

PROTEIN FOLDING WITH COARSE-GRAINED
OFF-LATTICE MODELS OF THE POLYPEPTIDE
CHAIN

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Marian Nianias

January 2006

© 2006 Marian Nanas

ALL RIGHTS RESERVED

PROTEIN FOLDING WITH COARSE-GRAINED OFF-LATTICE MODELS OF THE POLYPEPTIDE CHAIN

Marian Nancias, Ph.D.

Cornell University 2006

A hierarchical approach, together with the United Residue (UNRES) model of the polypeptide chain, is used to study protein structure prediction.

First, an efficient method has been developed as an extension of the hierarchical approach for packing α -helices in proteins. The results for 42 proteins show that the approach reproduces native-like folds of α -helical proteins as low-energy local minima. Moreover, this technique successfully predicted the structure of the largest protein obtained so far with the UNRES force field in the sixth Critical Assessment of Techniques for Protein Structure Prediction (CASP6).

Next, two popular methods of global optimization are coupled, and the performance of the resulting method is compared with that of its components and with other global optimization techniques. The Replica-Exchange Method together with Monte Carlo-Minimization (REMCM) was applied to search the conformational space of coarse-grained protein systems described by the UNRES force field. In summary, REMCM located global minima for four proteins faster and more consistently than two of three other global optimization methods, while being comparable to the third method used for comparison.

Finally, efficient methods for calculating thermodynamic averages were implemented with the UNRES force field, namely a Replica Exchange method (REM),

a Replica Exchange Multicanonical method (REMUCA), and Replica Exchange Multicanonical with Replica Exchange (REMUCAREM), in both Monte Carlo (MC) and Molecular Dynamics (MD) versions. The algorithms were applied to one peptide and two small proteins (with α -helical and $\alpha+\beta$ topologies). To compare the different methods, thermodynamic averages are calculated, and it is found that REM MD has the best performance. Consequently, free energy maps are computed with REM MD, to evaluate the folding behavior for all test systems.

BIOGRAPHICAL SKETCH

The author Marian Marko Nianas was born as the only child on 5. April, 1978 in Bratislava, the capital of the Slovak Republic, then part of Czechoslovakia. From the very young age, his parents Gabriela and Marian Nianas made sure that he received the best possible education available. This started with his early devotion to foreign languages at an excellent elementary school Halenárska in his home city Trnava. During his time in the elementary school, he experienced a life- changing event, the collapse of communism, which would shape the rest of his future career. Marko continued his studies at a secondary grammar school, known as Nový Gimpel, in Trnava, where he took every single opportunity to travel abroad in order to practice his newly acquired language skills and to learn about other cultures. Short trips to England and France were followed by a year-long stay as a foreign exchange student through AISE exchange program in the United States. His foreign exchange year with the Leggett family in western Kansas became one of the most important experiences in his life. Sixteen years old teenager, immersed in a completely new culture, with limited contact to the familiar world back home, had no choice but to adapt. Finding out that the language skills acquired in the classroom were not quite adequate for the real life in the host culture, beginnings were especially difficult. If it weren't for his host-father Bill Leggett, and his host-brothers Lance Leggett and Alek Bjornevik, the exchange year would not have ended as a successful and amazing experience.

After his exchange year, Marko returned to his home city Trnava, and finished his high school degree with highest honors. Accepting the stipend offered by a small liberal arts college, McPherson College, Marko decided to pursue his higher

education in the United States. While at McPherson, Marko was drawn to the mysteries of physics by his adviser Kent Noffsinger. Besides his passion for both physics and mathematics, Marko continued to pursue his interests in German. He managed to combine his foreign language studies with natural science by spending another year as a foreign exchange student, only this time studying physics at Philipps-Universität in Marburg, Germany. He concluded his undergraduate studies by graduating summa cum laude with a B.S. in May 2000.

Considering Marko's interest in physics, it was only a natural progression to continue his studies in graduate school. With the help of another McPherson alumni Dr. Edward Wolf, Marko was very fortunate to continue his studies at Cornell University, where after the first year of classes and teaching, Marko joined the group of Dr. Harold Scheraga, and began his computational work on off-lattice reduced models of polypeptide chains. Four years later, this thesis is the product of the late nights spent discussing many half-baked ideas, debugging immensely complex FORTRAN code without comments, and finally attempting to write out the results with as much clarity and precision as possible.

To my wife Vandulka who endured my Ph.D. studies.

To my parents who were always there for me.

ACKNOWLEDGEMENTS

In this section I would like to acknowledge the fantastic help and support of all of those who helped and supported me. This thesis is the end result of years of hard work while being influenced by many people in the process.

First and foremost, I would like to express my gratitude to Dr. Edward Wolf and his wife Marlene. Without them I would certainly not have had the privilege to learn from the best at such a fine institution as Cornell University. Their continuing support with invitations to family dinners, and boat trips on the Cayuga lake enabled me to experience life in upstate New York outside of campus, and made me feel as part of the family.

Next, I would like to thank my adviser Harold A. Scheraga for giving me the opportunity to learn in his laboratory. I would like to thank him for providing all the wonderful resources available to us, for his patience, and especially for giving me the discipline and clarity of thought.

Further, I would like to thank my committee Carl P. Franck and Erich J. Mueller for steering me in the right direction throughout my Ph.D. years at Cornell.

Throughout the first couple of years at Cornell, I spent most of my time in the teaching office located in Rockefeller 235. This office turned out to be an extremely social environment, and it was my office-mates Adam Kruger, Ethan Bernard, Chris Deufel, and Dan Hertz who made the early graduate school life (filled with never ending problem sets, and grading hours) enjoyable.

The person with who I spent perhaps the most time throughout my Ph.D. career is Maurizio Chinchio. Late night debugging, fun with scripting, experimentation with exotic Linux distros and modifying our Matrix cluster are just a few examples

of the fun times spent together. I also learned a lot from the senior members of our group. I'd like to thank Adam Liwo, Czarek Czaplewski and Stan Ołdziej for their patience and enthusiasm. I would like to thank Joanna and Mariusz Makowski for sharing the gallons of coffee consumed in hour office. Thanks goes to other (current or previous) members of our group Mey Khalili, Jorge Vila, Lena Arnautova, Anna Jagielska, Daniel Ripoll, Jarek Pillardy and Jeff Saunders. Extra thanks for the interesting points of view during our group meetings goes to our experimental group Mahesh Narayan, Robert Gahl, Lovy Pradeep, Ervin Welker and John Xu. Finally, special thanks to the amazingly efficient Elaine Redder, and Debra Hatfield.

Very little would be possible without the resourceful support from our favorite lady on the second floor, Shirley Rumsey. I especially appreciate all the toys available to us for "research" purposes. Special thanks to Red Barn Computers for providing affordable computing power with customer service. Much would be impossible without the help of our own cluster, Matrix, and the supercomputing centers, the Lemieux machine at the Pittsburgh Supercomputing Center and the Tungsten cluster at the The National Center for Supercomputing Applications in Illinois.

Skiing was an activity which completely changed the grey cold winters into a fun season of the year. I would like to thank my ski companions Stephan Braig, Ethan Bernard, Marcel Blais, Siddharth Alexander and Eric Rubin Schneiderman for the many beautiful white moments at the Greek Peak mountain.

Summers with no students around and not much going on in Ithaca would not have been the same without the softball league with the legendary Big Red Army team. I would like to thank the socialist oriented coach John C. Carasone for

his amazing vision in leading the team through both rough (winless) and great (almost undefeated) seasons. I would also like to thank the co-coach Marcel Blais for teaching the necessary softball skills to us foreigners, new to the American game of fame.

I would like to thank my gym partners Siddharth Alexander and Jesus Rodriguez. Many happy and unhappy moments were spent in the gym, and I appreciate all the interesting discussions which helped me broaden my horizons.

I would like to thank my wife, Vanda, whom I met the summer of my first year and who was my major motivation to finish my studies. Four years of trans-continental relationship was not easy, but it was certainly well worth waiting, and I am very happy that we can continue our journey through life together.

At last, I would like to thank my amazing parents, who have always been there for me, in both good and tough times. I would like to thank them for their encouraging support, for many sacrifices, which they undertook so that I could have a great education.

Ithaca, January 2006.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgements	vi
List of Tables	xii
List of Figures	xiii
List of Symbols	xiv
1 Introduction	1
1.1 Proteins	1
1.2 Protein Folding	4
1.3 Structure Prediction	5
1.3.1 Experimental Methods	5
1.3.2 Computational Methods	7
1.4 Hierarchical Approach to protein structure prediction	8
1.5 CASP	9
1.6 Summary of the present work	10
Bibliography for Chapter 1	13
2 Packing helices in proteins by global optimization of a potential energy function	16
2.1 Introduction	16
2.2 Methods	17
2.2.1 Protein Representation	18
2.2.2 Potential Energy Function	18
2.2.3 Global Optimization	19
2.2.4 Methods for Producing New Conformations.	21
2.2.5 Distance Measures	22
2.2.6 Protein Targets	22
2.3 Results	23
2.4 CASP6 Results	32
2.5 Discussion	34
Bibliography for Chapter 2	37
3 The UNRES Model of the Polypeptide Chain	38
3.1 The UNRES force field	38
3.2 Optimization of UNRES parameters	44
Bibliography for Chapter 3	51

4	Replica-Exchange Monte Carlo-with-Minimization as a global optimization method with the UNRES force field; comparison with MCM, CSA and CFMC	53
4.1	Introduction	53
4.2	Methods	55
4.2.1	Replica Exchange Monte Carlo (REM)	55
4.2.2	Replica Exchange Monte Carlo-with-Minimization (REMC)	59
4.3	Computational Details	60
4.3.1	Test Systems	60
4.3.2	REMC Implementation	60
4.3.3	Implementation of methods used for comparison	65
4.4	Results and Discussion	67
4.4.1	Production runs	67
4.4.2	Comparison of REMC to REM	73
4.4.3	Comparison of REMC to other methods	75
4.5	Conclusions	81
	Bibliography for Chapter 4	85
5	Replica Exchange and Multicanonical Algorithms with the coarse-grained UNRES force field	87
5.1	Introduction	87
5.2	Methods	90
5.2.1	The UNRES force field	90
5.2.2	Replica Exchange Method (REM)	91
5.2.3	Multicanonical Algorithm (MUCA)	92
5.2.4	Replica Exchange Multicanonical Algorithm (REMUCA)	96
5.2.5	Multicanonical Replica-Exchange Method (MUCAREM)	98
5.2.6	Replica Exchange Multicanonical with Replica Exchange Method (REMUCAREM)	101
5.3	Implementation Details	101
5.4	Results and Discussion	103
5.4.1	Poly-L-alanine	103
5.4.2	1BDD	110
5.4.3	1E0G	114
5.4.4	Free energy diagrams	118
5.5	Conclusions	124
	Bibliography for Chapter 5	127
6	Conclusion	131

Bibliography for Chapter 6	135
A Chapter 2: CFMC; Chapter 4: CSA, CFMC	136
A.1 Conformational Space Annealing (CSA)	136
A.2 Conformational Family Monte Carlo (CFMC)	140
Bibliography for Appendix A	146

LIST OF TABLES

2.1	Simulation results	25
2.2	Simulation results cont.	26
2.3	Stability of procedure with respect to different secondary structure assignment	30
4.1	Structures used in simulation	61
4.2	Simulation results	79
5.1	Parameters used in ala ₂₀ simulations	104
5.2	Parameters used in 1BDD and 1E0G simulations	111

LIST OF FIGURES

1.1	Description of an amino acid	2
1.2	Description of the polypeptide chain and the peptide bond	3
1.3	Trans (A) and cis (B) forms of a polypeptide chain	3
2.1	Simulation Results	27
2.2	1NFO superposition	28
2.3	Energy vs rmsd for 1LPE.	31
2.4	T0198 superposition	33
2.5	GDT analysis of T198	35
3.1	The UNRES model of polypeptide chains.	39
3.2	UNRES force field	41
3.3	UNRES force field, continued	42
3.4	Hierarchical optimization	47
4.1	Energy of REMCM simulation of 1E0L at a particular temperature with respect to MC step number.	68
4.2	Energy vs. RMSD plots for 1E0L, 1GAB, and 1BDD.	70
4.3	Energy vs. RMSD plots for 1CLB, and 1IGD.	72
4.4	Structure Superpositions.	74
4.5	Performance comparison of REM and REMCM for 1GAB.	76
4.6	Performance comparison of MCM and REMCM.	78
4.7	Performance comparison of different global optimization methods.	82
5.1	The parameters used for multicanonical simulations.	105
5.2	Histogram curves for simulations with alanine.	106
5.3	Thermodynamic quantities calculated by various methods for ala ₂₀ , 1BDD and 1E0G	108
5.4	Histogram curves for simulations with 1BDD.	112
5.5	Simulation results for 1BDD.	113
5.6	Simulation results for 1E0G.	116
5.7	Free energy surfaces calculated from REM MD simulations	120
A.1	CSA algorithm	137
A.2	CFMC algorithm	145

LIST OF SYMBOLS

All chapters

Notation	Description
r	Boldface font indicates vector quantity
<i>r</i>	Italic font indicates scalar (absolute value for vectors)
\dot{x}	Time derivative of variable x
$\partial_x, \partial/\partial x$	Partial derivative with respect to x
$\langle \dots \rangle$	Probability average
$\langle \dots \rangle_\rho$	Probability average performed at constant ρ
$P(\dots)$	Probability distribution

Chapter 2

Symbol	Description
e_{ij}	contact energy associated with residues i and j
r_{mn}	Distance between C^α atoms of residues m and n
U	Potential Energy of pairwise interaction

Chapter 3

Symbol	Description
α_{SC}	Side-chain angle in the UNRES model
β_{SC}	Side-chain angle in the UNRES model
β	Boltzmann factor ($1/RT$)
C^α	α carbon atom in an amino acid
Δ_i^β	Minimum required free-energy gap between levels i and $i + 1$
ΔE	Difference between the mean energy of the native-like structures and the mean energy of the non-native structures
$E(\mathbf{X}; \mathbf{Y})$	All-atom ECEPP/3 energy function
Φ	Penalty function to be minimized in the optimization
F	Restricted free energy (RFE)
$F_i(\beta)$	Free energy of structural level i at reduced inverse temperature β
γ	Dihedral angle in the UNRES model
N_{nat}	Number of native-like structures
$N_{non-nat}$	Number of non-native structures
$\Omega_{\mathbf{Y}}$	The integration region of the \mathbf{Y} subspace
R	The gas constant
θ	Virtual-bond angle in the UNRES model
T	The absolute temperature
T_f	Folding temperature
T_g	Glass-transition temperature
$V_{\mathbf{Y}}$	Volume of region \mathbf{Y}
$U_{SC_i SC_j}$	Mean free energy of the interactions between the side chains
$U_{SC_i p_j}$	Side-chain – peptide-group interactions
$U_{p_i p_j}$	Peptide-group interaction potential
U_{tor}	Virtual-bond dihedral angle torsional terms
U_{tord}	Virtual-bond dihedral angle double-torsional terms
U_b	Virtual-bond angle bending terms
U_{rot}	Side-chain rotamer terms
$U_{corr}^{(m)}$	Correlation contributions from the coupling between backbone-local and backbone-electrostatic interactions
$U_{turn}^{(m)}$	Correlation contributions for m consecutive peptide groups
w	Weights associated with the corresponding energy terms
w_i^β	Weight assigned to the deviation of the actual ($F_i(\beta) - F_{i+1}(\beta)$) and the requested (Δ_i^β) at reduced inverse temperature β
X	Set of UNRES degrees of freedom
\mathbf{Y}	Set of degrees of freedom which are averaged out
Z	Z-score of the deviation of energy

Chapter 4

Symbol	Description
β_m	Inverse temperature defined as $1/(k_B T_m)$
$E(X)$	Energy of conformation X
f	Number of degrees of freedom
k_B	Boltzmann constant
M	Number of replicas
P_{all}	Joint probability distribution for all replicas
$P_m(X)$	Probability of conformation X in replica m at temperature T_m
T_m	Absolute temperature of replica m
T_{min}	Temperature at which the protein is in its native state
T_{max}	Temperature at which the protein is unfolded
W	Transition probability that conformation X in replica m is exchanged with conformation Y in replica n
Z_m	Partition function $\int \exp(-\beta_m E(X)) dX$ of replica m

Chapter 5

Symbol	Description
A	Thermodynamic observable
β_m	Inverse temperature defined as $1/(k_B T_m)$
C_V	Specific heat capacity at constant volume
δt	Time step in Molecular Dynamics
δa_{cut}	Cutoff change of acceleration in UNRES MD (to reduce the time step)
$E(X)$	Energy of conformation X
E_{mu}	Multicanonical energy
$\epsilon_{mu}^0(E)$	Multicanonical energy in REMUCA
$\epsilon_{mu}^m(E)$	Multicanonical energy of m th replica in MUCAREM
\mathbf{f}_k	Force on the k th atom
f_m	Dimensionless free energy
$F(E_i, T)$	Microcanonical free energy
$F(r, \rho, T)$	Restricted canonical free energy
G	Generalized mass matrix
k_B	Boltzmann constant
M	Number of replicas
N	Number of residues
$n(E)$	Density of states
$N_m(E)$	Histogram at temperature T_m
N_{mu}	Histogram obtained from the multicanonical simulation
$P(E)$	Probability of occurrence of a state with energy E
$P(x)$	Probability of occurrence of a conformation x with energy E
P_{all}	Joint probability distribution for all replicas
$P_m(X)$	Probability of conformation X in replica m at temperature T_m
\mathbf{p}_k	Generalized momentum of the k th atom
\mathbf{q}_k	Generalized coordinate of the k th atom
ρ	Radius of gyration
r	Root mean square deviation (RMSD)
$S(E)$	Entropy of the state with energy E
τ_m	Integrated autocorrelation time at temperature T_m
T_m	Absolute temperature of replica m
T_{min}	Temperature at which the protein is in its native state
T_{max}	Temperature at which the protein is unfolded
T_0	Reference temperature
U	UNRES potential energy
W	Transition probability that conformation X in replica m is exchanged with conformation Y in replica n
Z_m	Partition function $\int \exp(-\beta_m E(X)) dX$ of replica m

Chapter 1

Introduction

The work presented in this thesis is applied to proteins and, therefore, this chapter is devoted to a brief overview of the following topics. First, a short overview of proteins and protein folding is given. Next, techniques for protein studies, including both experimental and computational methods, are described. This is followed by a section explaining a procedure for protein structure prediction, which was developed in our laboratory. Next, a description is provided as to how our algorithms are tested in community-wide blind-test experiments. The final section in this chapter discusses the author's contributions to the topics mentioned above.

