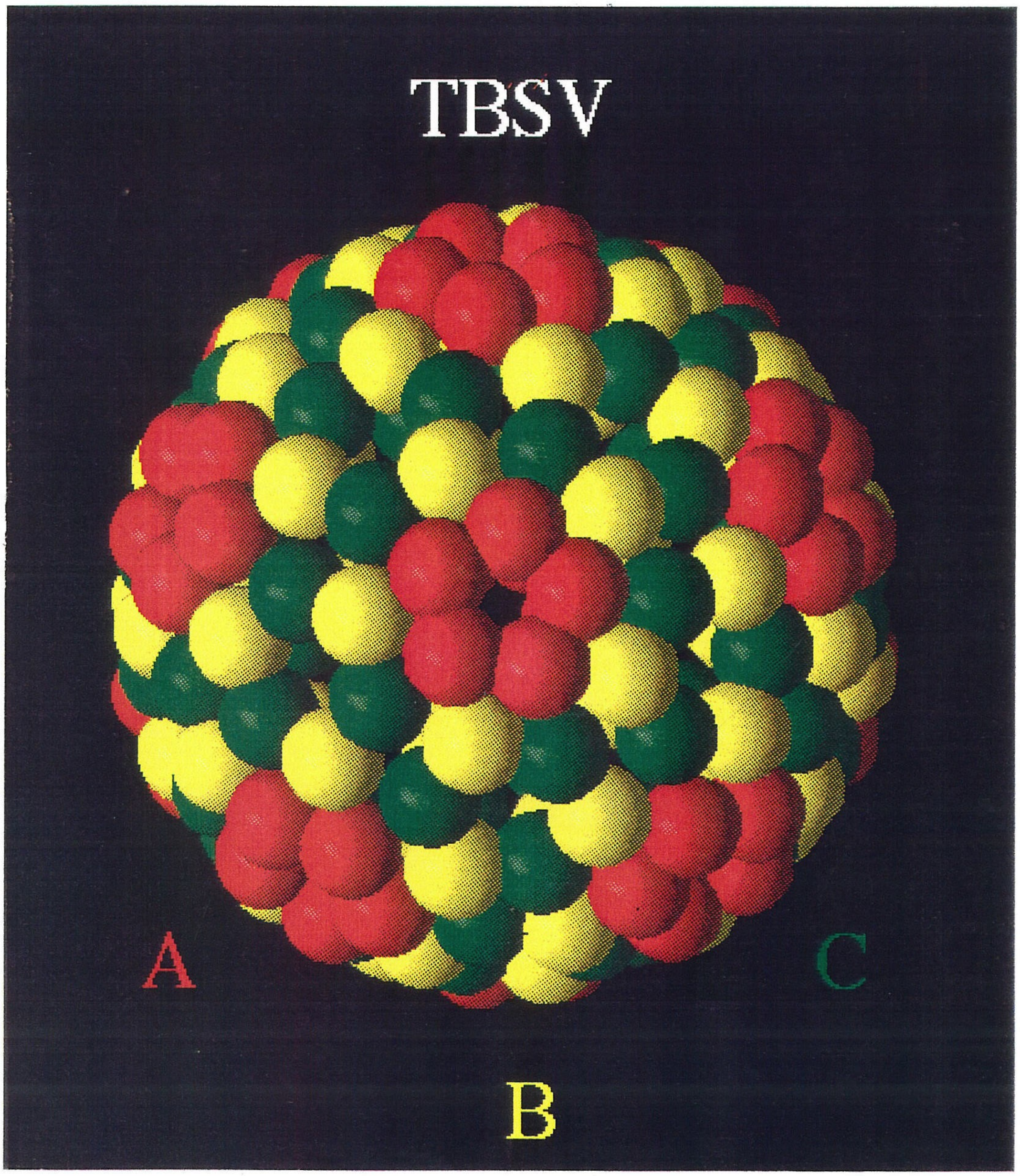


Figure 3.4. A second view of the TBSV coat, with the outer P domains removed to emphasize the symmetry of the virus.

ions as they appear in the protein database coordinate file (2TBV)[3], with hydrogen atoms added to nitrogen, oxygen and sulfur atoms to allow for hydrogen bonding. In addition, Na^+ and Cl^- ions were added to balance the charges of unpaired acidic and basic residues, respectively. This structure is termed the "PH7" model. The second representation is the "NOCA" model, in which the six Ca^{2+} ions were removed and the free aspartic acid residues were allowed to form salt bridges with basic residues, or were given Na^+ counterions. In this model, the 15 Asp residues are no longer held together by interactions with Ca^{2+}, but are free to move independently. This is believed to be the major factor in the expansion of the virus particle[1].

Inclusion of explicit waters in a calculation of this type can improve its accuracy but can also greatly expand the computational cost. We do not, therefore, have waters included in the system. However, we are able to compensate for this partially in two ways: 1) by using a distance-dependent dipole, and 2) by using counterions to balance lone charges. The first mimics the charge-shielding capacity of water by using an effective dielectric constant between charges i and j proportional to the distance between the two charges, r_{ij}

$$E(q_i, q_j) = \frac{q_i q_j}{\epsilon_{eff} r_{ij}} = \frac{q_i q_j}{\epsilon r^2}. \tag{3.1}$$

Inclusion of counterions in the simulation can provide charge stabilization for lone charged groups. In nature, such stabilization is provided both by water dipoles and free ions. The 893 residues of the simulated triad include 69 acidic (48 aspartic acid and 21 glutamic acid) and 58 basic (38 arginine and 20 lysine) residues. Of the 48 aspartic acid residues, 15 are involved in the binding of the 12 Ca^{2+} ions. At each protein interface, two calcium ions and five aspartic acids are present; at the A/B interface Asps 181, 183, and 186 from A are present along with Asps 153 and 225 from B. A basic residue, Lys 232 from A, is also present, approximately 3.1 Å from Asp 183. Excluding these six residues per protein, there were a total of 109 free charged residues.

Not every charged residue requires a counterion, since many are involved in salt bridges. We eliminated (+,-) paired sidechains by looking at all pairwise distances between the central atoms of oppositely charge sidechains. These atoms are C_γ in aspartic acid, C_δ in glutamic acid, C_ζ in arginine, and N_ζ in lysine. We checked not only for (+,-) pairs within the three proteins of the asymmetric unit but expanded the viral coat into its entire 180 proteins and checked for pairs between residues in different triads. Pairs less than 10 Å apart were considered to stabilize one another, and were not given countercharges. There were 30 such pairs in addition to the Lys 232–Asp 183 pair mentioned in the preceding paragraph. Of these, 23 occurred within the asymmetric unit and 10 occurred between residues of different triads. As can be seen in Figure 3.3, each P domain is closely paired with a P domain from a different triad, and several salt bridges occur in this region. After eliminating charge-paired residues, 25 Cl^- were still needed to balance 10 arginines and 15 lysines while 24 Na^+ were needed for 18 aspartic acids and 6 glutamic acids. These counterions were placed in idealized locations, determined by previous calculations on individual sidechains. In the NOCA model, the Ca^{2+} ions were removed, and the need for counterions was recalculated. In this model, the Asp 53 sidechains from B formed weak (+,-) pairs with Lys 182 of B, so counterions were not needed for these residues. This was the case at all three interfaces. Therefore, the NOCA model required a total of 33 Na^+ counterions and 22 Cl^- and there was no net change in the total number of atoms in the system.

The asymmetric unit of the TBSV viral coat contains three copies of the coat protein in slightly different conformations. The S (residues 102–274) and P (275–387) domains are resolved crystallographically in all three conformations. In each case, the two domains were built independently, so mismatches exist in the hinge region (residues 273-275). The crystal structure (2TBV)[3] lists alternate S and P coordinates for the residues in the overlap region. For these calculations, the two alternates were averaged and re-optimized by energy minimization. The R domains are not resolved, except for residues 67–101 in the C conformation. Only these are included in the calculations. The RNA is also not included in the current calculations, since no structure is available for it. The simulated PH7 system contains 893 protein residues having a total of 8083 atoms in addition to six Ca^{2+}, 24 Na^{+} and 25 Cl^{-} for a total of 8138 atoms. As explained above, the NOCA model has zero Ca^{2+} but 33 Na^{+} and 22 Cl^{-}, so the total number of atoms remains the same.

In order to accurately model the capsid environment, the nonbonded forces acting on the asymmetric unit included interactions with the other 177 proteins. This was made possible by the Cell-Multipole Method (CMM)[2], an extremely fast and accurate method for calculating nonbonds in large systems. CMM divides the simulation space into a hierarchy of cubic cells, the smallest of which contains, ideally, 4 or 5 atoms and the largest of which contains the entire system. Each cell at level c contains eight cells from level $c + 1$. For the triad alone, four levels are used. There are 4096 (8^4) cells at this level, measuring 6.397 Å on a side. Since the system is much flatter than it is cubic, 81.4% of these cells are empty, leaving 762 populated cells with an average of 10.7 atoms per cell. When the triad is expanded into the full 180 protein capsid, the cell multipole method uses six levels for the 488,820 atom system. At level 6, there are 262,144 cells measuring 5.340 Å on a side. 87.5% of these are empty, leaving 8.0 atoms per populated cell. Both the dimension and the average population in this case are better than those in simulations of the triad, alone.

In CMM calculations, the total charge, dipole, and quadrupole are calculated for each cell at each level. Exact pairwise interactions, like those in Equation (3.1) are only calculated for atoms in adjacent cells. Interactions with distant cells are calculated as interactions with the cumulative charge, dipole and multipole of those cells. For nearby cells, just beyond the nearest neighbors, interactions are calculated with the lowest-level cells, e.g., level six cells in the capsid simulation. Interactions at increasing distance are calculated with larger, higher-level cells, up to level 1. This hierarchical approach makes the nonbond calculation proportional to n, the number of atoms, rather than n^2, as would be the case for traditional methods, yet is highly accurate because the errors introduced are proportional to the strength of the interaction.

All calculations were performed on one processor of a Silicon Graphics 4D/380 workstation using the BIOGRAF software from Molecular Simulations, Inc.[6]. Energies were calculated using the DREIDING forcefield[5]. NEIMO dynamics calculations were performed using software written at JPL and Caltech, and interfaced to the BIOGRAF program.

III. Results

Timing results for the CMM molecular dynamics calculations are shown in Figure 3.5, in terms of cpu seconds on one processor of an SGI 4D/380 workstation. The total charge, dipole, and quadrupole of each cell, collectively termed the "farfield," can be recalculated every step or can be considered constant for a number of steps. These two cases are labeled "Update1" and "Update50," the latter referring to calculations in which the farfield was updated only every 50 steps. Also shown is the difference between calculations using only the nonbonds of the three-protein asymmetric unit, labeled "NB3," and those including interactions with the entire 180-protein capsid, labeled "NB180." It is interesting to note that the Update50

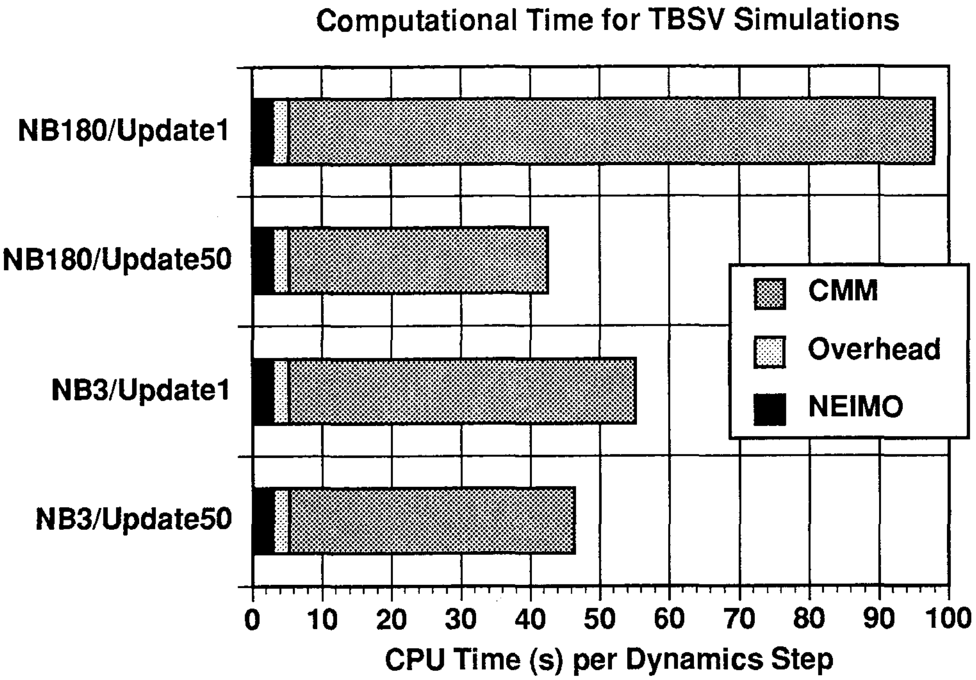


Figure 3.5. CPU times for CMM calculations, NEIMO acceleration calculations, and overhead, including coordinate updating for NEIMO.

calculations are actually faster for NB180 than NB3 (37.2 versus 41.0 s). This is because the calculation is dominated by the nearfield interactions. As noted above, the average cell in the NB180 case has 5.4 atoms versus 6.4 for the NB3 case. This means fewer pairwise interactions need to be calculated. The farfield interaction takes longer to calculate in the NB180 case, but the effect is not significant because of the hierarchical approach described above. Only when the farfield itself must be updated every step, i.e., the Update1 calculations, does the size of the system make a significant difference. In such calculations, including the entire capsid increases the time of the nonbond calculations from 49.8 s to 92.6 s.

Figure 3.5 also shows the total time required for NEIMO calculations, including the time required to calculate accelerations and velocities, and to update the system coordinates. Unlike other internal-coordinate methods, NEIMO dynamics are computationally proportional to $\mathcal{N}$, the number of internal degrees of freedom. Therefore, NEIMO can be used to calculate internal coordinate dynamics for a system this large, an almost impossible task for other internal-coordinate methods, which typically have computational costs proportional to $\mathcal{N}^3$. As is clear in Figure 3.5, the linear scaling with respect to $\mathcal{N}$ makes NEIMO entirely feasible for simulations of TBSV. NEIMO calculations for the three protein chains in the asymmetric unit ($\mathcal{N} = 4335$) require 5.3 s per dynamics step. The primary advantage of internal-coordinate methods is that large timesteps can be used; whereas Cartesian dynamics simulations typically require timesteps of 1 fs for accuracy, timesteps as large as 15 fs can be used reliably for small peptides. Although the current NEIMO implementation does not allow timesteps of this magnitude for large proteins, timesteps larger than 1 fs are practical. As is indicated in Figure 3.5, the overhead for NEIMO calculations is very small compared to the time required for the nonbond calculations. Therefore, an increase in step size just to 2 fs will nearly double the speed of the simulation with respect to a 1 fs Cartesian simulation.

Figure 3.6 shows the average scaled energy fluctuation, $\langle\mathcal{E}\rangle^*$, versus timestep for calculations on the PH7 model of TBSV. $\langle\mathcal{E}\rangle^*$ is the average value of $\mathcal{E}$ during a simulation, divided by the number of degrees of freedom. $\mathcal{E}$ is determined over 0.100 ps intervals using the equation:

$$\mathcal{E} = \frac{\langle E^2\rangle - \langle E\rangle^2}{k_B T}, \tag{3.2}$$

where E is the total energy (potential plus kinetic), k_B is the Boltzmann constant, and T is the temperature of the simulation. The figure indicates that no choice for nonbonds gives superior results at all timesteps. Fluctuations using the entire capsid, with the farfield updated every step, i.e., NB180/Update1, were not measured but are unlikely to provide a substantial improvement. In every case, there was typically a 4- to 5-fold increase in the fluctuation for each 1 fs increase in the timestep. The

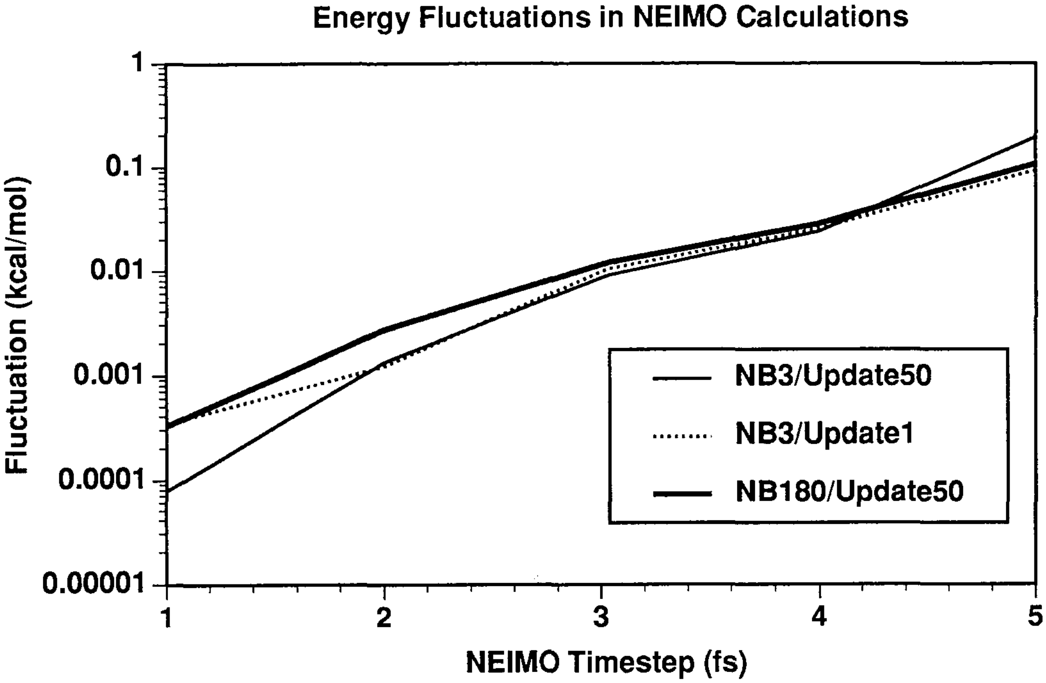


Figure 3.6. $\langle \mathcal{E} \rangle^*$ during 1.0 ps NEIMO simulations.

average scaled fluctuation, $\langle \mathcal{E} \rangle^*$, shown in Figure 3.6, is approximately the same for 1 fs NEIMO and 1 fs Cartesian dynamics. However, the unscaled fluctuation, i.e., the average value of $\mathcal{E}$ *not* scaled by the number of degrees of freedom, is of the same order of magnitude for 2 fs NEIMO and 1 fs Cartesian dynamics.

The two model systems, PH7 and NOCA, were initially optimized using Cartesian space conjugate gradients minimization. The nonbond calculations included the entire capsid, with the farfield updated every 50 steps. The radius of gyration of the entire 180 protein system was calculated every 50 steps, when the farfield was updated. The radius of gyration is defined as:

$$R_{gyr} = \left[\frac{\sum_i^n m_i[(x_i - x_{cm})^2 + (y_i - y_{cm})^2 + (z_i - z_{cm})^2]}{\sum_i^n m_i} \right]^{1/2} \qquad (3.3)$$

where the coordinates and mass of each particle i are (x_i, y_i, z_i) and m_i, respectively, and the coordinates of the center of mass are (x_{cm}, y_{cm}, z_{cm}). This is shown in

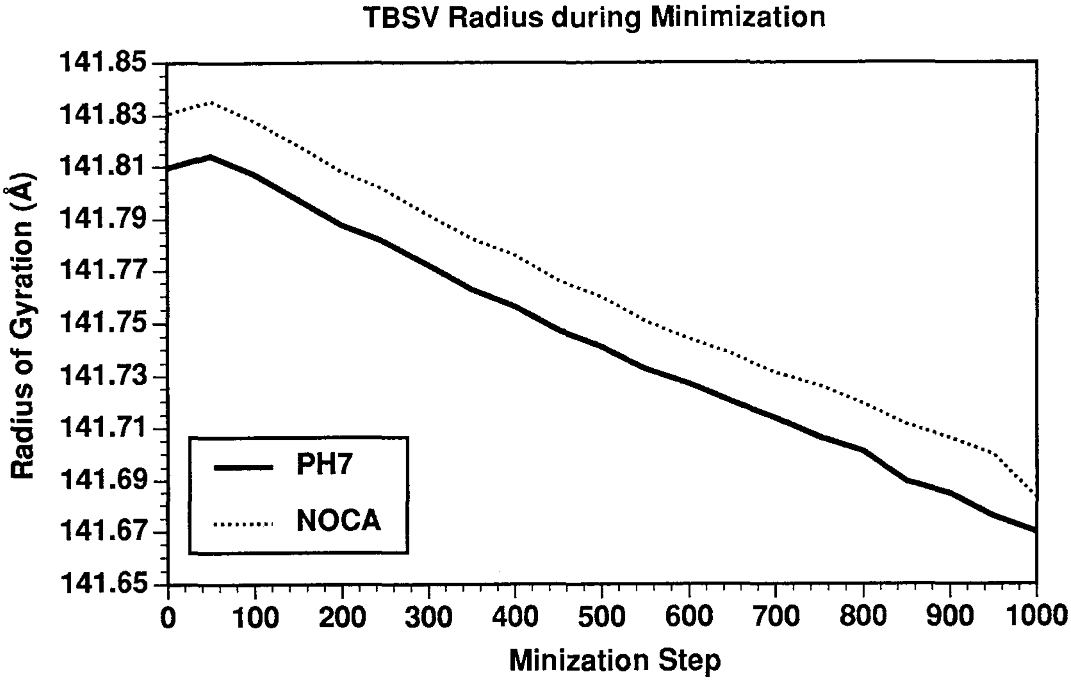


Figure 3.7. The radius of gyration during energy minimization.

Figure 3.7 for the first 1000 steps of minimization. PH7 and NOCA actually took 1300 and 1176 steps to minimize, respectively. Both structures contracted by about 0.08% during the minimization with their radii following almost identical curves. The contraction rate remained constant after the first 100 steps, indicating it would have continued had the minimization gone longer. However, curves of the potential energy during minimization, shown in Figure 3.8, clearly indicate that the energy had converged. The PH7 and NOCA energies also had similar curves. However, the PH7 structure is almost 700 kcal/mol lower in energy than NOCA, despite containing identical numbers of atoms. This energy difference is almost entirely due to the electrostatic energy term and indicates the large stabilizing energy of the Ca^{2+} ions.

Molecular dynamics calculations were carried out on the different minimized structures, again using nonbonds from the full capsid. The farfield was updated

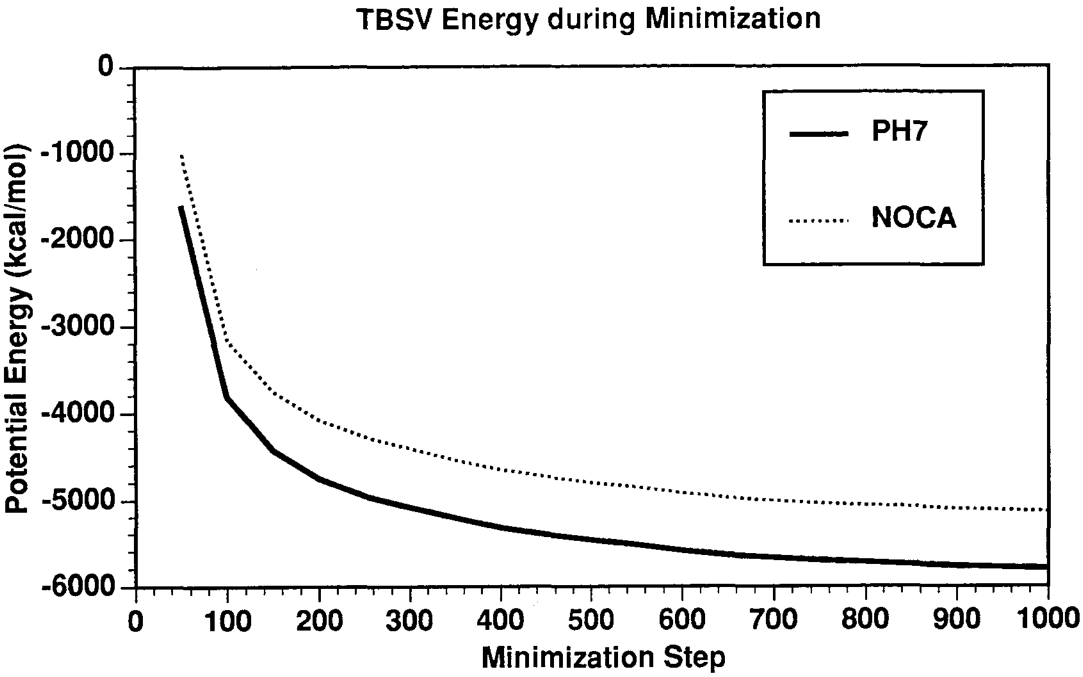


Figure 3.8. The potential energy of the TBSV triad, including CMM nonbonded interactions with the full viral coat, during energy minimization.

every 0.100 ps. Two different Cartesian dynamics methods were tried: microcanonical ensemble (NVE) with temperature scaling and canonical ensemble (NVT). In addition, NEIMO dynamics, which are NVE, were used with a timestep of 2 fs. Figure 3.9 shows the radius of gyration of PH7 during the first 2.0 ps of dynamics. All three methods give different curves but there are some similarities. Both Cartesian simulations have an initial expansion phase followed by a longer contraction. The NEIMO simulation has no expansion phase, but its contraction phase closely resembles that of the Cartesian NVE simulation, with roughly the same slope, contracting until approximately 1.8 ps, when it levels out. The canonical dynamics simulation, in contrast, shows no similar leveling out through the first 2.0 ps.

Longer simulations were run using Cartesian canonical dynamics and NEIMO dynamics on both the PH7 and NOCA models. The NEIMO dynamics simulations

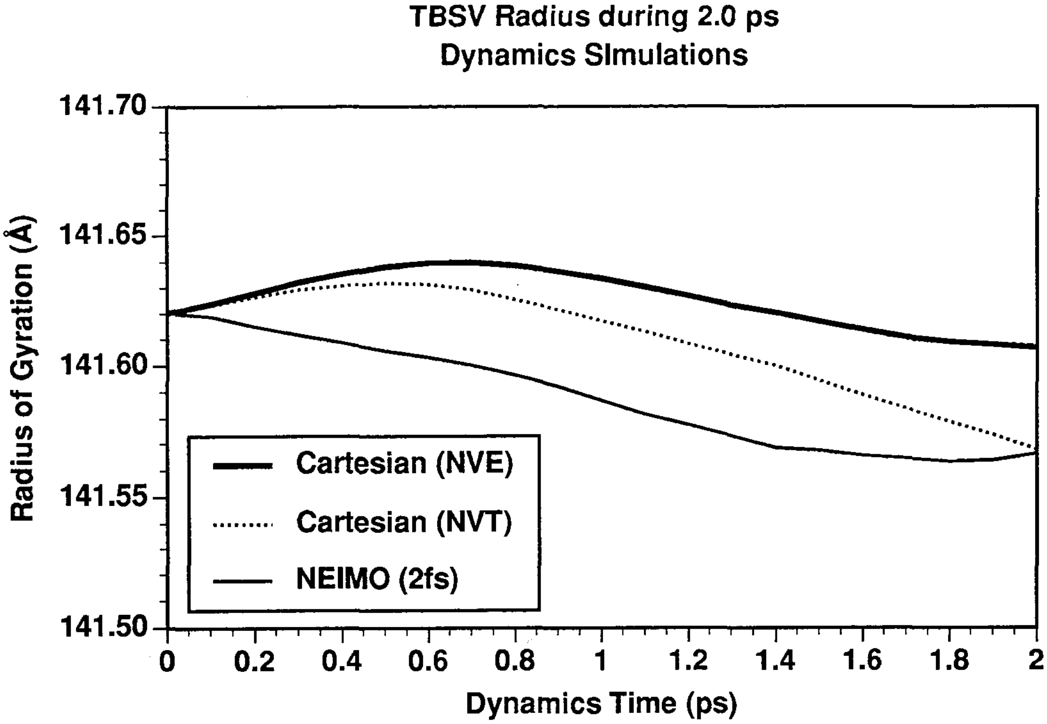


Figure 3.9. The TBSV radius during 2.0 ps Cartesian and NEIMO dynamics simulations.

were twice as fast, since 2 fs timesteps were used. Therefore, longer simulations could be run. Figure 3.10 shows the radius of gyration of the PH7 and NOCA models during the first 4.0 ps of NEIMO and Cartesian NVT simulations. In the NEIMO simulations, the NOCA model undergoes a rapid expansion during the first 2.0 ps, then an even sharper contraction. The PH7 model has no such expansion phase but does have a gradually increasing contraction rate. Both of these simulations show far more variation in their radius of gyration than the corresponding Cartesian simulations. However, in both simulations, the NOCA model initially has a larger radius of gyration than for PH7, but eventually becomes smaller. Although the curves of the radii cross after about 3.9 ps in the Cartesian simulation, the energy curves do not cross, as shown in Figure 3.11. The energy plot indicates that the NOCA model is less stable, as it is undergoes larger energy fluctuations after 3.0 ps. These fluctuations continue until the end of a 5.0 ps simulation (data not

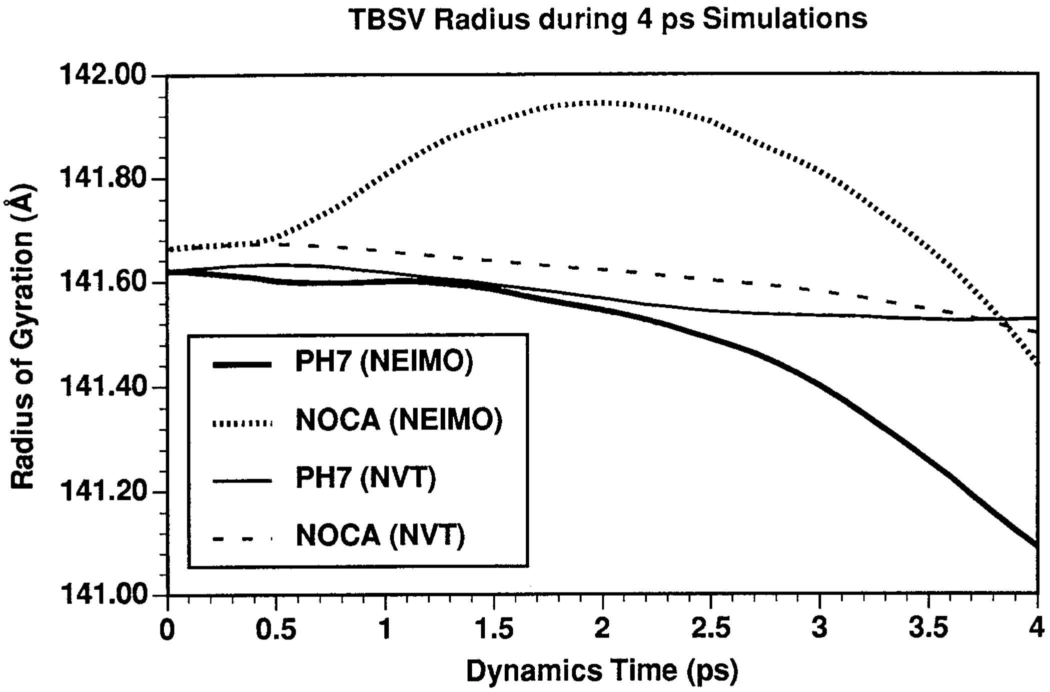


Figure 3.10. The radius of the PH7 and NOCA models of TBSV during 4.0 ps of Cartesian (NVT) and NEIMO dynamics.

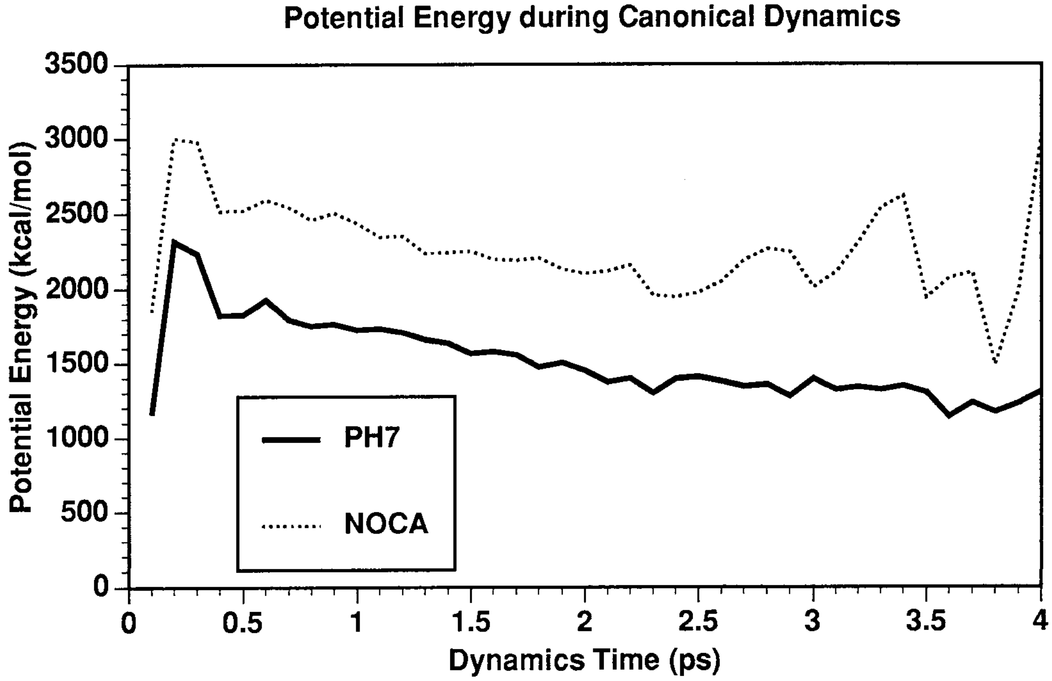


Figure 3.11. Potential energy during the 3.0 ps Cartesian canonical dynamics simulations.

shown). The PH7 model is relatively stable. For the NEIMO simulations, the contraction rate is comparatively exaggerated, but the energies do not show such large fluctuations. The NOCA model has a potential energy around -3850 kcal/mol while the PH7 model's potential remains near -4550 kcal/mol. Note that these energies are substantially lower than in the Cartesian simulation because the numerous bond and angle degrees of freedom are held at their minimum potential energy values. Therefore, the approximately 700 kcal/mol differential between PH7 and NOCA is relatively constant, even though the radii change significantly.

IV. Conclusions

The current simulations have not been able to reproduce the 10% expansion expected for the NOCA model on the basis of the experimental data[1]. However, the NOCA model is substantially higher in energy (700 kcal/mol), indicating that it might be driven to expand in more extensive calculations. The NEIMO simulations show substantial contraction of both the PH7 and NOCA models, but the nature of this phenomenon is unknown. The RNA in the interior of the virus may be needed to prevent contraction of the viral coat. Other models of TBSV should be investigated, including those in which all Ca^{2+} ions are removed, but are not replaced by Na^{+}.