1.1 Proteins

Proteins play important roles in virtually all biological processes. They are responsible for the molecular design of life by performing diverse functions such as enzymatic catalysis, mechanical support, immune protection, or generation and transmission of nerve impulses.

Proteins are composed of basic structural units called amino acids. An α -amino acid (shown in Figure 1.1) consists of an amino group, a carboxyl group, a C^α H group, and a unique R group. All the groups are bonded to an α carbon atom. There are 20 types of amino acids commonly found in proteins, differing only in their unique R group, which is also known as a side chain. The remarkable range

of functions performed by proteins is due to the diversity in these twenty amino acids.

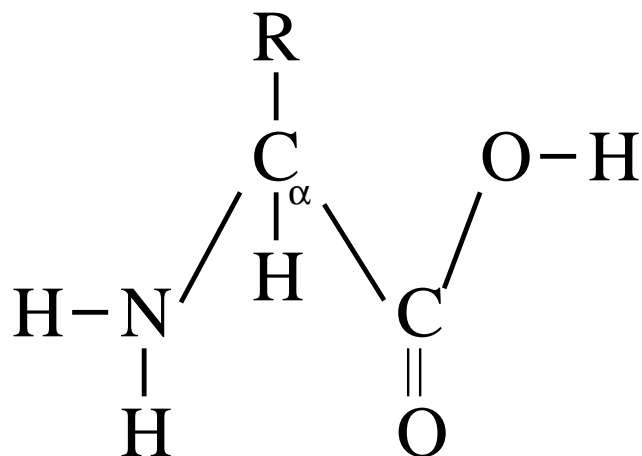


Figure 1.1: Description of an amino acid

To form a protein, amino acids are linked together by peptide bonds, where the α -carboxyl group of one amino acid is joined together with the α -amino group of another amino acid. The resulting chain is referred to as a polypeptide chain, and when the amino acids are parts of a polypeptide chain they are referred to as amino-acid residues. The peptide plane (which is described by the four atoms N, H, C, and O of Fig. 1.2) is fairly rigid and planar. This arises because the link between the carbonyl carbon atom and the nitrogen atom has partial double bond character. The peptide group can therefore exist in cis and trans forms, with small variations of the torsional angle around this bond in both forms. The trans (from latin meaning "across") form of the peptide group is that in which the consecutive R-groups are locked on the opposite sides of the CO-NH peptide group, whereas the cis (from latin meaning "on this side") form is when they are locked on the same side (Fig. 1.3). Figure 1.2 shows the peptide plane in a polypeptide chain.

An important feature of proteins is that they have well-defined unique three-dimensional structure. This unique 3D structure (also termed the native struc-

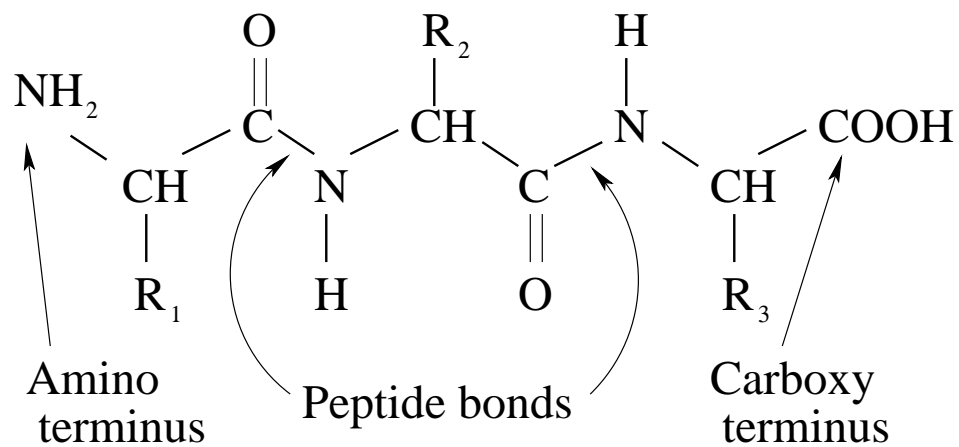


Figure 1.2: Description of the polypeptide chain and the peptide bond

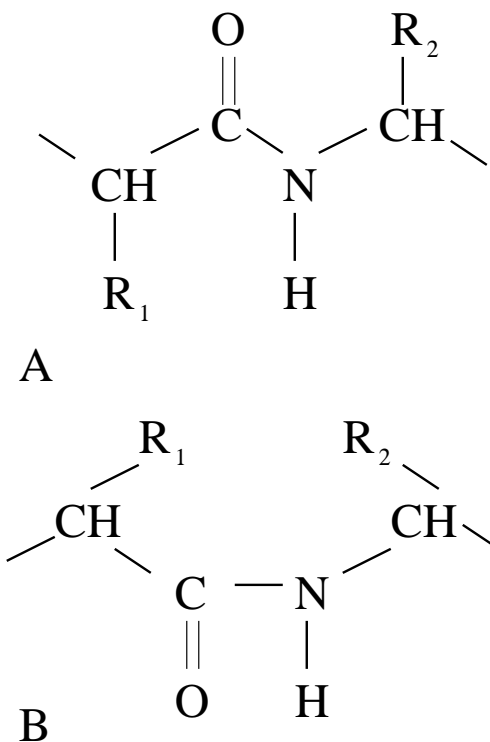


Figure 1.3: Trans (A) and cis (B) forms of a polypeptide chain

ture) is the biologically active form of the protein. In summary, four basic levels of protein structure exist. Primary structure refers to the amino acid sequence. Secondary structure corresponds to the spatial arrangement of neighboring amino acid residues within a chain. Some of these arrangements are fairly regular and give rise to periodic structure. The two main secondary structure components are the α -helix and the β -strand. Tertiary structure refers to the spatial arrangement of residues far from each other in the sequence. This is the three-dimensional shape which determines a protein's function. In addition to tertiary structure, proteins containing multiple subunits possess quaternary structure, which is the spatial arrangement of the subunits and the nature of their contacts.

1.2 Protein Folding

Essentially the protein folding problem can be summarized in one sentence: Given a sequence of amino acids, what is the tertiary (3D) structure of the protein, and how does it get there from the newly synthesized polypeptide chain.

The Protein Folding problem was first investigated experimentally by Anfinsen.¹ His experiment demonstrated that reduced and unfolded bovine pancreatic ribonuclease A (RNase A), could be spontaneously refolded by oxidation of all of its sulfhydryl groups to produce four disulfide bonds, which produced a native biologically-active structure. This result forms the underlying thermodynamic hypothesis, which states that the native conformation is expected to have the lowest free energy of the system (i.e., the protein plus its solvent environment).

Another underlying aspect concerning the protein folding problem is known as the Levinthal Paradox.² The paradox states that protein cannot find its native state by an exhaustive search through all possible conformations. This comes from

realization that a full enumeration of all possible conformations would require an unrealistic amount of time, and thus proteins would never fold in real time; therefore, physical interactions must play an important role in forming the appropriate native fold.

1.3 Structure Prediction

Structure prediction has been one of the most important tasks in computational structural biology, with the goal of being able to predict the nature of the inter-residue interactions that lead to three-dimensional protein structures and their folding pathways from their amino acid sequences. Motivation for protein structure prediction, i.e., prediction of relevant interactions, stems from vastly different fields such as:

1. Medicine: helping to understand biological functions, since binding of proteins with ligands and with other proteins, nucleic acids, carbohydrates and lipids constitute much of the cellular activity of living organisms.
2. Drug Design: Screening target libraries for docking drugs.
3. Agriculture: genetic engineering of richer and more resistant crops.
4. Industry: Synthesis of enzymes (e.g. those that can be incorporated in a mixture with detergents).

1.3.1 Experimental Methods

In practice there are two experimental methods used for protein structure determination. X-ray Crystallography and nuclear magnetic resonance (NMR) spec-

trospectroscopy.

X-ray Crystallography³ requires protein crystals, which are formed by vapor diffusion from purified protein solutions under optimal conditions. The crystals are subjected to X-ray radiation and the resulting diffraction pattern can be interpreted as a reflection of the primary beam source from sets of parallel planes in the crystal. The amplitudes and phases of the diffraction data are used to calculate electron density maps. The corresponding protein structure can then be obtained by fitting the amino acid sequence to the electron density maps.

NMR,⁴ on the other hand, does not require a protein crystal, but treats the protein in solution. Subjecting the solution to a powerful external magnetic field and high frequency radiation results in the splitting of the degenerate energy levels of nuclear spin states. The environment of the component atoms of the proteins determines the magnitude of the energy level splitting and can be used to identify resonance frequencies with particular atoms in the protein. The result is a network of distances involving pairs of spatially-proximate hydrogen atoms. The distances are derived from the Nuclear Overhauser Effects (NOEs) between neighboring atoms. The resulting distances together with other experimental information are converted to a 3D structure with a computational procedure in which an energy function is minimized and structure coordinates which conform to the experimental data are found. Recently, an extra step has been added to the NMR procedure, in which the resulting models are used again to calculate the spectra, and by matching of the calculated spectra to the experimental one, an iterative procedure for improvement is pursued.

1.3.2 Computational Methods

The computational approach to protein structure prediction can be classified into two main categories: Comparative modeling, and Ab initio approach.

Comparative modeling uses the existing database of experimentally determined protein structures⁵ as starting points. This class can be further split into two main subclasses: Homology modeling, and Threading. Homology modeling is based on the assumption that two homologous proteins (proteins that share similar amino-acid sequences) will presumably contain similar 3D structures.⁶⁻¹¹ The sequence of the solved structure is modified to that of the unknown structure and the resulting optimized conformation is the predicted three-dimensional model of the unknown structure. Threading¹²⁻¹⁸ scans the amino acid sequence of the unknown structure against a database of experimental structures.¹⁹⁻²³ A scoring function is evaluated for each comparison to assess the compatibility of the sequence to the structure, thereby producing plausible three-dimensional models.

The Ab initio (also known as De novo) approach is based on the physical principles governing the interactions of amino acids in a polypeptide chain and the surrounding solvent. This approach, which is composed of two key components, is described in more detail in section 1.4. First, an accurate model of the physical interaction within the polypeptide chain is necessary. This is captured in a potential energy function which describes the interatomic physical interactions. The potential energy function must be accurate enough to capture the important interactions yet simple enough, so that calculations can be performed with today's computational power in real time. Force fields of different resolutions (from all-atom to highly simplified coarse grained models) have been developed. Second, assuming an accurate energy function is available, the native fold of the protein populates its

global energy minimum, based on Anfinsen’s hypotheses, which must be located. This task is carried out by a variety of global optimization techniques ranging from energy minimization,^{24–27} to Monte Carlo-based methods²⁸ to Molecular Dynamics procedures.^{29–31}

1.4 Hierarchical Approach to protein structure prediction

As mentioned above, Anfinsen’s thermodynamic hypothesis states that the native structure of a protein is the global minimum of the free energy of a protein plus the surrounding solvent. Global optimization of a potential-energy function is therefore a first-choice approach to physics-based protein-structure prediction. However, it is computationally impossible at present to search the conformational space of an all-atom protein plus explicit water even with the aid of modern global-optimization techniques. Therefore, a hierarchical approach was developed in our laboratory for the computation of protein structure. The approach consists of the following stages:

1. A virtual-bond representation of the polypeptide chain, described by a united-residue (UNRES)^{32–42} potential, and an efficient procedure (Conformational Space Annealing, CSA),^{26, 43, 44} are used to search the conformational space of the virtual-bond chain rapidly. The combination of UNRES and CSA narrows the region of conformational space in which the global minimum is likely to lie, which can be achieved at this stage with the simplified virtual-bond model but not with the all-atom model. A cluster analysis of the resulting ensemble of conformations is carried out, and the lowest-energy conformations are selected for the next stages of the procedure.

2. Next, the lowest-energy conformations obtained in stage 1 are converted to all-atom chains.^{45, 46}
3. The all-atom energy of the chains is searched with the Electrostatically-Driven Monte Carlo procedure (EDMC),^{47, 48} and its energy, expressed by the Empirical Conformational Energy Program for Peptides (ECEPP/3) force field,⁴⁹ is minimized with a Secant Unconstrained Minimization Solver (SUMSL)⁵⁰ subject to the C^α distance constraints from the parent united-residue models.
4. Final energy refinement is carried out with the ECEPP/3 force field⁴⁹ plus the Solvent Radii Fixed with atomic solvation parameters OPTimized (SR-FOPT)⁵¹ surface-hydration model and the EDMC^{47, 48} method as a search technique, with gradual reduction of the $C^\alpha \dots C^\alpha$ distance constraints of the parent model (until they vanish at the end of the procedure).

This approach has been successfully implemented and tested in blind tests of protein structure predictions, as described in section 1.5.

1.5 CASP

CASP (Critical Assessment of Techniques for Protein Structure Prediction)⁵² is a blind test in the Protein Structure Community that takes place every two years. Its goal is to assess the abilities of computational models to predict structures of proteins based solely on their amino acid sequence. Computational groups from all over the world are presented with amino acid sequences of proteins whose structures are not yet publicly known. The number of structures and the length of sequences to be predicted increase in every event. Computational groups have

approximately three months to complete their calculations and submit their top five predicted models for each protein for evaluation. Our group has successfully participated in these exercises and the results from recent CASP5 and CASP6 events are summarized in reference 53.

1.6 Summary of the present work

This thesis describes the author's contribution to the hierarchical approach to protein folding.

First, chapter 2 describes an efficient method, which has been developed for packing α -helices in proteins. It treats α -helices as rigid bodies and uses a simplified Lennard-Jones potential with Miyazawa-Jernigan contact-energy parameters to describe the interactions between the α -helical elements in this coarse-grained system. Global conformational searches to generate packing arrangements are carried out rapidly with a Monte Carlo-minimization type of approach. The results for 42 proteins show that the approach reproduces native-like folds of α -helical proteins as low-energy local minima of this highly-simplified potential function. These results are based on the work published in reference 54. This method can be considered as an extra level of the hierarchy in the hierarchical procedure (it is even a more coarse-grained model than UNRES, i.e., it would be used before point 1 in section 1.4).

Next, because the work in chapters 4 and 5 in this thesis is based on the united residue (UNRES) model of the polypeptide chain, chapter 3 provides a brief overview of all the aspects of the UNRES force field, and its parameter optimization.

In chapter 4, two popular methods of global optimization are coupled, and its

performance is compared with its separate components and with other global optimization techniques. The Replica-Exchange Method together with Monte Carlo-Minimization (REMCM) was applied to search the conformational space of coarse-grained protein systems described by the UNRES force field. The method consists of several noninteracting copies of Monte Carlo simulation, and minimization was used after every perturbation to enhance the sampling of low-energy conformations. REMCM was applied to five proteins of different topology, and the results were compared to those from other optimization methods, namely Monte Carlo-Minimization (MCM), Conformational Space Annealing (CSA) and Conformational Family Monte Carlo (CFMC). In summary, REMCM located global minima for four proteins faster and more consistently than either MCM or , and it converged faster than CSA on three of the five proteins tested. A performance comparison was also carried out between REMCM and the traditional Replica Exchange method (REM) for one protein, with REMCM showing a significant improvement. Moreover, because of its simplicity, it was easy to implement, thereby offering an alternative to other global optimization methods used in protein structure prediction. This chapter is based on work in reference 55.

In chapter 5 efficient methods for calculating thermodynamic averages were implemented with the united residue (UNRES) force field, namely a Replica Exchange method (REM), a Replica Exchange Multicanonical method (REMUCA), and Replica Exchange Multicanonical with Replica Exchange (REMUCAREM), were implemented with the coarse-grained UNRES force field in both Monte Carlo and Molecular Dynamics versions. The MD algorithms use the constant-temperature Berendsen thermostat, with the velocity Verlet algorithm and variable time step. The algorithms were applied to one peptide (20 residues of Alanine with free ends;

ala₂₀) and two small proteins, namely an α -helical protein of 46 residues (the B-domain of the staphylococcal protein A; 1BDD), and an $\alpha+\beta$ -protein of 48 residues (the E. Coli MltD LysM Domain; 1E0G). Calculated thermodynamic averages, such as canonical average energy and heat capacity, are in good agreement among all simulations for poly-L-alanine, showing that the algorithms were implemented correctly, and that all three algorithms are equally effective for small systems. For protein A, all algorithms performed reasonably well, although some variability in the calculated results was observed whereas, for a more complicated $\alpha + \beta$ -protein (1E0G), only Replica Exchange was capable of producing reliable statistics for calculating thermodynamic quantities. Finally, from the Replica Exchange molecular dynamics results, we calculated free energy maps as functions of RMSD and radius of gyration for different temperatures. The free energy calculations show correct folding behavior for poly-L-alanine and protein A while, for 1E0G, the native structure had the lowest free energy only at very low temperatures. Hence, the entropy contribution for 1E0G is larger than that for protein A at the same temperature. A larger contribution from entropy means that there are more accessible conformations at a given temperature, making it more difficult to obtain an efficient coverage of conformational space to obtain reliable thermodynamic properties. At the same temperature, ala₂₀ has the smallest entropy contribution, followed by protein A, and then by 1E0G. This work is based on reference 56.

Since previously-developed methods for global optimization are either implemented (is utilized in chapter 2), or used for comparison (and CSA are used in chapter 4), appendix A gives a brief overview of these methods.

BIBLIOGRAPHY FOR CHAPTER 1

- [1] Anfinsen, C. B., *Science* 1973, 181, 223.
- [2] Levinthal, C., *J. Chim. Phys. PCB* 1968, 65, 44.
- [3] Kendrew, J.; Perutz, M., *Annul Rev BioChem.* 1957, 26, 327.
- [4] Wüthrich, K.; Wider, G.; Wagner, G.; Braun, W., *J. Mol. Biol.* 1982, 155, 311.
- [5] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M., *J. Mol. Biol.* 1977, 112, 535.
- [6] Baker, D.; Sali, A., *Science* 2001, 294, 93.
- [7] Chothia, C.; Lesk, A. M., *EMBO J.* 1986, 5, 823.
- [8] Vitkup, D.; Melamud, E.; Moulton, J.; Sander, C., *Nat. Struct. Biol.* 2001, 8, 559.
- [9] Jones, T. A.; Thirup, S., *EMBO J.* 1986, 5, 819.
- [10] Sali, A., *Curr. Opin. in Struct. Biol.* 1995, 6, 437.
- [11] Johnson, M. S.; Srinivasan, N.; Sowshamini, R.; Blundell, T. L., *Crit. Rev. Biochem. Mol. Biol.* 1994, 29, 1.
- [12] Wodak, S. J.; Rooman, M. J., *Curr. Opin. in Struct. Biol.* 1993, 3, 247.
- [13] Jones, D.; Thornton, J., *J. of Comput. Aided Mol. Des.* 1993, 7, 439.
- [14] Bowie, J. U.; Eisenberg, D., *Cur. Opin. Struct. Biol.* 1993, 3, 437.
- [15] Lemer, C.; Rooman, M. J.; Wodak, S. J., *Proteins: Struct., Funct., Genet.* 1996, 23, 337.
- [16] Godzik, A.; Kolinski, A.; Skolnick, J., *J. Mol. Biol.* 1992, 227, 227.
- [17] Wilmanns, M.; Eisenberg, D., *Prot. Eng.* 1995, 8, 627.
- [18] Jones, D. T.; Taylor, W. R.; Thornton, J. M., *Nature* 1992, 358, 86.
- [19] Orengo, C. A.; Jones, D. T.; Thornton, J. M., *Nature* 1994, 372, 631.
- [20] Orengo, C. A.; Flores, T. P.; Jones, D. T.; Taylor, W. R.; Thornton, J. M., *Curr. Biol.* 1993, 6, 131.

- [21] Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C., *Prot. Sci.* 1992, 1, 409.
- [22] Boberg, J.; Salakoski, T.; Vihinen, M., *Proteins: Struct., Funct., Genet.* 1992, 14, 265.
- [23] Fischer, D.; Tsai, C. J.; Nussinov, R.; Wolfson, H., *Prot. Eng.* 1994, 8, 981.
- [24] Li, Z.; Scheraga, H. A., *Proc. Natl. Acad. Sci., U. S. A.* 1987, 84, 6611.
- [25] Li, Z.; Scheraga, H. A., *J. Molec. Str. (Theochem)* 1988, 179, 333.
- [26] Lee, J.; Scheraga, H. A.; Rackovsky, S., *J. Comput. Chem.* 1997, 18, 1222.
- [27] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.
- [28] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., *J. Chem. Phys.* 1953, 21, 1087.
- [29] Duan, Y.; Kollman, P. A., *Science* 1998, 282, 740.
- [30] Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B., *Biopolymers* 2003, 68, 91.
- [31] Elber, R.; Ghosh, A.; Cárdenas, A., *Acc. Chem. Res.* 2002, 35, 396.
- [32] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *Protein Sci.* 1993, 2, 1697.
- [33] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *Protein Sci.* 1993, 2, 1715.
- [34] Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *J. Comput. Chem.* 1997, 18, 849.
- [35] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A., *J. Comput. Chem.* 1997, 18, 874.
- [36] Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A., *J. Comput. Chem.* 1998, 19, 259.
- [37] Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A., *J. Chem. Phys.* 2001, 115, 2323.
- [38] Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A., *J. Phys. Chem. B* 2001, 105, 7291.

- [39] Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Ołdziej, S.; Arnautova, Y. A.; Scheraga, H. A., *J. Phys. Chem. B* 2001, 105, 7299.
- [40] Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H.A., *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 1937.
- [41] Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A., *J. Phys. Chem. A* 2003, 107, 8035.
- [42] Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 9421.
- [43] Lee, J.; Scheraga, H. A., *Int. J. Quant. Chem.* 1999, 75, 255.
- [44] Czaplewski, C.; Liwo, A.; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [45] Kazmierkiewicz, R.; Liwo, A.; Scheraga, H.A., *J. Comput. Chem.* 2002, 23, 715.
- [46] Kazmierkiewicz, R.; Liwo, A.; Scheraga, H.A., *Biophys. Chem.* 2003, 100, 261.
- [47] Ripoll, D. R.; Scheraga, H. A., *Biopolymers* 1988, 27, 1283.
- [48] Ripoll, D. R.; Scheraga, H. A., *J. Protein Chem.* 1989, 8, 263.
- [49] Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A., *J. Phys. Chem.* 1992, 96, 6472.
- [50] Gay, D. M., *ACM Trans. Math. Software* 1983, 9, 503.
- [51] Vila, J.; Williams, R. L.; Vásquez, M.; Scheraga, H. A., *Proteins Struct. Funct. Genet.* 1991, 10, 199.
- [52] <http://predictioncenter.org/>.
- [53] Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nancias, M.; Vila, J.A.; Khalili, M.; Arnautova, Y.A.; Jagielska, A.; Makowski, M.; Schafroth, H.D.; Kazmierkiewicz, R.; Ripoll, D.R.; Pillardy, J.; Saunders, J.A.; Kang, Y.K.; Gibson, K.D.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 7547.
- [54] Nancias, M.; Chinchio, M.; Pillardy, J.; Ripoll, D.R.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2003, 100, 1706.
- [55] Nancias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H.A., *J. Comp. Chem.* 2005, 26, 1472.
- [56] Nancias, M.; Czaplewski, C.; Scheraga, H.A., *J. Chem. Theo. Comp.* 2005, in press.

Chapter 2

Packing helices in proteins by global optimization of a potential energy function *

2.1 Introduction

The problem of determining the structure of a protein starting from its amino-acid sequence has been approached from many different directions. Knowledge-based methods cannot predict entirely new folds, while ab-initio methods have this capability but are generally less accurate and more computationally intensive. One class of ab-initio methods is based on the minimization of a potential energy function. This immediately presents the challenge of producing a potential function that identifies the native fold as the lowest-energy structure, yet remains simple enough to permit adequate sampling of the conformational space.

If the secondary structure is known, the space that needs to be searched becomes much smaller, but still contains a very large number of incorrect packing arrangements. The secondary structure can either be predicted from the sequence

*Published as Nianias, M.; Chinchio, M.; Pillardy, J.; Ripoll, D.R.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2003, 100, 1706. Copyright (2003) National Academy of Sciences, U.S.A.

(using programs such as Jpred/Jnet^{1,2} Psipred³ etc.) or it can be extracted from the preliminary output of another method. Here, we demonstrate the feasibility of using a highly simplified energy-based method to pack secondary-structure elements in which the positions of residues within these elements are fixed. Each residue is represented by just one interaction center and the potential employed is much simpler than in previous work.⁴ Because helical structures have a simple geometry, the procedure is applied to 42 mainly α -helical proteins. It is shown that, for most structures with six or fewer helices, a limited number of plausible conformations can be identified that contain native-like structures, while completely wrong folds are eliminated. The resulting ensemble of conformations can then be used as a starting point for a search with a more detailed model and potential, such as UNRES,⁵ to refine and rank the predicted conformations. Some of the proteins investigated are 100-200 residues long (which overcomes a limitation of some previous studies⁶), but this does not seem to present any problems.

2.2 Methods

Our procedure uses an energy-driven Monte-Carlo-like search to generate an ensemble of plausible structures, and consists of three main parts. First a simplified representation of a protein is constructed. Then, a potential function is developed to assign an energy to a given conformation. Finally, a search is carried out to find the optimal (lowest-energy) arrangement of secondary structure elements.

2.2.1 Protein Representation

Given a sequence of amino acids and the corresponding secondary-structure assignment, we represent a protein only by its C^α atoms. Coordinates for loop residues are left unspecified (see the following section), while coordinates for residues in α -helical regions are constructed using ideal parameters,^{5, 7} namely, 3.6 residues per turn, 1.5 Å per residue along the helix axis, and 3.8 Å virtual C^α - C^α bond-length. Helices are then treated as rigid objects, simply described by the positions of their centroids and their orientations, while the relative positions of the residues within a given α -helix are fixed.

2.2.2 Potential Energy Function

The energy function is the pairwise interaction between two residues, m and n , of amino acid type i and j :

$$U(r_{mn}) = e_{ij} \left[\frac{q \left(\frac{r_0}{r_{mn}} \right)^p \pm p \left(\frac{r_0}{r_{mn}} \right)^q}{q \pm p} \right] \quad (2.1)$$

where p , q ($q < p$) and r_0 are adjustable parameters, r_{mn} is the distance between the C^α atoms of residues m and n , and e_{ij} is the contact energy associated with residues of types i and j . The signs are chosen to obtain a repulsive interaction if $e_{ij} > 0$, or negative if $e_{ij} < 0$, and to ensure that $U(r_0) = e_{ij}$, as in a Lennard-Jones potential. The main purpose is to capture the tendency for nonpolar residues to be buried in the cores of proteins.⁷ The contact potential developed by Miyazawa and Jernigan⁸ has been shown to represent the properties of nonpolar residues accurately,⁹ and it also provides interaction energies for the polar residues. The matrix of contact energies provided by Miyazawa and Jernigan¹⁰ is used for the

parameters e_{ij} . In their treatment, Miyazawa and Jernigan consider two residues to be in contact if the distance between their side-chain centroids is less than 6.5\AA . In eq. 2.1, the interaction is smoothed and equals e_{ij} only at the special contact distance r_0 (even when the interaction is purely repulsive).

The energy for a multi-helical structure is then calculated by summing over the interactions between all residue pairs belonging to distinct α -helices. There is no interaction between residues within an α -helix (since the relative coordinates are fixed), nor with residues belonging to loops. For this reason, coordinates for residues in loops are not necessary. The only contribution that loops make to the energy is a penalty if the distance between the ends of two helices connected by a loop becomes greater than the maximum length allowed for that loop (the number of bonds times the virtual C^α - C^α bond length, 3.8\AA).

2.2.3 Global Optimization

To search the conformational space of a particular structure, an efficient global optimization method, Conformation-Family Monte Carlo (CFMC),¹¹ previously developed in our laboratory, was employed with small modifications. This search is based on a conformational family database, which is an ensemble of conformations clustered into families.

The starting point for the search is the sequence and secondary-structure information. Helices are then built, using values of ideal α -helices as mentioned above.

The procedure clusters structures into families, in which each structure is similar to at least one other conformation within its family. A structure is said to

be similar to another structure or a family if a distance measure provides a value which is smaller than a chosen cutoff. The same is true for two structures being identical except that the cutoff values are stricter. The two distance measures used are explained in a following section.

To control the computational expense, the number of families and the number of structures within one family have a limit of N_f and N_c , respectively. The ensemble is initialized with N_f non-redundant structures selected randomly and then energy-minimized with the SUMSL algorithm.¹² This defines the initial phase after which the actual search starts. In each iteration of the search, a conformation is selected with a probability according to its Boltzmann weight. This structure is subsequently perturbed, its energy is minimized, and similarity, energy and metropolis tests are carried out to determine whether it will be kept in the ensemble and/or it forms a new family. The temperature was adjusted to maintain a reasonable fraction of new generating families. Thus, the conformations are improved iteratively, and the search is biased to investigate the regions of the lowest-energy families while trying to explore different areas of conformational space effectively. In every iteration, the perturbed structure is checked quickly to determine whether loops could be constructed without clashes. This is done by treating the C^α atoms of the loops as spheres with diameter set to the bond length. Using a soft-sphere potential [cubic in the extent (distance) of overlap] and subject to bond-length constraints, the energies of these residues are then minimized and checked to determine whether any clashes within each loop or between loops and α -helices occurred.

Since CFMC was originally applied to a united-residue model, it had to be modified for a rigid-body treatment of secondary-structure elements; i.e., a differ-

ent method for producing new conformations, described in the following section, was applied. Also, a new distance measure was devised to suit the objective of finding an ensemble of different folds.

2.2.4 Methods for Producing New Conformations.

Two major classes of moves were used for producing new conformations. The first one, called *Global Move*, produces radically different structures. This involves moves, such as randomizing the positions and orientations of all helices, by translational and rotational motions of any number of helices. Helices are allowed to flip upside down or have the positions of any two of them swapped while keeping the relative orientation unchanged. Moves are chosen randomly and can be combined in any number of ways to perturb the generating structure.

The second class, called *Local Move*, is designed to produce very similar structures. Like global moves, it also involves translations and rotations of α -helices, but only by much smaller distances and angles. The values by which the helices are translated and rotated are chosen randomly but they are bound by an upper limit which is different in global and local moves (Global: translation up to 15Å, rotation up to 360°; Local: translation up to 4Å, rotation up to 50°). Local moves can also rotate a helix (up to 180°) or shift it (up to 3Å) along its axis. The idea behind these moves is that, if a conformation has correct packing but wrong relative orientation, a local move should try to improve it.

2.2.5 Distance Measures

Two methods were used to describe the similarity of two structures.

1. *rmsd between C^α atoms in helices*. Unfortunately, the C^α rmsd does not provide an unambiguous measure to determine if the correct (i.e., native-like) fold is obtained. For example, if the alignment is not very good, the rmsd will be high but the folded protein might have correct orientation of secondary-structure elements. Also this number grows with the size of the protein; therefore, comparison of performance of the method for two proteins of different size is not straightforward. This measure was used only to present the results.
2. *Center of Mass rmsd and Maximum Angle (CMrmsd & MaxAngle)*. This distance measure was devised as a replacement for the C^α rmsd. The method works as follows. The centers of mass of each helix in the two conformations to be compared are superimposed. The angle between the axes of every pair of corresponding helices is calculated and the maximum angle taken. The center of mass rmsd and the maximum angle are the two values used to determine similarity. This measure works better for differentiating the correct orientation of helices from the wrong ones, and thus was used in the search for the definition of the families.

2.2.6 Protein Targets

Three main sources of target α -helical proteins were used in the simulations, namely, all 24 α -helical proteins from Zhang et al.,¹³ a set of α -helical proteins obtained from other simulations in our laboratory, and a set extracted from the

SCOP database¹⁴ (version 1.61), in which only proteins from the α -class and belonging to different families were considered. All three sources provided 42 proteins (36-188 residues long), which were a representative and diverse pool of target structures. The secondary structure information used in our simulations was determined by applying the dssp algorithm¹⁵ to the native structure.

2.3 Results

To produce a set of consistent results, most of the adjustable parameters were kept uniform for all the proteins tested. The potential parameters p , q and r_0 were set to 15, 14 and 7.5\AA , respectively. While a different set of parameters could perform slightly better for a particular protein, the values used were chosen for best performance over the entire set of 42 proteins, particularly the smaller ones (up to 5 helices).

The computations were carried out primarily on dual AMD Athlon MP 1800+ based machines (although only one processor was used). The searches for all 42 searches consisted of 10,000 iterations each, which kept the time for a complete search between 1 and 10 hours, depending on the protein size. Primarily, one such run was carried out for each protein, although several runs were carried out for a few models to check reproducibility. The similarity between structures was determined according to the CMrmsd & MaxAngle measure described above (to belong to the same family, the MaxAngle cutoff was 60° and the CMrmsd cutoff was between 2.5\AA and 4.5\AA , depending on the protein size and complexity, i.e., number of α -helices). To generate diverse packing arrangements, 75% of the moves were global, and only 25% were local. The size of the ensemble was increased with

protein complexity (from 100 families, each containing 4 structures, to 250 families, each containing 6 structures). At the end of each search, the entire ensemble was reclustered according to a stricter criterion: each structure within a family had to be similar to the lowest-energy member, not just to any other structure in that family. This was done to strengthen the link between a given structure and its family number (which is determined by sorting families according to the energy of their lowest member). Naturally, this increases the number of families, but it also makes the family number a more relevant property of a structure.

Tables 2.1 and 2.2 present the results of the simulations. 1dv5 had the structure closest to the native fold, with $\text{rmsd} = 2.2\text{\AA}$, which was also found as the global minimum (i.e. the lowest-energy structure in the lowest-energy family). 1i6z and 1a6s also had native-like global-minimum structures, 1kdx and 1dlw had structures resembling the native-like fold within the lowest-energy family.

Figure 2.1 shows the difficulty of obtaining structures with native-like folds for proteins with increasing numbers of helices. The three graphs are plots for the percentage of all proteins with the corresponding number of helices in the 20, 60 and 130 lowest-energy families, respectively, for which the method retrieves a fold within the rmsd indicated in the inset. For example, the structures of all three-helix proteins were within 4.5\AA rmsd from their native, where the computed structures were ranked in the 20 lowest-energy families of the final ensemble. It is important to note that, as the number of helices increases, the percentage of successful computations within the same rmsd decreases.

Figure 2.2 shows a superposition of a computed structure for 1ifo with its native structure. The superimposed structures agree to within 4.8\AA rmsd and show that the overall orientation of all helices is qualitatively correct. This is not

Table 2.1: Protein name (pdb id), followed by the number of helices, the total number of residues (excluding the non-helical residues at the N- and C-termini), and the number of residues only in helices. The last three columns show the best results obtained for the 20, 60 lowest-energy, and all families, respectively. The entry indicates the rmsd value in Å measured on C $^{\alpha}$ atoms of helices from the native, followed by the corresponding family number (in parentheses). The empty fields indicate that the value to the left is not improved by including more families. (b) The following are fragment proteins: 1lbu: 1lbu₁₋₈₃; 1ffh: 1ffh₂₋₈₈; 1aisB: 1aisB₁₁₀₈₋₁₂₀₅; 1b0nA: 1b0nA₁₋₆₈; 1bmtA: 1bmtA₆₅₁₋₇₄₀.