References

[1] I.K. Robinson and S.C. Harrison, *Nature*, 297, 563-568 (1982).

[2] H.Q. Ding, N. Karasawa, and W.A. Goddard III, *J. Chem. Phys.*, 97, 4309-4315 (1992).

[3] S.C. Harrison *et al.*, *Nature*, 276, 368-373 (1978).

[4] A. Jain, N. Vaidehi, and G. Rodriguez, *J. Comp. Phys.*, submitted.

[5] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[6] BIOGRAF/POLYGRAF. Copyright by Molecular Simulations, Inc. (1992).

Chapter 4

Probability Grid Monte Carlo

Abstract

We have devised a Monte-Carlo method that employs importance sampling of dihedral angles to model peptide and protein conformations. This new method, which we call Probability Grid Monte Carlo (PGMC), modifies amino acid residue backbone and/or sidechain dihedrals according to probability grids derived from the Brookhaven Protein Database. We have used this method to study peptide conformations and have successfully adapted it to a number of important problems in protein modeling, including the prediction of all-atom protein conformations from C_α coordinates alone, and the prediction of the conformations of protein loops. Here, the method is applied to a study of the low energy conformations of the peptide Met-enkephalin.

I. Introduction

Probability Grid Monte Carlo (PGMC) is a method developed for predicting the conformations of peptides and proteins by searching their torsional degrees of freedom. The PGMC method combines two of the best features from other torsion-space conformational search methods which have been developed to study peptide conformations; Monte Carlo importance sampling and grid searching. Like the importance sampling method of Lambert and Scheraga[1], the method described here assigns probabilities to different ϕ, ψ conformations, and conformations are generated according to those probabilities, rather than completely at random or through an exhaustive search of all possibilities. However, unlike Lambert and Scheraga, our probabilities are derived to work within the framework of a grid search method, i.e., only discrete values are chosen for the dihedral angles. There are two primary advantages to using discrete values for dihedral angles, rather than sampling from a continuum: the conformational space is reduced to a finite number of possible conformations per dihedral angle and the probabilities can be generated to reflect known ϕ, ψ distributions more accurately. In addition, the method is easily extended to sidechain (χ) dihedrals. Because no functional form is necessary to specify the probabilities, grids can be developed for any necessary dimensionality. They range from one-dimensional grids for small sidechains to five-dimensional grids for arginine.

Grid searches have been employed in many conformational studies, such as those designed to predict protein loop structures[2] and those employed in the study of organic molecules[3]. The conformational space in a grid method is still large, as each dihedral can assume $360/S$ conformations, where S is the grid spacing. Therefore, these methods usually employ sophisticated schemes for eliminating combinations which cause steric overlap. The PGMC method, in contrast, implicitly includes a great deal of steric information through the use of probability grids: probabilities are assigned to different protein backbone (ϕ, ψ) and sidechain (χ) dihedrals according to their distributions in known protein structures. Conformations with significant steric overlap are not found in nature and, therefore, have extremely low probabilities of being sampled.

The Probability Grid Monte Carlo method has evolved into a general tool for protein modeling. Its conformational search methodology has been adapted to several problems in addition to the study of peptide conformations. The first of these is the prediction of all-atom conformations of proteins from C_α coordinates, discussed in Chapter 5. The second is the study of loop conformations in proteins, as applied to immunoglobulin hypervariable loops in Chapter 6. Both of these methods use the fundamental PGMC algorithms in conjunction with geometric constraints: the C_α coordinates or the loop endpoints, respectively. The results from both of these applications are encouraging: the C_α-based modeling gives results comparable to or better than other published methods, while the loop modeling is nearly as good as other methods, even though they employ surface-area corrections in order to choose more native-like conformations. Success in these modeling studies indicates that the method can be applied to cases where experimental information is lacking, such as the conformational states of small peptides.

II. Methodology

In the PGMC method, conformations of a polypeptide are generated by rotating backbone (ϕ, ψ) and/or sidechain (χ) dihedral angles for individual amino acids. These dihedral angles are shown in Figure 4.1 for arginine, which has the largest number of freely-rotating χ dihedrals. The conformations are not chosen randomly, but are selected from probability grids calculated from a selected subset of proteins from the Brookhave Protein Database. Each grid is an N_d-dimensional matrix, where N_d is the number of dihedrals involved. For instance, backbone sampling involves two-dimensional grids, and each point on the grid is the probability of chosing a particular ϕ, ψ pair. The grids have $S°$ spacing, where S = 5, 10, 15, 30, or 60. Therefore, ϕ, ψ grids have N_S points, where $N_S = (360/S) \times (360/S)$. The probabilities were derived by partitioning every ϕ, ψ pair in a set of high-resolution protein crystal structures into S-degree bins. The probabilities $(P(\phi, \psi))$ are normalized so that

$$\sum_{i=1}^{360/S} \sum_{j=1}^{360/S} P(\phi_i, \psi_j) = 1. \tag{4.1}$$

Sidechain probability grids have varying dimensionality, depending upon the number of dihedrals needed to specify the conformation. This ranges from $N_d = 1$ for small sidechains like serine and threonine, to $N_d = 5$ for arginine. For alanine and glycine, $N_d = 0$.

Our standard approach for doing Monte Carlo simulations uses these probability grids to generate trial conformations and assesses these conformations using the Metropolis criterion[4]. The protein is assigned an initial conformation, usually by rotating all backbone and sidechain dihedrals to the highest-probability grid values. A new conformation is generated by modifying one amino acid, which is chosen at random. Depending on the nature of the simulation, either a new main-chain conformation is chosen from the ϕ, ψ probability grid, or a new side-chain conformation is chosen from the χ probability grid. The potential energy of the new conformation $(i+1)$ is calculated, using a standard forcefield such as DREIDING[5] or AMBER[6] and is compared to the energy of the previous conformation (i). Conformation $i+1$ is either accepted or rejected according to the Metropolis criterion[4]. If the new structure is lower in energy ($\Delta E < 0$, where $\Delta E = E_{i+1} - Ei$), it is accepted. If it

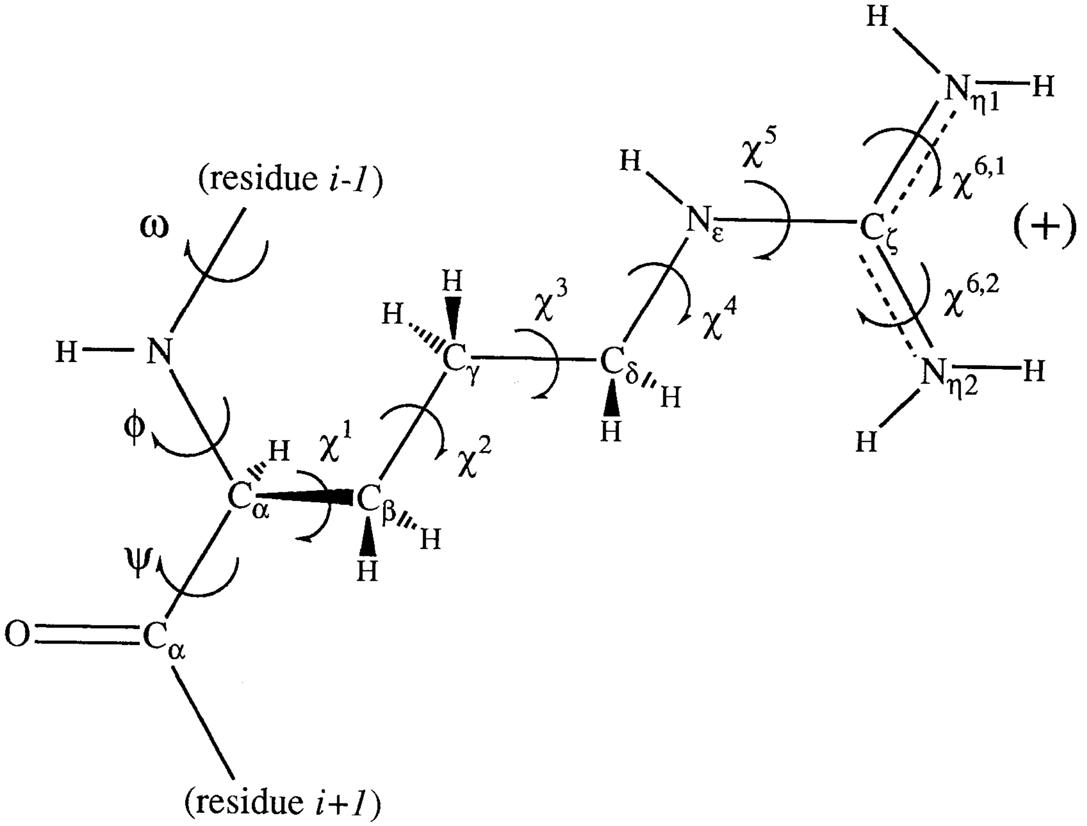


Figure 4.1. The backbone (ϕ, ψ, and ω) and sidechain (χ) dihedrals of arginine. The two outermost sidechain dihedrals, $\chi^{6,1}$ and $\chi^{6,2}$ rotate hydrogens, only, so they are not varied in the Probability Grid Monte Carlo method.

is higher in energy, the probability of acceptance, P, is defined by:

$$P = e^{-\Delta E/k_B T}, \tag{4.2}$$

where k_B is the Boltzmann constant and T is the simulation temperature. A random number n_r is generated between 0.0 and 1.0. If $P > n_r$, the new conformation is accepted. Otherwise, the structure is rejected and the previous structure is restored. This is the Metropolis criterion for accepting or rejecting a new structure, designed to ensure a Boltzmann distribution of conformations[4]. This criterion means, for example, that there is a 50% chance of accepting a new structure if $\Delta E = -k_B T \ln(0.5)$. At 300 K, this value is 0.413 kcal/mol. At higher temperatures, the probability of accepting a bad structure increases.

We are generally interested in finding the lowest-energy conformation of a given peptide. This can be done, theoretically[7], by the simulated annealing method: starting at a high simulation temperature, and slowly cooling the system in a process called simulated annealing. However, it is usually not possible to know beforehand exactly what cooling rate is necessary to achieve the global minimum. Considerable work has been done in optimizing the heating and cooling process in Monte Carlo simulations of peptides[8, 9]. We have employed both constant temperature and simulated annealing in our studies of protein conformations.

II.A. The Protein Database Subset

The dihedral probabilities which are integral to our method require a judicious choice of structural data. Therefore, we sought to use a subset of protein structures from the Brookhaven Protein Database (PDB) which was both diverse and accurate. The Brookhaven PDB contains more than 500 protein crystal structures, even excluding structures with only C_α coordinates. However, there are many proteins which are represented numerous times or are highly homologous to other proteins in the PDB dataset. Such identical, or nearly identical, structures would tend to distort our probabilities in favor of geometries found in those particular proteins. In order to eliminate highly redundant structures, we carried out pairwise sequence comparisons among 503 proteins in our initial PDB dataset, using the "align" program from W.R. Pearson's FASTA sequence analysis package[10]. Any protein with greater than 25% sequence identity with another protein of higher resolution was eliminated. This homology-elimination process reduced our dataset from 503 proteins to 121. This dataset of 121 proteins, which we call U121, is useful for a wide variety of statistical analyses. However, geometric analyses such as those required here require high resolution data, so we further reduced the dataset to 64 crystal structures which had 1.5 Å resolution data or better, or had better than 2.0 Å resolution and R-factors below than 20%. This dataset, which we call H64, was used to create our probability grids. The 64 crystal structures comprising this dataset are listed in Table 4.1.

II.B. Backbone (ϕ, ψ) Probability Grids

During a Monte Carlo step, either the backbone or side-chain conformation of one amino acid residue, selected at random, is altered. If the backbone conformation is to be changed, a new ϕ, ψ pair is selected for the residue. The ϕ, ψ pair is chosen from a grid of probabilities where the spacing between the gridpoints is S°. The grid, therefore, contains N_S gridpoints, where $N_S = (360/S) \times (360/S)$. The third backbone dihedral angle, ω, is fixed at 180° during Monte Carlo simulations, except where it occurs before proline residues. For prolines, there is a 7% chance of flipping to the *cis* conformation ($\omega = 0°$). However, even for proline the ω is treated independently and not as a third-dimension in the probability grid.

PDB	Res.(Å)	R	PDB	Res.(Å)	R	PDB	Res. (Å)	R
1AMT	1.5	.155	1UBQ	1.8	.176	3BLM	2.0	.163
1BP2	1.7	.171	1UTG	1.34	.23	3CLA	1.75	.157
1CRN	1.5	N.A.	1XY1	1.04	.088	3DFR	1.7	.152
1CSC	1.7	.188	256B	1.4	.164	3GRS	1.54	.186
1CSE	1.2	.178	2AZA	1.8	.157	3RNT	1.8	.137
1CTF	1.7	.174	2CA2	1.9	.176	451C	1.6	.187
1ECA	1.4	N.A.	2CCY	1.67	.188	4CPV	1.5	.215
1FB4	1.9	.189	2CDV	1.8	.176	4FD1	1.9	.192
1GD1	1.8	.177	2CPP	1.63	.19	4FXN	1.8	.200
1GMA	0.86	.071	2CYP	1.7	.202	4INS	1.5	.153
1GP1	2.0	.171	2ER7	1.6	.142	4PTP	1.34	.171
1HOE	2.0	.199	2GBP	1.9	.146	5CPA	1.54	N.A.
1I1B	2.0	.189	2LTN	1.7	.177	5CYT	1.5	.171
1L19	1.5	.153	2MHR	1.7	.158	5PTI	1.0	.200
1LZ1	1.5	.177	2MLT	2.0	.198	5RXN	1.20	.115
1MBA	1.6	.193	2OVO	1.5	.199	5TNC	2.0	.155
1MBD	1.4	N.A.	2RSP	2.0	.144	6TMN	1.6	.171
1NXB	1.38	N.A.	2SGA	1.5	.126	7RSA	1.26	.15
1PAZ	1.55	.18	2SNS	1.5	N.A.	9PAP	1.65	.161
1PCY	1.6	.17	2WRP	1.65	.180	9WGA	1.8	.175
1PPT	1.37	N.A.	3B5C	1.5	.16			
1THB	1.5	.196	3BCL	1.9	.189			

Table 4.1. Crystal structures used in the H64 dataset.

The probability grids were determined by partitioning every ϕ, ψ pair in the proteins comprising the H64 dataset into bins of size $S^\circ \times S^\circ$ and normalizing. We have determined separate probability grids for each amino acid, but it is sufficient to use individual grids for the three major residue types: glycine, which has no sidechain, proline, whose sidechain forms a closed loop with the backbone, and the other 18 "standard" amino acids. The ϕ, ψ probabilities are significantly different for these three residue types, as can be seen in Figure 4.2. The shape of the grid depends not only on the data, but on the grid spacing, S, as can be seen in Figure 4.3. A narrower spacing allows for much greater conformational flexibility, which is especially important in simulations of constrained systems. However, the total coverage of conformational space is somewhat reduced for narrower grid spacings. For instance, for standard residues, 110 of the 144 possible 30° gridpoints are populated (76.4%), while only 1114 out of 5184 gridpoints (21.5%) are populated on a 5° grid. Of course, the number of populated gridpoints, and their probabilities, depends on the size and quality of the dataset. Therefore, in order to evaluate the grids produced from the H64 dataset, we have also constructed grids using the U121 dataset.

The number of each type of residue found in the two datasets is shown in Table 4.2. The U121 dataset contains nearly three times as many residues as the H64 dataset. Although it is advantageous to have a larger sample size when doing statistical analyses, this advantage is mitigated for the U121 dataset because of the inclusion of low-quality structures. This problem is made clear in Table 4.3, where the number of non-zero gridpoints is listed for the three residue types at various grid spacings. The inclusion of data from all structures in the U121 dataset greatly increases the number of gridpoints which are populated. This is the case for all three residue types at all five spacing levels, but is particularly notable at grids spacings of 15° and less. Clearly, far more areas of ϕ, ψ conformational space have at least

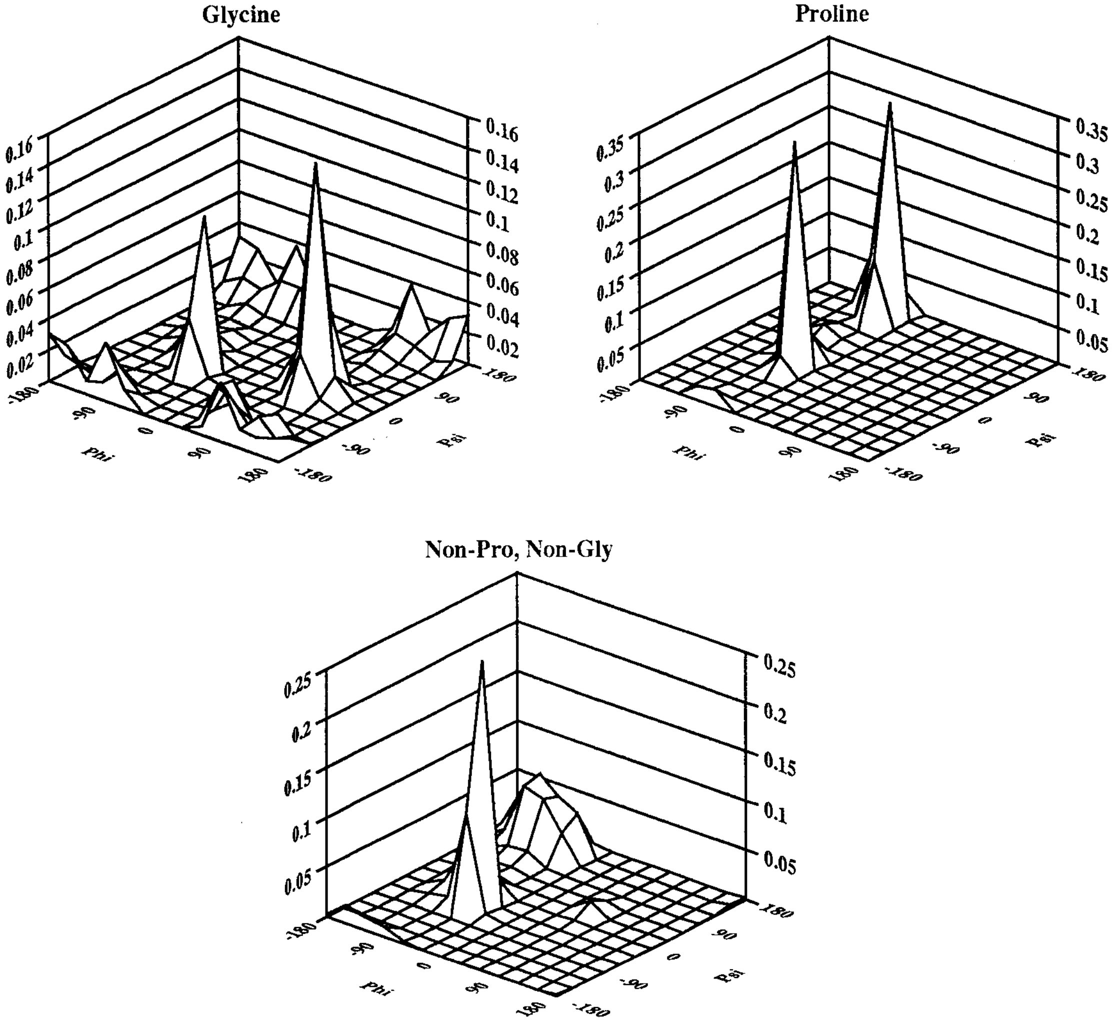


Figure 4.2. 30° ϕ, ψ grids for the three standard residue types.

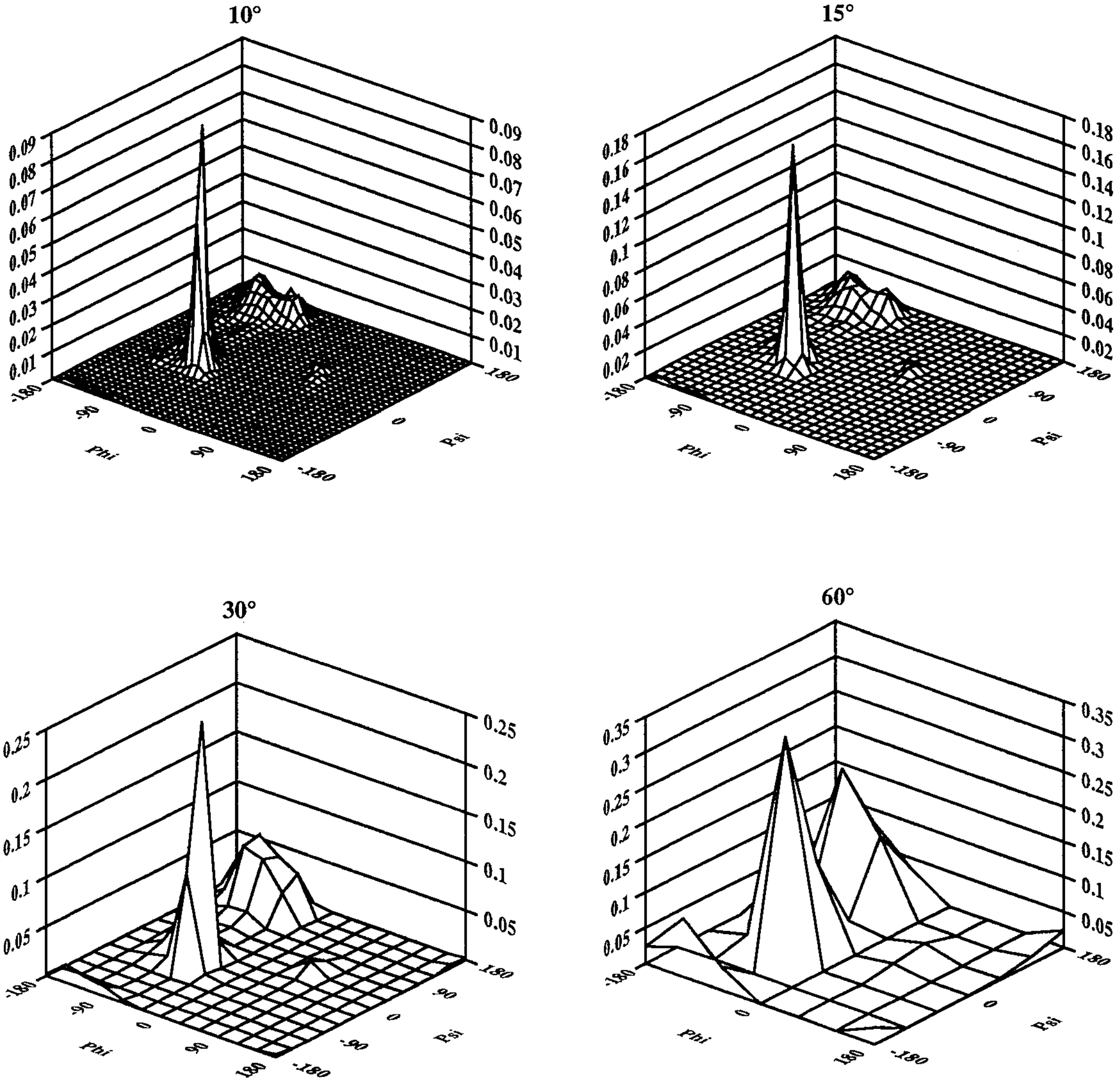


Figure 4.3. ϕ, ψ grids for standard (non-Proline, non-Glycine) residues at grid spacings of 10, 15, 30, and 60°.

	Residue Type		
Dataset	Standard	Glycine	Proline
H64	11052	1076	562
U121	33722	2787	1581

Table 4.2. The number of ϕ, ψ samples of the three residue types in the protein datasets.

one representative in the U121 dataset. However, it is difficult to say whether this is due to the larger sample size or reflects the fact that low-resolution structures are included in the U121 data. Unusual conformations in these low-resolution structures may be due to poor crystallographic data and might even be a cause of bad fits to data (high R-factors). A more interesting analysis is the number of high-probability gridpoints ($P(\phi, \psi) > \langle P \rangle$), as shown in Table 4.4. Because of the large number of gridpoints with $P(\phi, \psi) = 0$, the percentage having $P(\phi, \psi) > \langle P \rangle$ is substantially less than 50%. This number is very consistent across different grid spacings and is far more consistent between the datasets. This indicates that the U121 dataset has a large number of very rare ϕ, ψ conformations, and it should not be detrimental to exclude them from the probability grids used for our simulations. This is especially true for the standard residues and for the larger grid spacings of glycine and proline. For the ultrafine 5° grids, there clearly is insufficient data for proline and glycine conformations. The sample sizes for glycine and proline are less than the number of 5° gridpoints, so every nonzero gridpoint automatically has $P(\phi, \psi)$ greater than $\langle P \rangle$. This problem is particularly acute for the H64 dataset, where the percentage of high-probability conformations drops off dramatically at 5°. This dataset is probably inadequate for glycine and proline conformation sampling at a 5° resolution.

Non-zero Gridpoints

S	N_S	DS	Standard		Glycine		Proline	
60°	36	H64	33	91.7%	30	83.3%	10	27.8%
30°	144	H64	110	76.4%	83	57.6%	24	16.7%
15°	576	H64	253	43.9%	198	34.4%	48	8.3%
10°	1296	H64	429	33.1%	312	24.1%	76	5.9%
5°	5184	H64	1114	21.5%	593	11.4%	164	3.2%
60°	36	U121	36	100.0%	36	100.0%	18	50.0%
30°	144	U121	143	99.3%	131	91.0%	51	35.4%
15°	576	U121	519	90.1%	376	65.3%	118	20.5%
10°	1296	U121	1004	77.5%	627	48.4%	193	14.9%
5°	5184	U121	2400	46.3%	1280	24.7%	420	8.1%

Table 4.3. This table lists the number of (and percentage of the maximum possible) ϕ, ψ gridpoints which have non-zero values for each of the three residue types at different grid spacings, S.

High-probability Gridpoints

S	$\langle P\rangle$	DS	Standard		Glycine		Proline	
60°	2.778%	H64	7	19.4%	12	33.3%	4	11.1%
30°	0.694%	H64	23	16.0%	33	22.2%	11	7.6%
15°	0.174%	H64	78	13.5%	127	22.0%	48	8.3%
10°	0.077%	H64	172	13.3%	312	24.1%	76	5.9%
5°	0.019%	H64	630	12.2%	593	11.4%	164	3.2%
60°	2.778%	U121	7	19.4%	12	33.3%	4	11.1%
30°	0.694%	U121	25	17.4%	37	25.7%	18	12.5%
15°	0.174%	U121	89	15.5%	144	25.0%	63	10.9%
10°	0.077%	U121	194	15.0%	282	21.8%	121	9.3%
5°	0.019%	U121	788	15.2%	1280	24.7%	420	8.1%

Table 4.4. This table lists the number of high-probability gridpoints: the number which have probabilities above $\langle P\rangle$.

Table 4.4 confirms what can be seen in Figure 4.2: the grids are substantially different for the three residue types. Glycine is clearly more flexible, having a much larger number of high-probability conformations. Proline, in contrast, is far less flexible. There are far fewer high-probability conformations for proline, as would be expected from geometrical considerations. The closed ring formed by its backbone and sidechain severely restrict the ϕ angle to angles near -60°. The highest probability peak for each type of residue is shown in Table 4.5. For standard residues, the alpha-helical peak predominates. For every spacing level, the alpha helical conformation is the highest peak, even though the probability of picking the peak gridpoint decreases as the total number of gridpoints increases. The intra-strand hydrogen bonding of alpha-helices greatly favors conformations near ($\phi = -57, \psi = -47$). Therefore, the peak is very sharp, as becomes increasingly clear for the finer grids in Figure 4.3. In contrast, the beta sheet region of the ϕ, ψ grid, centered about ($\phi = -115, \psi = 130$), is much broader. No individual gridpoint in the beta sheet region is as high as the alpha helical peak, even though the beta sheet quadrant (I) has nearly the same overall probability as the alpha helix quadrant (II) (47.8% vs. 49.4% - see Table 4.6). Proline grids have two sharp peaks, as is seen for 30° in Figure 4.2. The two peaks are so similar that the identity of the highest peak depends on both the grid spacing and the dataset. There is little probability of proline conformations outside of the two peak regions; there is almost no chance that the conformation is in quadrant III or IV. The opposite is true for the third major residue type, glycine. Glycine's great flexibility is clearly seen in Table 4.6. The four quadrants are almost equally populated, since there is no sidechain to sterically hinder quadrant III and IV conformations. Because of this flexibility, no single peak has a particularly high probability (Table 4.5).

We have also used the secondary structure designators in the protein database

Highest-probability Gridpoints

S	DS	Standard		Glycine		Proline	
60°	H64	-60,-60	33.6%	60,0	17.2%	-60,120	30.2%
30°	H64	-60,-30	26.1%	90,0	14.2%	-60,-30	33.8%
15°	H64	-60,-45	17.6%	-60,-45	7.1%	-60,-30	15.1%
10°	H64	-60,-40	9.8%	-60,-40	3.7%	-60,-30	8.2%
5°	H64	-60,-45	3.3%	90,-5	1.5%	-60,-35	3.2%
60°	U121	-60,-60	29.8%	-60,-60	13.4%	-60,120	31.0%
30°	U121	-60,-30	20.1%	90,0	9.4%	-60,150	24.2%
15°	U121	-60,-45	12.5%	-60,-45	5.9%	-60,-30	9.6%
10°	U121	-60,-40	6.4%	-60,-40	2.7%	-60,-30	5.1%
5°	U121	-60,-45	2.2 %	-60,-45	1.0%	-60,-35	1.9%

Table 4.5. This highest-probability gridpoint of each residue type for each grid spacing.

Quadrant Populations

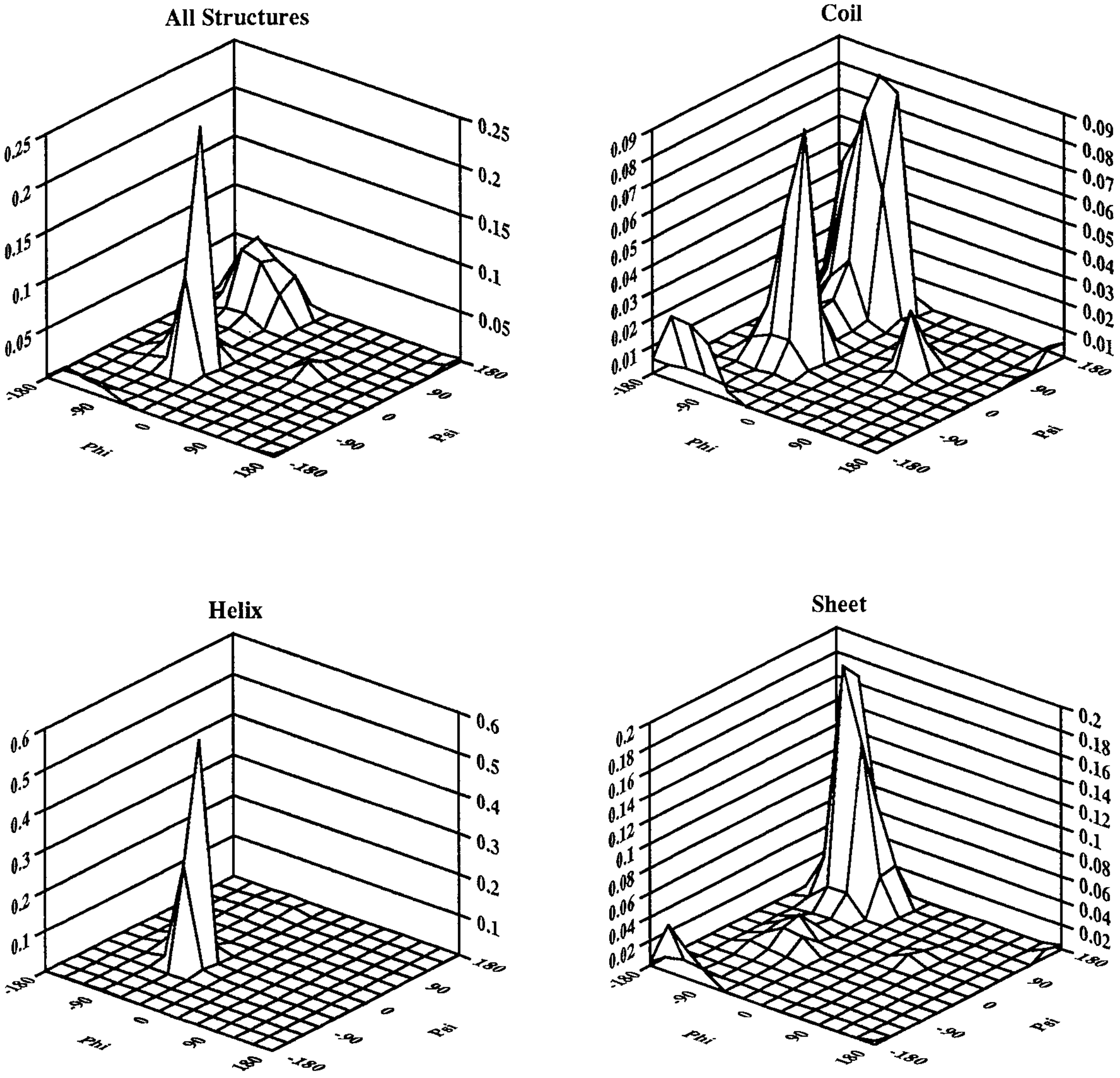
Quadrant	ϕ	ψ	DS	Standard	Glycine	Proline
I	< 0	> 0	H64	47.8%	14.8%	54.4%
II	< 0	< 0	H64	49.4%	29.9%	45.6%
III	> 0	> 0	H64	2.4%	29.1%	0.0%
IV	> 0	< 0	H64	0.4%	26.1%	0.0%
I			U121	49.0%	19.7%	56.5%
II			U121	46.8%	29.0%	42.9%
III			U121	3.0%	27.6%	0.3%
IV			U121	1.1%	23.6%	0.2%

Table 4.6. The percentage of sample ϕ, ψ's falling within each quadrant of ϕ, ψ conformational space.