Protein	N	Nres		Best Result rmsd _{min}		
		tot	hel	low 20	low 60	all
1cktA	3	61	47	3.6(9)		
1dv5	3	75	34	2.2(1)		
1fex	3	50	31	3.4(6)		
1g2h	3	36	28	3.4(20)		
1gab	3	42	35	2.9(6)		
1hdp	3	44	33	3.7(11)		
1i6z	3	114	102	2.5(1)		
1kdxA	3	66	50	2.6(1)		
1lbu ^b	3	60	32	3.9(6)		
1lea	3	48	39	3.1(7)		
1lre	3	66	55	3.4(10)		
2occH	3	53	42	4.0(15)	3.0(21)	
1a04	4	56	45	4.9(19)	4.7(31)	
1a6s	4	85	46	4.4(1)		
1bw6	4	43	29	4.1(17)	3.6(25)	2.7(93)
1c5a	4	61	46	4.6(7)	4.4(23)	
1eij	4	59	41	4.8(5)	4.6(21)	3.7(159)
1ffh ^b	4	83	63	3.7(11)	3.7(11)	3.0(75)
1hdj	4	61	40	5.2(16)	3.9(22)	
1unkA	4	67	48	4.7(18)	3.7(28)	3.2(146)

Table 2.2: (a) Protein name (pdb id), followed by the number of helices, the total number of residues (excluding the non-helical residues at the N- and C-termini), and the number of residues only in helices. The last three columns show the best results obtained for the 20, 60 lowest-energy, and all families, respectively. The entry indicates the rmsd value in Å measured on C $^{\alpha}$ atoms of helices from the native, followed by the corresponding family number (in parentheses). The empty fields indicate that the value to the left is not improved by including more families. (b) The following are fragment proteins: 1lbu: 1lbu₁₋₈₃; 1ffh: 1ffh₂₋₈₈; 1aisB: 1aisB₁₁₀₈₋₁₂₀₅; 1b0nA: 1b0nA₁₋₆₈; 1bmtA: 1bmtA₆₅₁₋₇₄₀.

Protein	N	Nres		Best Result rmsd _{min}		
		tot	hel	low 20	low 60	all
2abd	4	79	49	6.9(16)	4.1(28)	
1aisB ^b	5	88	67	6.7(4)		
1b0nA ^b	5	60	42	5.4(3)		
1b0x	5	62	43	4.0(7)	3.3(29)	
1beg	5	91	55	6.2(11)	6.2(11)	5.5(83)
1bmtA ^b	5	79	61	6.6(2)	6.6(2)	3.7(65)
1ctj	5	82	46	8.3(20)	7.4(35)	5.4(230)
1flf	5	85	48	5.9(8)		
1f68	5	100	66	8.8(13)	8.2(37)	6.2(93)
1lpe	5	138	117	3.4(6)		
1nfo	5	136	110	3.0(9)		
1nkl	5	70	54	5.2(14)	4.0(25)	
1qc7A	5	74	58	8.1(14)	6.6(34)	5.5(145)
2ezyA	5	83	54	6.7(17)	6.0(46)	5.4(129)
1bxm	6	92	50	7.0(4)	7.0(4)	6.4(229)
1fio	6	188	162	10.3(12)	6.1(25)	
1ngr	6	71	49	7.3(18)	5.4(59)	
1rzl	6	71	49	7.1(7)	5.7(32)	4.8(123)
1a0b	7	109	87	11.1(4)	8.4(24)	8.0(140)
1dlw	7	112	72	6.1(1)		
1emy	7	145	107	11.4(9)	8.4(57)	8.1(281)
1ezt	8	125	89	12.6(13)	11.2(59)	11.0(175)

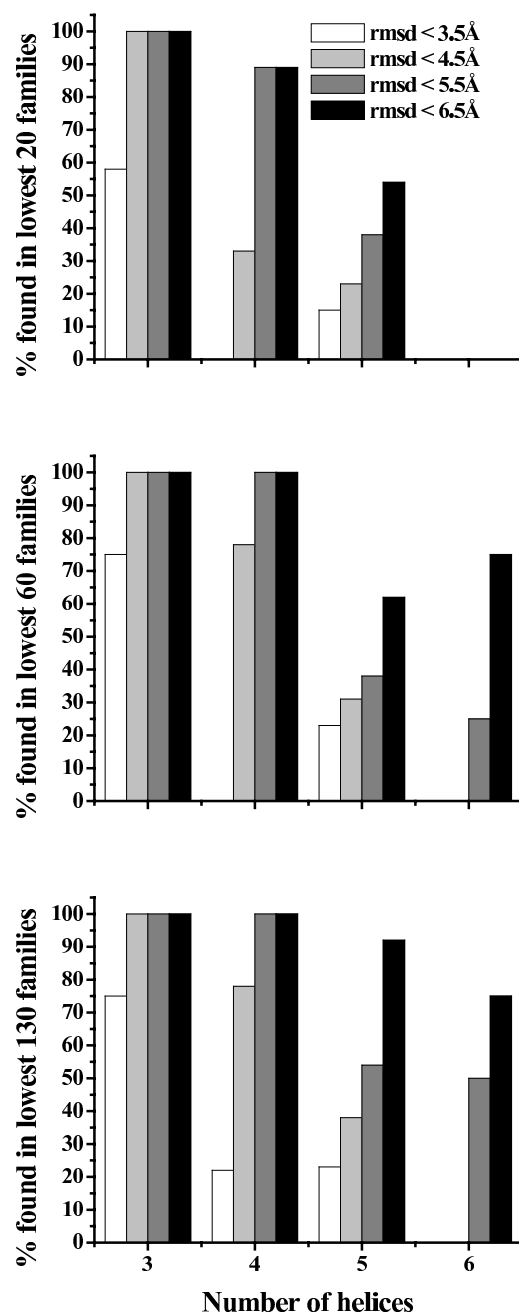


Figure 2.1: Percentage of all proteins with corresponding number of helices for which at least one structure was generated within the rmsd from the native indicated in the inset. The graphs correspond to the 20, 60, and 130 lowest-energy families respectively.

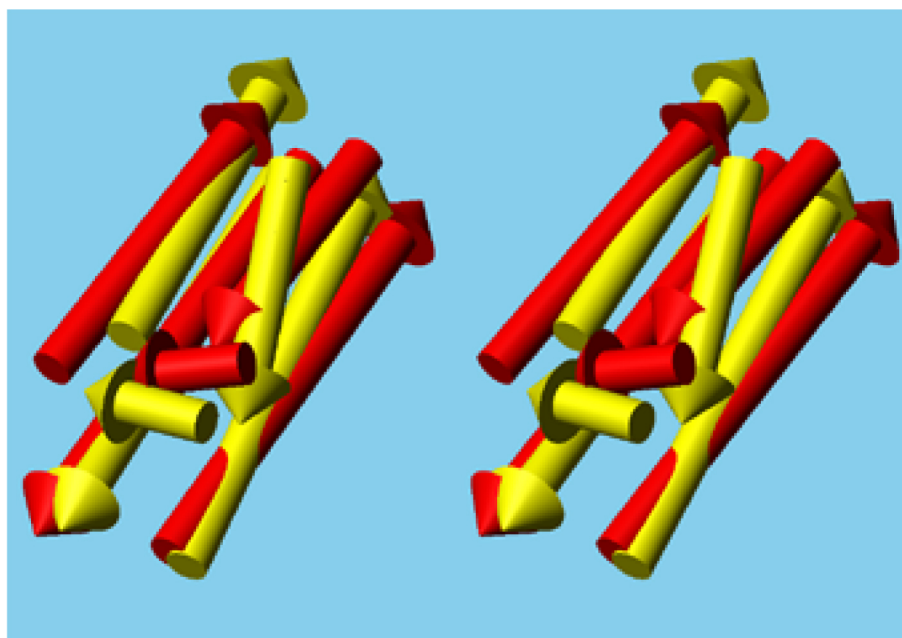


Figure 2.2: Stereo-view of the superposition of a generated structure of the five-helix protein 1fo (not the best) on the experimental structure. The C^α atoms agree to within an rmsd of 4.8\AA . The native structure is yellow, whereas the generated structure is portrayed in red.

the best conformation obtained; the rmsd of the best one is 3.0Å (see Table 1).

To determine the stability of the procedure with different positions of secondary structure elements in the sequence, several simulations were carried out on 6 of the 42 proteins (pdb codes: 1lre, 2abd, 1a6s, 1g2h, 1hdp and 1ctj) with different assignments of secondary structure, according to dssp and JNET/JPRED, respectively. The results are shown in Table 2.2 and are quite comparable with the ones from Table 2.1; thus, it seems that our procedure is stable with respect to secondary-structure assignment.

From Figure 2.1, it is clear that, as the number of helices grows, the performance of the method decreases. One source of difficulty is the imperfection of the potential function itself. Given all the simplifications in this approach, it would be unreasonable to expect the global-energy minimum to identify the native structure in all cases. For example, loops can play a role in determining the structure,¹⁶ but are neglected here. Also, some of the proteins examined are only parts of larger structures, the effects of which are also neglected. However, native structures ideally should always be present among the low-energy conformations, as shown in Fig 2.3. This has been confirmed for 41 of the above proteins (the exception being 1ais) by performing searches restricted to the neighborhood of the native structure. Native-like structures with low energies are generally present, even when searches without such restrictions fail to find them (examples being 1a0b, 1emy, 1ezt). The reason for this is the complexity of the fold and the large number of local energy minima in the search space. Even with a simplified potential, searches for proteins with 6 or more helices are not complete in 10,000 steps. In these cases, models within 6.0Å from the native are found within the final ensemble only if two helices are omitted from the comparison (i.e., 5- instead of 7-helix fragments for 1a0b

Table 2.3: Stability of procedure with respect to different secondary structure assignment. Protein name, H_{dssp} , number of helices according to dssp. H_{JNET} , number of helices predicted by JNET/JPRED. Q3, percentage of correctly predicted secondary structure. $rmsd_{min}$, lowest rmsd in Å (corresponding family number in parentheses) from the native structure in the whole ensemble, and in the 10 lowest energy-families, respectively.

Protein	H_{dssp}	H_{JNET}	Q3	$rmsd_{min}$	
				all	10
1lre	3	3	76	3.6(77)	5.5(1)
2abd	4	4	86	3.9(145)	4.9(3)
1a6s	4	4	68	4.7(9)	4.7(9)
1g2h	3	4	53	5.1(25)	5.2(7)
1hdp	3	3	82	2.3(3)	2.3(3)
1ctj	5	4	83	5.2(56)	8.3(4)

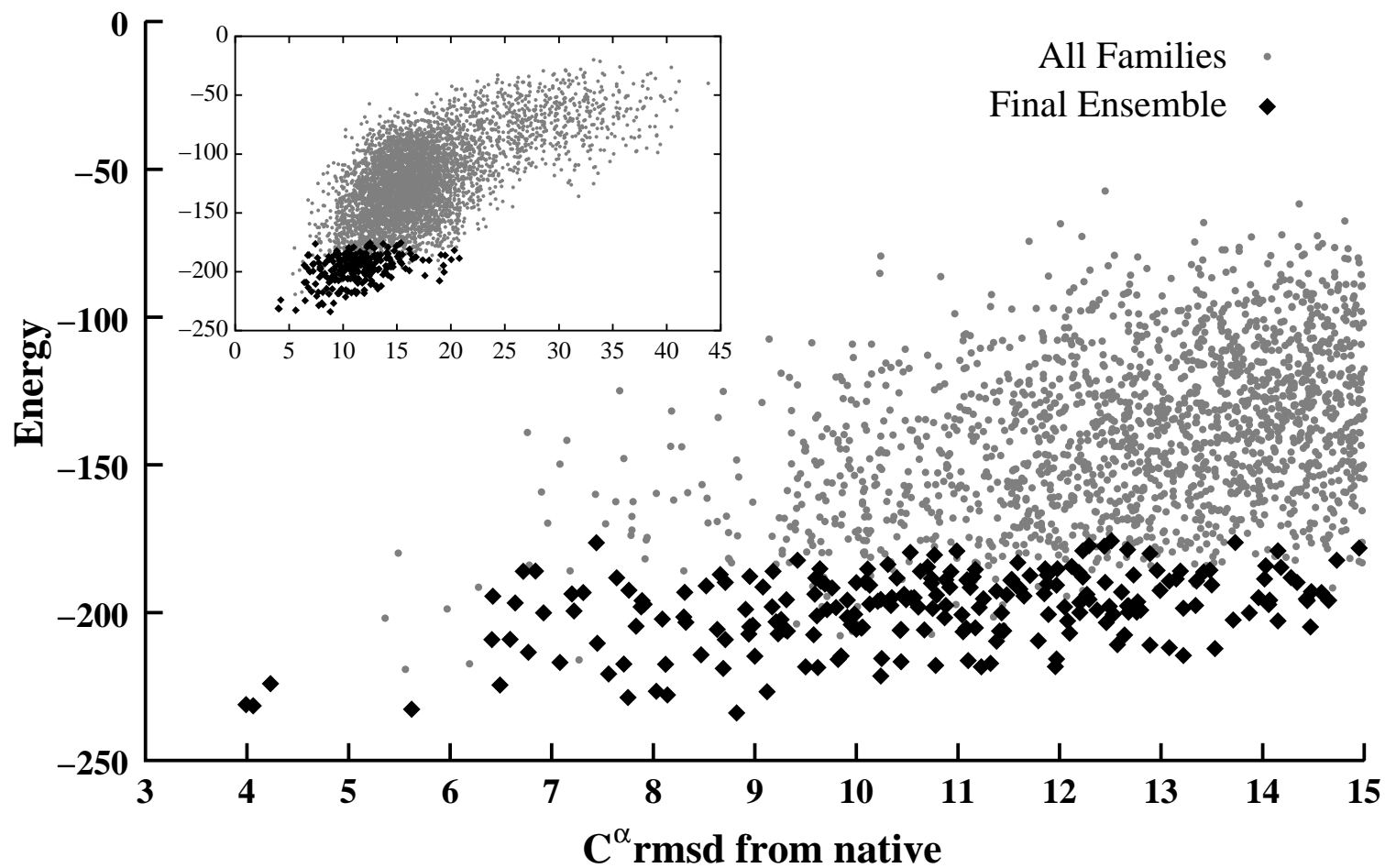


Figure 2.3: Energy vs rmsd for 1LPE. Black points are families in the final ensemble, whereas grey points are families encountered during the search of 10,000 iterations.

and 1emy, and 6- instead of 8-helix fragments for 1ezt). 1ais is the only protein for which native-like structures have significantly higher energies than the global minimum. Closer examination reveals that this structure is much more compact than the others, and in fact the results are improved by decreasing the parameter r_0 from 7.5Å to 6.0Å.

2.4 CASP6 Results

The repacking algorithm was employed in the latest CASP6 exercise, which took place between April and September 2004. Sixty-two targets whose structures were determined by experimental techniques, were available for prediction, out of which our group submitted predictions for 32 targets. Due to the fact that the repack algorithm works only on α -helical proteins, its use was limited to a few targets, which were predicted to have only an α -helical fold. This procedure produced great results for target T0198. Target T0198 is a large 235-residue α -helical protein (phosphate transport system regulator PhoU from *T. maritima*; PDB ID code 1SUM), classified by CASP assessors as a fold recognition/analogy target. The protein is composed of six α -helices, which form a bundle and a small C-terminal β -hairpin. After secondary structure prediction showed that this protein is mainly α -helical, the repack algorithm was applied in the computation of its structure. The lowest-energy conformations resulting from repack were converted to the UNRES representation and were subjected to a local search by using the UNRES potential.¹⁷ The structures obtained from this two-stage procedure were clustered and energy-ranked together with structures resulting from the regular UNRES/CSA search. One model resulting from the two-stage procedure, ranked as model 5,

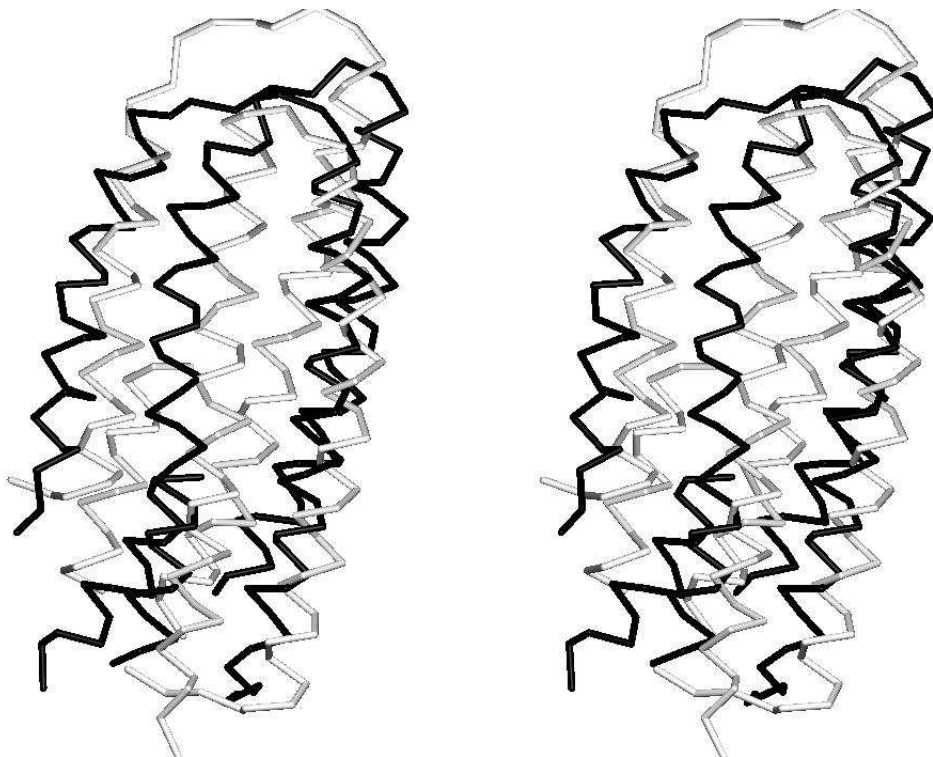


Figure 2.4: Fragment superposition of the submitted model 5 of T0198 with the experimental structure (1SUM). 153 residues superimpose within RMSD of 5.9 Å. Experimental structure is shown in white, whereas the submitted model is shown in black.

was submitted as a prediction. This model was our best prediction with overall rmsd 9.8 Å over all C α s, but with correct topology of the bundle (see Fig. 2.4 in which 153 nonconsecutive residues superimpose within RMSD of 5.9 Å). Also, 139 residues (62% of the sequence; the first three α -helices of the six-helix bundle) fit a 6.0-Å rmsd cut-off (data not shown), and 203 residues (86% of the sequence) that constitute the whole protein, except that the C-terminal part contains mainly a β -hairpin, with a 8.0-Å rmsd cut-off (data not shown).