(HELIX, SHEET, and TURN) to obtain separate probability grids for alpha helix, beta sheet, and coil regions. We decided not to create grids for beta turn residues because the four residues involved in a turn usually have completely different ϕ, ψ conformations and it would be counterproductive to treat them identically. Presumably, eight-dimensional probability grids generated for sequences of four consecutive ϕ, ψ pairs would have peaks for particular turn conformations as well, but the total number of turns in our set of crystal structures is tiny compared to the immense number of gridpoints on an eight-dimensional grid. Such grids would have little advantage over a method which simply tries all known turn configurations. We do have separate probability grids for coil residues, however. We define coil residues as all those not involved in any of the three major secondary structure types. Six proteins in the H64 database had no HELIX, SHEET, or TURN designators, and we excluded these from secondary structure analyses. We did not want to assume a complete lack of secondary structural elements for these proteins. The remaining 58 proteins with secondary structure designators comprise the SS58 dataset, which we used to create the probability grids shown in Figure 4.4. Table 4.7 lists the total number of samples of each residue type for each structural class. While the coil population is large for all residue types, it is particularly high for proline residues. The backbone nitrogen of proline is bonded to the C_δ of the sidechain, so it is not available for hydrogen bond formation. Prolines therefore cannot participate in the hydrogen bonds which stabilize α helices, β sheets, and turns. The coil grid in Figure 4.4 contains significant probabilities for both α helix and β sheet conformations, but the probabilities are much lower than those in the "all-structures" grid. Presumably, residues in the coil regions are not participating in the extended hydrogen-bonding networks or involved in the large-scale dipole-dipole interactions of α helices and β sheets. Therefore, the coil probability grids are more indicative of the inherent conformational energies of individual residues and, therefore, are the grids which most closely resemble classic Ramachandran plots[11] and ϕ, ψ potential energy maps[12]. These secondary structure-specific grids are useful only when the secondary structure is known beforehand. This is not the case for an *ab initio* prediction of protein conformation, but is for simulations used in conjunction with C_α coordinates, homology modeling, or secondary structure prediction algorithms.

II.C. Sidechain (χ) Probability Grids

While every amino acid backbone can be specified by the same three dihedral angles, ϕ, ψ and ω, there is a far greater diversity among sidechain dihedrals, χ. At the extremes are glycine, which has no sidechain, and tryptophan, which has 12 χ dihedral angles. Our simulations do not modify dihedral angles which affect only hydrogen positions, or those involved in rings, so the number of dihedrals is significantly re-



All Structures

Coil

Helix

Sheet

Figure 4.4. ϕ, ψ grids of different structural types for standard (non-Proline, non-Glycine) residues in the SS58 dataset.

Secondary structure distribution in the SS58 dataset

Structure	Standard		Glycine		Proline	
Total	11061	100.0%	1076	100.0%	562	100.0%
α-Helix	4255	38.5%	245	22.8%	129	23.0%
β-Sheet	2279	20.6%	167	15.5%	61	10.9%
Turn	1912	17.3%	300	27.9%	153	27.2%
Coil	2615	23.6%	364	33.8%	219	39.0%

Table 4.7. The distribution of secondary structure designations in the crystal structures of the SS58 dataset.

duced. Both alanine and glycine have zero PGMC sidechain dihedrals ($N_\chi = 0$), while tryptophan, tyrosine, phenylalanine and histidine have only two, despite being very large sidechains. The values of N_χ for the common amino acids, excluding alanine and glycine, are given in Table 4.8. Although proline is a ring, we allow χ^1 to vary while holding the C_δ atom fixed. This enables reasonable conformations of χ^1 - χ^4 to be sampled by modifying only a single dihedral, χ^1.

Table 4.8 also lists the number of occurrences of each amino acid in the H64 dataset as well as the number of populated (non-zero) gridpoints at each spacing level. These numbers can be compared to the total number of gridpoints at each spacing level, listed in Table 4.9. It is clear that there is insufficient data for the multidimensional grids ($N_d \geq 3$) at the fine spacings. For these cases, the number of populated gridpoints approaches the sample size. In other words, almost every conformation occupies a different gridpoint and the probability grid is extremely flat. This extreme variability is due primarily to the enormous number of possible conformations available for these structures (Table 4.9), rather than unusual flexibility in these particular sidechains. The χ^1,χ^2 distributions of these residues makes this

			Populated Gridpoints					
Amino Acid	N_χ	Samples	120°	60°	30°	15°	10°	5°
CYS	1	283	3	4	10	15	21	34
PRO	1	568	2	3	5	9	13	22
SER	1	925	3	6	12	24	35	70
THR	1	791	3	6	12	23	32	54
VAL	1	991	3	6	12	23	30	51
ASN	2	634	9	28	82	198	282	465
ASP	2	728	9	31	84	200	296	485
HIS	2	317	9	27	66	125	170	253
ILE	2	603	8	23	56	89	134	238
LEU	2	1025	9	27	67	135	191	343
PHE	2	491	8	23	51	119	175	318
TRP	2	179	8	19	39	73	98	141
TYR	2	453	9	22	52	107	172	294
GLU	3	699	26	116	200	528	528	*688*
GLN	3	409	24	83	312	331	331	*404*
MET	3	241	20	54	120	185	218	*240*
LYS	4	858	67	288	580	775	*834*	*858*
ARG	5	438	116	195	322	*421*	*429*	*436*

Table 4.8. The number of PGMC χ dihedrals (N_χ) for each amino acid, the number of occurrences of each amino in the H64 crystal structures, and the number of populated (non-zero) gridpoints at each grid spacing. $N_\chi = 0$ for alanine and glycine. The numbers in italics are more than 95% of the sample size, indicating that nearly every conformation occupies a different gridpoint.

Possible conformations

N_χ	120°	60°	30°	15°	10°	5°
1	3	6	12	24	36	72
2	9	36	144	576	1296	5184
3	27	216	1728	13824	46656	3.7×10^5
4	81	1296	20736	3.3×10^5	1.7×10^6	2.7×10^7
5	243	7776	2.5×10^5	8.0×10^6	6.0×10^7	1.9×10^9

Table 4.9. The total number of gridpoints for N_χ dihedrals at different grid spacings.

more clear (see Table 4.10): only lysine has an unusually large number of populated conformations when only χ^1 and χ^2 are considered.

One- and two-dimensional χ probability grids are shown in Figures 4.5 and 4.6, respectively. It is not possible to show the higher dimensional grids in their entirety. These figures make it clear that there is a great deal of variety even among residues with the same number of significant χ dihedrals. One caveat about these grids: some of the χ's actually have a period of 180°, rather than 360° as shown. This arises when two branches are the same, as in Asp, where the two carboxylate oxygens are chemically identical but only one is labeled O$_\delta$1 and is used to specify χ^2. This labeling is not always done the same way, hence there are separate peaks at 150°and -30°. Note that this does not affect our Monte Carlo simulations, since the orientations will be simulated identically with a total probability equal to the sum of the individual probabilities. The great variety in sidechain conformations can also be seen in Table 4.11, which lists the highest probability sidechain conformation for each amino acid.

χ^1,χ^2 Distribution

		120°	60°	30°	15°	10°	5°
ASN	634	9	28	82	198	282	465
ASP	728	9	31	84	200	296	485
HIS	317	9	27	66	125	170	253
ILE	603	8	23	56	89	134	238
LEU	1025	9	27	67	135	191	343
PHE	491	8	23	51	119	175	318
TRP	179	8	19	39	73	98	141
TYR	453	9	22	52	107	172	294
GLU	699	9	31	86	192	267	435
GLN	409	9	27	67	125	175	275
MET	241	8	18	48	79	118	174
LYS	858	9	33	98	229	326	515
ARG	438	9	27	60	128	178	275

Table 4.10. The number of populated gridpoints in the χ^1,χ^2 distribution of amino acids with $N_\chi \geq 2$. Lysine appears to be the most flexible while tryptophan is the least flexible.

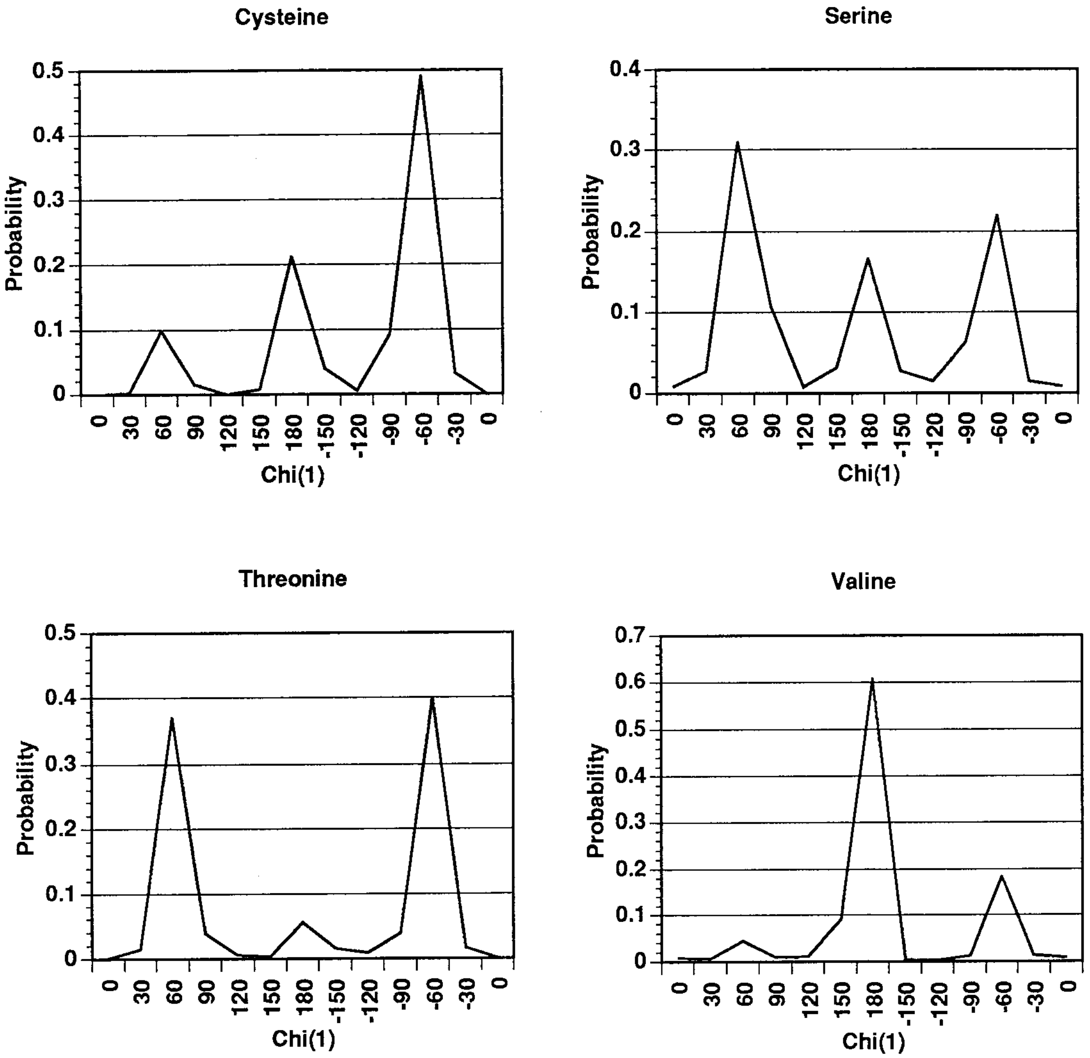


Figure 4.5. χ grids for sidechains with one PGMC dihedral.

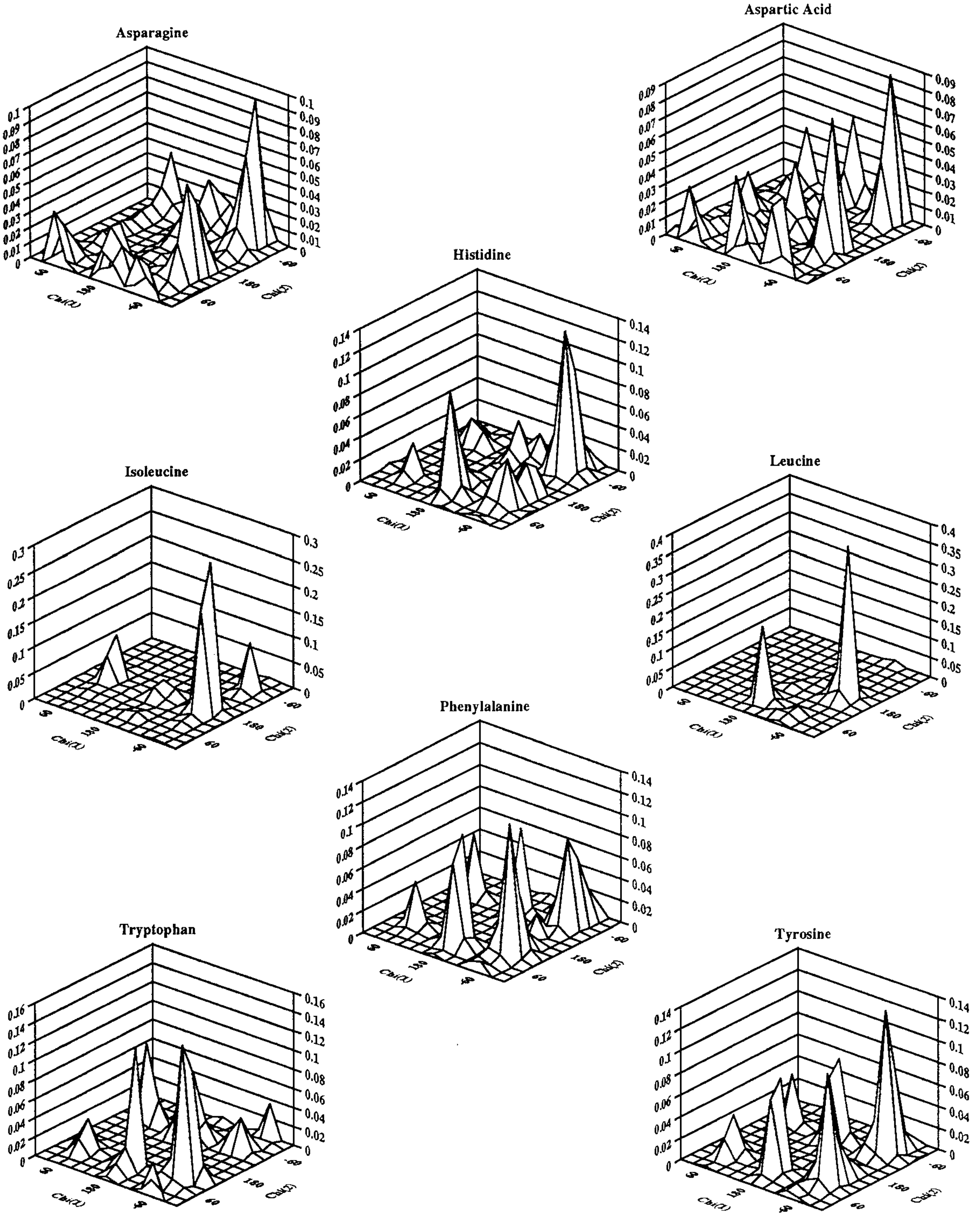


Figure 4.6. χ grids for sidechains with two significant dihedrals.

II.D. Note on Computations

The Probability Grid Monte Carlo programs are interfaced to the BIOGRAF program from Molecular Simulations, Incorporated[13]. Basic functions such as graphical displays and energy evaluations are performed by the routines provided by the commercial version of BIOGRAF. All calculations reported here were performed on Silicon Graphics workstations, primarily on a model 4D/380 VGX.

III. Conformations of Met-Enkephalin

Small peptides are an important class of molecules to study with computational methods because it is extremely difficult to obtain experimental information about them. Polypeptides of less than 50 residues in length rarely assume a single conformation in solution[14]. This flexibility makes them virtually impossible to characterize structurally with either X-ray crystallography or multidimensional NMR[15]. Even if a crystal or solution structure could be determined, its relevance would be questionable since these molecules usually act by binding to a receptor, and it is their bound conformation which is of utmost importance. There is evidence that the bound conformations of such peptides are different from their solution structures[16]. Computational studies of such molecules are extremely important, therefore, because they may produce not only a prediction for the global minimum energy conformation but an ensemble of low energy conformations which may include the receptor-bound conformation.

The small neuropeptide Met-enkephalin has become a standard test case for peptide structure prediction programs. This pentamer, with sequence Tyr-Gly-Gly-Phe-Met, has been studied by a number of groups using different conformational search methodologies [9, 17, 18, 19, 20]. Li and Scheraga believe they have found

	P_{max}	χ^1	χ^2	χ^3	χ^4	χ^5
CYS	49.1%	-60				
PRO	38.2%	0				
SER	31.0%	60				
THR	42.4%	-60				
VAL	60.6%	180				
ASN	9.5%	-60	-30			
ASP	9.3%	-60	-30			
HIS	13.2%	-60	-90			
ILE	31.5%	-60	180			
LEU	38.8%	-60	180			
PHE	12.6%	-60	90			
TRP	12.3%	-60	90			
TYR	13.9%	-60	-90			
GLU	3.8%	-60	180	150		
GLN	2.9%	-60	180	-30		
MET	5.8%	-60	-60	-60		
LYS	4.7%	-60	180	180	180	
ARG	2.1%	-60	180	180	-150	0

Table 4.11. The highest-probability gridpoint for each amino acid for $S = 30°$. P_{max} is the probability of this particular conformation.

the global minimum [17] of Met-enkephalin, and this claim is supported in the work of von Freyberg and Braun[9], who found the same minimum-energy conformation several times during their simulations. It is important to note that this conformation is merely the lowest found so far using the ECEPP/2 forcefield[21]. It is not computationally possible, currently, to evaluate all possible conformations so it remains a possibility that lower-energy conformations do exist even using the ECEPP/2 forcefield. It is also very possible that different forcefields have different global minima. The ECEPP/2 global minimum is by no means guaranteed to be the DREIDING or AMBER global minimum.

The purpose of our simulations is to determine low-energy conformations of Met-enkephalin using the DREIDING forcefield. To this end, we have carried out numerous simulations on Met-enkephalin using the Probability Grid Monte Carlo method, at different grid spacings and different temperatures. These studies aim to answer the following questions:

1. What is the lowest-energy conformation of Met-enkephalin using the DREIDING forcefield?

2. How does the ECEPP/2 minimum-energy conformation compare to the DREIDING minima?

3. How well does the PGMC method work for conformational searching of peptides?

4. What parameters, such as grid spacing and temperature, produce the best results?

The simplest calculation using our probability grids is to build Met-enkephalin using standard geometries, and then rotate each ϕ, ψ pair and each sidechain to its peak conformation according to the probability grids. We used the "Peptide Builder" function of BIOGRAF to build an initial structure from a "Peptide Library" of amino acid structures. This starting structure was then modified by rotating the backbone and sidechain dihedrals to their highest-probability conformations. This process was repeated for each grid spacing, and the energy of each conformation was calculated using BIOGRAF and the DREIDING forcefield[5]. The results are shown in Table 4.12. We call these conformations "S°_i", where S is the grid spacing. Except for the 5°-peak conformation, 5°_i, the energy decreases as the grid spacing becomes finer – i.e., finer grids give better conformations. The energy of 5°_i is extremely high because of steric overlap between the sidechain of Tyr 1 and the backbone of Met 5. This overlap is caused by both the tyrosine sidechain conformation, and the fact the the glycines have $\phi = 90^\circ, \psi = -5^\circ$ conformations, thereby forming a turn in the backbone. The 30°-peak conformation, 30°_i has a very similar backbone conformation, but its sidechain conformations, most importantly that of Tyr 1, are completely different, so it avoids the steric overlap problem. The other conformations have completely different backbone conformations than 5°_i and 30°_i. The conformations 10°_i and 15°_i are almost identical α-helices.

We used the conjugate-gradients minimizer of BIOGRAF to minimize these conformations in order to find a stable local minimum in the DREIDING potential energy surface. The dihedral angles of the minimized conformations, S°_m, were no longer restricted to lattice values, and the bonds and angles were no longer fixed. The resulting structures were significantly different and had far lower energies than the un-minimized conformations, as can be seen in Table 4.13. The minimized conformations differ from their unminimized counterparts by 1.4 Å or more. Except for the 5°_i - 5°_m transition, minimization primarily modifies the backbone dihedrals (see Table 4.13). Minimization of 5°_i, however, substantially modifies the sidechain

Met-Enkephalin Conformations

Dihedral			60°_{i}	30°_{i}	15°_{i}	10°_{i}	5°_{i}
Tyr	1	ϕ	-60.0	-60.0	-60.0	-60.0	-60.0
		ψ	-60.0	-30.0	-60.0	-40.0	-45.0
		χ^1	-60.0	-60.0	-60.0	-70.0	180.0
		χ^2	-60.0	-90.0	-90.0	-90.0	90.0
Gly	2	ω	180.0	180.0	180.0	180.0	180.0
		ϕ	60.0	90.0	-60.0	-60.0	90.0
		ψ	0.0	0.0	-45.0	-40.0	-5.0
Gly	3	ω	180.0	180.0	180.0	180.0	180.0
		ϕ	60.0	90.0	-60.0	-60.0	90.0
		ψ	0.0	0.0	-45.0	-40.0	-5.0
Phe	4	ω	180.0	180.0	180.0	180.0	180.0
		ϕ	-60.0	-60.0	-60.0	-60.0	-60.0
		ψ	-60.0	-30.0	-45.0	-40.0	-45.0
		χ^1	180.0	-60.0	-60.0	-60.0	180.0
		χ^2	60.0	90.0	105.0	100.0	85.0
Met	5	ω	180.0	180.0	180.0	180.0	180.0
		ϕ	-60.0	-60.0	-60.0	-60.0	-60.0
		ψ	-60.0	-30.0	-45.0	-40.0	-45.0
		χ^1	-60.0	-60.0	-60.0	-60.0	-70.0
		χ^2	-60.0	-60.0	-180.0	-170.0	-175.0
		χ^3	-60.0	-60.0	75.0	70.0	-60.0
Energy			260.8	215.1	183.6	180.0	***

Table 4.12. Conformations generated for Met-enkephalin from peak ϕ, ψ, and χ gridpoints at different grid spacings. Because of steric overlap, the energy of the 5° conformation is greater than 1×10^6 kcal/mol.

dihedrals, especially χ^1 and χ^3 of Met 5. This removes the steric overlap and improves the energy of the 5° conformation to a value similar to the other minimized conformations. Interestingly, 60°_m and 30°_m are the best minimized conformations, even though 60°_i and 30°_i were substantially higher in energy than 15°_i and 10°_i. The turn-like conformations of the 30° and 60° are more compact than the helical 10° and 15° conformations. This compactness caused unfavorable overlap initially, but upon minimization led to favorable van der Waals packing as well and electrostatic interactions between the N- and C- termini.

It is interesting to compare these conformations to the ECEPP/2 global minimum conformation of Li and Scheraga[17]. This conformation, LS_i is described in Table 4.14 both before and after DREIDING minimization. Clearly, the Li and Scheraga dihedral angles do specify a global minimum using our geometries and the DREIDING forcefield. The energy of LS_i is substantially higher than many of the unminimized conformations in Table 4.13. After minimization, its energy is far lower but the conformation has changed substantially: the ϕ, ψ angles have changed by an average of 37.4°. Nevertheless, the minimization altered the initial conformation less than it altered the grid-peak conformations of Table 4.12, indicating that LS_i lies closer to a local minimum in the conformational energy space. It should be noted that the Li and Scheraga global minimum is the product of minimization using an internal-coordinate minimizer whereas the DREIDING minimizations use full Cartesian-coordinate minimization of all 3n-6 degrees of freedom. It is possible that the changes to its torsions upon minimization are due to the increased dimensionality, i.e., allowing bonds and angles to vary decreases the energy barriers for the torsional degrees of freedom as well, allowing them to move. This factor may be as important as differences between the DREIDING and ECEPP forcefields. In any case, with an energy of 67.7 kcal/mol, the minimized Li and Scheraga confor-

Met-Enkephalin Conformations after Minimization

Dihedral	60°_m	30°_m	15°_m	10°_m	5°_m
Tyr 1 ϕ	-53.4	-54.0	-49.0	-48.5	-55.1
ψ	-75.3	-68.2	-57.6	-57.0	-63.4
χ^1	-48.5	-64.2	-66.4	-66.2	-153.3
χ^2	-70.0	-74.9	-69.4	-69.4	69.9
Gly 2 ω	-179.6	179.4	176.0	174.8	173.8
ϕ	67.4	106.0	-58.7	-59.1	117.0
ψ	7.6	60.6	-48.2	-47.1	-9.7
Gly 3 ω	-170.8	-175.8	177.7	176.2	179.1
ϕ	94.6	76.7	-75.8	-74.4	104.8
ψ	-2.9	16.7	3.5	2.8	-9.7
Phe 4 ω	-170.8	-175.8	178.3	177.5	-176.3
ϕ	-121.3	-115.6	-113.4	-112.3	-124.2
ψ	-60.2	-57.9	-48.7	-49.4	-56.6
χ^1	-172.0	-54.7	-62.2	-60.5	-162.5
χ^2	65.3	115.2	114.0	114.7	70.1
Met 5 ω	-179.9	-179.0	-177.0	-177.4	-175.6
ϕ	-124.1	-114.4	-109.8	-110.5	-116.8
ψ	-66.5	-58.3	-60.6	-60.7	-59.8
χ^1	-65.1	-64.3	-70.0	-70.4	-163.1
χ^2	-67.1	-76.3	173.5	174.2	-172.1
χ^3	-88.5	-90.7	90.2	90.3	90.4
Energy	81.2	80.8	87.4	87.5	86.3
RMS (ϕ,ψ) (°)	34.3	39.8	31.9	31.1	47.7
RMS (χ) (°)	13.2	17.5	11.5	14.3	68.6
RMS (Å)	1.6557	1.5109	1.3953	1.5529	2.0174

Table 4.13. Conformations generated from peak ϕ, ψ, and χ gridpoints at different grid spacings, after minimization with the DREIDING forcefield. The energy of each conformation is given, along with the RMS deviation from its original (un-minimized) structure.

mation, LS_m, is more than 10 kcal/mol lower in energy than any of the minimized conformations listed in Table 4.13.

The goal of our simulations was to sample the conformational space of Met-enkephalin using the PGMC method, and to fully minimize the conformations to yield optimum conformations. Our simulations produced approximately 50,000 conformations per hour of cpu time on a single processor of a Silicon Graphics 4D/380 workstation. The energy of each conformation was calculated using the DREIDING forcefield with no nonbond cutoffs – all van der Waals and electrostatic pairs were included. Short runs of 10,000 Monte Carlo steps were carried out first, in order to optimize the parameters to be used in longer runs.

The key parameter in any Metropolis Monte Carlo simulation is the simulation temperature (Equation (4.2)), which controls the probability of accepting a new conformation generated at random. A higher temperature means that a conformational change which increases the energy is more likely to be accepted than at a lower temperature. The advantage is that energy barriers can be traversed more easily and conformational space can be searched more thoroughly. The corresponding disadvantage of high simulation temperatures is that the computation can spend a lot of time in bad areas of conformational space and not settle near optimal conformations. The effect of temperature can be seen in Figure 4.7, where the results from four separate 10,000-step runs are shown. At 100 step intervals during the simulation, both the current energy (Figure 4.7a) and the best energy to that point (Figure 4.7b) were recorded. The four simulations, run at temperatures of 0 K, 300 K, 1000 K, and 10,000 K, give very different results. As expected, the 10,000 K simulation has a far greater fluctuation in energy than the lower temperature simulations. However, although it samples a wider variety of conformations, it does not sample as many good conformations. As can be seen in (Figure 4.7b), it is the

Li and Scheraga Global Minimum

Dihedral			LS_i	LS_m
Tyr	1	ϕ	-86.0	-58.8
		ψ	156.0	108.9
		χ^1	-173.0	-153.9
		χ^2	-101.0	-104.2
Gly	2	ω	-177.0	174.2
		ϕ	-155.0	-123.8
		ψ	84.0	58.9
Gly	3	ω	169.0	-173.1
		ϕ	84.0	98.5
		ψ	-74.0	-61.2
Phe	4	ω	-170.0	179.5
		ϕ	-137.0	-117.3
		ψ	19.0	-47.9
		χ^1	59.0	73.7
		χ^2	95.0	108.7
Met	5	ω	-174.0	-177.7
		ϕ	-164.0	-117.7
		ψ	160.0	121.2
		χ^1	53.0	60.8
		χ^2	175.0	-177.7
		χ^3	180.0	-179.7
Energy			236.5	67.7
RMS (ϕ, ψ) (°)			0.0	37.4
RMS (χ) (°)			0.0	11.3
RMS (Å)			0.0000	1.0574

Table 4.14. LS_i is the structure produced by rotating the dihedrals of our starting structure to the values reported for the ECEPP/2 global minimum [17]. LS_m was obtained by a Cartesian-coordinate conjugate-gradients minimization of LS_i using the DREIDING forcefield.

low-temperature simulations at 0 K and 300 K which obtain the lowest-energy conformations. All four simulations began with the 183.6 kcal/mol conformation 15°_i and produced substantially better conformations. The 300 K simulation produced the best conformation, which had an energy of 131.9 kcal/mol before minimization and 70.3 kcal/mol after minimization – nearly as low as the LS_m. In contrast, the 10,000 K simulation had a best conformation more than 10 kcal/mol higher, with an energy of 145.7 kcal/mol before minimization. After minimization, however, the energy dropped to 73.9 kcal/mol, nearly as low.

Monte Carlo simulations use random numbers both to produce new conformations and to determine whether new conformations are accepted or rejected. The use of random numbers ensures that simulations will proceed differently every time they are run, even when initial conditions are the same. Therefore, it is possible to run several simulations at the same grid spacing, with the same starting structure, and achieve different final conformations. Single runs like those plotted in Figure 4.7 are not sufficient for establishing optimum parameters. In order to obtain optimal parameters, we carried out numerous simulations at different grid spacings and temperatures. One such series is show in Figure 4.8. These calculations used a 15° grid spacing and a temperature of 1000 K. Forty simulations were run, using different random numbers, and the lowest-energy reached during each simulation was recorded. Twenty simulations, labeled (I), began with the same initial conformation, 15°_i and twenty began with conformations generated at random from the backbone and sidechain 15° probability grids. All the simulations progressed in the same way, with new conformations generated at every step by selecting one sidechain or one ϕ, ψ pair at random and choosing a new conformation from the probability grids. However, a different series of random numbers was used each time, so different conformations were generated and different conformations were accepted or rejected.

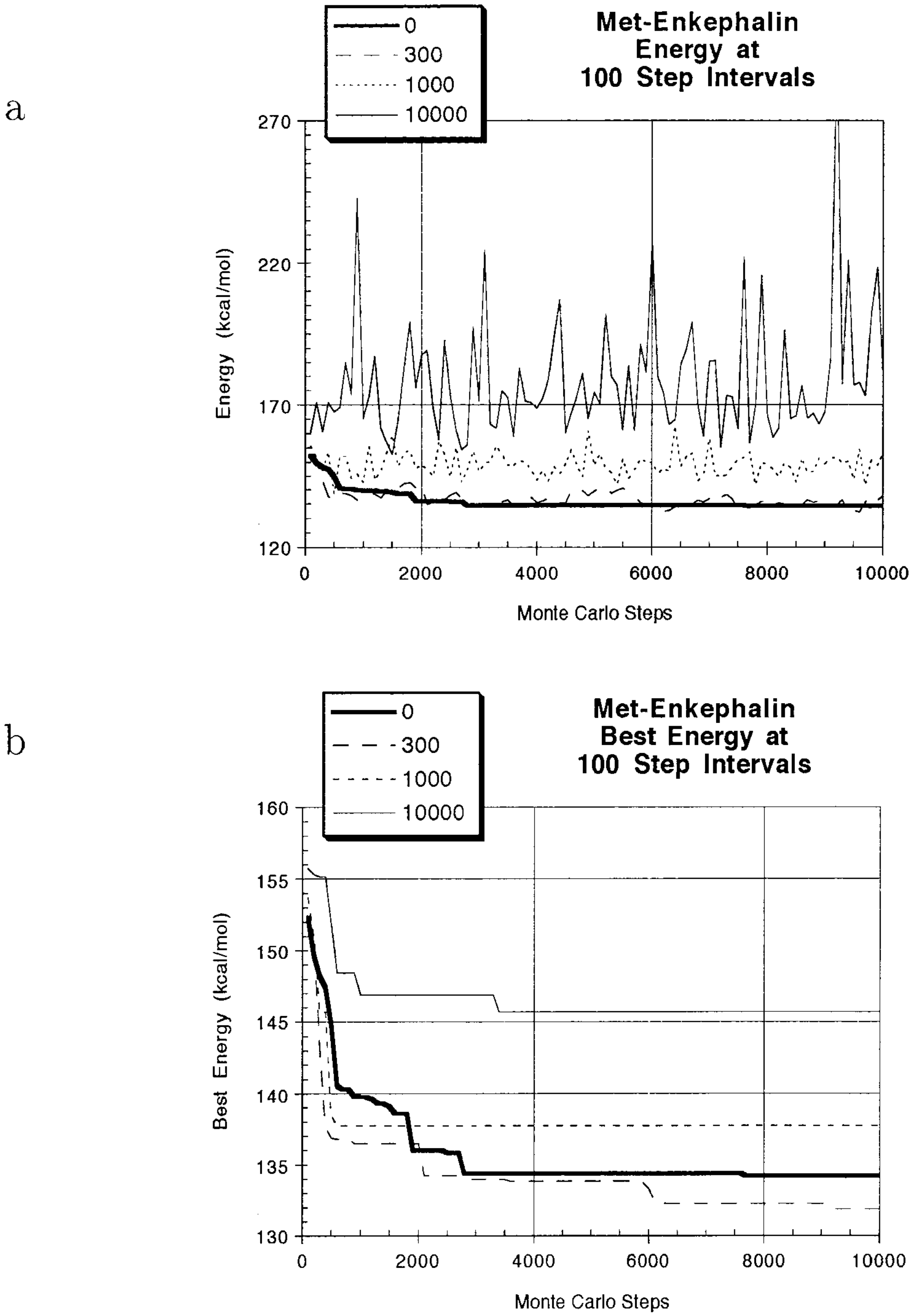


Figure 4.7. Results from several Monte Carlo simulations of Met-enkephalin. The simulations were identical except for the temperature used. The starting structure of each simulation was the 15° peak conformation and 15° ϕ, ψ and χ grids were used for conformational sampling. At 100 step intervals, the energy of the current conformation and the best overall conformation were recorded. The two graphs plot the current(a) and best energy(b) vs. Monte Carlo step.

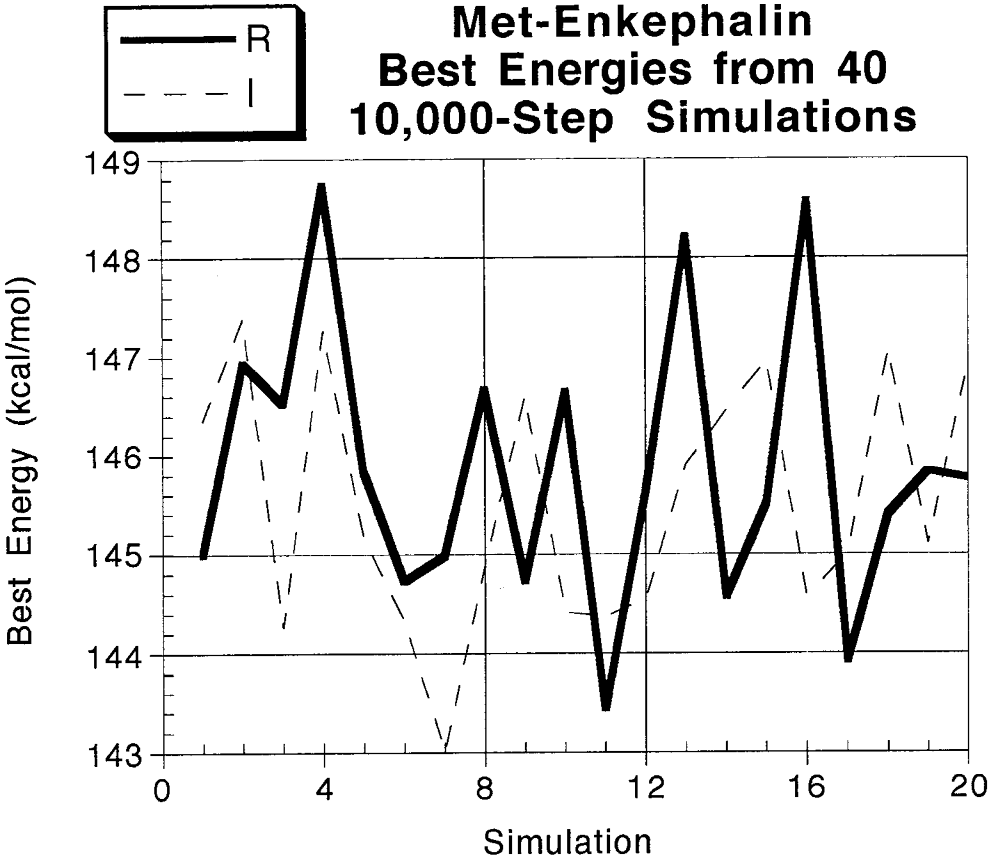


Figure 4.8. The best energy from 40 simulations of Met-enkephalin using 15° grid spacing at 1000 K. 20 simulations began with conformations generated at random (R) and 20 began with the peak 15° conformation, 15°_i, (I).

Apparently, there is little advantage to starting with the peak conformation rather than one generated at random. The best energy from the twenty simulations (I) begun with 15°_i was 143.0 kcal/mol. The average best energy from the twenty simulations was 145.5 ± 1.3 kcal/mol. This compares to an overall best of 143.4 kcal/mol and an average best of 145.9 ± 1.4 kcal/mol for the twenty simulations (R) begun with random conformations.

The next series of simulations was carried out in order to determine which combinations of temperature and grid spacing gave the best results. For each grid spacing (5, 10, 15, 30, and 60 degrees), 10 simulations of 10,000 steps were run at each of five temperatures (0 K, 300 K, 600 K, 1000 K, and 5000 K). Each simulation began with ϕ, ψ and χ conformations chosen at random from the probability grids. For each of the 250 simulations, the best overall conformation was saved and its energy recorded. We also recorded the overall acceptance rate during each run. This number, the percentage of new conformations which are accepted, depends directly upon the simulation temperature and indirectly upon the probability grids, which determine the conformational sampling and, therefore, the energy range of conformations which are generated. For each of the 25 temperature/spacing combinations, the overall best energy, the average best energy and the acceptance rate were calculated for its 10 10,000-step runs. The acceptance rates are shown in Figure 4.9. As expected, the rate is highly temperature-dependent, with roughly a 70% acceptance rate at 5000 K, 33% at 1000 K, 20% at 600 K, 10% at 300 K, and 0.5% at 0 K. The latter figure actually represents the percentage of time a new minimum energy conformation is reached, since at 0 K, there is no probability of accepting a conformation which is higher in energy than the preceding one. At nonzero temperatures, the acceptance rate includes instances when $\Delta E < 0$ and when $\Delta E > 0$, but meets the Metropolis criterion (see Section 4.II).

The energy minima for the 25 temperature/spacing combinations are shown in Figure 4.10. Figure 4.10a shows the lowest energy obtained during all ten runs combined while Figure 4.10b shows the average energy minimum for the ten separate runs. The overall lowest-energy conformation gives a good indication of how effective a set of parameters is, but it only requires one good conformation to be sampled over 100,000 steps, so it depends on good luck as well as on good parameters. A low value for the average minimum energy, on the other hand, requires that a set of parameters give consistently good answers. This is, therefore, the more useful number. The best overall energy obtained during the 250 runs was 127.9 kcal/mol, sampled during a

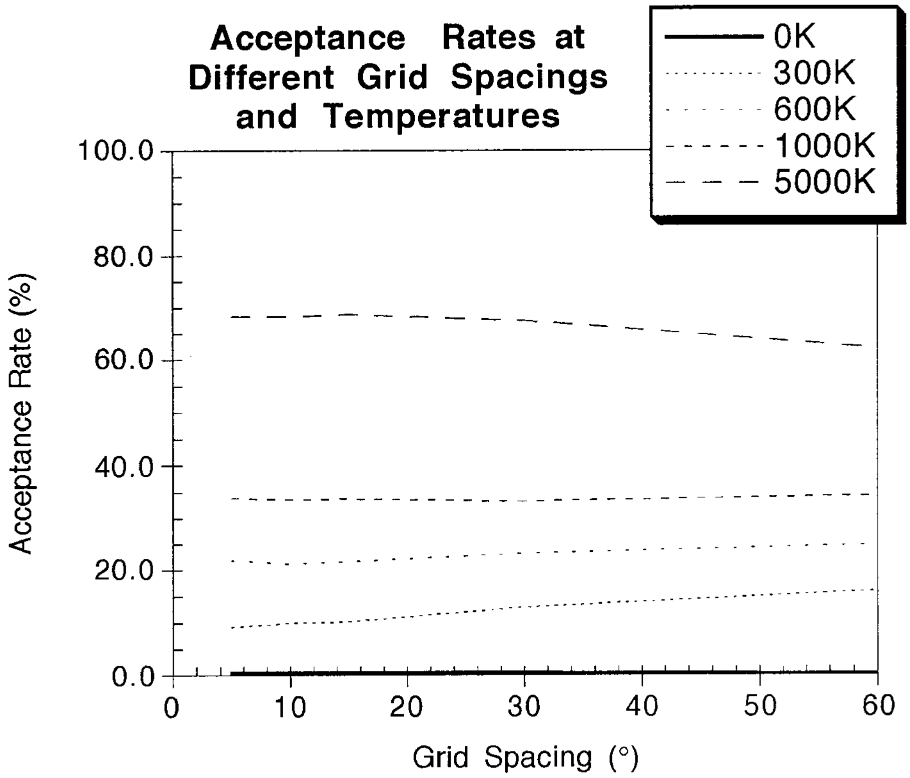
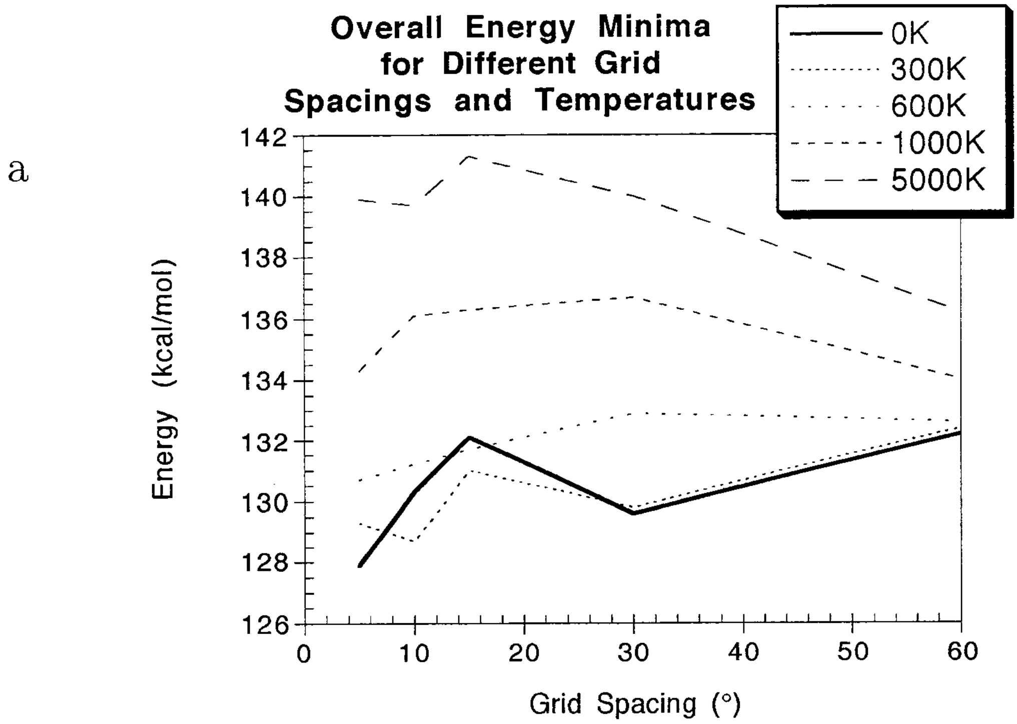


Figure 4.9. PGMC simulations were run at temperatures of 0 K, 300 K, 600 K, 1000 K and 5000 K, for each grid spacing (5°, 10°, 15°, 30°, and 60°). For each spacing/temperature pair, ten simulations of 10,000 steps each were run. Here, the overall acceptance rate for the ten runs is shown for each spacing/temperature combination.

0 K simulation using 5° probability grids. However, the best average energies were obtained during 300 K simulations. The average over 10 simulations at 300 K using both 5° and 10° probability grids was essentially the same: 131.8 kcal/mol. This is slightly less than the 5°/0 K average of 131.9 kcal/mol and significantly less than the next closest, the 15°/300 K simulations, which averaged 132.6 kcal/mol.

It is interesting to note that the energies appear to converge as the grid spacing increases. At low temperatures, finer grids give better results, but at high temperatures, the opposite result is obtained. Figure 4.9 indicates that this is related to the acceptance rate, which is much higher for 60° than 5° grids at 300 K (16.3% vs. 9.3%), but lower for 60° than 5° grids at 5000 K (62.2% vs. 68.3%). Figure 4.10 and Figure 4.9 show a strong negative correlation between high acceptance rates and low energies. This is likely to be the case because low acceptance rates require that the simulation take a more downhill path through conformation space. This may prevent sampling of low energy conformations in distant areas of conformation space, but it leads to lower energies on average. The relationship between grid spacing and acceptance rates is more subtle. The fine 5° and 10° grids have far more possible conformations which vary slightly from one another. Therefore, at high temperatures it may be difficult to make progress towards an energy minimum, because so many conformations are nearby in energy and are accepted. The 60° conformations are likely to vary much more widely in energy, and fewer would be acceptable. At low temperatures, however, the far greater flexibility of the fine-grained grids provide much easier pathways to minima in the potential energy surface. Rapid descent into an energy minimum may lock the structure into areas of conformational space where few low-energy alternatives exist, so the acceptance rate drops rapidly. Indeed, acceptance rates for the first few hundred steps are usually much higher than for later steps. The fact that grids play a significant role in the acceptance rate may mean



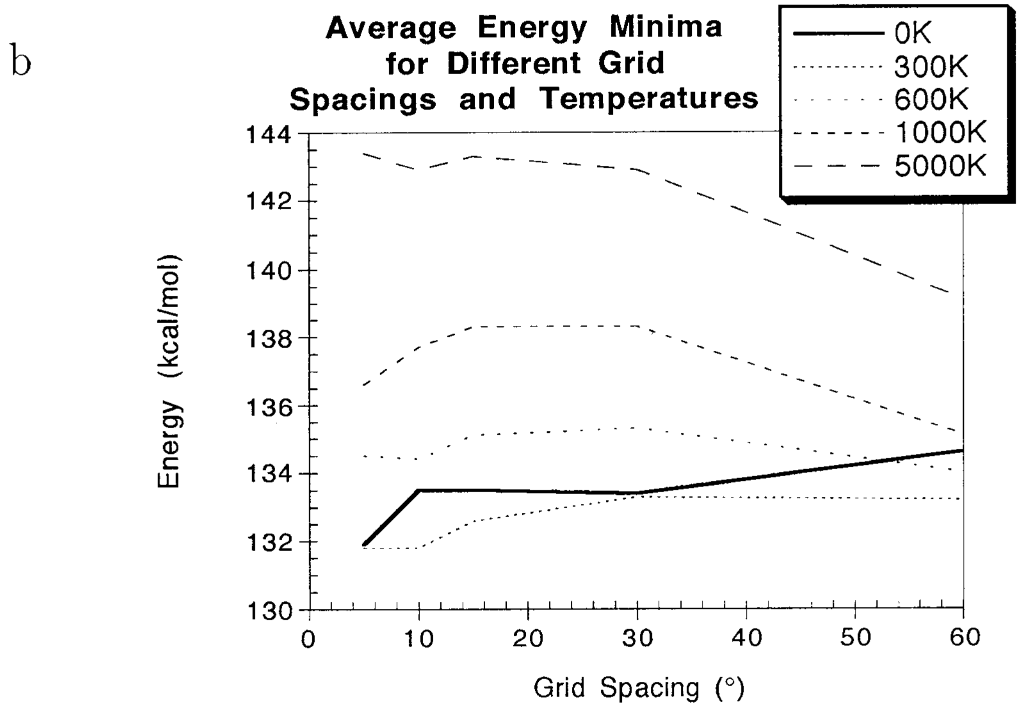


Figure 4.10. For each of 25 temperature/grid spacing combinations, ten 10,000-step simulations were run and the minimum energy from each run was recorded. The plots show both the overall lowest energy for the ten runs (a) and the average minimum energy (b) for the ten runs.

that a grid-based annealing scheme may be effective, as temperature-base annealing has been shown to be. For instance, one might start with 60° grids in order to sample broader regions of conformational space, but then slowly decrease the grid spacing as the simulation proceeds, to give added flexibility in favored subregions of the potential energy surface.

We chose the most successful spacing/temperature combinations from the above study (5°/300 K and 10°/300 K) to use in a more thorough second set of simulations. In this set of calculations, twenty 50,000-step simulations were run for each of the two spacing/temperature choices. The best conformation sampled during each run was saved and then minimized. The energies of the best conformations, before and after minimization, are shown in Figure 4.11. The results are fairly consistent from one simulation to the next, and there is not a large difference between the simulations which used 10° probability grids and those which used 5° grids. The best energy, on average, for the 10°-grid simulations, was 129.4 kcal/mol before minimization and 71.4 kcal/mol after minimization. For the 5° simulations, the averages were 128.0 kcal/mol before minimization, and 73.0 kcal/mol afterwards. It is interesting that the 10° conformations are, on average, better than the 5° conformations after minimization, even though they are worse before minimization. This implies that the 5° conformations are closer to their local minima and are optimized less dramatically (by an average of 55 kcal/mol vs. 58 kcal/mol) during the full Cartesian-coordinate minimization.

Although the results are very consistent for all 40 simulations, no two simulations produce the same energy minimum. Several are extremely close, differing by only 5° or 10° at a single dihedral, but in general there is a fairly good diversity of conformations represented. This can be clearly seen in Figure 4.13, where the dihedral angles from the various energy minima are plotted. The numbering for the dihedrals

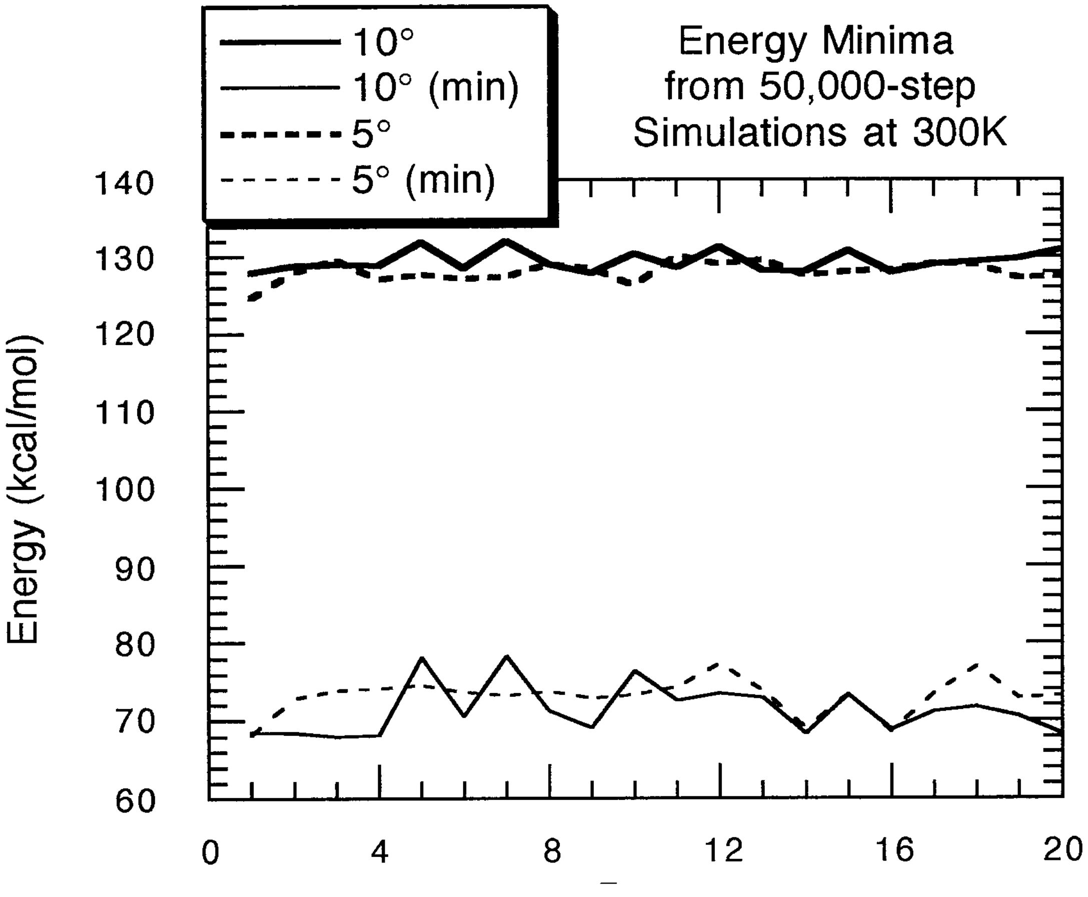


Figure 4.11. 20 simulations of 50,000 steps each were run at 300 K using grid spacings of 5° and 10°. The best conformation from each run was energy minimized and the energy before and after minimization was recorded. The top two lines plot the unminimized energies while the bottom two, labeled "(min)", show the minimized energies.

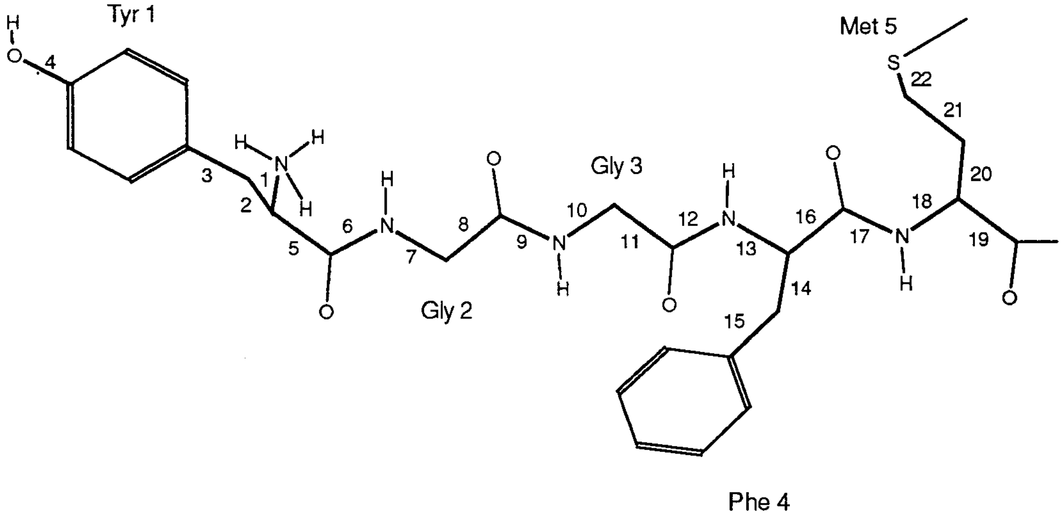


Figure 4.12. The peptide Met-enkephalin, with the dihedrals numbered as in Table 4.15.

is shown in Table 4.15 and Figure 4.12. Although there are definite trends among the different minima, only the five dihedrals unsampled by the PGMC method are the same in each conformation. These dihedrals are χ^6 of Tyr 1 (dihedral 4), and the four backbone ω dihedrals, (6, 9, 12, and 17). All other dihedrals are represented by at least two different conformations and some by as many as 11. The dihedrals which show the greatest variability are the ϕ and ψ of the two glycine residues, dihedrals 7, 8, 10, and 11. This is an important factor to consider in understanding the active conformations of Met-enkephalin. The great flexibility of the glycine backbone means that many conformations are possible which differ radically in their backbone conformation but which are near each other in energy. The active conformation may require one particular glycine conformation which, in the absence of a receptor, is not especially favorable.

Of the twenty optimized conformations created by the 10° simulations, six have energies within 1.0 kcal/mol of LS_m (Table 4.14). Two of the 5° conformations also had energies within 1.0 kcal/mol. None of these eight conformations, however, was

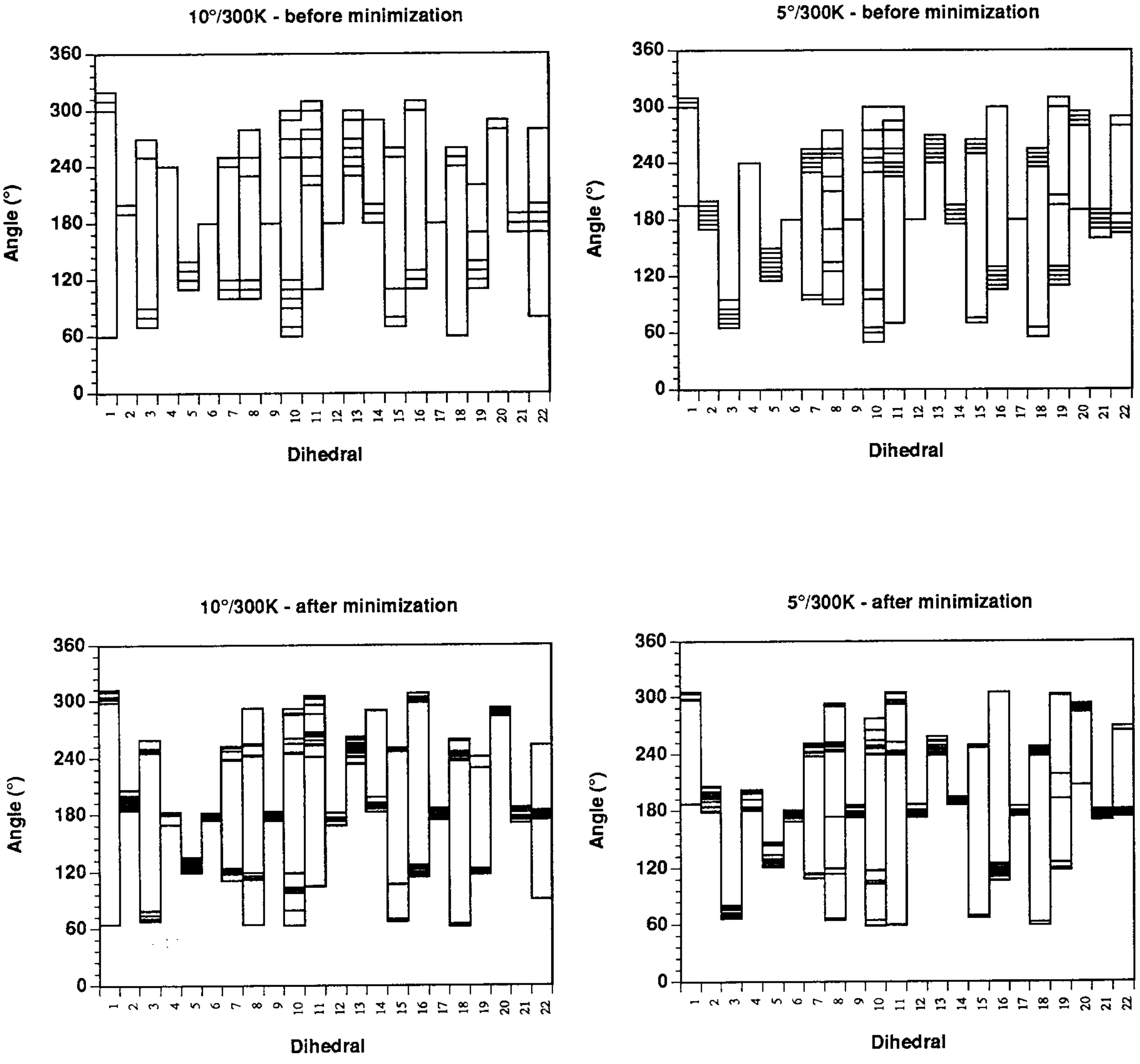


Figure 4.13. The dihedral angles from each of the 20 minima produced by 50,000-step simulations at 300 K. Only the dihedrals unsampled by the PGMC method, χ^1 (dihedral 4) and the ω's (6, 9, 12, and 17) are the same in every conformation. Even these degeneracies are broken during minimization.

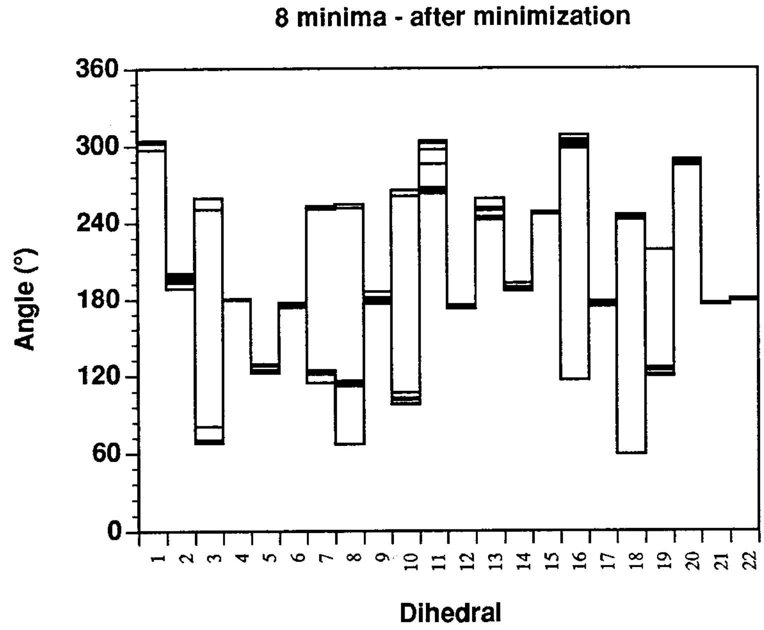
Tyr 1					Gly 2			Gly 3			Phe 4					Met 5					
ϕ	χ^1	χ^2	χ^3	ψ	ω	ϕ	ψ	ω	ϕ	ψ	ω	ϕ	χ^1	χ^2	ψ	ω	ϕ	ψ	χ^1	χ^2	χ^3
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

Table 4.15. Dihedral numbering used in Figure 4.13 and Figure 4.14.

actually lower in energy than LS_m; the closest had an energy of 68.0 kcal/mol, only 0.3 kcal/mol higher. It is not known whether any conformation close to the Li and Scheraga minimum was sampled during any of the simulations, since its energy is so high before minimization. The eight lowest energy conformations predicted here are extremely similar to one another, particularly at residues Phe 4 and Met 5, where they are all virtually identical. Only one of the 5° conformational minima, which had an energy of 68.6 kcal/mol, differed significantly in this region. As can be seen in Figure 4.14b, the other seven minima are identical in this region, and are distinguished primarily at χ^2 of Met 1, the ϕ and ψ of Gly 2, and the ϕ of Gly 3. The Li and Scheraga minimum, LS_m, not shown, is also very similar, but differs in the sidechain dihedrals, χ^1 and χ^2 of Phe 4 and χ^1 of Met 5.

IV. Conclusions

The Probability Grid Monte Carlo method is an effective strategy for sampling the conformation space of peptides. Even without the benefit of simulated annealing or other more elaborate temperature adjustment schemes, it regularly produced very low energy conformations for Met-enkephalin. These conformations are quite similar to the published global minimum of ECEPP/2. In addition, a much wider variety of conformations was generated at a slightly higher energy. This ensemble



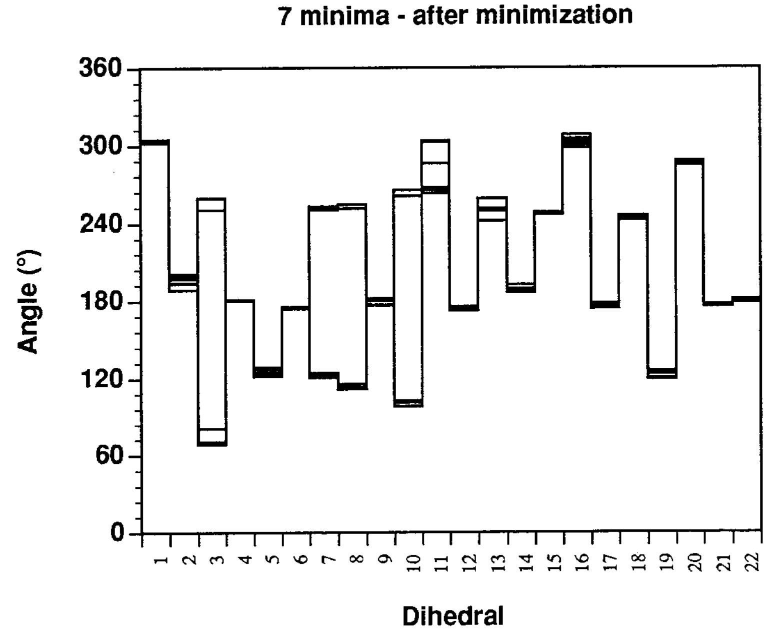


Figure 4.14. The top plot shows the distribution of dihedral angles for the eight conformations with energies within 1 kcal/mol of LS_i, the conformation with the lowest known energy . The bottom plot shows the same distribution minus the one conformation which differs significantly from the others at dihedrals 16, 18, and 19. The other seven conformations have virtually identical conformations for residues Phe 4 and Met 5.

of conformations may be important for understanding the bound conformations and activity of Met-enkephalin.

References

[1] M.H. Lambert and H.A. Scheraga, *J. Comp. Chem.*, 10, 817-831 (1989).

[2] R.E. Bruccoleri and M. Karplus, *Biopolymers*, 26, 137-168 (1987).

[3] M. Lipton and W.C. Still, *J. Comp. Chem.*, 9, 343-355 (1988).

[4] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller, *J. Chem. Phys.*, 21, 1087-1092 (1953).

[5] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[6] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P. Weiner, *J. Am. Chem. Soc.*, 106, 765-784 (1984).

[7] E.H.L. Arts, J.H.M. Korst, and P.J.M. van Laarhoven, *J. Stat. Phys.*, 50, 187-206 (1988).

[8] J.K. Shin and M.S. Jhon, *Biopolymers*, 31, 177-185 (1991).

[9] B. von Freyberg and W. Braun, *J. Comp. Chem.*, 12, 1065-1076 (1991).

[10] W.R. Pearson and D.J. Lipman, *Proc. Natl. Acad. Sci., USA*, 85, 2444-2448 (1988).

[11] G.N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.*, 23, 283-437 (1968).

[12] S.S. Zimmerman *et al.*, *Macromolecules*, 10, 1-9 (1977).

[13] BIOGRAF/POLYGRAF. Copyright by Molecular Simulations, Inc. (1992).

[14] T.E. Creighton, *Proteins*, W.H. Freeman and Company, New York (1984).

[15] R.W. Woody in *Conformation in Biology and Drug Design*, V.J. Hruby, Ed., Academic Press, Orlando (1985).

[16] K.C. Garcia *et al.*, *Science*, 257, 502-507 (1992), and references therein.

[17] Z. Li and H.A. Scheraga, *Proc. Natl. Acad. Sci., USA*, 84, 6611-6615 (1987).

[18] H. Kawai, T. Kikuchi, and Y. Okamoto, *Protein Engineering*, 3, 85-94 (1989).

[19] J.K. Shin and M.S. Jhon, *Biopolymers*, 31, 177-185, (1991).

[20] A. Nayeem, J. Vila, and H.A. Scheraga, *J. Comp. Chem.*, 12, 594-605 (1991).

[21] M.J. Sippl, G. Némethy, and H.A. Scheraga, *J. Phys. Chem.*, 88, 6231-6233 (1984).

Chapter 5

Protein modeling from C_α coordinates

Abstract

We present a method for predicting the complete conformation of a protein from its C_α coordinates based on the Probability Grid Monte Carlo (PGMC) Method described in Chapter 4. Unlike most methods designed to solve this problem, the PGMC Method does not attempt to fit known polypeptide conformations onto the C_α framework. Rather, conformational propensities for individual residues are used to guide conformational searches while the protein is built from the amino-terminus to the carboxy-terminus. Therefore, no structural homology to other known structures is required. We present results for a number of proteins and show that both the backbone and sidechain can be accurately modeled using the PGMC method. Backbone atoms can generally be predicted to within 0.6 Å of their X-ray crystal structure coordinates, while the total rms for all atoms can be predicted to 1.7 Å or better. The method is also used to build all-atom protein models from C_α coordinates derived from lattice-based methods of protein structure prediction, through the use of a "C_α Forcefield."

I. Introduction

The global conformation of a protein can be well approximated by a trace drawn through the coordinates of its C_α atoms. As the central atom of each amino acid residue – the point at which the sidechain branches off from the main chain – the C_α atom is the best choice to represent the amino acid as a whole. Figure 5.1 shows the C_α trace of the small protein crambin, as well as a picture of the backbone atoms and a picture of all atoms in the structure, from the crystal structure by Hendrickson and Teeter[1] (Brookhaven protein database (PDB) structure 1CRN). Because of their central location, C_α coordinates usually form the starting point for the process of building a protein model from X-ray crystallographic data[2]. In addition, purely theoretical schemes to predict tertiary-structure often use a simplified protein model containing only C_α coordinates[3, 4]. And C_α coordinates can form a template for homology-based molecular modeling [5]. However, the C_α coordinates do not provide sufficient information for understanding the most critical aspects of proteins such as binding and catalysis, which are determined by the chemical and steric properties of the protein backbone and sidechains. It is therefore necessary to provide a means of obtaining all atomic coordinates for proteins when the C_α coordinates alone are known.