Figure 2.5 shows a global distance test (GDT) analysis of T0198. GDT analysis is used in the CASP exercise to evaluate and compare the quality of predictions of different groups. The graph shows the largest set of C α atoms that can fit under a distance (not rmsd) cutoff (i.e., all residues from the native and the submit-

ted model are compared sequentially and the number of residues that are within the specified distance cutoff is reported). The global distance test total score (GDT_TS) provides a reasonable single value approximation of the quality of the tertiary structure prediction. It is defined as the average of four separate GDT calculations identifying maximal sets of residues at 1, 2, 4 and 8 Å distance cutoffs. The blue curve shows the results of model 5 predicted with repack and UNRES, the green curves are the results of our traditional UNRES/CSA approach, whereas the brown curves represent the results of other groups. The repack model for T0198 had the correct topology with 141 residues (62%) fitting a 8-Å distance cut-off, with a GDT_TS score of 31.78 % corresponding to 12th place in the ranking of 454 models and is thus far the largest protein predicted correctly by our physics-based approach.

2.5 Discussion

Packing of secondary structure elements is one of the important steps in achieving the ultimate goal of predicting a structure from sequence. We have developed an energy-based method to generate a variety of folds by treating α -helices as rigid bodies, applying a simple potential and searching the conformational space with a Monte Carlo-type search. Despite the simplicity of our model, we were able to produce native-like folds ranked in low-energy families for many proteins.

Although the method provided good results for proteins with a small number of helices, there is considerable room for improvement in our procedure. It is important to note that it is the number of helices rather than the size of the protein that seems to cause difficulties. A more systematic approach to generate diverse topologies would increase the probability of locating native-like folds.¹⁸ Further

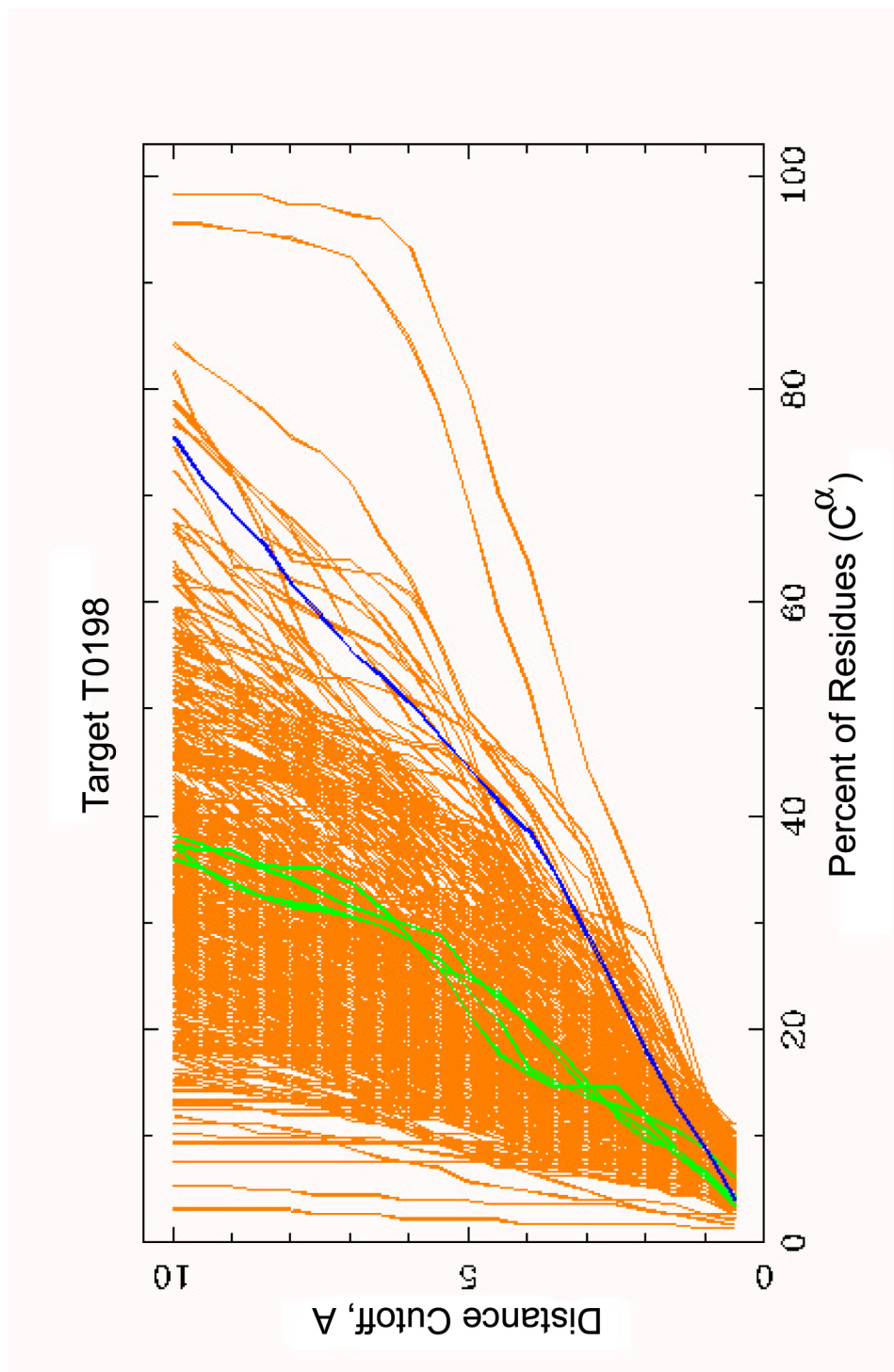


Figure 2.5: GDT analysis of T198. Largest set of C^α atoms (percent of the modeled structure) that can fit under DIS-TANCE cutoff. Blue: Model 5 (predicted with Repack, local search, and subsequently converted to UNRES representation), Green: Models predicted by UNRES and CSA, Brown: Models submitted by other groups.

improvements could come from modifications to the contact energies that take into account the environment of a residue¹⁹ (i.e., the kind of secondary structure element to which it belongs), or by carrying out a systematic optimization procedure for the potential parameters.²⁰ Another possibility is the improvement of the functional form of the potential or the protein representation, which could be further simplified to reduce the large number of local minima in our conformational space.

Although generating folds is an important step, the main purpose of this exercise is to continue with the refinement of the generated models by using them as input for an algorithm with a more detailed representation of the polypeptide chain, such as the united-residue model.⁵ The procedure described here greatly reduces the number of helical conformations that have to be explored with the united-residue model.

Currently only α -helices are treated by this simple procedure, but inclusion of β -strands and sheets in the model is a natural extension. For this, it will be necessary to address the issue of hydrogen bonds which is currently not treated.

BIBLIOGRAPHY FOR CHAPTER 2

- [1] Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J., *Bioinformatics* 1998, 14, 892.
- [2] Cuff, J. A.; Barton, G. J., *Proteins Struct. Funct. Genet.* 2000, 40, 502.
- [3] Jones, D. T., *J. Mol. Biol.* 1999, 292, 195.
- [4] Eyrich, V.; Standley, D.; Friesner, R., *J. Mol. Biol.* 1999, 288, 725.
- [5] Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Ołdziej, S.; Arnautova, Y. A.; Scheraga, H. A., *J. Phys. Chem. B* 2001, 105, 7299.
- [6] Huang, E. S.; Samudrala, R.; Ponder, J. W., *J. Mol. Biol.* 1999, 290, 267.
- [7] Branden, C.; Tooze, J., *Introduction to Protein Structure*, Garland, New York 1999.
- [8] Miyazawa, S.; Jernigan, R. L., *Macromolecules* 1983, 18, 534.
- [9] Wingreen, N. S.; Tang, C.; Li, H., *Physical Review Letters* 1997, 79, 765.
- [10] Miyazawa, S.; Jernigan, R. L., *J. Mol. Biol.* 1996, 256, 623.
- [11] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.
- [12] Gay, D. M., *ACM Trans. Math. Software* 1983, 9, 503.
- [13] Zhang, C.; Hou, J.; Kim, S. H., *Proc. Natl. Acad. Sci., USA* 2002, 99, 3581.
- [14] Murzin, A. G.; Brenner, S. E.; Hubbard, T., *J. Mol. Biol.* 1995, 247, 536.
- [15] Kabsch, W.; Sander, C., *Biopolymers* 1983, 22, 2677.
- [16] Chou, K.; Maggiora, G. M.; Scheraga, H. A., *Proc. Natl. Acad. Sci., USA* 1992, 89, 7315.
- [17] Chinchio, M.; Scheraga, H. A., *J. Comp. Chem.* 2005, To be submitted for publication.
- [18] Fain, B.; Levitt, M., *J. Mol. Biol.* 2001, 305, 191.
- [19] Zhang, C.; Kim, S. H., *Proc. Natl. Acad. Sci., USA* 2000, 97, 2550.
- [20] Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H.A., *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 1937.

Chapter 3

The UNRES Model of the Polypeptide Chain

3.1 The UNRES force field

In this section, the UNRES model of polypeptide chains and the corresponding force field is described briefly. In the UNRES model,¹⁻¹¹ a polypeptide chain is represented by a sequence of α -carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p). Each united peptide group is located in the middle between two consecutive α -carbons. Only these united peptide groups and the united side chains serve as interaction sites, the α -carbons serving only to define the chain geometry, as shown in Figure 3.1. All virtual bond lengths (i.e. $C^\alpha \dots C^\alpha$ and $C^\alpha \dots SC$) are fixed; the distance between neighboring C^α 's is 3.8 \AA corresponding to *trans* peptide groups, while the side-chain angles (α_{SC} and β_{SC}), and virtual-bond (θ) and dihedral (γ) angles can vary.

The UNRES force field has been derived as a Restricted Free Energy (RFE) function of an all-atom polypeptide chain plus the surrounding solvent, where the all-atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (i.e., the degrees of freedom of the solvent, the dihedral angles χ for rotation about the bonds in the side chains,

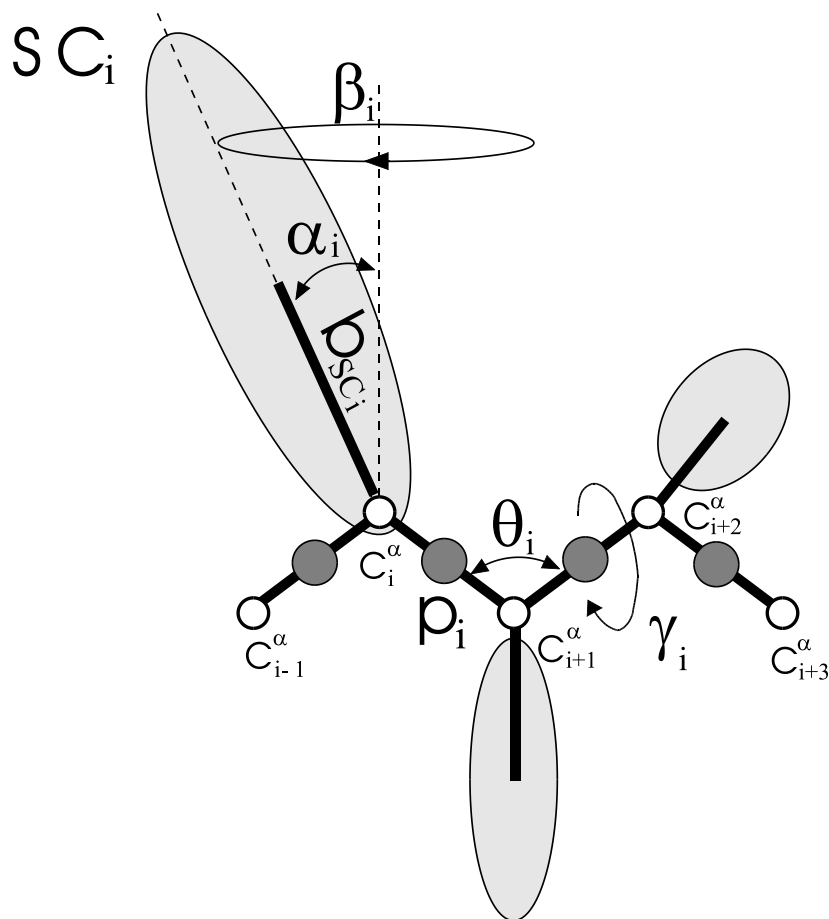


Figure 3.1: The UNRES model of polypeptide chains. The interaction sites are side-chain centroids of different sizes (SC), and peptide-bond centers (p) indicated by shaded circles, whereas the α -carbon atoms (small empty circles) are introduced only to assist in defining the geometry. The virtual $C^\alpha \cdots C^\alpha$ bonds have a fixed length of 3.8 \AA , corresponding to a trans peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α -carbon with a fixed “bond length”, b_{SC_i} , variable “bond angle”, α_{SC_i} , formed by SC_i and the bisector of the angle defined by C_{i-1}^α , C_i^α , and C_{i+1}^α , and with a variable “dihedral angle” β_{SC_i} of counterclockwise rotation about the bisector, starting from the right side of the C_{i-1}^α , C_i^α , C_{i+1}^α frame.

and the torsional angles λ for rotation of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual bonds).^{5, 6, 12} The RFE is

$$F(\mathbf{X}) = -RT \ln \left(\frac{1}{V_{\mathbf{Y}}} \int_{\Omega_{\mathbf{Y}}} \exp[-E(\mathbf{X}; \mathbf{Y})/RT] dV_{\mathbf{Y}} \right) \quad (3.1)$$

where $E(\mathbf{X}; \mathbf{Y})$ is the all-atom ECEPP/3 energy function, X is the set of UNRES degrees of freedom, \mathbf{Y} is the set of degrees of freedom over which the average is computed (e.g., the positions and orientations of solvent molecules, the side-chain dihedral angles, etc.), R is the gas constant, T is the absolute temperature, $\Omega_{\mathbf{Y}}$ is the region of the \mathbf{Y} subspace over which the integration is carried out, and $V_{\mathbf{Y}}$ is the volume of this region.

The RFE is further decomposed into factors arising from interactions within and between a given number of united interaction sites.⁶ Expansion of the factors into generalized Kubo cumulants¹³ facilitated the derivation of approximate analytical expressions for the respective terms,^{5, 6} including the *multibody* or *correlation* terms, which are derived in other force fields from structural databases or on a heuristic basis.¹⁴ The theoretical basis of the force field is described in detail in reference.⁶ The energy of the virtual-bond chain is expressed by eq. (3.2).

$$\begin{aligned} U = & \sum_{i < j} U_{SC_i SC_j} + w_{SCp} \sum_{i \neq j} U_{SC_i p_j} + w_{el} \sum_{i < j-1} U_{p_i p_j} + w_{tor} \sum_i U_{tor}(\gamma_i) + \\ & + w_{tord} \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + \\ & + w_{corr}^{(3)} U_{corr}^{(3)} + w_{corr}^{(4)} U_{corr}^{(4)} + w_{turn}^{(3)} U_{turn}^{(3)} + w_{turn}^{(4)} U_{turn}^{(4)} \end{aligned} \quad (3.2)$$

The different terms of the UNRES force field are depicted graphically⁶ in Figures 3.2 and 3.3. The term $U_{SC_i SC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contains the contributions from the interactions of the side chain with the solvent. The term

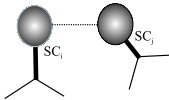
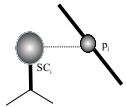
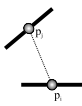

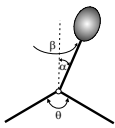
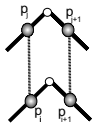
Energy term(s)	Illustration	Type of expression	Parameterization
U_{SCSC}		Empirical	Fitting to distributions derived from the PDB
U_{SCp}		Empirical	Adjusting to reproduce local structure
U_{pp}		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{tor}(\gamma)$		Empirical (Fourier series)	Fitting to averaged free energy surfaces
$U_b(\theta)$ $U_{rot}(\alpha, \beta)$		Empirical	Fitting distributions of angles derived from the PDB
$U_{corr;el}^{(4)}$		Analytical (cumulant expansion)	Directly from analytical expressions

Figure 3.2: UNRES force field⁶

Energy term(s)	Illustration	Type of expression	Parameterization
$U_{\text{el-loc}}^{(3)}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{\text{el-loc;turn}}^{(3)}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{\text{el-loc;turn}}^{(4)}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{\text{el-loc}}^{(5,6)}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{\text{el-loc;turn}}^{(6)}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces

Figure 3.3: UNRES force field, continued⁶

$U_{SC_i p_j}$ denotes the excluded-volume potential of the side-chain – peptide-group interactions. The peptide-group interaction potential ($U_{p_i p_j}$) accounts mainly for the electrostatic interactions (i.e., the tendency to form backbone hydrogen bonds) between peptide groups p_i and p_j . U_{tor} , U_{tord} , U_b , and U_{rot} are the virtual-bond dihedral angle torsional terms, virtual-bond dihedral angle double-torsional terms, virtual-bond angle bending terms, and side-chain rotamer terms; these terms account for the local propensities of the polypeptide chain. The terms $U_{corr}^{(m)}$ represent *correlation* or *multibody* contributions from the coupling between backbone-local and backbone-electrostatic interactions and the terms $U_{turn}^{(m)}$ are correlation contributions involving m consecutive peptide groups; they are, therefore, termed turn contributions. The correlation contributions were derived^{5, 6} from a generalized-cumulant expansion¹³ of the restricted free energy (RFE) of the system consisting of the polypeptide chain and the surrounding solvent. The multibody terms are indispensable for reproduction of regular α -helical and β -sheet structures.