Several methods for modeling complete protein structures from C_α coordinates have been published in recent years[2,6-10]. The primary purpose for such methods is to speed and automate the process of building a protein model from crystallographic data[2], but several other uses have been suggested. Holm and Sander[9] describe how correct and incorrect protein folds can be evaluated by such a method, while Rey and Skolnick mention that their procedure may enable complete protein structures to be built from the C_α coordinates of a lattice representation[9]. The work reported here has been motivated by both of these factors: the desire to build full protein

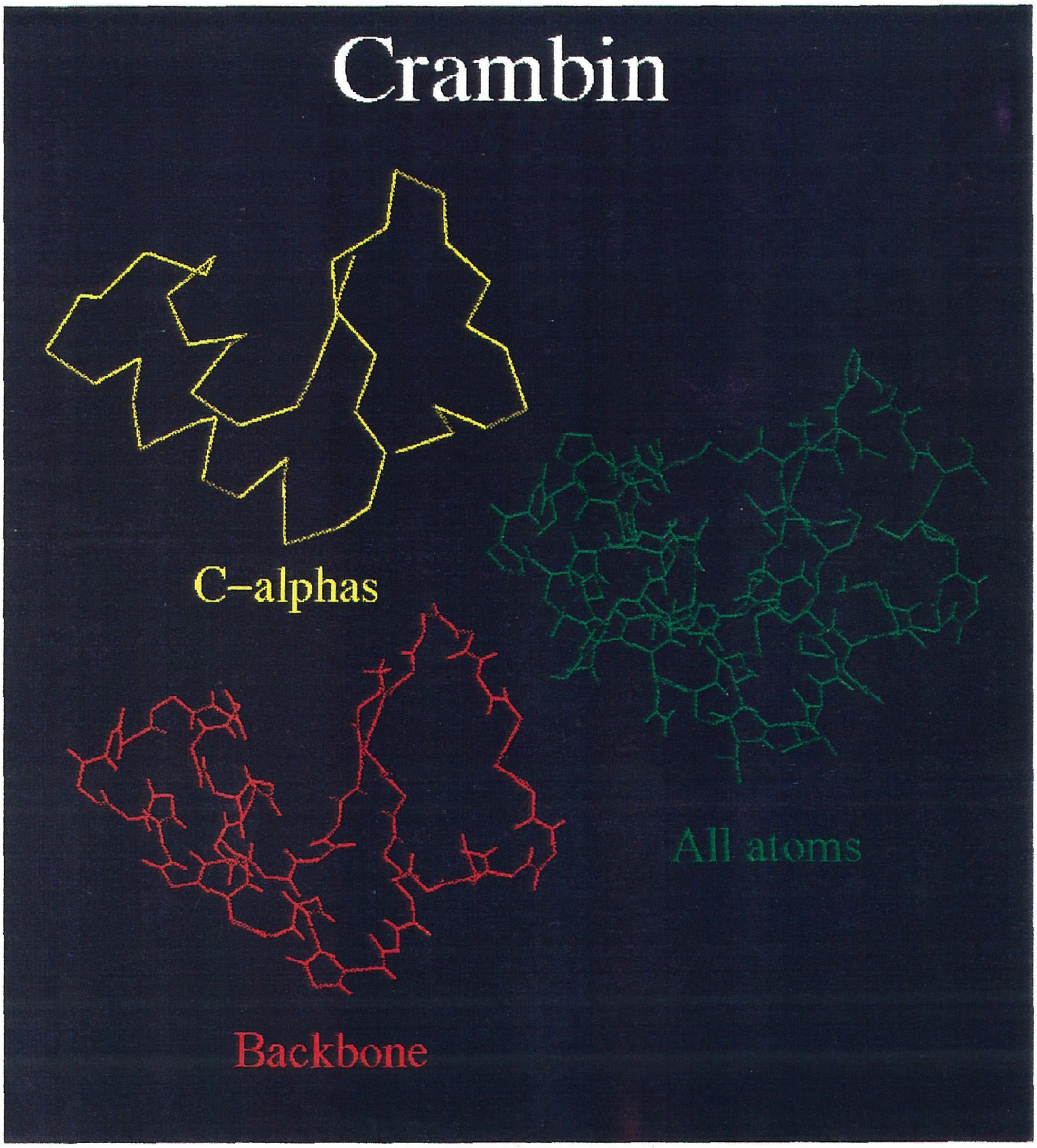


Figure 5.1. Three views of crambin: the C_α's (46 atoms), the peptide backbone (185 atoms), and the all-atom structure (402 atoms). From the crystal structure by Hendrickson and Teeter[1].

structures from lattice structures, and to provide a means for evaluating different lattice conformations. In addition, we have found that the "C_α Builder" described here has been useful for homology modeling, as it allowed us to build a model of Hin recombinase from the C_α coordinates of λ Cro[5].

The process of building full protein conformations from C_α coordinates requires success in two areas: prediction of backbone conformations in the presence of explicit geometric constraints (the known C_α coordinates) and prediction of sidechain conformations constrained only by the conformation of the backbone and the presence of other sidechains. Our method provides a consistent approach to solving the two problems. Based primarily on Monte Carlo conformational searching, our technique differs significantly from previously published techniques, which range from the purely geometric[6, 10] to methods based primarily on database searches of several consecutive residues[7, 2, 9] or molecular mechanics[8].

Our procedure for building protein structures from C_α coordinates uses the conformational probabilities of individual residues, rather than groups of residues and, therefore, does not depend upon the prior existence of particular conformations in the protein database. The process uses the Probability Grid Monte Carlo (PGMC) method to build, first, the backbone conformation then, second, the sidechains. The PGMC method, described fully in Chapter 4, modifies protein conformations one residue at a time, by choosing either new backbone (ϕ, ψ) or sidechain (χ) dihedral angles from probability matrices. In the first phase of the PGMC C_α Builder, the backbone is built one residue at a time. As the protein chain grows, the conformational space of the backbone is sampled by the PGMC method using ϕ, ψ probability grids. The DREIDING forcefield[11] is used to evaluate the energy of each structure, with additional harmonic constraint terms added between the template C_α coordinates and the C_α coordinates of the growing chain. After the entire backbone is built in this way, sidechain positions are optimized during a second PGMC simulation. This second simulation uses χ probability grids to modify one sidechain conformation at a time. Because the PGMC method uses random numbers both to determine whether new conformations are accepted or rejected and to choose new conformations, each run produces different results. Therefore, it is general practice to generate numerous backbone conformations and select those with the best energy to use in the second stage. Likewise, for each backbone conformation, several Monte Carlo simulations are run to optimize the sidechains, and the structure with the best overall energy is selected as the optimized model.

II. Methodology

II.A. General Methodology

The PGMC C_α Builder was developed as an extension of the BIOGRAF program from Molecular Simulations, Incorporated[18]. All calculations reported here were run on Silicon Graphics Power Series and Indigo workstations; all timing numbers were obtained from simulations run on a single processor of an SGI 4D/380. During the first stage of the model-building procedure, the protein is created one residue at a time until the entire protein has been built. As each residue i is added, its geometry is initially built from the standard peptide geometries in the BIOGRAF peptide library, then the backbone (ϕ, ψ) and sidechain (χ) dihedrals are rotated to their most probable conformations according to the relevant probability grids. A Monte Carlo simulation using ϕ, ψ probability grids is then used to search the conformational space of a "pulse" of residues: the last p residues of the current chain (residues $l - p + 1$ through l). The residues preceding the pulse are held fixed and are not included in the energy calculations. Simulations in which these early residues are held fixed, but included in the energy calculation, are considerably slower and give worse results. The sidechains are also ignored during the chain-building phase; they are added in the second stage after the backbone conformation has been built. The energy used during the Monte Carlo simulations is essentially the DREIDING energy of the backbone atoms of the pulse, plus harmonic terms constraining the pulse C_α coordinates to the true coordinates. The best conformation sampled during the Monte Carlo simulation is saved and then optimized by conjugate gradients minimization. This process proceeds sequentially, with each new residue being involved in several optimization cycles before finally being held in its final position as the pulse moves beyond it.

The backbone Monte Carlo simulations are aided by pre-determination of the secondary structure, where possible. There is a high correlation between the ϕ, ψ dihedrals of a protein and its C_α coordinates, so knowledge of the C_α coordinates can limit the possible ϕ, ψ values. The most common secondary structural elements, α helices and β sheets, have very specific C_α configurations, as described by the virtual angle ζ and virtual dihedral γ, shown in Figure 5.2. Analysis of the ζ, γ distributions of the proteins in our H64 dataset showed that HELIX and SHEET residues almost always have ζ and γ values within the ranges specified in Table 5.1. Residues with ζ, γ distributions in one of these two regions are assumed to have ϕ, ψ values common to that secondary structure type; when their ϕ and ψ conformations are sampled during the chain-building process, the ϕ, ψ grids determined for HELIX or SHEET residues are used. Residues with ζ, γ values falling outside this region are sampled using the generic ϕ, ψ probability grids. 85% of the residues in the H64 dataset having ϕ and ψ values within the high-probability β sheet region listed in Table 5.1, also have ζ, γ values within the specified region. The correlation is even higher for α helices, where 88% of the residues with α helix ϕ, ψ values have ζ, γ val-

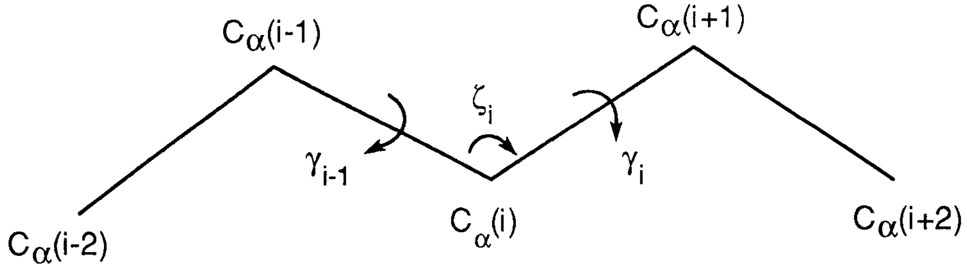


Figure 5.2. Definition of virtual angle, ζ, and dihedral, γ, for residue i.

ues within the corresponding range. If there were no variation in bond lengths and angles in the protein backbone, the ζ, γ angles would provide almost completely sufficient information to determine the ϕ, ψ angles, according to the method developed by Purisima and Scheraga[6]. Unfortunately, the variability in real conformations is too high for this exact method to work, and ϕ, ψ angles must be derived from simulation methods such as the one presented here. Nevertheless, the correlation between ζ, γ and ϕ, ψ angles is sufficient to determine which residues should be sampled using the HELIX and SHEET ϕ, ψ grids. The use of these grids for the appropriate residues improves our results significantly.

Because Monte Carlo simulations depend on random numbers, each time the calculation is run, it produces a different backbone conformation. However, an exhaustive search is much more computationally intensive, even if only a few conformations were allowed for each residue. A complete sampling of just the top 20 ϕ, ψ conformations for each residue in a three-residue pulse would require evaluation of 8000 different conformations. In contrast, we are able to obtain excellent results from only 200 Monte Carlo steps. The Metropolis criterion (see Section 1.III and Reference [12]) rejects conformations which produce very bad energies, allowing the conformational sampling to focus on low-energy conformations. It is therefore

2° Structure	γ_i	ζ_i
α helix	$25° < \gamma_i < 75°$	$80° < \zeta_i < 110°$
β sheet	$160° < \gamma_i, \gamma_i < -75°$	$100° < \zeta_i < 145°$
	ϕ_i	ψ_i
α helix	$-90° < \phi < -30°$	$-60° < \psi < 0°$
β sheet	$-165° < \phi < -45°$	$100° < \psi < 180°$

Table 5.1. ζ, γ regions indicating residue i is likely to be in an α helix or β sheet conformation; i.e., its ϕ and ψ fall within the corresponding ϕ, ψ region listed in the lower table. ζ_i and γ_i are defined in Figure 5.2.

possible to quickly build backbone conformations. A typical simulation takes approximately 15 seconds per residue on one processor of an Silicon Graphics 4D/380 workstations, or less than 12 minutes for the 46 residue protein, crambin. Speed is crucial for simulations where different C_α conformations are being evaluated, for instance when numerous conformations are generated by a lattice-based protein structure prediction method[3]. In cases where a single set of C_α coordinates is being used, it may not be necessary to limit the calculations to a matter of minutes. In these cases, several simulations can be run, using different random numbers for the Monte Carlo calculation. Each will produce a slightly different backbone conformation. From these, the lowest energy conformations are selected for the second stage of the calculation.

The best-energy conformations generated in Phase 1 were evaluated without regard to their sidechain positions. During the chain-building process, energies were determined for only a small pulse of residues; all previous residues were ignored. However, after the chain is built, the energy of the entire backbone is evaluated and this value is used to determine which backbone conformations are used in Phase 2.

The sidechain conformations are optimized by a PGMC simulation using χ probability grids. In this stage, the backbone atoms are held fixed, but are included in the energy calculation. Because the backbone is held fixed, constraints to the C_α coordinates are removed. In these calculations, at every Monte Carlo step, one sidechain is selected at random, and a new sidechain conformation is chosen from it according to the residue-specific χ probability grid. The energy of the new conformation is calculated, and the Metropolis criterion is used to accept or reject this structure. Since the Metropolis acceptance probability (Equation (4.2)) is dependent upon the ΔE, the change in energy, only the energy of the sidechain being modified needs to be evaluated; all interactions not involving the sidechain being modified can be considered constant and do not need to be evaluated. This results in a huge speed increase over calculations which re-evaluate the entire energy of the protein at every step. Using this method, the second stage can be quite rapid. For the small protein crambin, which has 46 residues and 396 atoms in the DREIDING calculations, 1000 Monte Carlo steps requires seven minutes of cpu time, while plastocyanin, with 98 residues and 857 atoms, requires 22 minutes for 1000 steps. Like the backbone-building process, the sidechain-modeling process is a stochastic simulation, dependent upon random numbers. Therefore, it is useful to run the simulation several times, using different random number seeds, and to use the lowest-energy structures for further studies.

II.B. Variables

There are a considerable number of variables which affect the efficiency of the PGMC C_α builder. Several of these are listed in Table 5.2. In order to determine which combination of parameters were most effective, we ran numerous simulations using crambin[1] as a model. This protein was chosen because of its small size, which

Phase 1 Variables

Variable	Description
Pulse	The number of residues used in Monte Carlo sampling.
Constraint	Force constant of harmonic C_α constraint.

Variables for Phases 1 and 2

Spacing	The dihedral increment used: 5°, 10°, 15°, 30°, or 60°.
Temperature	The constant controlling the Monte Carlo acceptance probability.
Steps	The number of conformations sampled by the PGMC calculation.

Table 5.2. Variables used in PGMC C_α Builder. For Phase 1, "steps" refers to the number of conformations sampled as each residue is added. For Phase 2, it refers to the total number of conformations sampled.

allowed for rapid calculations, and because it contained α helix, β sheet, and β turn regions. Phase 1 parameters were evaluated by running 20 simulations for each set of parameters, building the complete crambin backbone from its C_α coordinates. The efficacy of the parameters was determined by averaging, over the 20 runs, the root-mean-square (rms) deviations from the crystal structure for the backbone atoms of the models produced. This average correlated very well with a second measure of the accuracy of the backbone model: the rms deviations in the ϕ, ψ dihedrals. Not every variable had a large impact on the results. In particular, the simulation temperature and the grid spacing had smaller effects than did the pulse size, the harmonic constraint, or the number of Monte Carlo steps.

The average rms deviations from twenty Phase 1 simulations are shown in Figure 5.3 for several temperatures and pulse sizes. These simulations were run using 200 Monte Carlo steps for each pulse, a grid spacing of 10°, and a C_α constraint of 1000 (kcal/mol)/Å^2. There are no consistent trends with respect to temperature.

For pulse lengths of three or four, the best results are obtained at a temperature of 1000 K. However, for longer pulses, higher temperatures are more favorable. The pulse length, itself, has a much bigger impact on the results. There is a consistent trend favoring shorter pulse lengths at all temperatures except 5000 K, where a pulse of six is better than a pulse of five. It was clear from numerous other simulations that a pulse length of three gave the best results, with four residues being slightly worse and large numbers significantly worse. The number of possible ϕ, ψ conformations grows exponentially with the number of residues in the pulse, so smaller pulse lengths are clearly favored in that a larger percentage of their conformational space can be searched during the Monte Carlo calculation. This makes up for the fact that important hydrogen bonding interactions occur between residues i and $i+4$ in α helices, a fact that would favor a pulse length of at least four. In addition, the time of the simulation is roughly proportional to p, so a pulse length of three is preferable from the standpoint of speed, as well.

Another important variable in these simulations is the force constant of the harmonic constraint between the C_α's of the protein chain being built and the input C_α coordinates. The energy of each constraint is given by the expression

$$E_c = \frac{1}{2} K_c (r_i)^2, \tag{5.1}$$

where K_c is the force constant and r_i is the distance between the C_α coordinate of residue i in the model and in the template. There is a constraint of this type for each residue in the pulse. There is an additional constraint, with a weak force constant of $K_c/10$ and an offset of 2.0 Å, between the carbonyl carbon of the most recently added residue, l, and the template C_α of residue $l+1$. This helps to orient the final residue of the growing chain. Figure 5.4 shows the effect of the constraint on the average rms errors in the backbone atoms (RMSB) and the C_α coordinates (RMSC). These simulations were run at a temperature of 1000 K, using a grid spacing

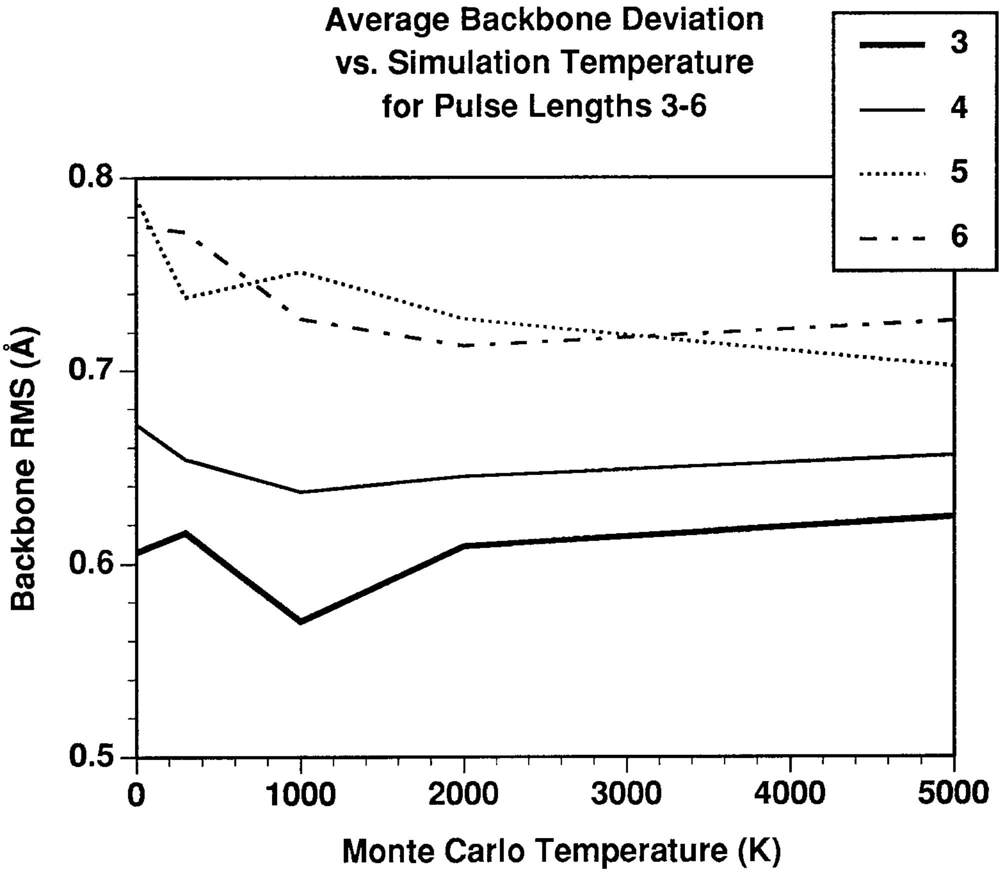


Figure 5.3. The average rms deviation from the crystal structure for models of the crambin backbone built using various temperatures and pulse sizes.

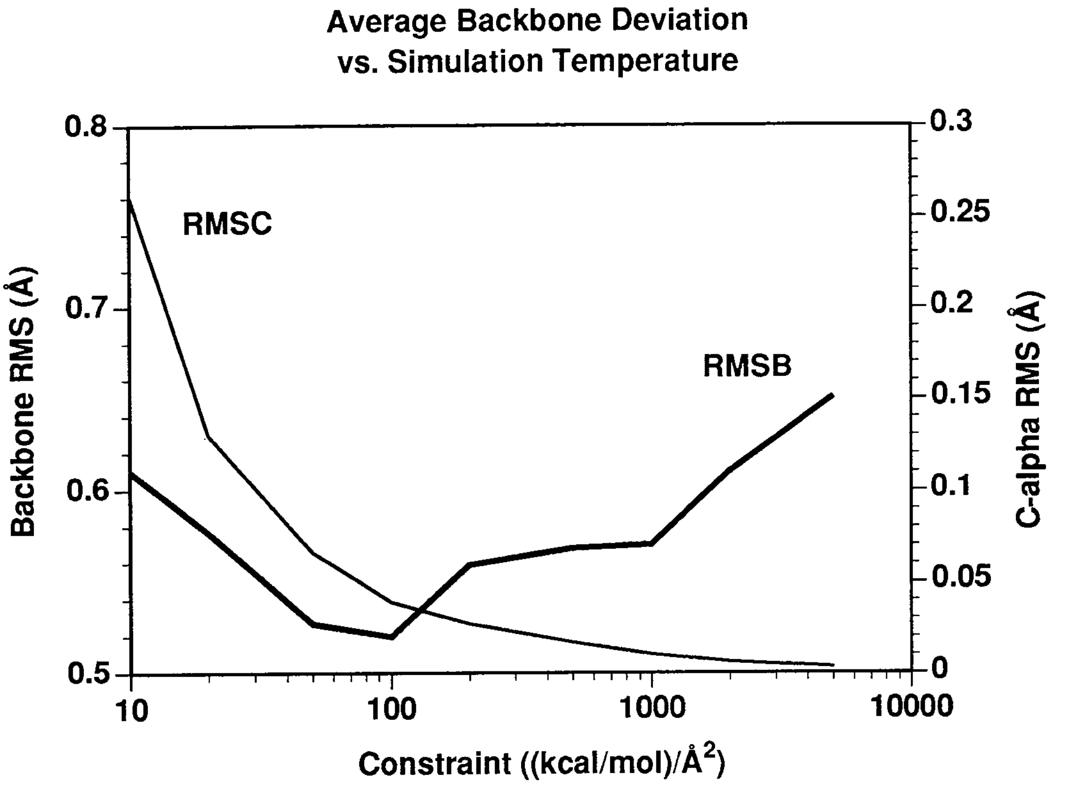


Figure 5.4. The average rms deviation in backbone atoms (RMSB) and C_α coordinates (RMSC) for crambin backbone models built using different C_α constraint force constants.

of 10° and a pulse length of three. As should be expected, the deviations for the C_α coordinates decrease exponentially as the force constant increases. However, the fit of the entire backbone has a minimum of 0.520 Å when $K_c = 100$ (kcal/mol)/Å^2. This is substantially less than a typical DREIDING force constant of 700 (kcal/mol)/Å^2 or more for bondstretches. Therefore, the C_α constraints do not cause distortions in the geometries during the conjugate gradients minimization stage which follows the Monte Carlo.

As each new residue is added, the pulse of residues is optimized first by the Monte Carlo conformational search, then by 100 steps of conjugate gradients minimization. Both stages are important. The minimization process is necessary to provide flexibility in the bond lengths and angles of the protein model, in order to match closely the specific C_α geometry of the protein being built. Although the minimization process makes only small adjustments in the conformation of the pulse residues, it makes a substantial difference in the results. With no minimization, the errors in the backbone model built up very quickly. Using the same parameters which produced an average backbone deviation of 0.52 Å when minimization was included, the C_α Builder produces crambin backbone models with an average rms deviation of 1.32 Å when no minimization is involved. The parameters were optimized for simulations including minimization and probably do not represent the best possible results for simulations without minimization. Nevertheless, it is clearly preferable to include the minimization process. It is also important to include the Monte Carlo conformational search. The results using different numbers of Monte Carlo steps are shown in Figure 5.5. Simulations with one step correspond to simply using the highest probability conformation from the ϕ, ψ grids for each residue; no other conformations are sampled. Although the results for this case are good (0.60 Å rms), the results are clearly improved by the use of even a small number of Monte Carlo steps, and get better as the number of steps increases. The standard error in these averages is typically 0.01 Å, so there is little statistical significance to the improvements above 50 steps. Nevertheless, in order to increase the number of conformations sampled while keeping the simulation time to 10 minutes per crambin backbone conformation, we chose to use a value of 200 Monte Carlo steps for most simulations.

The choice of grid spacing was based upon simulations of the pentapeptide Met-enkephalin (see Section 4.III), which found that the best results were obtained using a grid spacing of 10°. The 10° dihedral spacing appears to provide the best balance between conflicting trends which arise as the grid spacing becomes smaller: there are far more possible conformations, so the protein can assume more low-energy

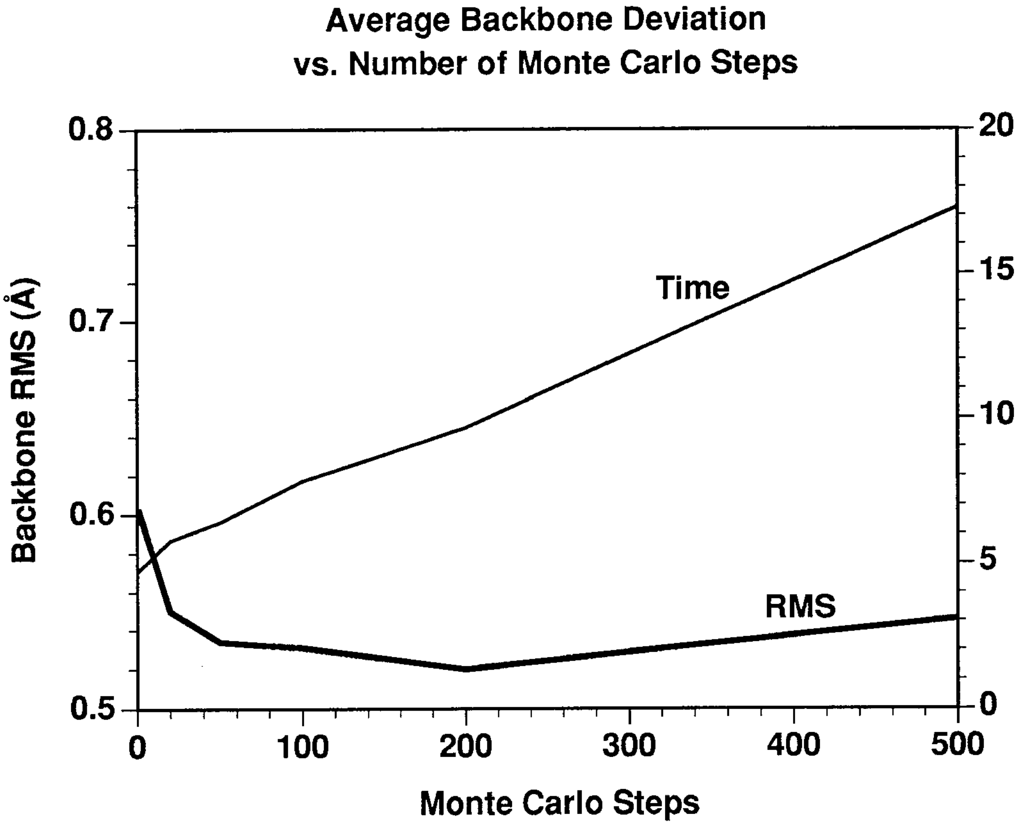


Figure 5.5. Average backbone rms vs. the number of Monte Carlo steps. Also shown is the average time to build each backbone conformation.

conformations, but the fraction of the total conformational space that can be sampled during a given number of Monte Carlo steps decreases.

After backbone models have been developed in Phase 1, the sidechains are optimized in Phase 2. In these calculations, the backbone is held fixed while the sidechains are modified by randomly choosing new conformations according to the χ probability grids. The most important variables for these simulations are the grid spacing, the temperature, and the number of Monte Carlo steps. A grid spacing of 10° was selected for these calculations in order to be consistent with the grid spacing chosen for Phase 1. Results improved consistently as the number of Monte Carlo steps was increased, but improvement slowed after about 500 steps; therefore, a value of 1000 was used for the calculations reported below. As discussed below, this number may be insufficient for large proteins, but for crambin it represents more than 25 conformations per residue for the 37 non-alanine, non-glycine optimized during these simulations.

In order to determine the best simulation temperature for Phase 2, ten PGMC calculations were run at several temperatures between 0 K and 5000 K. The starting structure for these calculations was the crambin crystal structure, with its sidechains rotated to their most probable conformations according to the 10° χ probability grids. This structure had an rms deviation from the crystal structure of 1.52 Å; the deviation for sidechain atoms alone was 2.34 Å. For each simulation, 1000 Monte Carlo calculations were run, after which the lowest energy conformation was saved and its overall rms deviation from the crystal structure was recorded. The average for the ten simulations at each temperature is shown in Figure 5.6. As was found for the backbone Monte Carlo simulations in Phase 1 (see Figure 5.3), there is not a large variation with respect to temperature. This is the case despite the fact that the acceptance rate for new structures rises from 7.7% at 0 K to 46.8% at 5000 K.

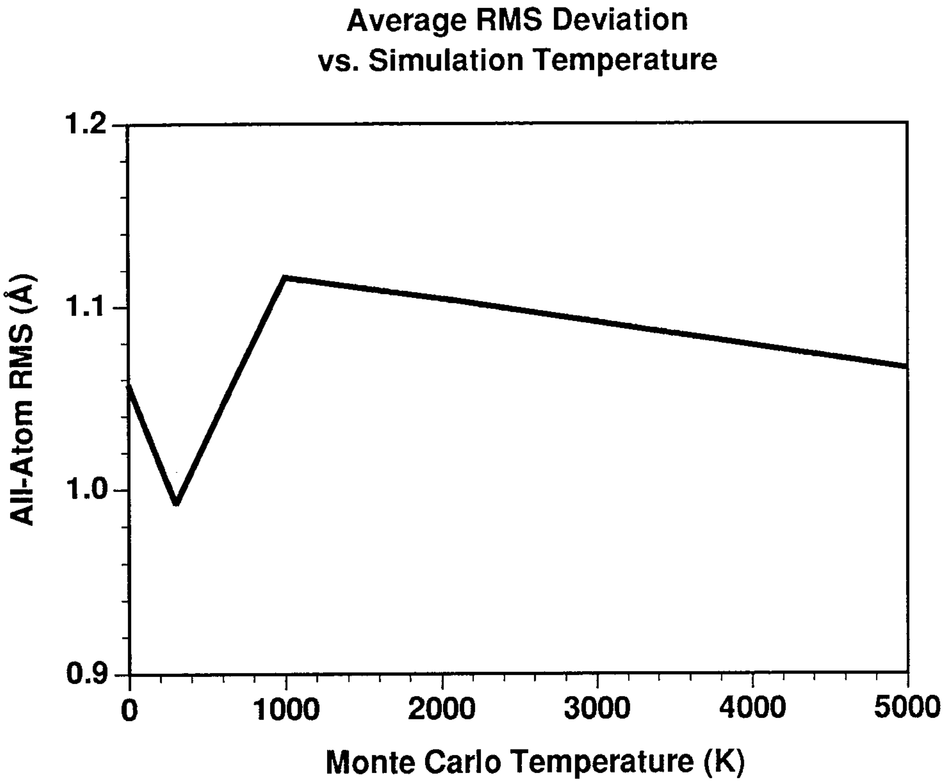


Figure 5.6. Average all-atom rms deviations for sidechain Monte Carlo simulations of crambin at different temperatures.

Apparently, the much greater acceptance rate of new structures does translate directly into the creation of more low-energy conformations. The simulations at 300 K were more consistently accurate, so this temperature was used in the simulations reported below. Table 5.3 lists the values used for Phase 1 and Phase 2 simulations reported in the following sections.

III. Results

III.A. Crambin

The values listed in Table 5.3 were used in an attempt to reproduce the structure of crambin using the C_α coordinates from the crystal structure[1]. Twenty different

Variable	Phase 1	Phase 2
Spacing	10°	10°
Temperature	1000 K	300 K
Steps	200	1000
Constraint	100 (kcal/mol)/Å^2	–
Pulse	3	–

Table 5.3. Values used for production runs of the C_α Builder.

backbone conformations were generated by using different random numbers to control the selection of ϕ, ψ dihedrals as well as to determine which conformations would be accepted and which rejected. The conformational energy of the backbone, the rms deviations in backbone atoms and ϕ, ψ dihedrals from each of these structures is listed in Table 5.4, ranked by energy. The average backbone rms deviation for these 20 simulations was 0.527 Å, in close agreement with the previous result of 0.520 Å mentioned in the preceding section. The average all-atom deviation was 1.696 Å. It is apparent that there is only a small correlation between the backbone energy and the rms fit to the crystal structure backbone. The backbone of the crystal structure itself has an energy of 759.8 kcal/mol, higher than 12 of the 20 model conformations. This is likely to be due both to errors in the crystal structure and in the limitations of the forcefield approach: no forcefield can be optimal for every crystal structure, even when such factors as crystal packing and solvation are considered. Nevertheless, in cases where the crystal structure is unknown, the backbone energy is the best criterion for selecting model structures. Other possible selection criteria, including C_α constraint energy and total energy including sidechain atoms, had even worse correlation with the deviation in the backbone coordinates.

Energy (kcal/mol) and RMS Deviations from Phase 1					
Energy	RMSB (Å)	RMSD (°)	Energy	RMSB (Å)	RMSD (°)
335.3	0.494	22.05	597.6	0.481	31.15
338.4	0.430	19.43	652.7	0.572	33.77
363.3	0.543	25.75	796.9	0.588	33.13
363.8	0.495	26.00	797.1	0.430	21.49
366.4	0.515	28.69	822.7	0.498	27.47
376.9	0.576	29.40	850.3	0.505	27.74
377.6	0.545	29.88	872.4	0.595	33.38
393.2	0.582	32.96	1445.3	0.589	32.08
465.5	0.668	42.27	5266.2	0.447	27.67
577.1	0.483	28.94	5700.5	0.513	34.44

Table 5.4. The energy, rms deviation in backbone atoms (RMSB), and rms deviation in ϕ, ψ dihedrals (RMSD) for each of the 20 backbone conformations generated for crambin.