The internal parameters of $U_{p_i p_j}$, U_{tor} , U_{tord} , $U_{corr}^{(m)}$, and $U_{turn}^{(m)}$ were derived by fitting the analytical expressions to the RFE surfaces of model systems computed by quantum mechanics at the MP2/6-31G** *ab initio* level,^{10, 11} while the parameters of $U_{SC_i SC_j}$, $U_{SC_i p_j}$, U_b , and U_{rot} were derived by fitting the calculated distribution functions to those determined from the PDB.⁴ The w 's are the weights of the energy terms, and they were determined (together with the parameters within each cumulant term) by optimization of the potential-energy function, as described in the next section.

3.2 Optimization of UNRES parameters

To properly represent the physical features of proteins, it is necessary that the weights and the parameters in the UNRES energy function (Eq. 3.2) be optimized. Following Anfinsen’s thermodynamic hypothesis¹⁵ and also works of other authors,^{16–18} the first procedure^{7, 8} to optimize the UNRES energy function was based on using a set of training proteins to maximize the energy gap between the lowest-energy native-like structure and the lowest-energy non-native structure (ΔE) and/or the Z-score (Z) defined as the difference between the mean energy of the native-like structures and the mean energy of the non-native structures divided by the standard deviation of the energy of the non-native structures:^{7, 8}

$$\Delta E = \min_{i \in \text{nat}} E_i - \min_{i \in \text{non-nat}} E_i \quad (3.3)$$

$$Z = \frac{(1/N_{\text{nat}}) \sum_{i=1}^{N_{\text{nat}}} E_i - (1/N_{\text{non-nat}}) \sum_{i=1}^{N_{\text{non-nat}}} E_i}{\sqrt{(1/N_{\text{non-nat}}) \sum_{i=1}^{N_{\text{non-nat}}} E_i^2 - [(1/N_{\text{non-nat}}) \sum_{i=1}^{N_{\text{non-nat}}} E_i]^2}} \quad (3.4)$$

where *nat* and *non-nat* indicate the sets of native-like and non-native conformations, respectively, and N_{nat} and $N_{\text{non-nat}}$ denote the number of native-like and non-native structures, respectively. In the second-order cumulant expansion of the free energy in temperature,¹⁸ the negative of the Z-score is approximately equal to the ratio of the folding temperature (T_f) to the glass-transition temperature (T_g); the bigger this ratio the lower the glass-transition temperature compared to the folding temperature which prevents trapping a system in one of the local minima before the folded structure can be thermally accessed.

In the initial approach, the parameters to be optimized were the energy-term weights [the w ’s of Eq. (3.2)]. To optimize the energy gap and Z-score simultaneously and also to treat many training proteins, the Vector Monte Carlo (VMC)

method⁸ was used. The complete algorithm involves iterations consisting of the following three steps:^{7, 8} (i) updating the decoy set by a global search using the current weights, (ii) a local search in the neighborhood of the experimental structure with the current weights, in order to locate the lowest-energy native-like structure corresponding to the current set of parameters of the energy function, and (iii) determination of new weights by making ΔE and Z as negative as possible, by using the VMC method. In the present version of the hierarchical optimization procedure, both global and local conformation search steps are carried out with the Conformational Space Annealing (CSA)¹⁹⁻²¹ algorithm. Steps (i) - (iii) are iterated until the global CSA search finds the native-like structure as the lowest-energy structure. This procedure was successful in optimizing the energy landscape of proteins with simple topology, such as the 10-55 residue N-terminal domain of staphylococcal protein A (a three-helix bundle), and betanova (a 20-residue designed β -sheet peptide); using these two proteins simultaneously UNRES was capable to fold $\alpha+\beta$ proteins.⁸ However, the approach failed for more complex proteins, such as 1IGD (a 61-residue $\alpha+\beta$ -protein): the resulting force field could not locate native-like structures of 1IGD (used as a training protein) in global CSA searches despite the large energy and Z-score gaps achieved in optimization.

By carrying out model studies on 12-bead cubic-lattice protein models where all conformations can be enumerated²² it was concluded that optimizing the energy gap and Z-score is generally insufficient to obtain a searchable potential. It was demonstrated²² that energy functions characterized by similar energy gap and Z-score values can correspond to both excellent and very poor folders, even for the simple models studied, and that the foldability depends strongly on the energy-ordering of non-native structures with some native elements according to native

likeness, i.e., their energy should decrease with increasing native-likeness (Figure 3.4). If this condition is not satisfied and only the native-like structures have distinctively low energy, the resulting energy landscape can be compared to a golf course, while it should resemble a funnel-like landscape^{18, 23} in which native-likeness increases with decreasing energy. Therefore a hierarchical method of force-field optimization^{9, 22, 24, 25} was designed in our laboratory, which is directed at lowering the energy with increasing number of native-like elements. The conformational space is divided into levels, each level containing conformations with similar degree of native-likeness. Level 0 contains no native-like elements, level 1 contains single native secondary-structure elements, and higher levels contain gradually increasing native-like segments. The composition and sequence of levels is termed a structural hierarchy. The construction of the hierarchy is depicted in Figure 3.4.

Both by model studies with lattice chains²² and by test optimization of the UNRES force field,²⁴ it was found that, for the optimization to succeed, the hierarchy should follow the folding pathway. Our group implemented the experimental information of folding whenever it was available,^{24, 25} otherwise the most probable folding pathways were constructed, which also gave good results.²⁵ It should be noted that using the experimental information about the folding pathway(s) of training protein(s) in force field calibration is conceptually the same as using experimental bond lengths, bond angles, formation heats, etc., in the calibration of all-atom force fields or even the semiempirical methods of quantum mechanics and does not introduce knowledge-based elements into the procedure because the folding-pathway information is not directly implemented in the conformational search procedure for the prediction of the structures of proteins with unknown structure.

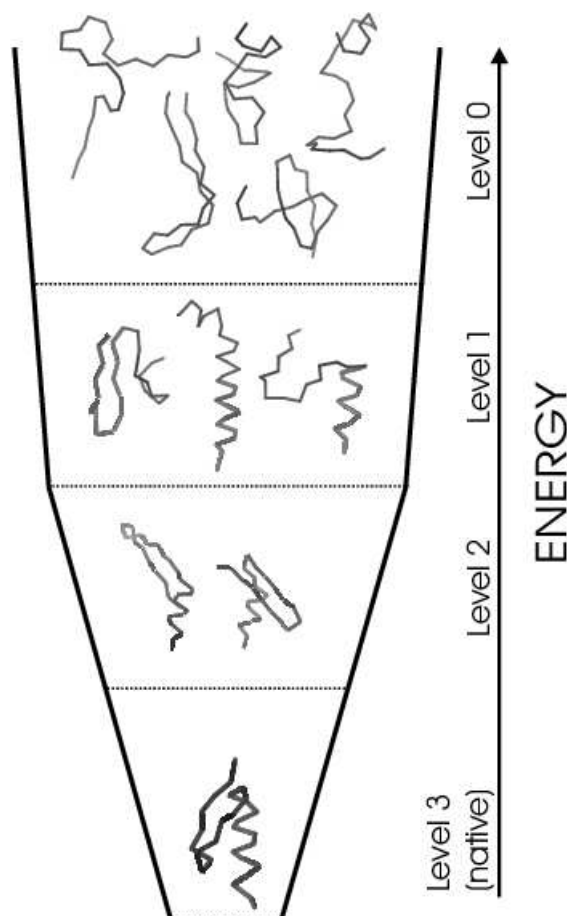


Figure 3.4: Schematic illustration of energy-ordering of structures with increasing native-likeness. The highest energy level (Level 0) is occupied by structures with either no or non-native secondary structure. Next (Level 1) is occupied by the structures with one native secondary structure element (the N-terminal β -hairpin or the C-terminal α -helix; the native-like structure fragments are indicated by thicker lines). Yet lower energy (Level 2) have structures with both α -helix and β -hairpin, but no or incorrect packing of these two substructures and/or shifted turn in the β -hairpin. Finally, the native-like structures, with α -helix and β -hairpin packed correctly occupy the lowest energy level (Level 3). Because the number of structures with more and more defined native-like elements decreases, such ordering of structures leads to diminishing conformational entropy following the energy decrease, which is highly desirable in order to find the native structure quickly in a spontaneous energy-driven search of the conformational space.

The hierarchical optimization algorithm is composed of the same steps as the energy-gap and Z-score optimization algorithm outlined at the beginning of this section except that, instead of the differences of the lowest energies of the conformations of the ensembles [Eq. 3.3], the differences between their configurational free energies (the free-energy gaps) are considered. This modification prevents optimization focusing on a conformation with accidentally outstandingly low energy. The free energy of structural level i is defined as a direct Boltzmann average over all conformations that belong to this level.

$$F_i(\beta) = -\frac{1}{\beta} \ln \sum_{k \in \{i\}} \exp(-\beta E_k) \quad (3.5)$$

where $\{i\}$ denotes the set of conformations of level i , E_k denotes the energy of the k th conformation of this level, and β can be identified with $1/RT$, T being the absolute temperature, or treated as a parameter of the method. A classification scheme was developed²⁴ based on the similarity of elementary fragments and larger portions (including the complete molecule) of a given conformation to those of the experimental structure in terms of secondary structure, contact pattern, and RMSD in which each conformation is represented by a binary number; this enabled the conformations to be assigned automatically to the pre-specified structural levels.²⁴ To increase the efficiency of optimization, the VMC method was replaced^{9, 22, 24, 25} by minimization of a penalty function containing the differences between the actual and the target free-energy and Z-score gaps between structural levels whose most important part is defined by Eq. 3.6. This last modification led to the optimization of the coefficients of the cumulant expansion of the correlation terms and the well-depths of the side-chain interaction parameters in addition to

the energy-term weights in Eq. 3.2.^{9, 24, 25}

$$\Phi = \frac{1}{4} \sum_{\substack{\text{training} \\ \text{proteins}}} \sum_{\beta} \sum_{i=0}^{n-1} w_i^{\beta} \begin{cases} [(F_i(\beta) - F_{i+1}(\beta)) - \Delta_i^{\beta}] & \text{if } F_i(\beta) - F_{i+1}(\beta) \leq \Delta_i^{\beta} \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

where Δ_i^{β} is the minimum required free-energy gap between levels i and $i+1$ and w_i^{β} is the weight assigned to the deviation of the actual $(F_i(\beta) - F_{i+1}(\beta))$ and the requested (Δ_i^{β}) free-energy gap between levels i and $i+1$ (the number of levels being $n+1$) at reduced inverse temperature β . Other penalty terms such as, e.g., the penalty for deviating from correct local geometry of α -helices and β -sheets, Z-score terms, etc., can also be present in Eq. 3.6; this is discussed in detail in reference 24.

Using the hierarchical algorithm, the UNRES force field was first optimized using 1IGD as the training protein²⁴ and, finally four training proteins:²⁵ 1GAB²⁶ (a 47-residue α -protein), 1E0L²⁷ (a 28-residue β -protein), 1E0G²⁸ (a 48-residue $\alpha + \beta$ protein), and 1IGD²⁹ [a 61-residue $(\alpha + \beta)$ -protein]. The force field obtained with the four training proteins, hereafter referred to as the 4P force field, was tested on a set of 66 proteins [26 α -, 15 β -, and 25 $(\alpha + \beta)$ -proteins with chain length from 28 to 144 amino-acid residues].²⁵ The average length of a continuous segment matching the corresponding segment of the experimental structure within 6 Å RMSD and the percentage of correctly predicted chain length are 54 (67 %), 34 (45 %), 42 (55 %), and 45 (58 %) for the α , β , $\alpha + \beta$, and all proteins, respectively, and the length of the longest predicted continuous fragment is 96, 49, and 70 residues for the α -, β -, and the $\alpha + \beta$ -proteins, respectively. These results were achieved without using ancillary knowledge-based information from sequence similarity, threading, secondary-structure prediction or fragment coupling. The

two force fields mentioned above, obtained by hierarchical optimization were also tested with success in the CASP5 and CASP6 experiments,³⁰ respectively.

BIBLIOGRAPHY FOR CHAPTER 3

- [1] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *Protein Sci.* 1993, 2, 1697.
- [2] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *Protein Sci.* 1993, 2, 1715.
- [3] Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A., *J. Comput. Chem.* 1997, 18, 849.
- [4] Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A., *J. Comput. Chem.* 1997, 18, 874.
- [5] Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A., *J. Comput. Chem.* 1998, 19, 259.
- [6] Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A., *J. Chem. Phys.* 2001, 115, 2323.
- [7] Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A., *J. Phys. Chem. B* 2001, 105, 7291.
- [8] Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Ołdziej, S.; Arnautova, Y. A.; Scheraga, H. A., *J. Phys. Chem. B* 2001, 105, 7299.
- [9] Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H. A., *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 1937.
- [10] Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A., *J. Phys. Chem. A* 2003, 107, 8035.
- [11] Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 9421.
- [12] Nishikawa, K.; Momany, F. A.; Scheraga, H. A., *Macromolecules* 1974, 7, 797.
- [13] Kubo, R., *J. Phys. Soc. Japan* 1962, 17, 1100.
- [14] Kolinski, A.; Skolnick, J., *J. Chem. Phys.* 1992, 97, 9412.
- [15] Anfinsen, C. B., *Science* 1973, 181, 223.
- [16] Sali, A.; Shakhnovich, E.; Karplus, M., *Nature* 1994, 369, 248.

- [17] Meller, J.; Elber, R., *Proteins: Struct. Funct. Genet.* 2001, 45, 241.
- [18] Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G., *J. Chem. Phys.* 2002, 117, 4602.
- [19] Lee, J.; Scheraga, H. A.; Rackovsky, S., *J. Comput. Chem.* 1997, 18, 1222.
- [20] Lee, J.; Scheraga, H. A., *Int. J. Quant. Chem.* 1999, 75, 255.
- [21] Czaplewski, C.; Liwo, A.; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [22] Liwo, A.; Arlukowicz, P.; Oldziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16918.
- [23] Wolynes, P. G., *Phil. Trans. Royal Soc. London A* 2005, 363, 453.
- [24] Oldziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16934.
- [25] Oldziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nancias, M.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16950.
- [26] Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drakenberg, T.; Bjorck, L., *J. Mol. Biol.* 1997, 266, 859.
- [27] Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H., *Nat. Struct. Biol.* 2000, 7, 375.
- [28] Bateman, A.; Bycroft, M., *J. Mol. Biol.* 2000, 299, 1113.
- [29] Derrick, J. P.; Wigley, D. B., *J. Mol. Biol.* 1994, 243, 906.
- [30] Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nancias, M.; Vila, J.A.; Khalili, M.; Arnautova, Y.A.; Jagielska, A.; Makowski, M.; Schafroth, H.D.; Kazmierkiewicz, R.; Ripoll, D.R.; Pillardy, J.; Saunders, J.A.; Kang, Y.K.; Gibson, K.D.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 7547.

Chapter 4

Replica-Exchange Monte

Carlo-with-Minimization as a global

optimization method with the UNRES

force field; comparison with MCM, CSA

and CFMC *

4.1 Introduction

Computation of the three-dimensional structures of proteins from their amino acid sequence has been a formidable problem in structural biology and theoretical chemistry. One class of methods, known as physics-based ab initio, relies solely on physical principles to obtain the three-dimensional protein structure.¹

Structure determination involves two components, namely an accurate potential energy function to describe the interactions between amino acids and thereby distinguish the native structure from non-native ones, and a procedure for global optimization of the potential energy. Much research has been devoted to this problem. In particular, our laboratory has developed a hierarchical procedure the first

*Published as Nianias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H.A., *J. Comp. Chem.* 2005, 26, 1472. Copyright (2005) John Wiley & Sons Inc.

step of which is the parameterization of a united-residue (UNRES) energy function; this parameterization has been carried out on several proteins simultaneously,^{2, 3} based on the assumption that the native structure lies in the global minimum of the energy hyper-surface,⁴ it is required that the global optimization method locates this minimum. Global optimization is an extremely difficult procedure because protein energy landscapes encompass a vast rugged area with many local minima (traps), so that the search for the global minimum is nontrivial.

Global optimization has been at the center of many research fields, and a variety of different methods have been used. Among those methods developed in our laboratory, three are compared here. The first class includes modifications of the Metropolis Monte Carlo procedure,^{5, 6} viz., Monte Carlo-with-Minimization (MCM),^{7, 8} electrostatically-driven Monte Carlo (EDMC),^{9, 10} and Conformational Family Monte Carlo (CFMC).¹¹ The second class includes deformation-based methods, such as the diffusion-equation method (DEM),¹² the distance-scaling method (DSM),¹³ and the self-consistent basin-to-deformed-basin method (SCBDBM).^{14, 15} The third class includes genetic algorithms such as the Conformational Space Annealing (CSA) method.¹⁶⁻¹⁸ A new method, Replica Exchange Monte Carlo-with-Minimization (REMCM) combining the traditional Replica Exchange Method (REM) with MCM, is introduced here. CSA, CFMC and MCM have been used with the UNRES force field (described in section 3.1), and are compared with REMCM in the current work. CSA is a hybrid method which combines genetic algorithms, essential aspects of the build-up method and a local gradient-based minimization. It evolves the population of conformations through genetic operators (mutations, and crossovers) to a final population optimizing their conformational energy. CFMC maintains a database of low-energy conformations that are clus-

tered into families. They are consequently improved iteratively by a Metropolis-type Monte Carlo-with-local minimization, while annealing both in temperature and in the number and size of the conformational families.

The Replica Exchange method (also known as Exchange Monte Carlo,¹⁹ or Parallel Tempering²⁰) was originally developed by Swendsen et al.²¹ for spin-glass systems. This method has been used extensively in protein-folding simulations using lattice models.²²⁻²⁵

This chapter applies REMCM to a coarse grain protein system described by the UNRES model. It is not a review of global optimization methods used in protein-structure prediction but is rather a description of the method and its application. REMCM expands the idea of MCM, wherein minimization finds local minima in a given basin, while replica exchange ensures the exploration of different regions of the energy surface. The advantage of Replica Exchange lies in its simplicity and, in contrast to other methods, it is not very sensitive to the few parameters involved therein (e.g., simulated tempering depends heavily on the cooling schedule, and generalized ensemble algorithms depend on successful estimation of weight factors). In this work, REMCM is applied to five proteins of different topology, and the performance is compared to those of three other methods, namely MCM, CSA and CFMC.