Energy (kcal/mol) and All-Atom RMS (Å) from Phase 2							
Energy	RMS	Energy	RMS	Energy	RMS	Energy	RMS
668.1	1.386	767.8	1.449	971.7	1.102	2225.6	1.272
669.2	1.367	793.9	1.430	1039.0	1.337	2576.8	1.393
688.2	1.132	801.3	1.278	1074.0	1.519	3023.2	1.486
691.6	1.259	823.0	1.243	1111.8	1.153	3077.1	1.391
706.6	1.313	860.7	1.297	1304.6	1.332	3105.8	1.487
757.3	1.170	947.9	1.111	1696.1	1.468	3334.5	1.221
						3383.6	1.484

Table 5.5. The energy and rms deviation in atomic coordinates for each of the crambin models produced by the PGMC C_α Builder.

The five lowest-energy backbone conformations from Phase 1 were used as a starting point for Phase 2. For each of the five backbone conformations, five Phase 2 simulations were carried out, again using different random numbers to produce different results. Each simulation involved 1000 Monte Carlo steps using 10° probability grids and a simulation temperature of 300 K. The 25 conformations produced are listed in Table 5.5. Again, there is only a small correlation between energy and rms fit to the crystal structure. Nevertheless, the fits are quite good, with an average rms deviation from the crystal structure of 1.323 Å. All five backbone conformations were represented throughout the list of of all-atom conformations, so the backbone energy was not the determining factor in the overall energy.

The best energy conformation from Phase 2 was chosen as the "model" conformation of crambin for detailed comparison to the "true" structure, the crystal structure[1]. Table 5.6 gives a breakdown of the rms deviation of the crambin model for different regions of the protein. Some of this information is shown graphically in

Region	Residues	Backbone	ϕ, ψ	All Atoms	Sidechains
All	1–46	0.543 Å	25.8°	1.386 Å	2.010 Å
No C-term	1–44	0.361 Å	23.0°	1.248 Å	1.841 Å
Helix 1	7–19	0.209 Å	13.7°	1.658 Å	2.347 Å
Helix 2	23–30	0.394 Å	22.3°	1.026 Å	1.454 Å
Sheet 1	1–4	0.417 Å	22.3°	1.146 Å	1.771 Å
Sheet 2	32–35	0.315 Å	19.2°	1.070 Å	1.530 Å
Turn 1	41–44	0.571 Å	32.9°	1.853 Å	1.853 Å
N-terminus	1–2	0.559 Å	31.1°	1.184 Å	1.688 Å
C-terminus	45–46	1.872 Å	67.9°	3.175 Å	4.682 Å
Coil	others	0.373 Å	28.5°	0.511 Å	0.728 Å

Table 5.6. RMS deviations for different regions of the crambin model.

Figure 5.7, where the backbone rms deviation of each residue is shown. The largest deviations occur at the carboxy terminus, where residues 45 and 46 are very poorly modeled. If these two residues are excluded, the backbone rms deviation drops from 0.543 Å to 0.361 Å. The carboxy terminal residues are generally the worst modeled residues because there are fewer constraints on the structure: they usually lie on the surface of the protein where there are fewer inter-residue contacts and there is no $l+1$ C_α to constrain the orientation of the terminal carboxyl group. In the crambin model, the Asn 46 sidechain and the terminal carboxyl group have reversed positions, giving rise to a large error even though the chemical significance is small. The backbone rms deviation is fairly consistent throughout the rest of the protein, with 34 of the 46 residues having deviations in the 0.1–0.4 Å range. The lowest backbone deviations are in the residues of the long α helix, Helix 1, where the de-

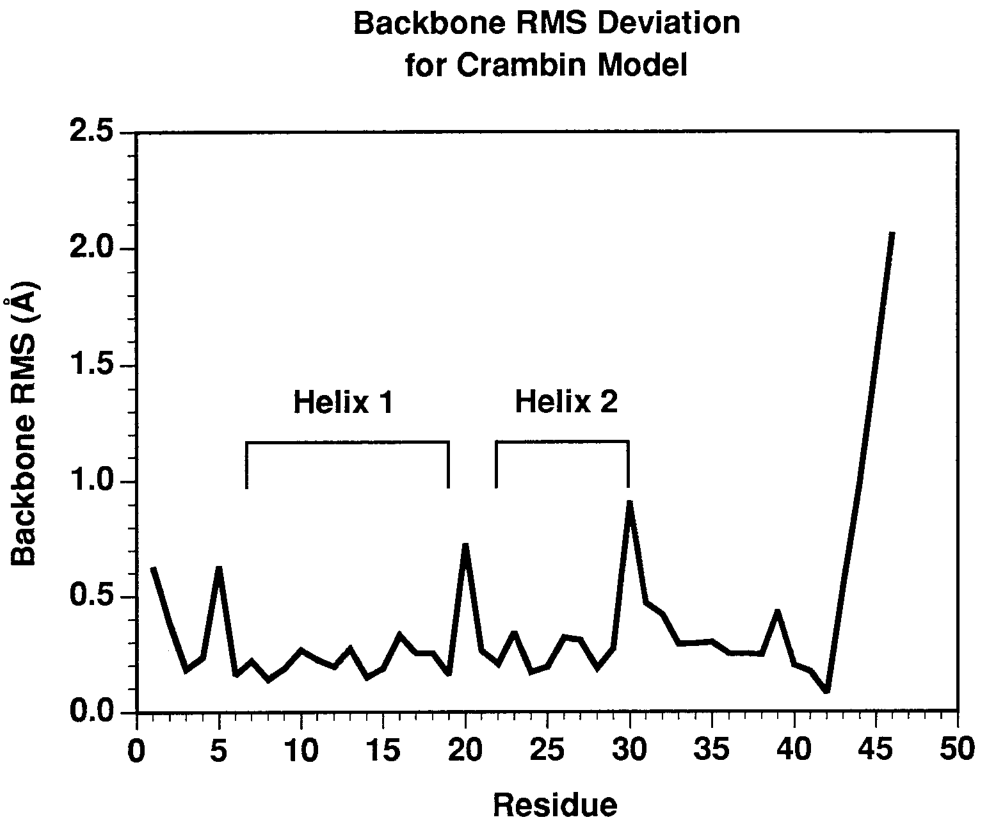


Figure 5.7. Backbone rms per residue for crambin model.

viation in atomic coordinates is 0.209 Å, and the deviation in ϕ and ψ dihedrals is only 13.7°. The deviations are equally low (0.232 Å and 13.1°) for the first seven residues of Helix 2. However, the last residue in the helix starts a turn, and is poorly modeled. In general, the turn regions before and after α helices are the most poorly modeled residues other than those at the C-terminus. This is very apparent from both the graph in Figure 5.7 and the picture in Figure 5.9. These regions (particularly residues 5, 20, and 30) have nonstandard ϕ, ψ values which have very low probabilities in the ϕ, ψ probability grids. No ϕ, ψ probability grids were specifically developed for turn regions, but these might prove very valuable.

The sidechain modeling is not as successful as the backbone modeling, with the average deviation in atomic coordinates being near 2.0 Å. This is not at all surprising, since each peptide unit in the polypeptide backbone is constrained at both ends by the positions of two consecutive C_α's while the sidechains are usually constrained only by their attachment to a single C_α. The constraints on the sidechain conformations are primarily steric in nature: sidechains in the interior of a protein can have considerable steric overlap and their conformations must be correlated to allow for closest packing. The rms deviation for the atomic coordinates may not be the best indication of modeling success, since it will be heavily weighted toward any poorly modeled large sidechain such as arginine. A better measure is the deviation in sidechain dihedral angles, χ, defined as the absolute value of the difference between the dihedral in the model and in the crystal structure. The deviations in χ^1 are shown for the crambin model in Figure 5.8. Most χ^1 dihedrals have high probabilities at 60°, -60°, and 180°, so deviations would be expected to be near 0° or 120°. Of the 37 χ^1's in crambin, 24 have deviations less than 30°, and 11 have deviations between 90° and 150°. Therefore, only two have deviations between 30° and 90°. It is important to note that five of the 11 poorly modeled sidechains are cystein residues involved

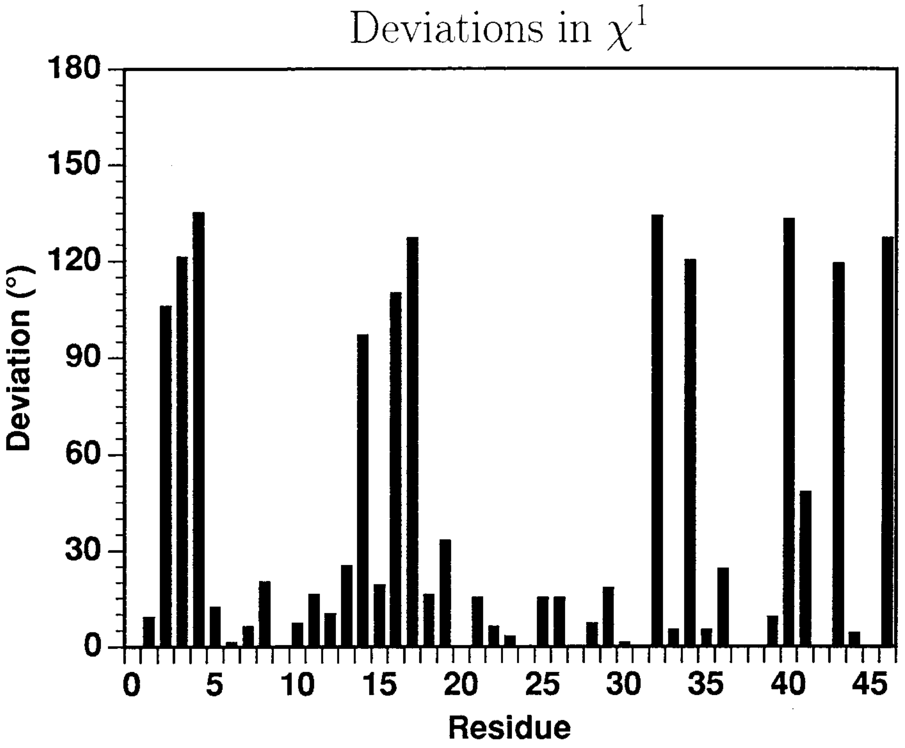


Figure 5.8. The deviation in χ^1 for each residue of the crambin model.

in disulfide bridges in the crystal structure. The C_α Builder does not currently predict the presence of disulfide bridges, so the disulfide bond is not included in the Monte Carlo energy evaluations. Such a term could be included and would certainly improve the results for these residues. RMS deviations for the different backbone and sidechain dihedrals are shown in Table 5.7. Although the sidechain dihedrals are not as well modeled as the backbone, the results are not discouraging with respect to other methods. As discussed below, our method provides results for flavodoxin χ dihedrals as good or better than other methods, and these results for crambin are even better.

The differences between the crambin model and the crystal structure are shown in detail in Figures 5.9, 5.10, and 5.11. Figure 5.9 shows the model and crystal structure backbones for the entire protein. For most of the protein, it is very

		Deviation		
Dihedral	Number	RMS	$< 30°$	$> 90°$
ϕ	45	22.3°	86.7%	0.0%
ψ	45	28.8°	75.6%	2.2%
ω	45	5.4°	100.0%	0.0%
χ^1	37	69.6°	62.2%	29.7%
χ^2	21	84.5°	38.1%	28.6%
χ^3	8	75.1°	25.0%	37.5%
χ^4	7	34.9°	71.4%	0.0%
χ^5	2	9.8°	100.0%	0.0%

Table 5.7. The rms deviations in various types of dihedrals for the crambin model and the percentage of each type of dihedral with deviations less than 30° or more than 90°.

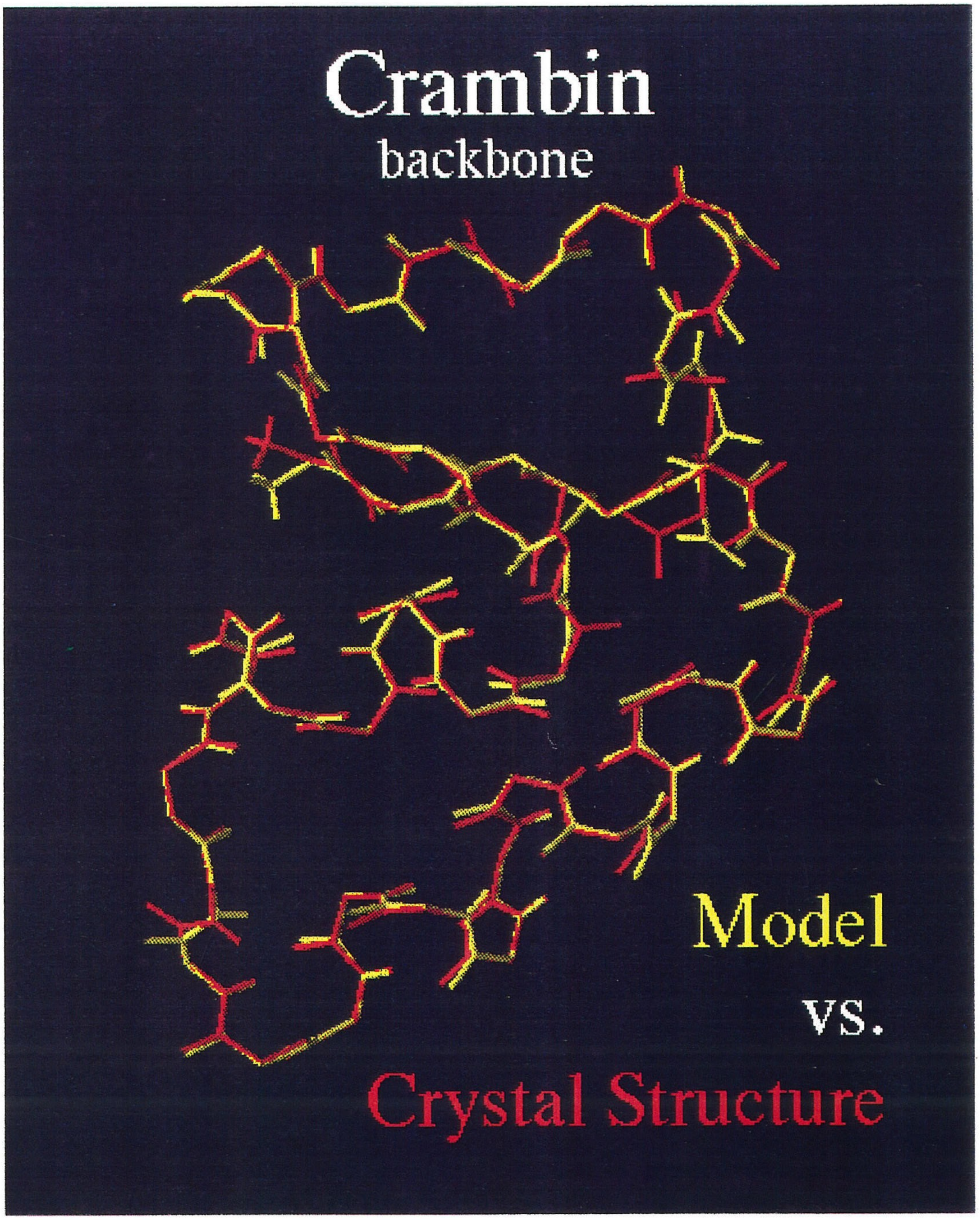


Figure 5.9. The peptide backbone of the model and crystal structures of crambin. The rms deviation is 0.538 Å.

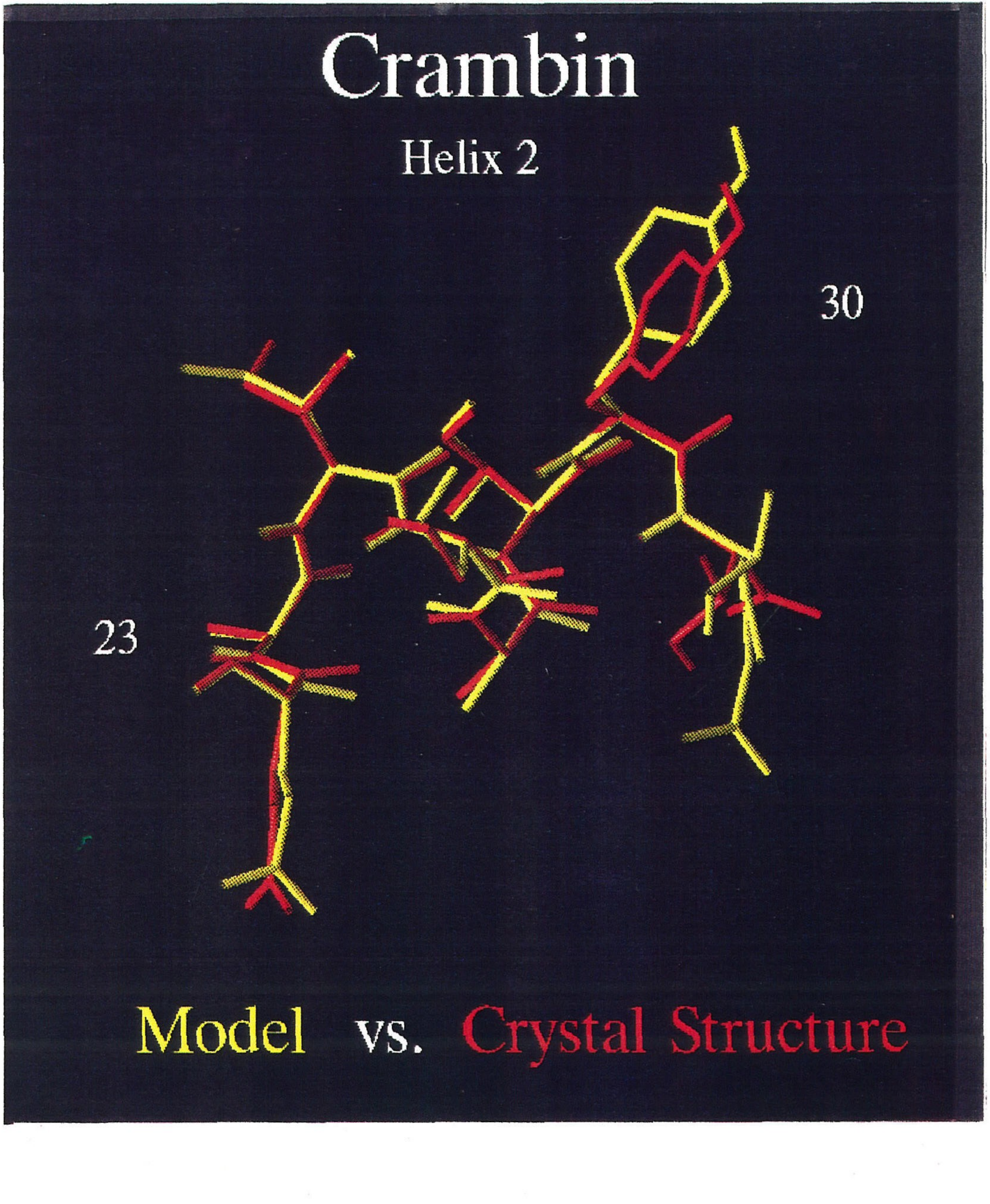


Figure 5.10. A comparison of helix 2 (residues 23 to 30) in the model and crystal structures. The rms deviation is 1.026 Å for all atoms and 0.394 Å for the backbone atoms.

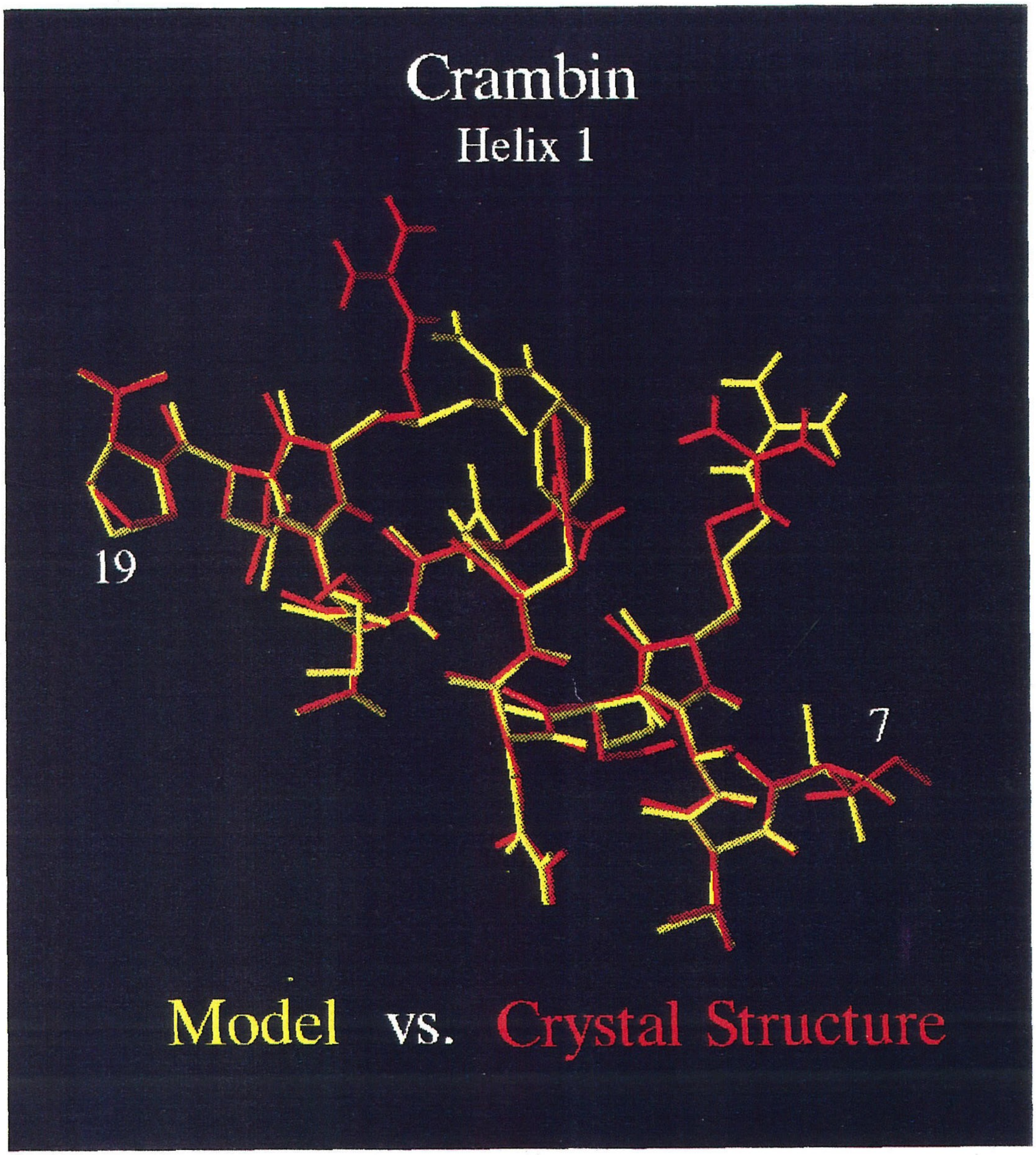


Figure 5.11. Helix 1 (residues 7 to 19) in the model and crystal structures. The rms deviation is 1.658 Å for all atoms and 0.209 Å for the backbone atoms.

difficult to distinguish between the two structures. Only in the turn regions after the two helices is the difference readily apparent. The two following figures show the all-atom structures of the two helices of crambin. Helix 2, shown in Figure 5.10, is very well modeled, with an rms deviation of 1.03 Å for all atoms. In terms of the all-atom deviation, it is the best modeled region of the protein (see Table 5.6). The picture shows this quite well, with both sidechain and backbone atoms showing little difference between the two structures, except for Thr 30 on the C-terminal (right) end of the helix. As explained above, this residue begins a turn in the backbone conformation and is poorly sampled during the Phase 1 backbone Monte Carlo. The Helix 1 backbone, in contrast, is modeled quite well throughout its length, including Pro 19 at its C-terminal (left) end. However, Helix 1 has many large sidechains which are difficult to model. Large errors can be seen in Asn 14 and Arg 17. The latter has a particularly large impact on the rms deviation. Excluding Arg 17, the crambin model has an rms deviation of 1.207 Å, rather than 1.386 Å. However, this incorrect conformation of Arg 17 may be energetically more favorable than other conformations more similar to the crystal structure. Of the next four lowest-energy conformations listed in Table 5.5, all five have more native-like conformations of Arg 17, but all are higher in energy.

The crambin model illustrates several general findings for simulations using the PGMC C_α Builder. The lowest-energy structures from Phases 1 and 2 are usually among the best models built, but are rarely the very best. Regardless, the backbone models from Phase 1 are consistently good, and almost any one of them provides an acceptable model of the true backbone. The model backbones are especially good in regions of regular secondary structure such as helices and sheets, but rather poor in turn regions. These results are obtained consistently in different simulations. There is a much larger variation among the results from Phase 2. This may be due to the constraints of time; the number of 1000 Monte Carlo steps was selected largely in order to keep the simulation time below ten minutes, so that large numbers of different conformations could be evaluated. Better and more consistent results might be obtained by substantially longer calculations. Nevertheless, between 40% and 60% of χ^1 dihedrals are modeled correctly.

III.B. Larger Proteins

Although the variables discussed in the previous section could be tuned to specific problems, the same values were used for six different proteins, ranging from the 46 residue crambin to myoglobin, which has 153 residues. These proteins are listed in Table 5.8. The proteins have widely different structures, as indicated by the percentages of their secondary structures which are α-helical and β sheet. Four of the six proteins are included in the subset of crystal structures used to develop the Monte Carlo probability grids (see Section 4.II). Of the other two, the flavodoxin structure used is merely a different form (oxidized) than the one used in the dataset (semiquinone), while the plastocyanin studied is homologous, but not identical, to the structure used in the dataset.

For each of these six proteins, the C_α coordinates from the listed crystal structure were used to rebuild the backbone conformation twenty times, as described in the preceding sections for crambin. In each case, all prosthetic groups, such as the myoglobin heme, were removed from the crystal structure, as were any cofactors or solvent molecules. Each of the twenty backbone conformations was compared to the crystal structure and the results were analyzed. Table 5.9 lists the average rms deviation as well as the standard deviation (σ) for the twenty structures. Also listed are the rms deviations for the lowest energy conformation and the conformation with the best fit. Again, it is seen that the lowest energy conformation is never the one

Protein	PDB	Ref.	Size	% Helix	% Sheet
Crambin	1crn	[1]	46	45.7	17.4
BPTI	5pti	[13]	58	27.6	25.9
Plastocyanin	7pcy	[14]	98	7.1	58.2
Ribonuclease A	7rsa	[15]	124	26.7	46.8
Flavodoxin	3fxn	[16]	138	37.7	26.8
Myoglobin	1mbd	[17]	153	79.1	0.0

Table 5.8. The proteins modeled by the PGMC C_α Builder. The reference crystal structure is given along with the number of residues in the protein and the percent of these which are in α helices and β sheets.

with the best fit to the crystal structure. However, it is encouraging that the lowest energy conformation was better than average for five of the six proteins.

Comparing Tables 5.8 and 5.9, it is clear that the size of the protein has little effect on the accuracy of Phase 1. In fact, the largest protein, myoglobin, is consistently modeled most accurately. This is not surprising considering the crambin results, where the average backbone deviations was approximately 0.2 Å for helical residues. The protein myoglobin, with almost 80% of its residues in α helices, is greatly benefited by the accuracy with which the method models helices. Plastocyanin is also modeled relatively well, even though it has a β sheet protein, with little helical content. The large β sheet content is probably also a favorable factor, as these conformations are also very well represented by the probability grids. It is proteins such as bovine pancreatic trypsin inhibitor (BPTI), with only about 50% α helix and β sheet content, which are relatively poorly modeled, though even for this case the rms deviation is badly distorted by poor modeling of the C-terminal residues. The average rms deviation for residues 1-54 is 0.501 Å.

Crystal	Backbone RMS Deviation			
Structure	Average	σ	Best E	Best Fit
1crn	0.527	0.062	0.494	0.430
5pti	0.610	0.065	0.582	0.506
7pcy	0.550	0.048	0.602	0.470
7rsa	0.601	0.052	0.551	0.530
3fxn	0.593	0.050	0.577	0.509
1mbd	0.453	0.033	0.451	0.366

Table 5.9. The results from Phase 1 constructions of the backbone conformations of several proteins.

Phase 2 simulations were carried out on flavodoxin and plastocyanin, building five complete structures from each of the top five backbone conformations from Phase 1. The same parameters were used for these simulations as were used for Phase 2 simulations of crambin. The energy and all-atom rms deviation for each of the 25 conformations was evaluated and the results were analyzed. Table 5.10 lists the results for these two proteins, along with those for crambin. Unlike Phase 1, the results for Phase 2 are highly dependent on the size of the protein, with the average deviation increasing substantially for larger proteins. In Phase 1 simulations, each residue was sampled the same number of times, regardless of the size of the protein. In the Phase 2, simulations, however, each simulation involved a total of 1000 Monte Carlo steps. For crambin, this meant that the average residue was varied 27 times during the simulation (alanine and glycine residues are not affected). For plastocyanin, the 73 relevant dihedrals were sampled an average of 14 times; for flavodoxin, the average was 8.5. Clearly, the sidechains of flavodoxin are not being adequately sampled. Unfortunately, the cpu time required for the simulations also

Crystal	All-Atom RMS Deviation		
Structure	Average	Best E	Best Fit
1crn	1.323	1.386	1.102
7pcy	1.483	1.398	1.299
3fxn	1.796	1.663	1.607

Table 5.10. The results from Phase 2 constructions of the sidechains of crambin, plastocyanin, and flavodoxin.

grows substantially as the size of the protein grows. While the 1000 Monte Carlo steps take seven minutes for crambin, they require nearly 20 minutes for plastocyanin and over 40 minutes for flavodoxin. Therefore, it is computationally expensive to increase the number of steps for flavodoxin. Nevertheless, the results for flavodoxin are comparable to or better than published results using other methods.

The lowest energy conformation of flavodoxin was chosen for comparison with other methods. This protein has become a standard test case for published methods of building all-atom conformations from C_α coordinates. This includes both methods based on molecular mechanics[8] and those using database searches to determine conformations for multiple-residue peptide fragments from the protein[7, 9]. Table 5.11 lists several measures of the accuracy of these models. "Peptide flips" refer to the number of peptide units (the planar backbone unit between the C_α coordinates) which are rotated by more than 90° degrees from the crystal structure. This occurs seven times in our model, compared to only 5 and 4 times in the fragment-matching methods of Reid and Thornton[7] and Holm and Sander[9]. This is the only measurement by which the PGMC method appears deficient. In most of the other measures, the PGMC method is comparable to, or better than, the other published methods. The PGMC C_α Builder is currently not quite as accurate as the

	Reference			PGMC
Atoms	[7]	[8]	[9]	Model
RMS All Atoms (Å)	1.73	1.64	1.57	1.66
RMS Main Chain (Å)	0.57	0.49	0.48	0.57
RMS Side Chain (Å)	2.41	–	2.19	2.31
Peptide Flips	5	–	4	7
Correct χ^1 (%)	40	–	44	41
Correct χ^1, χ^2 (%)	17	–	25	24

Table 5.11. A comparison of the results for flavodoxin vs. other methods. "Correct" refers to dihedrals predicted to within 20° of their crystal structure values.

method of Holm and Sander[9], but is comparable in most respects, even though it is based on a more general approach to protein modeling: Probability Grid Monte Carlo. The PGMC method is applicable to unconstrained systems as well as those constrained by *a priori* knowledge of the C_α coordinates.

IV. The C_α Forcefield

In recent years, lattice-based methods have become increasingly popular tools for theoretical studies of protein folding[3,19-21]. In these calculations, a protein is represented by points on a 2-D or 3-D lattice. Typically, each amino acid occupies a single lattice site[3], but some methods use other models, such as one backbone and one sidechain site per residue[19]. Conformations of a protein are represented by chains traced through the lattice, with consecutive residues occupying adjacent sites. Adjacent sites can also be filled if the chain folds back upon itself. Because positions are limited to points on a lattice, energy calculations are extremely fast.

Valence terms such as bond stretches can be eliminated entirely, since there are only a few possibilities. In addition, nonbonded forces can be calculated rapidly because distances between lattice sites are known in advance. Therefore, lattice simulations greatly speed the evaluation of a protein's conformational space in two ways: the size of conformational space is decreased by allowing only lattice conformations and evaluation of each conformation is greatly decreased through the use of simplified energy terms.

Despite the simplifications of the lattice methodology, there is still a huge number of possible conformations available to even a small protein. And while energy functions may give favorable values to the "correct" structure (the lattice conformation most closely resembling the native structure)[3], they are rarely sufficiently accurate to predict it outright. In order to evaluate lattice conformations more fully, and to enable construction of all-atom protein conformations from lattice models, we have developed a "C_α Forcefield" (C_αFF) for use in molecular mechanics simulations of C_α models of proteins. This forcefield is used to optimize lattice conformations, enabling them to have conformations more like true proteins. These optimized C_α conformations can then be used as templates for the PGMC C_α Builder. This process, termed the "Hierarchical Protein Folding Strategy" (HPFS), is shown in Figure 5.12. The method has a hierarchy of refinement levels:

1. The lattice C_α-only model.

2. The C_α model optimized using the C_αFF.

3. Backbone atoms added by the C_α Builder (Phase 1).

4. Sidechain atoms added by the C_α Builder (Phase 2).

5. All-atom conformation optimized by full Cartesian energy minimization.

The final level of optimization is not shown in Figure 5.12. It involves energy minimization of the all-atom protein conformation using a forcefield such as DREIDING. At this point, solvent molecules may be introduced into the system to reflect the protein environment more accurately.