4.2 Methods

4.2.1 Replica Exchange Monte Carlo (REM)

The Replica Exchange method is an extension of the Metropolis Monte Carlo method. The underlying idea is to run different copies (replicas) of the system

at different levels of a certain property (such as temperature). To summarize the method, the following procedures are performed in *each* cycle:

1. Select several temperatures and assign a different random protein conformation to each temperature.
2. Monte Carlo simulation is carried out on each selected conformation at its assigned temperature for a determined number of Monte Carlo steps by performing the following:
 - (a) Obtain a new conformation by perturbing the parent conformation.
 - (b) Accept or reject the new conformation at its corresponding temperature using the Metropolis acceptance criterion.
3. At a chosen interval, stop the MC simulation of each replica and attempt an exchange of whole conformations between neighboring replicas. The exchange acceptance criterion is described below.
4. Continue MC with each newly formed conformation at each new temperature as in step 2.
5. Iterate points 3 and 4 until the system converges to the lowest energy independent of the temperature.
6. At the end, select the lowest-energy conformation over all trajectories and temperatures.

Although different properties have been used in published work,^{26, 27} the property of change across different replicas in the current context is temperature.

In our computations, each replica is a Metropolis Monte Carlo simulation; hence, each replica produces an ensemble which obeys a Boltzmann distribution at each temperature. Thus, the probability of conformation X in replica m at temperature T_m is

$$P_m(X) = \frac{1}{Z_m} \exp \left[-\beta_m E(X) \right], \quad (4.1)$$

where β_m is the inverse temperature defined as $1/(k_B T_m)$, $E(X)$ is the energy of conformation X , and Z_m is the partition function $\int \exp(-\beta_m E(X)) dX$. Many replicas are treated at different temperatures T_m . The joint probability distribution of the whole system, can be represented by multiplying the probabilities of all replicas

$$P_{all} = \prod_m^M P_m(X_m) \quad (4.2)$$

where M is the number of replicas. The transition probability that conformation X in replica m is exchanged with conformation Y in replica n can be written as $W(X, \beta_m | Y, \beta_n)$. In order for the system to be in equilibrium, the detailed balance condition (also known as microscopic reversibility) has to be satisfied:

$$P_{all}(X, \beta_m; Y, \beta_n) W(X, \beta_m | Y, \beta_n) = P_{all}(Y, \beta_m; X, \beta_n) W(Y, \beta_m | X, \beta_n) \quad (4.3)$$

Combining the previous equations, one obtains

$$\frac{W(X, \beta_m | Y, \beta_n)}{W(Y, \beta_m | X, \beta_n)} = \exp \left[-(\beta_m - \beta_n) \{ E(Y) - E(X) \} \right] \quad (4.4)$$

Let

$$\Delta \equiv \left[(\beta_m - \beta_n) \{ E(Y) - E(X) \} \right] \quad (4.5)$$

If one adopts the Metropolis method, the replica-exchange transition probability can be expressed as

$$\begin{aligned} W(X, \beta_m | Y, \beta_n) &= 1 \text{ for } \Delta \leq 0 \\ &= \exp(-\Delta) \text{ for } \Delta > 0 \end{aligned} \quad (4.6)$$

i.e., if Δ is less than or equal to 0, the exchange is performed (since the probability is 1); otherwise a random number between 0 and 1 is generated and compared to the factor $\exp(-\Delta)$. If the value of this factor is smaller, the exchange is performed; otherwise the exchange is rejected.

An important requirement for the procedure to work correctly is the proper choice of the temperature range and spacing. The lowest T_{min} should be chosen such that the protein is stable in the native form, and the highest T_{max} should be high enough for the protein to be unfolded. The temperature spacing between replicas should be small enough so that the exchange would occur at reasonable probabilities. This temperature spacing can be satisfied by the condition that the energy fluctuation of a replica should be of the same order as the spacing of the mean values of the replica energies²⁶

$$\Delta E_m \sim \bar{E}_{m+1} - \bar{E}_m \quad (4.7)$$

By equipartition, we assume that the mean energy of a replica scales as $\bar{E}_m \sim k_B T_m f$, where f is the number of degrees of freedom, and the energy fluctuation scales^{26, 28} as $\Delta E_m \sim k_B T_m \sqrt{f}$. This leads to the following expression

$$f k_B (T_{m+1} - T_m) \sim k_B T_m \sqrt{f} \quad (4.8)$$

Solving this recurrence relationship for T_m one obtains $T_m \sim \exp(m/\sqrt{f})$ which suggests a choice of exponential dispersion of temperatures. Further, for a given temperature range (T_{min} and T_{max}) the expression for T_m quantifies a number of replicas necessary $M \sim \ln(T_{max}/T_{min})\sqrt{f}$.

4.2.2 Replica Exchange Monte Carlo-with-Minimization (REMCM)

The REMCM procedure is almost exactly the same as the REM procedure described in the previous section, with one modification. There is an extra step between 2(a) and 2(b), in which the perturbed structure is minimized.

Adding the minimization to relax the conformations after every perturbation changes the behavior of the classical Markov chain Monte Carlo. The simulations are no longer free to sample the entire conformational space but are rather restricted to the space of energy minima. This approach destroys the detailed balance condition, which is absolutely essential for the calculation of thermodynamic variables. However, since the objective of this work is to locate global minima, the idea is to sample the energy basins instead of spending time in higher energy regions. In practice, the effect of energy minimization in replica exchange presents itself through a different pattern of energy probability distributions which must overlap in order to obtain a non-zero probability of exchange. Although, on average, the higher temperature replicas sample higher-energy regions, they are still brought down to the local minima at each temperature just as their low-energy counterparts. However, since they have high temperature, they more readily accept new conformations. Thus, the high-temperature replicas sample different parts of the conformational space, whereas the low-temperature replicas focus more strongly on the area around the current conformation. By exchanging the replicas, a conformation from a high temperature can be swapped into a lower-temperature replica and, as in typical REM, it has a chance to explore the surroundings more properly. On the other hand, going from low to high temperature provides the simulation with a fresh starting point and, if the global energy basin is smooth, minimization

can help even high-temperature replicas to locate the global minimum. The power of the method lies in the fact that all the replicas have some reasonable chance of locating the global minimum.

4.3 Computational Details

4.3.1 Test Systems

REMCM was applied as a global optimization method with the UNRES 4P force field,³ and was tested on five proteins^{29–33} of which three were α -helical (1GAB, 1BDD, 1CLB), one consisted of a β -sheet (1E0L), and one was $\alpha+\beta$ (1IGD) (Table 1). These proteins were chosen so that basic α , β or $\alpha+\beta$ topologies were tested, and their size was reasonable with respect to the computational time. To be comparable with CSA results, all the proteins except 1IGD had their length modified from the original length in the PDB (Table 4.1) to the length as in reference.³

4.3.2 REMCM Implementation

Method Parameters

The performance of REMCM is affected by the following parameters: temperature distribution and range, frequency of replica exchange, number of replicas, length of simulation and frequency of individual move types. Section 4.2.1 suggests that the exponential temperature distribution should be adopted. The choice for the range and number of replicas, however, is not quite straightforward. The number of replicas required to cover the energy space depends on the size of the system. The lowest T_{min} should be chosen such that the protein is stable in its native

Table 4.1: Proteins used in the calculations, and their PDB id. Nres corresponds to the number of residues in the PDB native structure, whereas Nres' is the protein length in our calculations because some end-segments whose locations were not precise were removed.

Protein System	PDB id	Nres	Nres'	Reference
Fbp28Ww Domain	1E0L	37	28	29
Albumin-Binding Domain	1GAB	53	47	32
Protein A	1BDD	60	46	30
Apo calbindin D9k	1CLB	76	75	31
IgG domain (protein G)	1IGD	61	61	33

state, and the highest T_{max} should preserve an unbiased sampling of any part of the energy landscape (i.e., the sampling should ideally be able to overcome any barrier). In a typical replica exchange method, it is important to have enough replicas to maintain a sufficient overlap between neighboring replicas which will guarantee a nonzero probability of exchange. Section 4.2.1 shows a quantitative estimate for number of replicas required for traditional Replica Exchange Method. As mentioned in section 4.2.2, the distribution overlap behavior is altered slightly in REMCM, since the structures are relaxed through minimization. Minimization effectively shifts the energy distributions towards the global minimum and broadens the distributions for high-temperature replicas because both low and high-energy regions are sampled at high temperature. This also results in an overlap of the distributions for non-neighboring replicas, so that REMCM is even less sensitive to the choice of temperature range than regular REM, and REMCM requires fewer replicas than REM.

Due to the fact that REMCM is not as sensitive as REM to the choice of its parameters, and to produce a set of consistent results, the number of replicas, temperature range, and frequency of exchange were kept constant for all the proteins. The number of replicas was set to 10 (the number of degrees of freedom ranged from 112 for 1E0L to 304 for 1CLB, which would correspond to a range of 10-18 replicas necessary for REM), with kT_{min} , kT_{max} set to 1 and 100 respectively (where 1 and 100 corresponded to 20% and 60% acceptance rate, respectively), and the exchange occurred every 10 steps. The exchange was carried out according to the instructions described in reference³⁴ in which the exchange starts from the low to the high temperature, i.e., first we attempt to swap replicas 0 and 1, then replicas 1 and 2, ..., replicas n-1 with n with increasing temperature. Each

REMC simulation was carried out with a total of 100,000 minimization steps for all replicas (i.e., if 10 replicas were chosen, each replica would run for 10,000 MC steps), and the exchange was attempted every 100 Monte Carlo steps (i.e., each replica performed 10 MC steps before an exchange).

Moves

The Monte Carlo step in each replica consisted of a trial move followed by a minimization with the local minimizer SUMSL (Secant Unconstrained Minimization Solver),³⁵ a quasi-Newton method. The perturbation moves were attempted according to the protocol described below:

1. Backbone angles (θ , γ) were perturbed at random within a randomly selected fragment.
2. A local perturbation was applied to all the residues within a randomly selected fragment, while keeping the rest of the molecule fixed.³⁶
3. A helix or a strand was created within a randomly selected fragment.
4. If a hairpin or a nonlocal β -contact was detected in the parent conformation, it was extended.

The choice of moves was made as follows. First, the simulations were carried out with two basic moves: random perturbation of backbone angles (move #1), and local perturbation of selected regions (move #2). However, the native global minimum was not found in the test runs and, thus, the following topology-specific moves were used in addition. For helical proteins, the helix move (move #3) was added whereas, for β -strand proteins, the helix move was turned off and all the β -strand moves (moves #3, 4) were added. It might be argued that these

specific moves introduced a bias towards the experimental structure, and thus this approach is invalid for a search for the global minimum for an unknown protein. However, as in CSA, several simulations biased towards different topologies can be carried out, of which the correct simulation produces structures lowest in energy.

Because of the perturbative nature of these moves, many structures with side-chain clashes were produced with such bad geometry that even the local minimizer was unable to improve them. Therefore, carefully designed local side-chain moves were applied to relax the structures,³⁶ after which the minimization move was successful. The side-chain move relaxes the side-chains involved in a clash by minimizing the side-chain energy while keeping the backbone frozen.

Parallel Algorithm Implementation

Typically the implementation of the algorithm on a parallel machine involves tying a replica to a single processor. When the exchange between replicas occurs, one can either exchange the structure coordinates or the temperature; exchanging temperature is much more efficient because it involves only two variables whereas many more variables are involved in exchanging coordinates. Our approach to parallelization was slightly different for the following reason. Given the nature of minimization with SUMSL (each minimization step involves a different number of energy evaluations), and the fact that our Linux cluster is very inhomogeneous (i.e., the fastest processors operate at four times the speed of the slowest ones), tying a replica to a single processor for the entire run would result in different simulation lengths for each replica. Furthermore, one might have a smaller number of processors than replicas available or, on the other hand, one should be able to take advantage of more processors than the number of replicas. The most time-

consuming task is the energy minimization (i.e., the time to minimize the energy of a conformation is much longer than the time to transfer the coordinates from master to workers); thus, time to transfer coordinates causes no real overhead. Hence, in our approach the master is the only processor which updates the structures, exchanges the replicas, and passes the structures to workers. The latter are responsible for perturbing the structures and minimizing their energies. However, this approach poses one problem for successful scalability. When using many processors, the energies of a large number of structures are being minimized in several processors for each replica at any given time. After minimization, these conformations are typically not compared to the structure that generated them in the first place, which can lead to oscillations of high-low energy conformations in the Markov chain. To circumvent this problem, the acceptance procedure is modified as follows. When a worker returns an energy-minimized conformation to the master, it is then compared to its original parent conformation (if the original is different from the current one) and, if it is lower in energy, it is compared to the current parent conformation with a standard Metropolis criterion. Otherwise, the conformation is discarded. This reduces the high-low energy fluctuations and allows for smoother behavior in energy vs. step number graphs.

4.3.3 Implementation of methods used for comparison

In order to compare REMCM to the other methods of interest here, each method was implemented as described below.

The traditional Replica Exchange (REM) with UNRES was carried out as follows. All four UNRES angles in every residue of the protein were subject to a perturbation. One MC sweep consisted of updating all of these angles with a

Metropolis evaluation for each perturbation. Twenty replicas were used to cover the entire energy range for REM (REMCM used 10 replicas, but 10 replicas was not enough to cover the energy range of REM). The frequency of exchange (number of MC sweeps before an exchange) was kept the same as in the REMCM simulations. The temperature range was shifted to lower values from the range used for REMCM to make sure that the low-temperature replicas explored the low-energy regions. The kT values ranged from 0.1 to 30, distributed exponentially, which corresponded to MC acceptance rates from 5% to 70%.

Simulations with MCM were carried out with similar parameters as for REMCM. The number of steps for each Monte Carlo-with-Minimization run was set to the overall number of steps for REMCM. A total of five independent MCM simulations were carried out for each protein, each run at a different temperature so that the simulations were able to explore a reasonable range of the energy landscape, yet still have a good chance of obtaining the global minimum.

CSA simulations were carried out with the usual three different sets of moves. One set supported production of mainly α -helical structures, whereas the second set was biased towards exploring β -strand conformations. The third set, on the other hand, supported both types of structures equally likely. This focused the search on different areas of the conformational space; thus, by comparing the results, one could conclude whether the individual simulations converged to the same ensemble of global minima.

For comparison of CFMC with REMCM, the side-chain clashes, produced by perturbation moves, were checked before the actual minimization. As for REMCM, this decreased the number of structures rejected by the local minimizer, thereby decreasing the number of structures which had no chance of being accepted into

the CFMC bank, and hence making the overall procedure much more efficient. In addition, a short evaluation was carried out to find out how much the individual (local and global) moves contributed, by successively turning off these moves. The following three topological moves were added because the original CFMC had limited success with the proteins considered here. The first two moves consisted of choosing a fragment of n residues and creating either an ideal α -helix, or a β -strand. The third move tested the sequence for secondary structure, located the position of all the loops, and attempted to form a hairpin in one of the loops. This improved the performance of CFMC in some although not in all the cases, as will be seen in section 4.4.3.

4.4 Results and Discussion

4.4.1 Production runs

A total of five independent REMCM simulations were carried out for each protein with the parameters described in section 4.3.2. To summarize the results, the procedure was able to find the global minimum for four out of five tested proteins (namely for 1E0L, 1BDD, 1GAB, 1CLB). As an example (using 1E0L), figure 4.1 shows the nature of the individual simulations within each replica and the influence of temperature. The graphs show the energy of the simulation at a particular temperature with respect to the MC step number. It can be seen that the distribution at neighboring temperatures overlap and that the average energy and the fluctuation correlate with the increase in temperature. Further, although the higher temperature replicas sample higher energy regions, they still sample the local minima; the latter important feature is not encountered in traditional REM.