The simple C_α Forcefield which we have developed for lattice structure optimization has valence terms, only. Nonbonded interactions, such as van der Waals and electrostatic terms, are not included in the forcefield. Future enhancements of the C_αFF will include such terms and will be amino acid-specific. The current implementation, however, treats all amino acid types equally, and has the three terms:

$$V_b(b_{i,i+1}) = \frac{1}{2}K_b(b_{i,i+1} - b_{eq})^2, \tag{5.2}$$

$$V_\zeta(\zeta_i) = \frac{1}{2}K_{\zeta,\alpha\beta}(\zeta_i - \zeta_{eq,\alpha\beta})^2, \tag{5.3}$$

and

$$V_\gamma(\gamma_i) = \frac{1}{2}K_{\gamma,\alpha\beta}(\gamma_i - \gamma_{eq,\alpha\beta})^2. \tag{5.4}$$

The bond energy, V_b, is summed over all $C_\alpha(i)$-$C_\alpha(i+1)$ distances $(b_{i,i+1})$, while angle and torsion terms are summed over all virtual angles, ζ_i, and virtual dihedrals, γ_i, as defined in Figure 5.2. The $\alpha\beta$ subscripts denote that different angle and torsion force constants (K_ζ, K_γ) and equilibrium geometries $(\zeta_{eq}, \gamma_{eq})$ are used for α helix and β sheet conformations. These bond and angle terms are commonly found in atomic forcefields, but the torsion term is unlike a typical torsion forcefield, which uses an expansion of cosine terms (see Equation (1.6)). The present form was used because the virtual dihedrals do not have probability minima or maxima at $\gamma = 0$, so no cosine expansion could reproduce the known distribution. Unfortunately, problems arise for calculating atomic forces when $(\gamma_i \approx 180°)$, so alternate functional forms are being investigated.

Parameters for the C_α Forcefield have been determined from analyses of the C_α

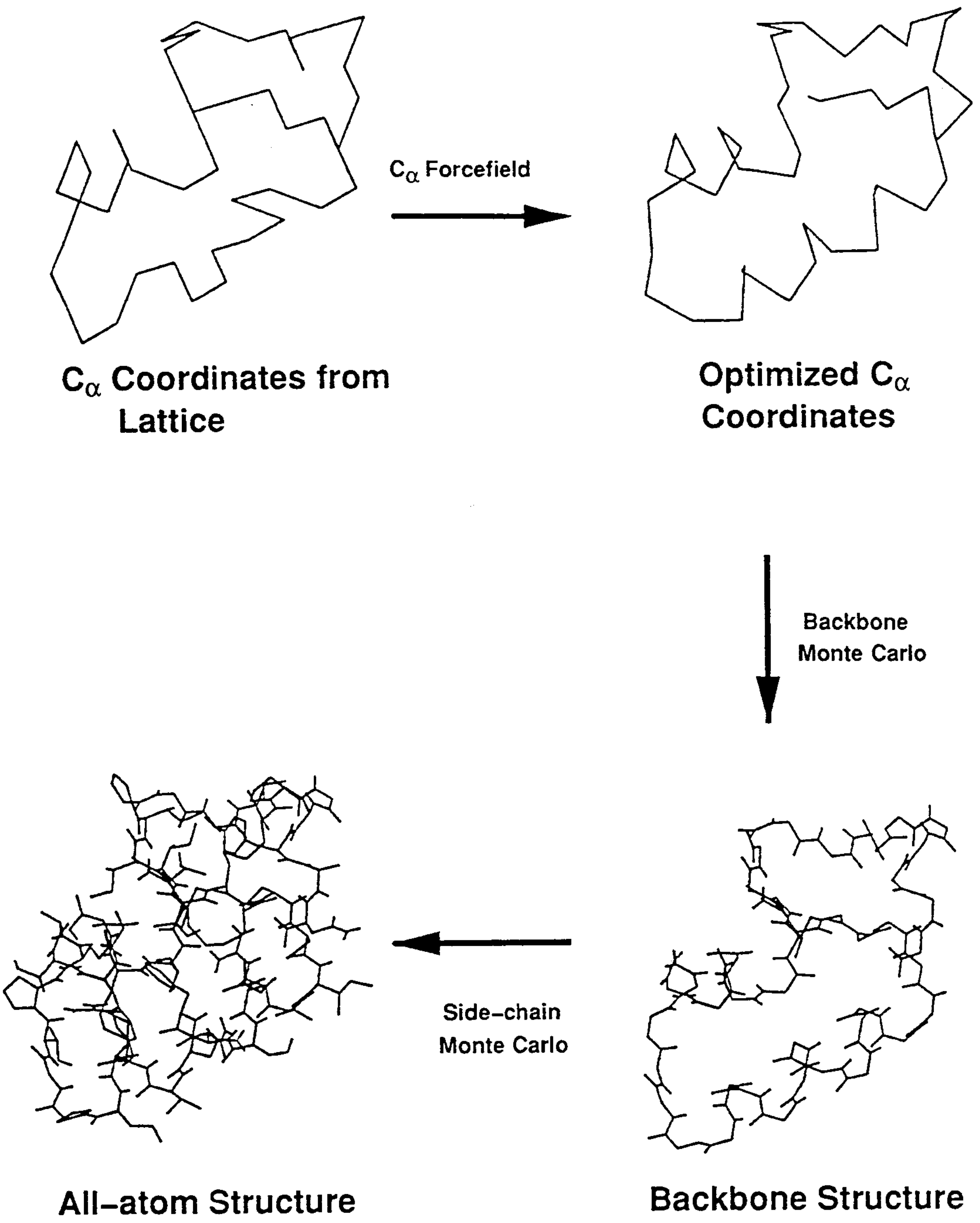


Figure 5.12. The Hierarchical Protein Folding Strategy, which converts lattice C_α coordinates into all-atom protein conformations.

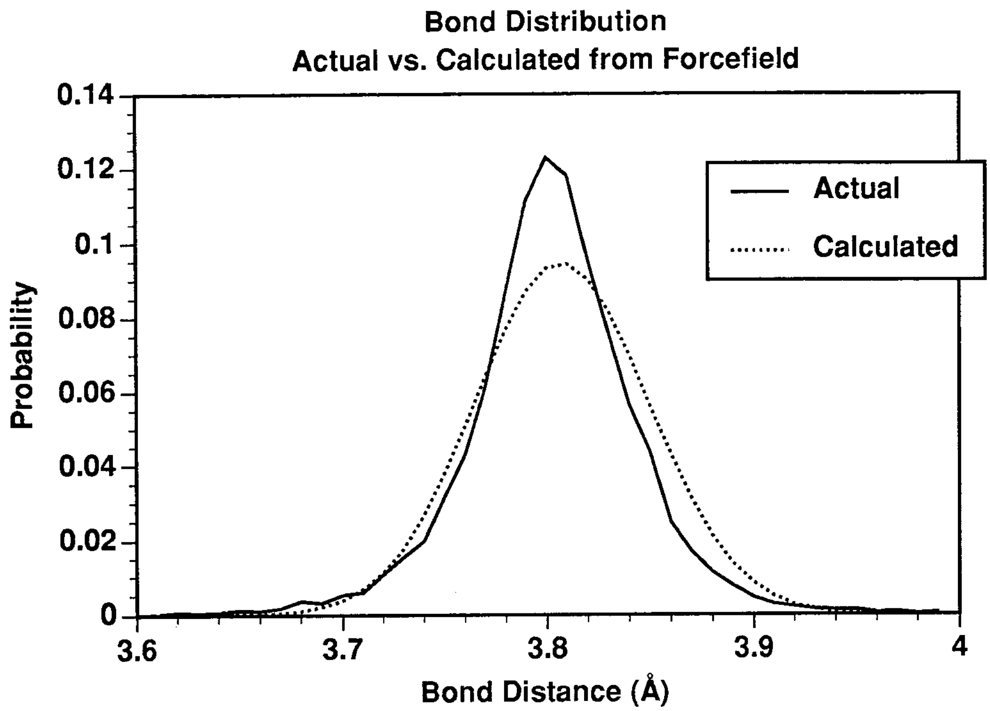


Figure 5.13. The probability distribution (0.01 Å resolution) for C_α-C_α bonds. The actual distribution is compared to that derived from the bond-stretch term of the C_αFF.

coordinates in the protein structures of the Brookhaven PDB. A subset of 64 of the protein structures was used. This "H64" dataset was also used for the development of ϕ, ψ and χ grids and is described in detail in Section 4.II. Figure 5.13 shows the distribution of C_α(i)-C_α(i+1) distances in the H64 dataset, using a 0.01 Å interval to determine probabilities. From this distribution, an average, b_{eq}, and standard deviation, σ, can be calculated. The average is used directly in Equation (5.2), while the force constant is derived from

$$K_b = \frac{kT}{\sigma^2}, \tag{5.5}$$

where k is the Boltzmann constant and T is the temperature. Using these parameters in Equation (5.2) gives a probability distribution very similar to that derived from

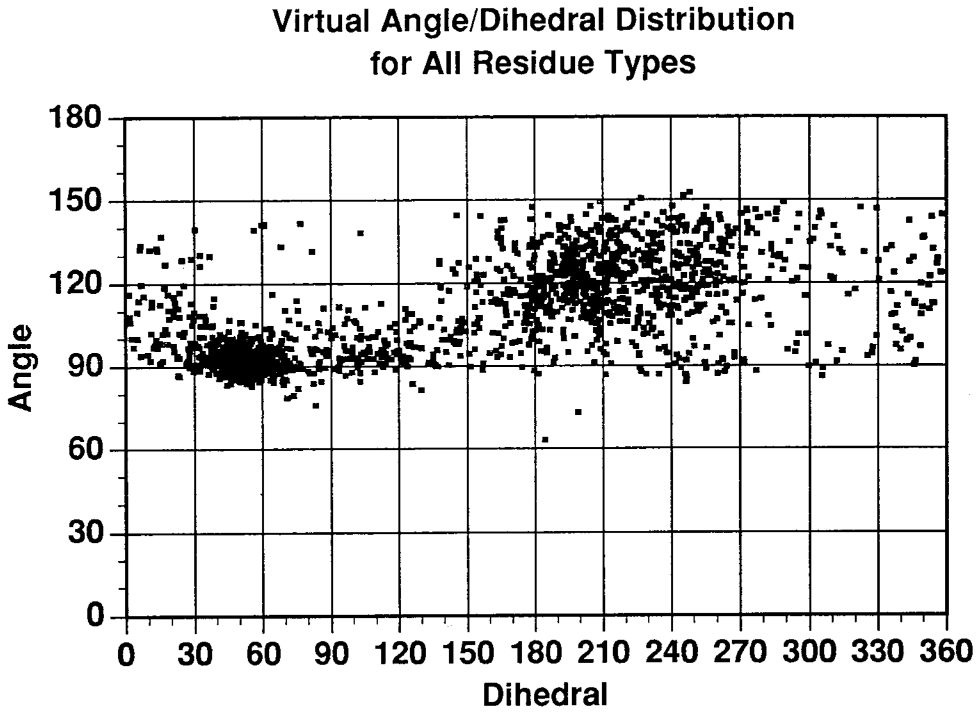


Figure 5.14. 2000 virtual angle, dihedral (ζ, γ) values.

the crystal structure. The probability distribution is determined from:

$$P(b) = \frac{e^{-V_b(b)/kT}}{\int_{b=0}^{\infty} e^{-V_b(b)/kT} db}. \tag{5.6}$$

Replacing the integral by a sum over 0.01 Å intervals gives the probability distribution in Figure 5.13.

Similar analyses can be made for the virtual angles (ζ) and dihedrals (γ). However, it should first be noted that there are strong ϕ, ψ propensities in protein backbones which lead to corresponding ζ, γ correlations. This is clearly seen in Figure 5.14, where 2000 randomly selected ζ, γ values from the H64 dataset are plotted. There are two high density regions. This can also be seen by binning the data. Figure 5.15 shows probability grids derived from determining the fraction of all points in the region ($\zeta_0 \pm 7.5°$, $\gamma_0 \pm 7.5°$) for ζ_0 and γ_0 intervals of 15°. There are two distinct peaks, which correspond to α helix and β sheet regions, as is made evident by the

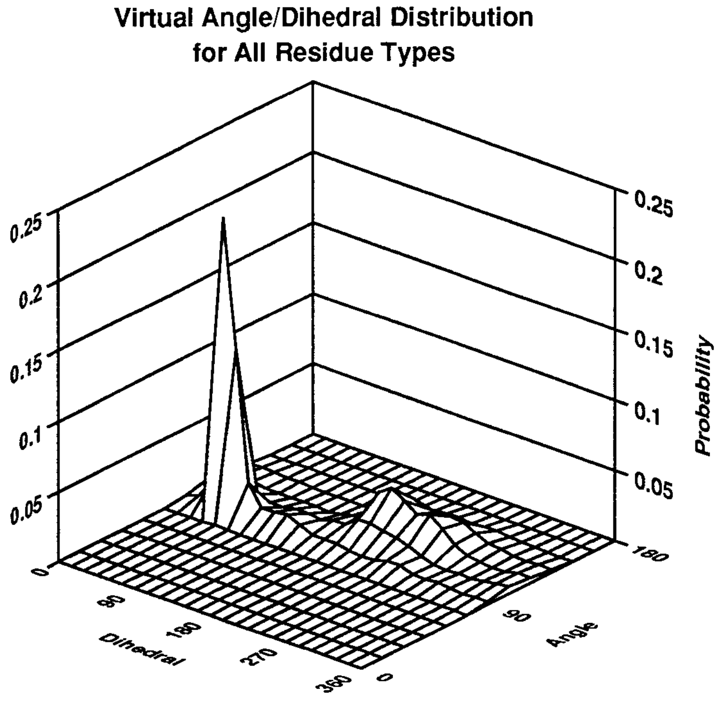
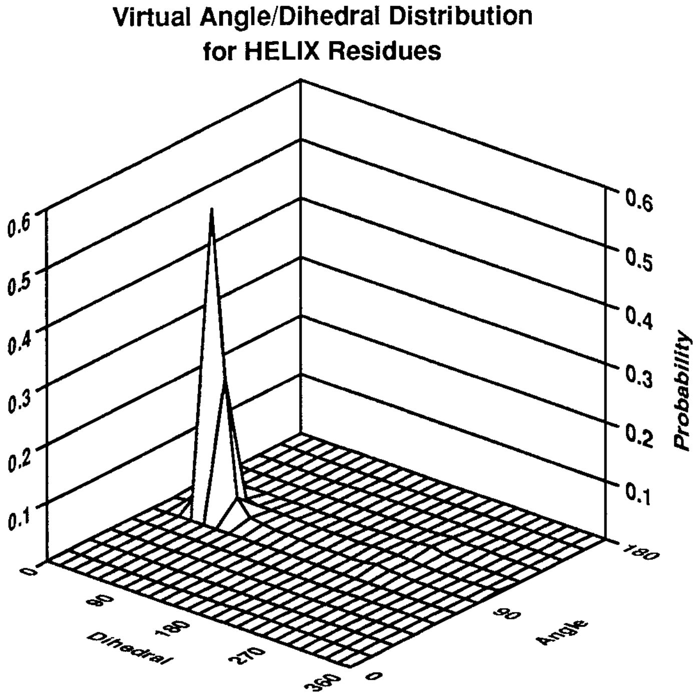


Figure 5.15. The ζ,γ probability grids for the entire H64 dataset, using 15° bins.

probability grids for HELIX and SHEET residues in Figure 5.16. The high probability regions for the two major secondary structure types are listed in Table 5.1. These regions account for 39.7% (α helix) and 34.7% (β sheet) of all ζ,γ points. The ζ,γ pairs which fell within the α helix or β sheet regions were used to calculate average values and standard deviations of ζ and γ for each of these regions. These, in turn, were used as equilibrium geometries and to calculate force constants as was done for bond lengths (Equation (5.5)). All such parameters for the C_αFF are listed in Table 5.12. Note that the α force constants are significantly higher than the β ones, reflecting the much sharper peak in the α helix region of the ζ,γ probability distribution.

This forcefield described above was used to optimize lattice conformations for several proteins. These lattice conformations were generated by finding the conformations on a face-centered cubic (fcc) lattice which best matched the crystal



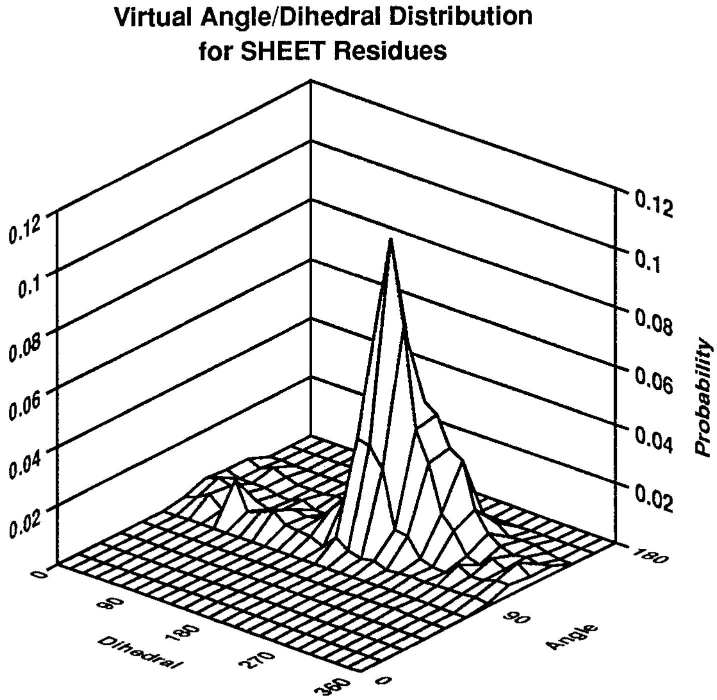


Figure 5.16. The ζ, γ probability grids for HELIX and SHEET residues in the H64 dataset.

Parameter	Equil. Geom.	Force Constant
b	3.807 Å	335 (kcal/mol)/Å^2
ζ_α	92.145°	117.15 (kcal/mol)/rad^2
γ_α	50.837°	25.07 (kcal/mol)/rad^2
ζ_β	121.976°	17.25 (kcal/mol)/rad^2
γ_β	-144.495°	2.20 (kcal/mol)/rad^2

Table 5.12. Equilibrium geometries and force constants in the C$_\alpha$FF.

Protein	PDB	Ref.	Residues	Lattice fit	Minimized
Crambin	1crn	[1]	46	1.93 Å	1.50 Å
BPTI	4pti	[22]	59	1.79 Å	1.66 Å
Cobratoxin	1ctx	[23]	71	1.99 Å	2.18 Å
Calmodulin	1cln	[24]	145	2.15 Å	2.02 Å

Table 5.13. Best-fit fcc lattice conformations of several proteins, before and after minimization with the C$_\alpha$FF.

structures. These conformations were then optimized by conjugate-gradients minimization using the C$_\alpha$FF. As shown in Table 5.13, C$_\alpha$ coordinates after minimization by the C$_\alpha$FF are usually much better than lattice conformations. Figure 5.12 displays this improvement more dramatically, by showing the lattice and minimized structures of crambin. Clearly, the lattice constraint imposes unnatural geometries on the C$_\alpha$ configuration, a problem remedied by the C$_\alpha$FF.

The utility of the C$_\alpha$FF is further displayed by the results in Table 5.14. In these simulations, several C$_\alpha$ coordinate sets for crambin were used as templates for the PGMC C$_\alpha$ Builder. The results are shown in the table after the final all-atom

Origin of C_α's	RMS-All	RMS-BB	RMS-C_α
Crystal Structure	1.39	0.50	0.20
Crystal Minimized	1.80	1.00	0.88
Best Fit to Lattice	3.04	1.99	1.96
Lattice Minimized	2.48	1.57	1.50

Table 5.14. Results from building all-atom conformations from various C_α conformations of crambin using the PGMC C_α Builder.

conformation is minimized with energy minimization using DREIDING. Naturally, the C_α coordinates from the crystal structure, itself, form the best template for the C_α Builder. Minimizing the crystal structure C_α atoms with the C_αFF causes them to diverge from their true coordinates, but a good model, with a backbone RMS deviation of only 1.0 Å, can still be built. Use of the lattice conformation, however, produced poor results, with a backbone RMS deviation of nearly 2.0 Å. The results are significantly improved through the use of the C_αFF, which reduces the error per atom by almost 0.5 Å.

The C_αFF is, therefore, able to assist significantly in the building of all-atom conformations of proteins from lattice models of their C_α coordinates. Other uses may include the evaluation of different lattice models by energy evaluation and/or minimization. This may ease the difficult task of determining which lattice conformations are native-like. In addition, future enhancements of the C_α Forcefield will include nonbond forces as well as residue masses, thereby allowing for the possibility of extremely fast molecular dynamics simulations of a C_α protein model.

V. Conclusions

Probability Grid Monte Carlo provides a new method for predicting all-atom protein conformations from C_α coordinates. Most of the previous methods [2, 7, 9] use database searches to find conformations for several consecutive residues which match the configuration of the C_α coordinates being used as a template. The PGMC method, in contrast, uses probabilities for individual residues to guide Monte Carlo searches. The method produces results as good as or better than the previously published methods for the protein flavodoxin. In general, backbone conformations are modeled accurately to within 0.6 Å rms deviation from the crystal structure. Most of the error comes at the C-terminal ends and in turns, while the extended secondary structures, α helices and β sheets are modeled much better, with a typical rms deviation of 0.3 Å or better. Sidechain conformations are not modeled as accurately. Sidechain rms deviations over 2.0 Å can be expected for large proteins where the computational cost of optimizing all sidechains concurrently is very large. The sidechain deviation for the small protein crambin was much better, averaging 1.87 Å for 25 models. Overall rms deviations are typically better than 2.0 Å, and depend primarily upon the amount of time spent optimizing the sidechain conformations.

The PGMC C_α Builder is an extremely fast, automatic method. For proteins the size of crambin, both the backbone and sidechain can be modeled accurately in less than 20 minutes on a standard workstation. This may enable the method to be used for evaluating numerous possible C_α conformations, such as those generated from a lattice-base protein folding simulation. To this end, a simple C_α forcefield has been developed which enables lattice conformations to be smoothed, thereby providing a template for the C_α Builder.

References

[1] W.A. Hendrickson and M.M. Teeter, *Nature*, 290, 107-113, (1981).

[2] T.A. Jones, J.-Y. Zou, S.W. Cowan, and M. Kjeldgaard, *Acta. Cryst.*, A47, 110-119 (1991).

[3] D.G. Covell and R.L. Jernigan, *Biochemistry*, 29, 3287-3294 (1990).

[4] M.S. Friedrichs and P.G. Wolynes, *Science*, 246, 371-373 (1989).

[5] K.W. Plaxco, A.M. Mathiowetz, and W.A. Goddard III, *Proc. Natl. Acad. Sci. USA*, 86, 9841-9845 (1989).

[6] E.O. Purisima and H.A. Scheraga, *Biopolymers*, 23, 1207-1224 (1984).

[7] L.S. Reid and J.M. Thornton, *Proteins*, 5, 170-182 (1989).

[8] P.E. Correa, *Proteins*, 7, 366-377 (1990).

[9] L. Holm and C. Sander, *J. Mol. Biol.*, 218, 183-194 (1991).

[10] A. Rey and J. Skolnick, *J. Comp. Chem.*, 13, 443-456 (1992).

[11] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[12] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller, *J. Chem. Phys.*, 21, 1087-1092 (1953).

[13] A. Wlodawer *et al.*, *J. Mol. Biol.*, 180, 301-329 (1984).

[14] C.A. Collyer *et al.*, *J. Mol. Biol.*, 211, 617-632 (1985).

[15] A. Wlodawer *et al.*, *Biochemistry*, 27, 2705-2717 (1988).

[16] W.W. Smith *et al.*, *J. Mol. Biol.*, 117, 195-225 (1977).

[17] S.E.V. Phillips, *J. Mol. Biol.*, 142, 531-554 (1980).

[18] BIOGRAF/POLYGRAF. Copyright by Molecular Simulations, Inc. (1992).

[19] J. Skolnick and A. Kolinski, *J. Mol. Biol.*, 221, 499-531 (1991).

[20] K.F. Lau and K.A. Dill, *Macromolecules*, 22, 3986-3997 (1989).

[21] D.A. Hinds and M. Levitt, *Proc. Natl. Acad. Sci., USA*, 89, 2536-2540 (1992).

[22] M. Marquart *et al.*, *Acta. Cryst.*, B39, 480-490 (1983).

[23] M.D. Walkinshaw *et al.*, *Proc. Natl. Acad. Sci., USA*, 77, 2400-2404 (1980).

[24] Y.S. Babu *et al.*, *Nature*, 315, 37-40 (1985).

Chapter 6

Prediction of Loop Conformations in Antibodies

Abstract

The Probability Grid Monte Carlo technique is an effective technique for many problems in protein modeling. Here, the technique has been adapted for modeling loop structures in proteins. The method has three phases: (1) generation of numerous loop backbone conformations using ϕ, ψ probability grids, (2) optimization of sidechain conformations of the best backbones from Phase 1 using χ probability grids, and (3) optimization of the best loop conformations from Phase 2 using conjugate gradients minimization. The method is applied to the six hypervariable loops of immunoglobulins, using the crystallized Fab fragments from HyHEL-5 and McPC603 as test cases. Conformations are predicted which are very similar to the crystal structure conformations – several backbone conformations have rms deviations of 1.0 Å or less from the crystal structure.

I. Introduction

In homology modeling studies, the X-ray crystal structure of one protein serves as a template for predicting the 3-dimensional structure of a second protein, which is similar in sequence but whose tertiary structure has not been determined experimentally. This method has been used successfully for a variety of systems, such as HIV-1 protease modeled from a protease of Rous sarcoma virus[1] and amyloid precursor protease inhibitor domain modeled from bovine pancreatic trypsin inhibitor[2]. If the sequences of the template protein (T) and the unsolved protein (U) are very similar in length and composition, the structure of protein U can be modeled simply by using nearly the entire 3-dimensional structure of T, modifying only the coordinates of the sidechains which differ in the two proteins. Replacement of sidechain geometries is a standard facility of most molecular modeling software. A more complex modeling task arises when regions of the proteins differ in both sequence and the number of residues. There is no standard method for replacing three residues in protein T with six residues from protein U. Even if such sequence length mismatches are localized to short segments of the protein, there are significant rearrangements in the backbone conformations which must be modeled by more sophisticated techniques.

The loop-modeling procedure described here provides a methodology for making such replacements by sampling the conformational space of the variable-length sequences. Regions of the two proteins which are highly similar in both sequence and length are termed "framework" regions. The regions of variable length are termed "loops." This is a broader use of the term "loop" than in the terminology of Rose and coworkers[3], who use the term to define regions of proteins which meet certain geometric criteria. Nevertheless, the variable regions described here are often loops in both senses. Modeling these loops requires success in two endeavors: determination of the loop backbone, which must meet the geometric constraints imposed by the loop "endpoints," where it attaches to the framework, and optimization of the sidechain positions. We have separated these two components into different phases. The first phase rapidly produces a large number of backbone conformations which meet the endpoint criteria while the second phase samples sidechain conformations for the best loops from Phase 1. A third phase optimizes the best structure from Phase 2 by using energy minimization of all atomic positions. This strategy allows increasing sophistication to be built into the model as the breadth of the conformational searches is decreased.

II. Methodology

Several methods have been reported for modeling loop conformations. These fall into two general categories: those using databases of known loop conformations[4, 5] and those using conformational searching[6, 7]. A combined approach has also been described[8]. The method here also combines features of both approaches, as it uses a conformational search method similar to Bruccoleri and Karplus[6], but with dihedral angles sampled from probability grids developed through an analysis of the Brookhaven Protein Database.

The Probability Grid Monte Carlo (PGMC) method is described in detail in Section 4.II. The method samples conformational space by modifying the dihedral angles of a polypeptide or protein according to probability matrices determined from an analysis of known protein structures. At each step of the simulation, one amino acid residue is selected for modification and either its backbone or sidechain conformation is modified. If its backbone conformation is to be modified, new values of ϕ and ψ are chosen from 2-dimensional grids, where each possible ϕ, ψ combination has been assigned a specific probability. The values of ϕ and ψ are confined to discrete values between -180°and 180°. The spacing between gridpoints is S, so there are $[360/S]^2$ possible conformations for each amino acid. Probabilities have been determined for grids with spacings of 5, 10, 15, 30, and 60°. Different probability grids were determined for three different residue types: glycine, proline, and the 18 standard residues, from the ϕ, ψ distributions found among residues in a selection of high-quality structures in the Brookhaven Protein Database (PDB). Additionally, different grids were developed for different secondary structure types: α helices, β sheets, and coil conformations. Separate grids were not determined for β turns because these conformations require specific four-residue conformations and are not represented well by single-residue probabilities. The probabilities for coil regions were derived from all residues not specified by the HELIX, SHEET, and TURN designators of the PDB files. These coil probability grids are the most pertinent to loop conformations and are, therefore, the ones used in these simulations.

Sidechain conformations are also chosen from probability grids, but these grids are 1- to 5-dimensional, depending upon the number of sidechain χ dihedrals that are sampled in a particular residue. Only χ dihedrals which affect the geometries of heavy atoms (non-hydrogen) and are not part of ring systems are included. The number of PGMC χ dihedrals varies from one (e.g., for serine and threonine) to five (for arginine).

The goal in loop modeling is to predict the native conformation of the loop. If a forcefield is used to evaluate conformations, and the forcefield is highly accurate, the minimum energy conformation should be very similar to the native conformation. Our simulations use the DREIDING forcefield[9] to evaluate structures. Although there is probably no forcefield in existence which guarantees that its global minimum is the native conformation, we have included the capacity to improve the results by increasing the sophistication of the calculations. Ideally, such factors as solvation and loop-protein and loop-substrate interactions would be included in the calculation. However, including such terms can increase computational time so much that only a few possible conformations can be evaluated. There must be a balance between speed and accuracy. We have addressed this need for balance by creating a hierarchical procedure, which increases in accuracy as the simulation proceeds. The first stage of the procedure creates backbone conformations which meet the endpoint criteria; these conformations are evaluated without regard to sidechain interactions. The second stage optimizes the positions of the sidechains of the loop residues. Interactions among all the sidechain and backbone atoms of the loop are considered, as are interactions with residues from other regions of the protein and, if possible, interactions with a substrate. The final stage is complete optimization of the best conformations from the second stage, using energy-minimization of all degrees of freedom of the loop. At this stage, solvent may be added to enhance the accuracy of the simulation.

The first stage of the simulation involves the generation of numerous backbone conformations which meet the endpoint criteria established by the constant framework. The conformation of each loop in a protein is predicted independently. The framework residues from the template protein are held constant, while all loop residues are removed. The loop being modeled is then constructed using standard geometries from the BIOGRAF[10] peptide libraries. Loop conformations are then generated which meet the criteria that the endpoint residues of the loop attach to the framework with the same geometry as in the template protein. There is, theoretically, no limit to the number of conformations of an n_l-residue loop which meet such endpoint criteria if $n_l > 3$, so there is no method for directly calculating all possible conformations. The seminal work of Gō and Scheraga[11], however, described a method for exactly determining the conformations of three consecutive residues which enable them to meet endpoint criteria. This is done by solving constraint equations for the three ϕ and ψ dihedrals while holding all other dihedrals, bonds, and angles fixed. The algebraic equations described cannot be solved for all cases; Gō and Scheraga[11] found the number of solutions varied from 0 to 8.

We have implemented the chain-closure algorithm to work in conjunction with our probability grids to generate n_l-residue loop structures which exactly meet the endpoint criteria. An initial conformation is generated by randomly selecting ϕ, ψ pairs from the probability grids for the "outer" loop residues – those besides the central three residues. New conformations are generated by randomly choosing one of the outer residues and choosing a new ϕ, ψ pair from the appropriate probability grid. After each new conformation is constructed, the chain-closure algorithm is used to determine whether any combination of ϕ's and ψ's for the central three residues can close the loop. If so, each of the solutions is constructed and the energy of the structure is calculated. If not, the process continues with a new loop residue selected at random and a new ϕ, ψ pair chosen from the probability grids. The process continues until a loop is successfully built. As each successful loop structure is saved, its energy is calculated. Because the first phase is only concerned with the generation of backbone conformations, the sidechain atoms are ignored in the energy calculations. A typical calculation produces and tests 2000 conformations per cpu minute on a single processor of a Silicon Graphics 4D/380 workstation. On average, 20 of these 2000 structures can form a successful loop.

The successful loops from Phase 1 are ranked by energy and the best are saved for sidechain positioning in Phase 2. This second phase uses the PGMC method, including selection of sidechain conformations from χ probability grids, calculation of the energy of each new conformation, and acceptance or rejection according to the Metropolis criterion[12]. For each of the backbone conformations saved from Phase

2, a Monte Carlo simulation is run in which sidechain conformations are initially randomly selected from the sidechain probability grids, and new conformations are built by modifying one sidechain at a time. As each new conformation is built, the energy of its sidechain and backbone are calculated, including its interactions with nearby atoms of the framework. The energy of the new conformation is compared to the previous energy. If the change in energy, ΔE, is less than 0, the new structure is saved. If the new structure is higher in energy, the probability of accepting it is $\exp(-\Delta E/k_B T)$, where k_B is the Boltzmann constant and T is the simulation temperature. The Monte Carlo simulation proceeds for a number of steps and the best energy conformation is saved. A similar simulation is run for each backbone structure.

The best conformations from Phase 2 are selected for full minimization. The lowest-energy conformation for each of the six loops is built onto the crystal structure framework. This new structure is then minimized using conjugate gradients minimization with the framework atoms included in the force calculations, but only the loop atoms allowed to move. Considerable refinement can be achieved even for the lowest-energy conformations from Phase 2 because bonds and angles need no longer be fixed and dihedral angles are no longer restrained to gridpoint values, i.e., multiples of the grid spacing S. At this point, solvation models may be introduced into the calculations. Initial work has been done to incorporate the solvation potential of Eisenberg and McLachlan[13], but without substantial improvement over the vacuum calculations reported here.

In summary, the three phases of the loop-building simulations presented here are:

1. **Phase 1.** Generate a large number of loop conformations which meet the endpoint criteria, using ϕ, ψ probability grids and the chain-closure algorithm.

2. **Phase 2.** Optimize the sidechains of the lowest-energy backbone conforma-

tions from Phase 1, using χ probability grids in the PGMC method.

3. **Phase 3.** Reconstruct the six loops from the lowest-energy conformation of each loop from Phase 2. Energy-minimize the resulting structure.

III. Antibody Hypervariable Loops

One of the most important classes of proteins to be studied by homology modeling is the immunoglobulins. These molecules are ideal candidates for homology modeling studies because each organism can produce a huge number of immunoglobulins, or antibodies, which differ dramatically in their binding specificities, but are nearly identical in sequence and structure. The specificity is due to the great variability of six small loop regions, the "complimentarity determining regions" (CDR's), also known as hypervariable loops. A reliable method for predicting the conformations of the six CDR's would essentially be a method for predicting the antigen binding sites of an enormous variety of immunoglobulins and would provide valuable information about the immune system, catalytic antibodies, and on a more fundamental level, molecular recognition.

The most common of the five closely related immunoglobulin classes is immunoglobulin G (IgG), which is Y-shaped and contains two antigen-binding sites. Figure 6.1 shows a schematic diagram of an IgG molecule, which contains two copies each of two different types of chains. The light chains contain a variable (V_L) and a constant (C_L) domain, while the heavy chains contain one variable (V_H) and three constant (C_H1, C_H2, C_H3) domains. All six domain types are similar in sequence and structure, containing 100-120 residues folded into two antiparallel β sheets. The great diversity of antibody specificity arises from the extreme variability of three loops in each variable domain. The three variable loops in V_H are called H1, H2,

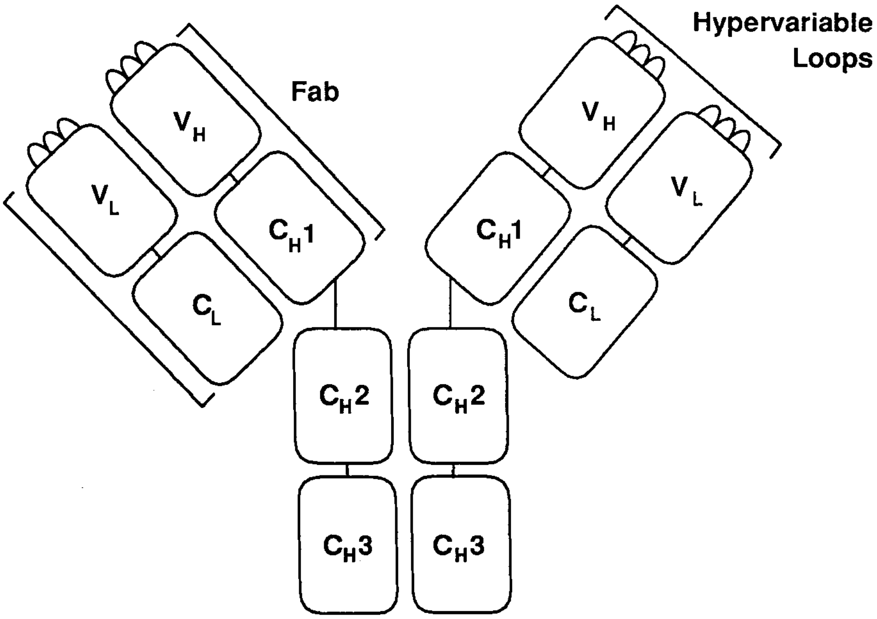


Figure 6.1. A schematic diagram of an IgG molecule.

and H3, while those in V_L are L1, L2, and L3. The spatial arrangement of the six loops is shown in Figure 6.2 for two immunoglobulins whose F_{ab} fragments have been crystallographically resolved: HyHEL-5 [14] and McPC603 [15].

Because the specificity of antibodies is determined by the six CDR's, considerable work has gone into understanding the structure-sequence relationships for these small loop regions. Several crystal structures of murine and human IgG's have been solved. Most are not complete IgG's, but just the Fab fragment (see Figure 6.1). Many are co-crystals which include the antigen or a hapten. After analyzing these crystal structures, as well as several hundred other immunoglobulin sequences, C. Chothia and coworkers[4, 16] proposed that a small number of "canonical structures" exist for each of the five loops other than H3. Most examples of these loops should conform to one of these structures. For example, as many as 95% of the L2 loops are believed to conform to a single canonical structure, while L1 has four canonical structures

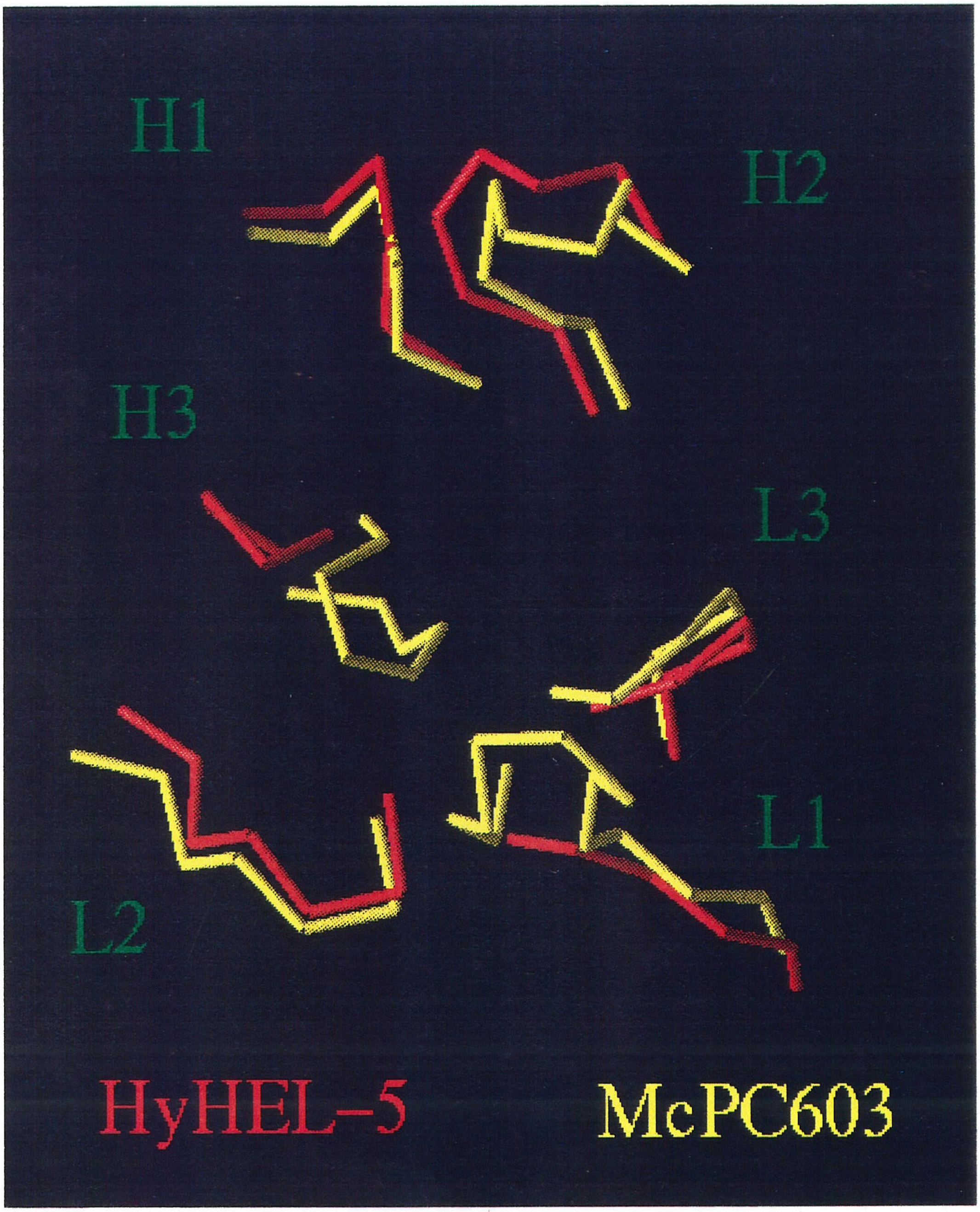


Figure 6.2. The six hypervariable loops of the immunoglobulins HyHEL-5[14] and McPC603[15], shown as C_α coordinates only.

which account for about 70% of the total. For these loops, antigen specificity is due primarily to the sidechain conformations, so it is of paramount importance to model these correctly. Additionally, no canonical structures have been identified for the H3 loops, which show extreme variation in size and sequence. Therefore, a major challenge in modeling the antigen binding site is to model both the backbone and sidechains of the H3 loop.

The studies reported here used the PGMC-based method described above to reproduce the conformations of the hypervariable loops from the murine IgG molecules HyHEL-5[14] (structure 2HFL from the Brookhaven Protein Database (PDB)) and McPC603[15] (PDB structure 1MCP). These two cases have been used to test methodologies which use canonical conformations[16], as well as those using conformational search methods [17, 18] and those which use a combination of database information and conformational searching[8]. Different published reports have used slightly different definitions of the loop regions. As our method is most similar to that of Bruccoleri and coworkers[6, 18], we used the loop definitions similar to those in Reference [18], as shown in Table 6.1. There are two differences: we include an additional residue in the L1 and L3 loops of HyHEL-5, making these six-residue loops. In these cases, increasing the size of the loops consistently improved our results.

In order to produce the largest variety of loop conformations, we used probability grids with $S = 5°$. As described above, all six loops were removed from the crystal structure, and each loop was modeled independently. In the first stage of the simulation, 1000 loops were created. The energy of the backbone atoms of each of these loops was calculated, and the top 100 were saved for Phase 2. Any repeat conformations were eliminated. These 100 best backbone conformations were each subjected to 50 steps of PGMC, using χ probability grids and a simulation temperature of 1000 K. The full energy of the loops, including the sidechain atoms, was used for

	HyHEL-5		McPC603	
Loop	Residues	Size	Residues	Size
L1	25 - 30	6	26 - 37	12
L2	49 - 54	6	56 - 61	6
L3	89 - 94	6	97 - 102	6
H1	28 - 32	5	28 - 32	5
H2	50 - 56	7	50 - 58	9
H3	100 - 103	4	102 - 109	8

Table 6.1. The loop residues of HyHEL-5 and McPC603.

the Monte Carlo phase. For each of the six loops, the overall lowest-energy conformation from the 100 Monte Carlo runs was saved for Phase 3. In this last stage, the six loops were assembled and were energy-minimized using the conjugate-gradients minimizer of BIOGRAF[10].

The results for McPC603 and HyHEL-5 are described in Tables 6.2 and 6.3. The Phase 2 root-mean-square (rms) deviations are given with respect to the crystal structures. The Phase 3 rms deviations are given with respect to DREIDING reference structures, in order to reduce the effects of the choice of forcefield. The DREIDING reference structures were created by minimizing the loop conformations from the crystal structure while holding the framework atoms fixed. For all cases, hydrogen atoms were not included in the determination of rms deviations.

Results were consistently good for small and medium-sized loops (4-6 residues), with the exception of the H1 loop of HyHEL-5. For this loop, 17 of the top 25 loops from Phase 2 had C_α rms deviations of less than 1.5 Å, but the best loop had a deviation of 2.48 Å. It is possible that increasing the sophistication of the

Predicted Loops for McPC603

		RMS deviations (Å)					
		Phase 2			Phase 3		
Loop	Size	All	BB	C_α	All	BB	C_α
L1	12	4.53	3.72	3.70	4.95	4.11	4.17
L2	6	1.79	1.15	1.09	1.14	0.82	0.62
L3	6	1.88	1.19	1.09	1.45	0.94	0.58
H1	5	1.86	1.41	1.22	1.90	1.44	1.17
H2	9	4.50	2.73	2.71	4.25	2.81	2.70
H3	8	3.05	1.72	1.66	3.16	1.80	1.75
All	46	3.45	2.48	2.44	3.49	2.64	2.59

Table 6.2. The rms deviations for all atoms, backbone atom (BB), and C_α coordinates are given for the predicted conformations of the McPC603 hypervariable loops.

Predicted Loops for HyHEL-5

		RMS deviations (Å)					
		Phase 2			Phase 3		
Loop	Size	All	BB	C_α	All	BB	C_α
L1	6	2.18	1.42	1.44	1.67	1.04	0.87
L2	6	2.07	1.59	1.30	1.57	1.23	0.92
L3	6	3.04	2.54	1.99	2.96	2.31	1.75
H1	5	3.77	1.98	2.00	4.06	2.31	2.48
H2	7	2.73	1.89	1.77	2.53	1.87	1.98
H3	4	2.77	0.92	0.32	2.64	0.98	0.27
All	34	2.84	1.83	1.62	2.73	1.74	1.59

Table 6.3. The lowest-energy loops created in phases 1 and 2 of simulations of HyHEL-5.

energy term, such as including terms for surface area or solvation, could improve the selection of more native-like conformations. As noted in other studies[18, 8], the energy from vacuum calculations correlates only modestly with improved fit to the crystal structure. For example, the bests C_α fits from the 1000 loops generated in Phase 1 are listed in Table 6.4, clearly indicating that better loops were sampled. Nevertheless, the vacuum calculations used here did quite well for most of the medium sized loops, with C_α deviations ranging from 0.27 Å for H3 of HyHEL-5 to 1.75 Å for L3 of HyHEL-5. It is important to note that the backbone conformations of the two H3 loops were modeled quite well, since these loops cannot be modeled using canonical structures[16]. Modeling of the L3 loop of McPC603 was very successful, not only in that the rms deviations were low, but that the *cis* peptide bond of Pro 101 was accurately predicted. Likewise, the *trans* peptide bond of the homologous Pro 94 of HyHEL-5 was predicted, as well. On the other hand, the large L1 and H2 loops of McPC603 were modeled poorly in comparison to the shorter loops. The conformational space of these loops is so large that even 1000 trial structures is probably too small a number for adequate sampling. Modeling loops greater than 10 residues in length is probably beyond the capabilities of both database and conformation-searching algorithms, the former because of the lack of example conformations meeting the necessary criteria and the latter because the huge conformational space cannot be adequately searched in a short period of time. Generally, the calculations here took two hours per loop. One hour was required to generate the 1000 loop conformations and one hour was required to optimize the sidechain conformations of the 100 initial conformations in Phase 2.

The C_α traces of the modeled and actual loops for HyHEL-5 and McPC603 are shown in Figures 6.3 and 6.4. The loops are shown in the same face-on view of the antigen-binding site as used in Figure 6.2. The excellent fits of the predicted loops

Best C_α Conformations from Phase 1

Loop	McPC603	HyHEL-5
L1	2.48	0.44
L2	0.41	1.12
L3	0.90	1.21
H1	0.38	0.44
H2	2.21	0.88
H3	1.34	0.26

Table 6.4. The best rms deviation of C_α coordinates from the crystal structure from among the 1000 loops generated in Phase 1.

for H3, L1, and L2 of HyHEL-5 and H1, L2, and L3 of McPC603 are very apparent, even in the simple C_α trace pictures. More detail can be seen in Figures 6.5 and 6.6, which show the entire backbone conformations of the model and actual loops in views perpendicular to the plane of the loops. The large errors in the backbone conformations of the L1 and H2 loops of McPC603 would prevent an adequate model of the antigen-binding pocket from being predicted, if this were a blind test where the conformation of the loops was unknown. L1 and H2 are both involved in binding the phosphocholine hapten[19] co-crystallized with the McPC603 Fab in the PDB structure 2MCP[20], so better modeling would be required to understand this interaction if no crystal structure were available. In contrast, the predicted backbone conformations of the HyHEL-5 loops may be sufficient to understand its binding to lysozyme. Sidechain modeling would still require improvement. This may be possible with concurrent optimization of the sidechain loops. Currently, the Phase 2 simulations are done on each individual loop, without the other loops present. Simulations run with all six loops present should improve the sidechain packing in the binding site model.

The results are quite comparable to the results of Bruccoleri *et al.*[18], who used a more brute-force conformational search algorithm for constructing possible loop structures and included surface area calculations in a variety of ways to choose particular loop conformations to use. That study predicted the HyHEL-5 loops to an rms deviation of 2.6 Å for all atoms and 1.4 Å for backbone atoms, compared to our results 2.73 Å and 1.74 Å, respectively. Our results for McPC603 were not as good, since the two large loops were modeled less successfully. Nevertheless, the results for the other four loops are quite good. The results from various methods are shown in Table 6.5. The PGMC method reported here, produces results similar to those of the published packages, despite being a highly generalized method, which does not require the loops being modeled to have any relationship to loops of known conformations, and without having taken into account any solvent effects. The method described here does not require any user input once the initial conformation and the sequence to be changed has been specified.

IV. Conclusions

An effective loop-modeling procedure has been developed which uses probability grid Monte Carlo (PGMC) to search the conformational space of the loop backbone and its sidechains. Although this method is completely general, applicable to any loop conformation and sequence, it produces results comparable to methods requiring database matching or canonical structure matching. Modeling of the hypervariable loops of the immunoglobulins HyHEL-5 and McPC603 showed that most loops can be modeled to within 2 Å (backbone) or 3 Å (all-atom) rms deviations from the crystal structures. Additional energy terms using solvent-accessible surfaces or other

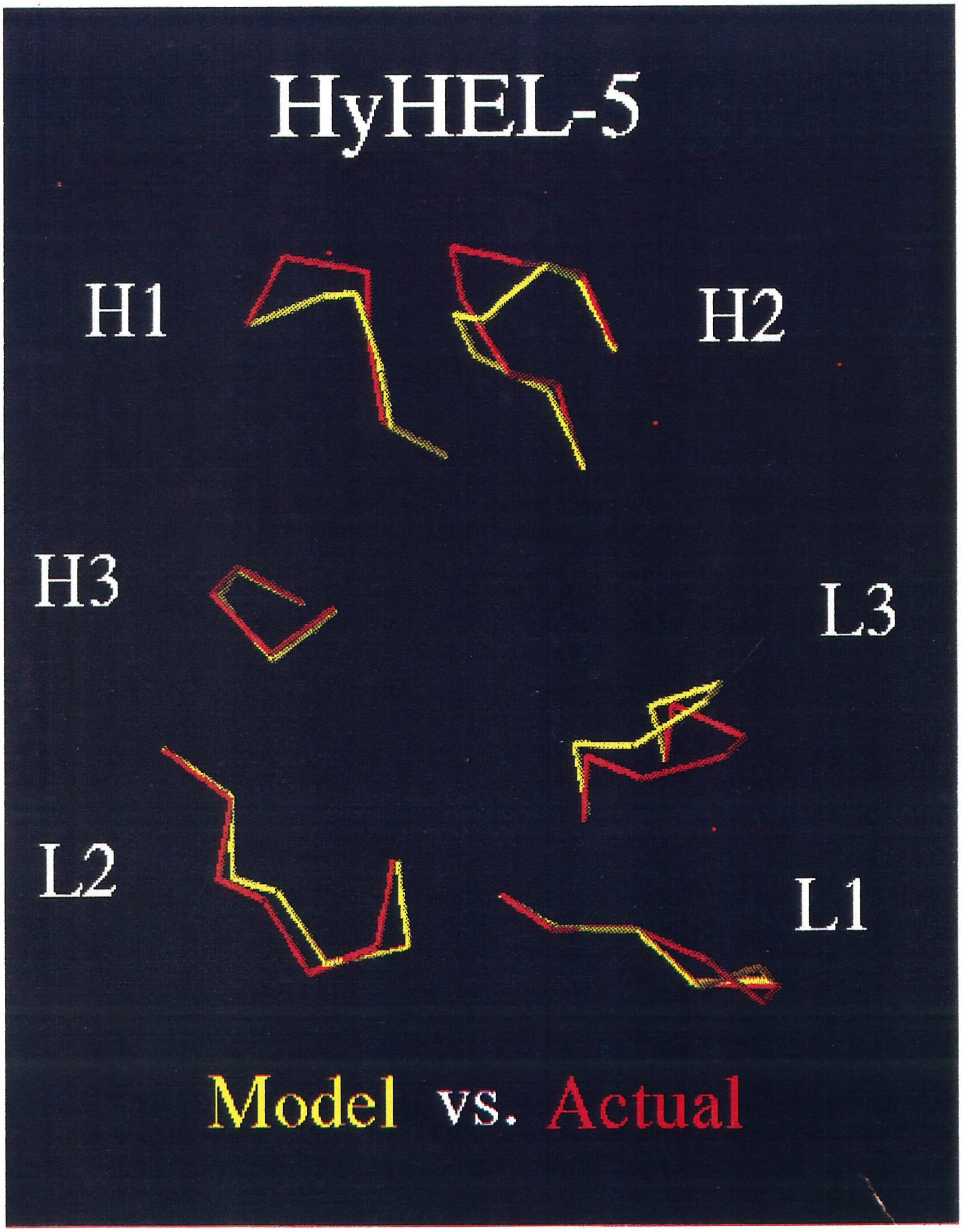


Figure 6.3. C_α traces of the predicted and actual loop conformations of HyHEL-5.

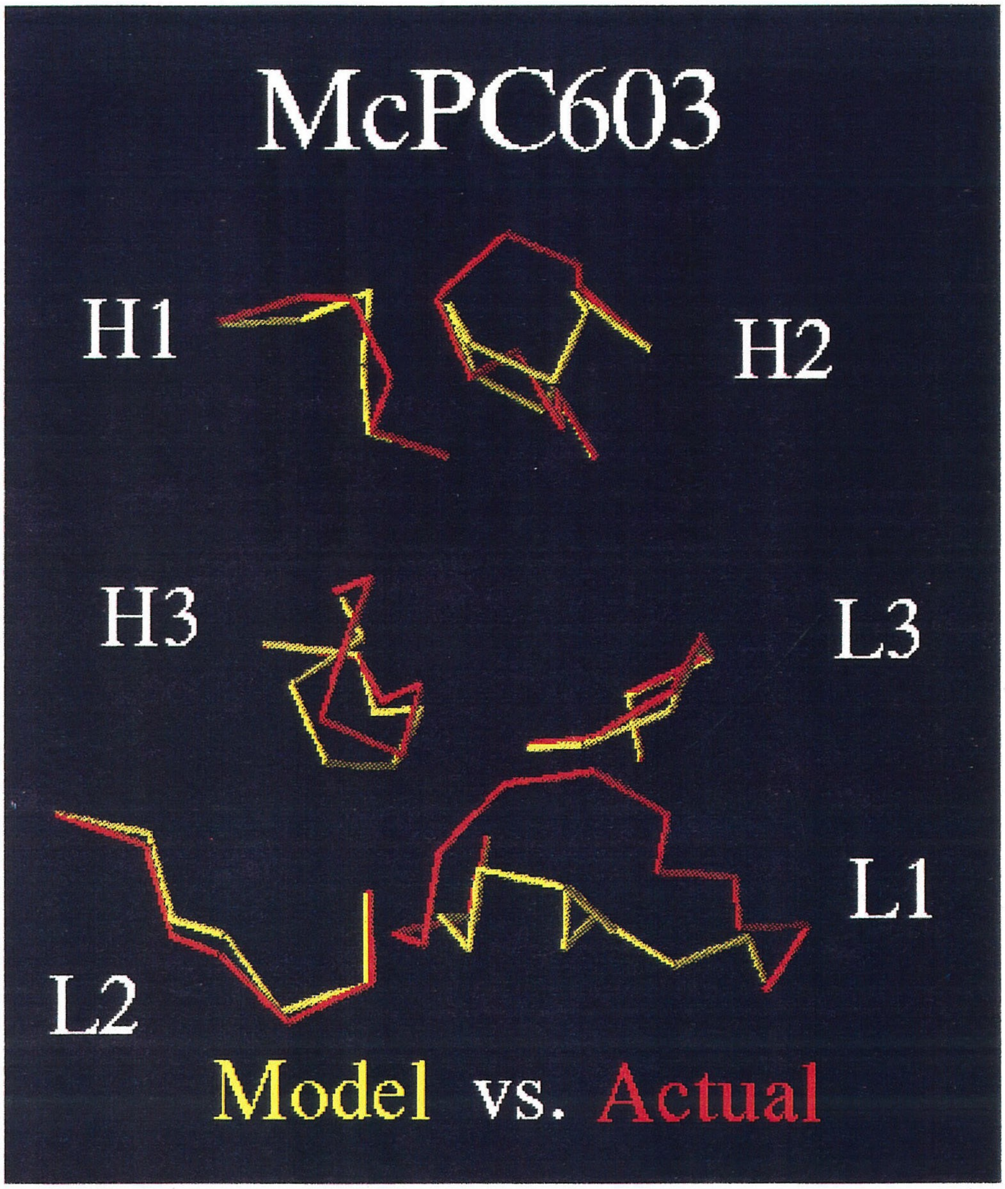


Figure 6.4. C_α traces of the predicted and actual loop conformations of McPC603.

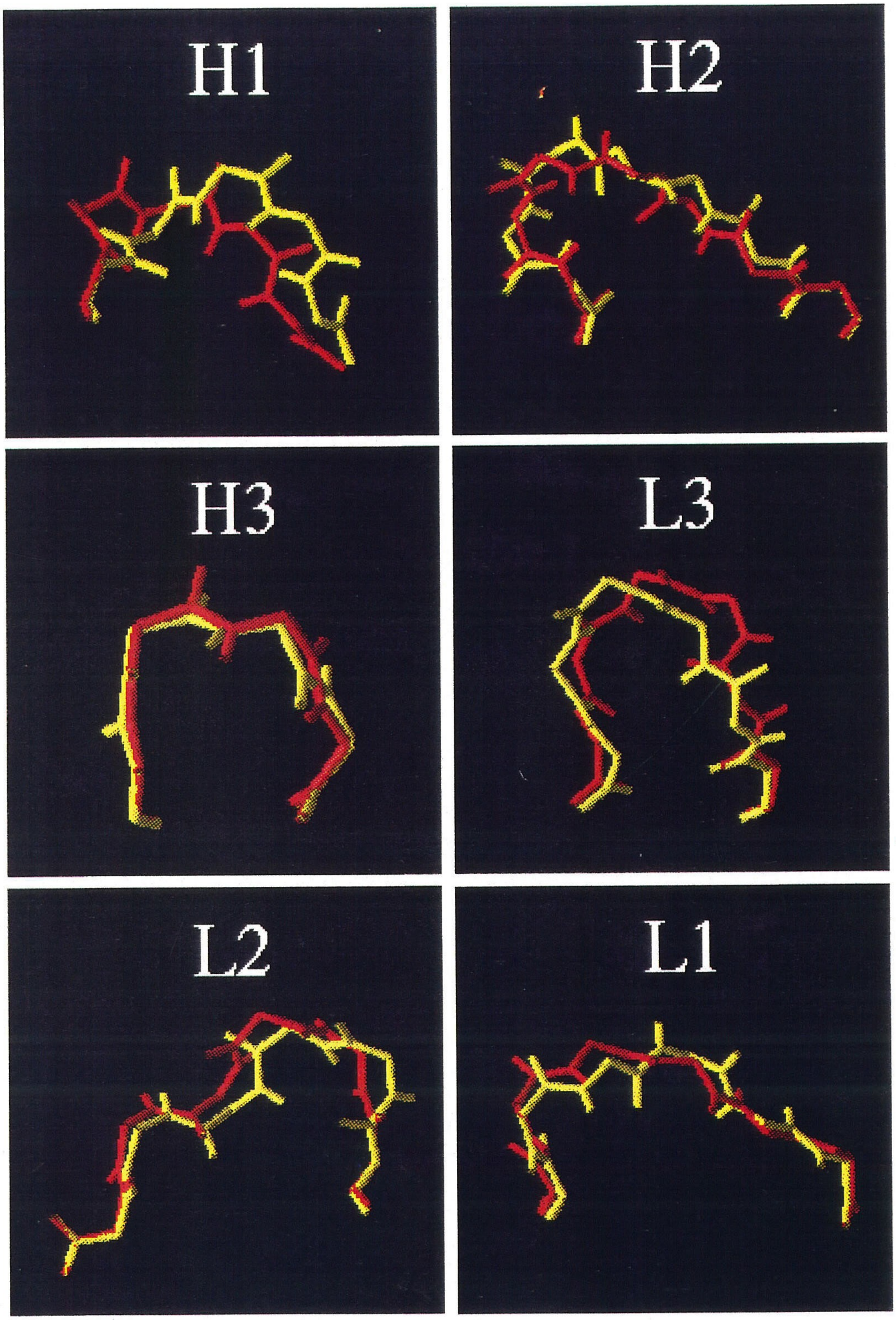


Figure 6.5. Backbone conformations of the predicted (yellow) and actual (red) loop conformations in HyHEL-5.

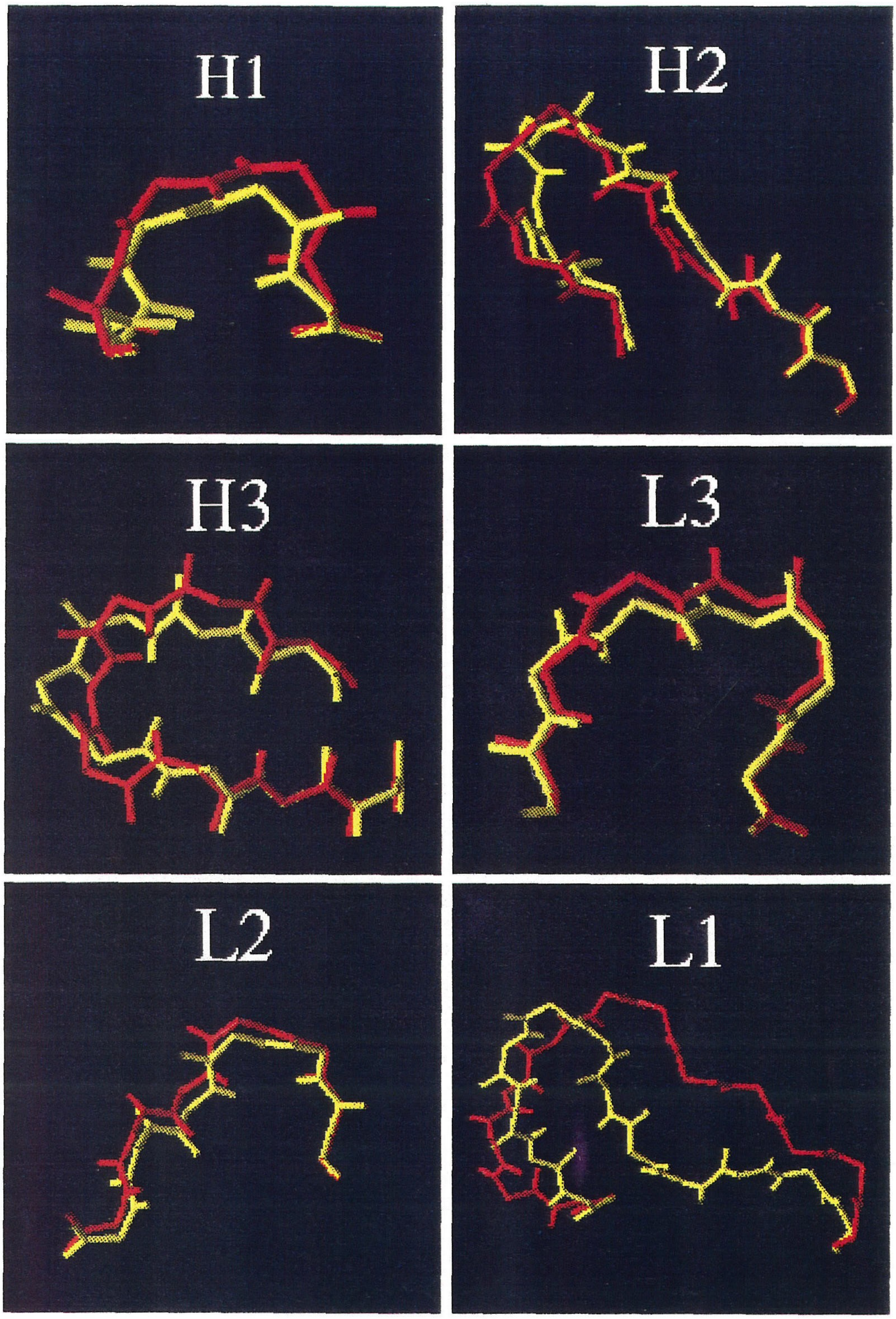


Figure 6.6. Backbone conformations of the predicted (yellow) and actual (red) loop conformations in McPC603.

All-atom and Backbone atom rms deviations (Å)								
HyHEL-5								
Loop	PGMC		Ref.[18]		Ref.[16]		Ref.[8]	
L1	1.67	1.04	1.8	0.6	–	0.8	1.53	1.02
L2	1.57	1.23	1.7	0.8	–	0.9	1.55	0.76
L3	2.96	2.31	4.1	1.1	–	–	2.50	1.37
H1	4.06	2.31	1.8	1.1	–	1.4	1.64	0.90
H2	2.53	1.87	3.1	2.1	–	1.1	1.83	1.13
H3	2.64	0.98	2.7	1.0	–	–	2.31	1.45
McPC603								
L1	4.95	4.11	3.0	2.6				
L2	1.13	0.82	1.9	1.6				
L3	1.45	0.94	1.4	0.8				
H1	1.89	1.44	1.7	0.7				
H2	4.25	2.70	2.1	1.6				
H3	3.16	1.80	2.9	1.1				

Table 6.5. The results from this work (PGMC) compared to results from three different methods: a conformational-search algorithm[18], a method which uses the canonical structures of loops other than H3 (L3 of HyHEL-5 also does not fit one of the canonical structures)[16], and a method which combines conformational searching with comparisons to database conformations[8].

solvation terms, may provide a means for improving the correlation between energy and rms fit to the crystal structure, thereby enabling backbone conformations to be regularly fit to near 1 Å or better. In addition, concurrent optimization of the sidechains of all six loops during Phase 2 should improve the packing of sidechains and the prediction of the shape of the antigen-binding site.

References

[1] I.T. Weber *et al.*, *Science*, 243, 928-931 (1989).

[2] R.S. Struthers, D.H. Kitson, and A.T. Hagler, *Proteins: Structure, Function, Genet.*, 9, 1-11 (1991).

[3] J.F. Leszczynski and G.D. Rose, *Science*, 234, 849-855 (1986).

[4] C. Chothia and A.M. Lesk, *J. Mol. Biol.*, 196, 901-917 (1987).

[5] T.A. Jones and S. Thirup, *EMBO J.*, 5, 819-822 (1986).

[6] R.E. Bruccoleri and M. Karplus, *Biopolymers*, 26, 137-168 (1987).

[7] P.S. Shenkin *et al.*, *Biopolymers*, 26, 2053-2085 (1987).

[8] A.C.R. Martin, J.C. Cheetham, and A.R. Rees, *Proc. Natl. Acad. Sci., USA*, 86, 9268-9272 (1989).

[9] S.L. Mayo, B.D. Olafson, and W.A. Goddard III, *J. Phys. Chem.*, 94, 8897-8909 (1990).

[10] BIOGRAF/POLYGRAF. Copyright by Molecular Simulations, Inc. (1992).

[11] N. Gō and H.A. Scheraga, *Macromolecules*, 3, 178-187 (1970).

[12] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller, *J. Chem. Phys.*, 21, 1087-1092 (1953).

[13] D. Eisenberg and A.D. McLachlan, *Nature*, 319, 199-203 (1986).

[14] Y. Satow *et al.*, *J. Mol. Biol.*, 190, 593-604 (1986).

[15] S. Sheriff *et al.*, *Proc. Natl. Acad. Sci., USA*, 84, 8075-8079 (1987).

[16] C. Chothia *et al.*, *Nature*, 342, 887-883 (1989).

[17] R.M. Fine *et al.*, *Proteins: Structure, Function, Genet.*, 1, 342-362 (1986).

[18] R.E. Bruccoleri, E. Haber, and J. Novotný, *Nature*, 335, 564-568 (1988).

[19] D.R. Davies, E.A. Padlan, and S. Sheriff, *Annu. Rev. Biochem.*, 59, 439-473, (1990).

[20] E.A. Padlan, G.H. Cohen, and D.R. Davies, *Ann. Inst. Pasteur/Immunol.*, 136C, 259-294 (1985).