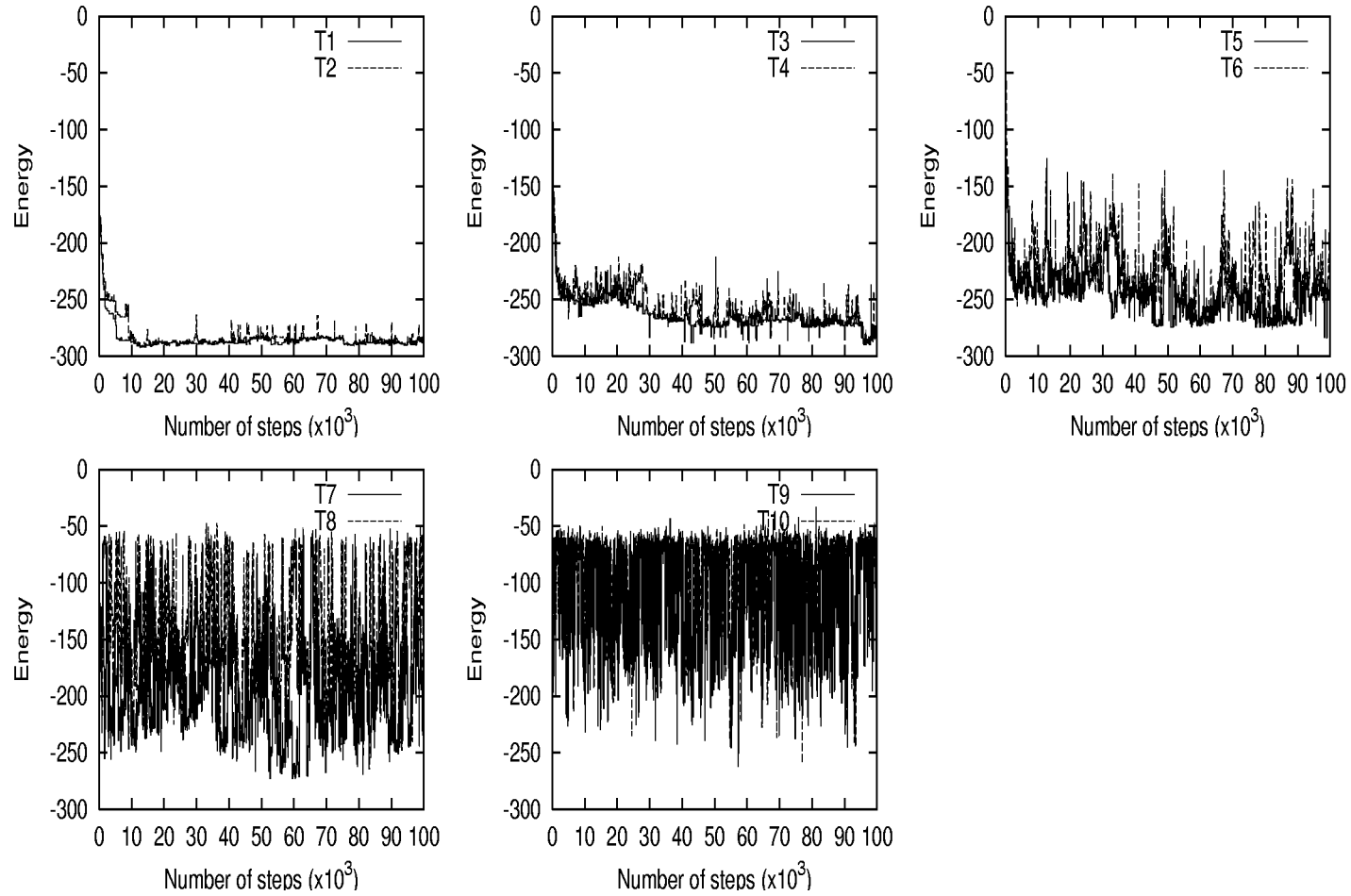
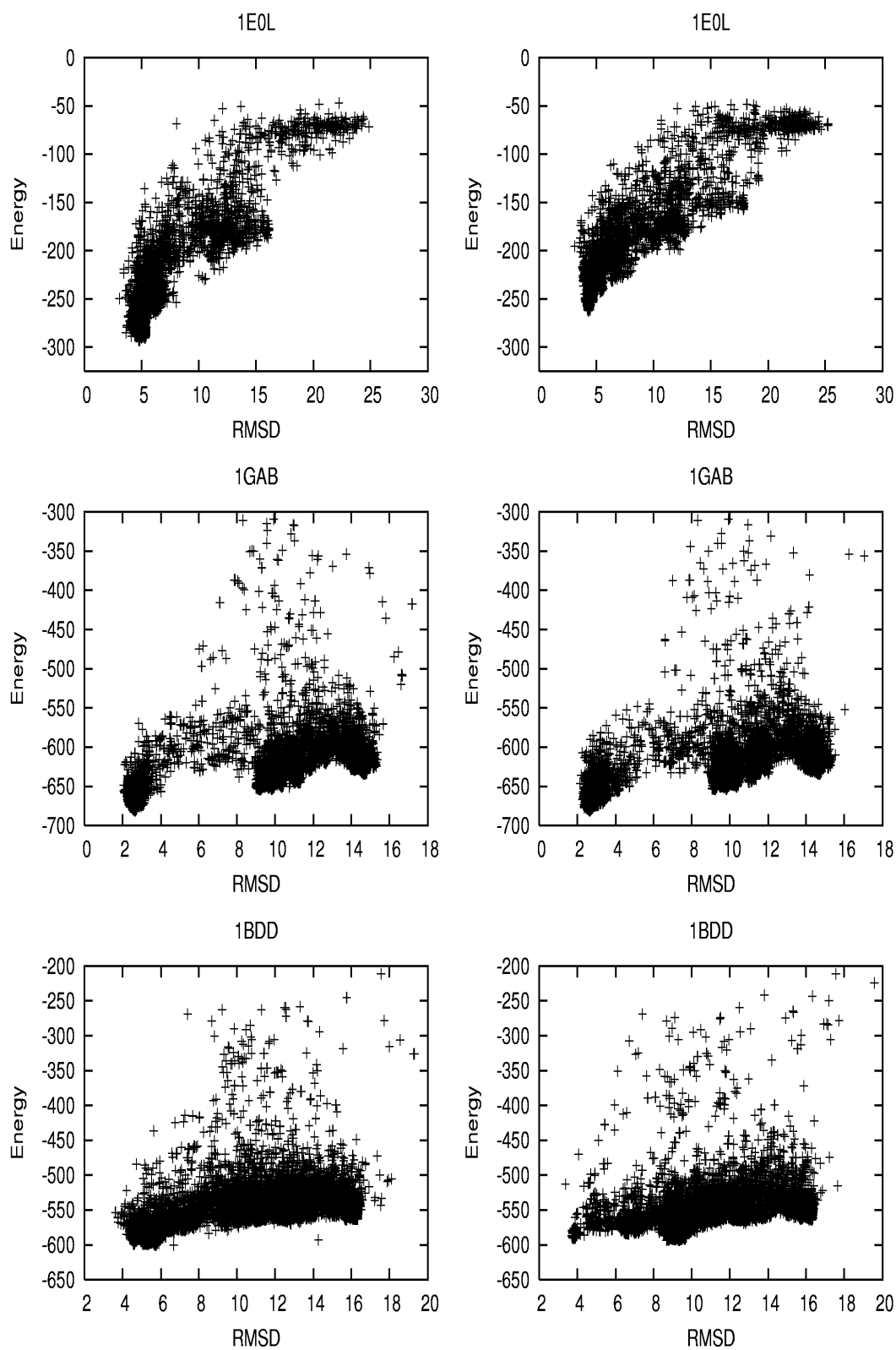


Figure 4.1: Energy of REMCM simulation of 1EOL at a particular temperature with respect to MC step number. Each plot contains two replicas at neighboring temperatures. The temperatures are labeled from lowest (T1) to highest (T10), and the average energy and its fluctuation increase with temperature as governed by the Metropolis criterion. The exchange between replicas occurs every 100 Monte Carlo Steps.

Figures 4.2 and 4.3 illustrate the energy vs RMSD profiles for all the proteins, where each graph corresponds to a particular REMCM run, and each point within a graph corresponds to a visited conformation in that particular run. Two plots are shown for each protein; the left plot is an example of a successful run, whereas the right plot shows an example of a not so successful simulation. Although only two runs are shown for each protein, the conformational space visited is similar for both simulations shown. 1E0L produced a native-like structure as a global minimum in all five simulations, although three runs took a considerably longer time to converge. Both 1E0L runs in Figure 4.2 show that the RMSD increases with increasing energy, but the one on the right does not reach the global-minimum energy. Overall, the energy correlated very well with RMSD in contrast to profiles of 1GAB, 1BDD, and 1CLB. The dominance of non-native structures in the unsuccessful runs leads to an almost uncorrelated nature of the Energy vs. RMSD plots for these proteins. Although the procedure consistently identified global minima for 1GAB and 1BDD as native like, it succeeded for only three simulations for 1CLB. Finally, Figure 4.3 shows a failure of REMCM for 1IGD where only one simulation (shown on the left in Fig. 4) sampled the space below 10 Å, and even these structures were high in energy. The expected global minimum is located at around -747 kcal/mol (this minimum was located by CSA as mentioned in section 5.3), but REMCM reached only the -700 kcal/mol energy levels.

Figure 4.4 shows the superposition of the native structures for 1GAB, 1E0L, 1BDD and 1CLB with the best structures obtained within 10 kcal/mol of the global minima. An outstanding superposition is obtained for 1GAB, with an RMSD of 2.3 Å and energy -673.2 kcal/mol (7.8 kcal/mol above the global minimum). The only discrepancy occurs in the loop between the first and second helix. It

Figure 4.2: Energy vs. RMSD plots for 1E0L, 1GAB, and 1BDD. The left column is an example of a good run, whereas the right column shows an unsuccessful simulation. In the left column, the global energy structures also have low RMSD values, whereas in the right column the simulation is trapped in a higher-than-native energy but low RMSD (1E0L), or the energy of the non-native conformations are on the same order as the native structures (1GAB, 1BDD).



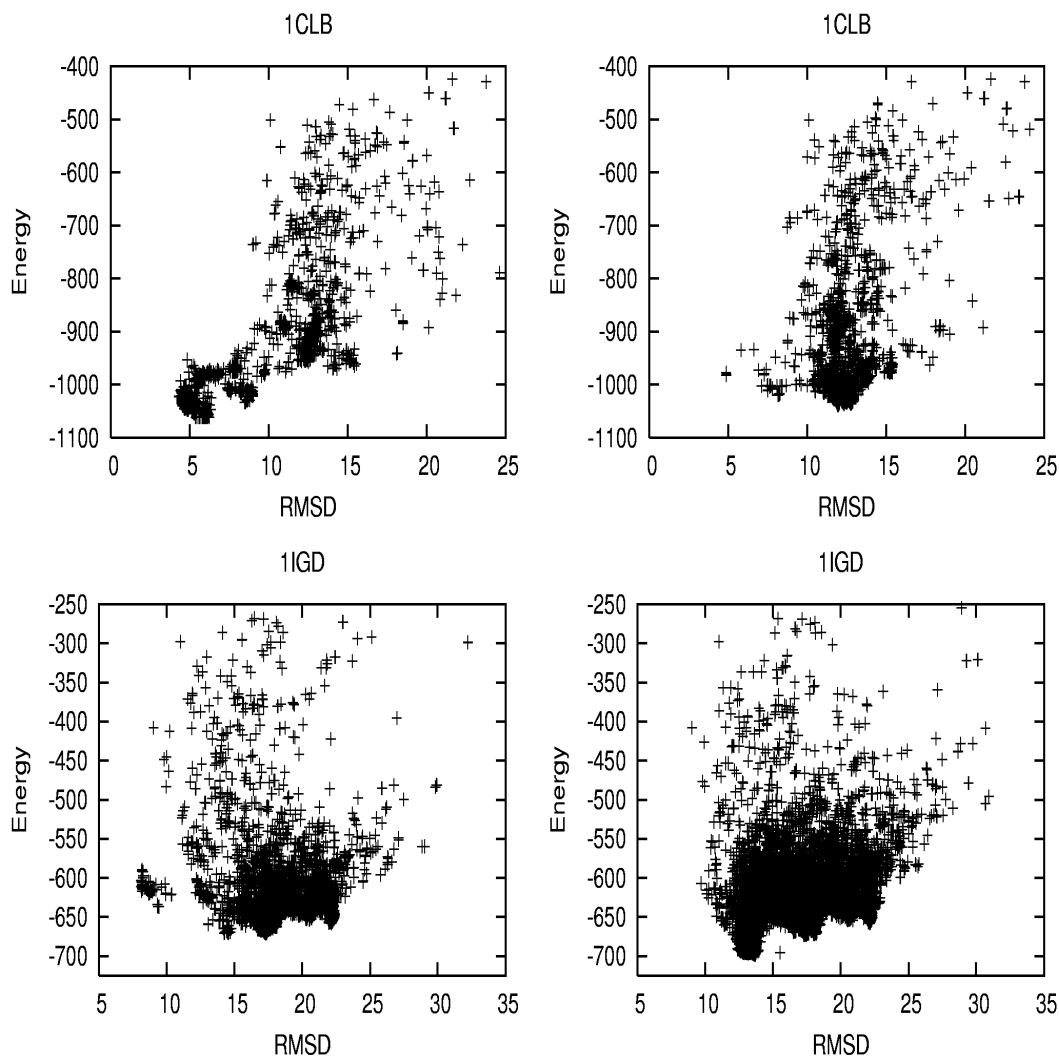


Figure 4.3: Energy vs. RMSD plots for 1CLB, and 1IGD. The left column is an example of a good run, whereas the right column shows an unsuccessful simulation. In the left column, the global energy structures of 1CLB also have low RMSD values, whereas in the right column the energies of some non-native conformations are on the same order as or lower than the native structures (1CLB). For 1IGD, none of the runs was successful, although the simulation visited structures closer to the native but high in energy (left plot).

can be seen that 1E0L deviates slightly at the C-terminus, where the computed conformation is missing the native interaction between residues 32 and 20. The energy of the low-energy structure of 1E0L is -286.1 kcal/mol (6.8 kcal/mol above the global minimum) and the best fit is 4.2 Å from the native. The best fit reported for 1BDD is 4.3 Å (10.5 kcal/mol above the global minimum), which, instead of forming a second helix as in the native, forms two short ones with the second helix in place of a loop. For 1CLB with an RMSD of 5.4 Å (9.2 kcal/mol above the global minimum), the overall topology is correct, although helix 3 is longer and more regular than in the native. These values are shown in Table 4.2. Finally, a comparison of the native structure of 1IGD with the best structure obtained within 10 kcal/mol is shown. It can be seen that the simulated conformation lacks both β hairpins and, instead, is a three-helix bundle.

4.4.2 Comparison of REMCM to REM

Since REMCM is based on the REM-without-minimization method, an important question arises as to whether the new modified method is faster and more consistent in locating the global minimum than the traditional Replica Exchange method. REM has been used mainly for calculations of thermodynamic properties; nevertheless, it has also been employed as a global optimization method.³⁷ Experience with Monte Carlo-with-Minimization has shown that addition of minimization to traditional Monte Carlo improves the search procedure considerably,^{7, 8} suggesting that adding minimization to the Replica Exchange method might show a similar effect.

Fig 4.5 compares the performance of REM and REMCM showing the lowest energy obtained for the given number of energy evaluations. The comparison was

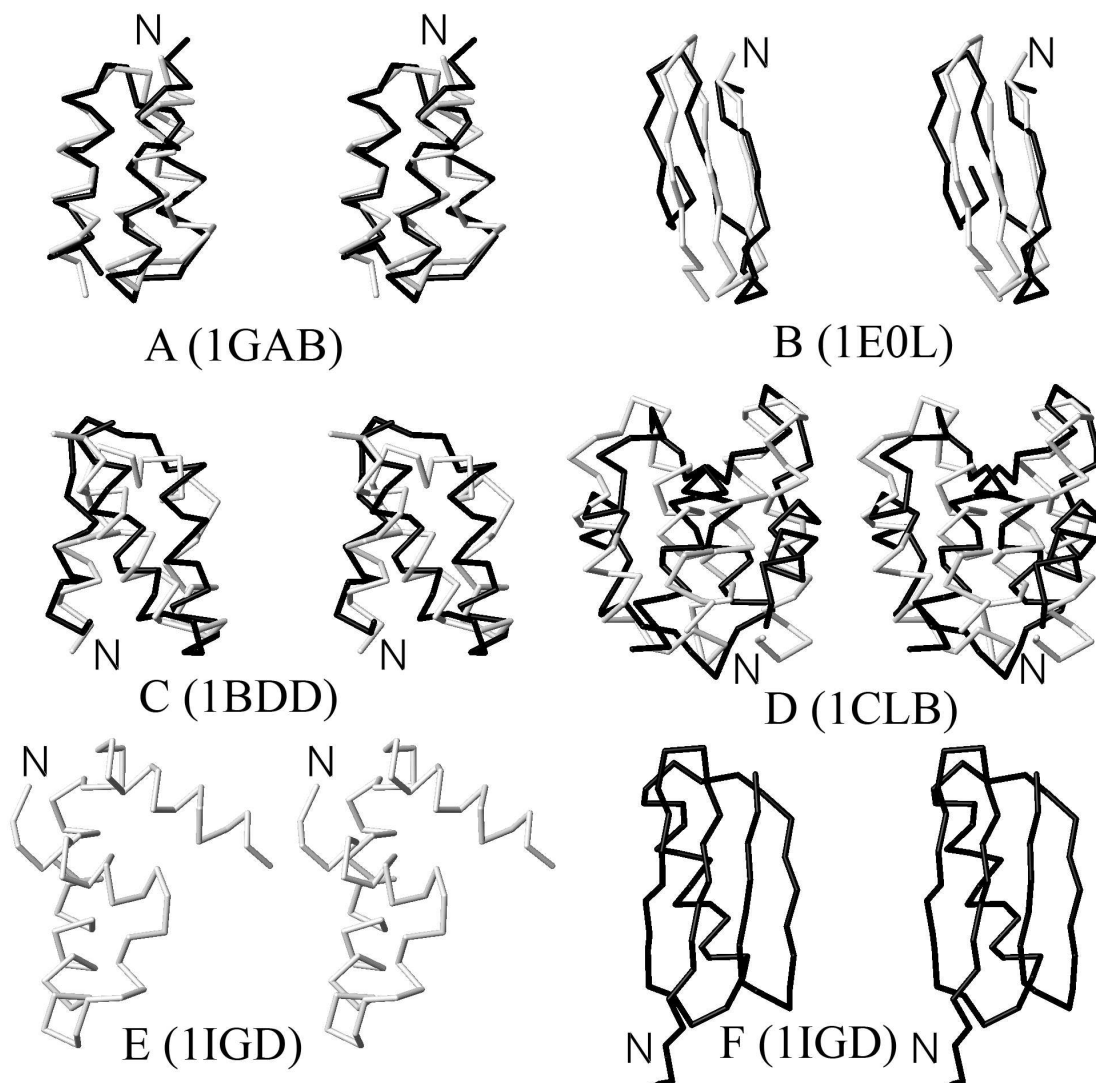


Figure 4.4: Stereo-views of the superposition of the experimental (black) and best predicted structures (gray) within 10 kcal/mol energy cutoff from the global minimum obtained by REMCM. (A) 1GAB (7.8 kcal/mol above the global minimum, RMSD = 2.3 Å); (B) 1E0L (6.8 kcal/mol above the global minimum, RMSD = 4.2 Å); (C) 1BDD (10.5 kcal/mol above the global minimum, RMSD = 4.3 Å); (D) 1CLB (9.2 kcal/mol above the global minimum, RMSD = 5.4 Å); (E) 1IGD (structure obtained by REMCM, 7.0 kcal/mol above the global minimum, RMSD = 12 Å); (F) 1IGD (native structure). Because of the large RMSD between the low-energy structure and the native structure, they are displayed separately in (E) and (F).

made for 1GAB, for which the energy landscape presumably is smooth, because the folded state was obtained in *all* five independent REMCM simulations with this protein; thus, REM might be expected to perform very well. It can be seen that the lowest energies obtained with REM are higher by about 70 kcal/mol than the lowest energies found by REMCM. A similar observation was also seen in Molecular Dynamics simulations of several proteins using UNRES,³⁸ where the lowest energy obtained by Molecular Dynamics was much higher than the lowest energy obtained by CSA. This effect might be explained by thermal motion which is neglected when using minimization-based methods such as CSA or REMCM. The present method of hierarchical optimization of protein energy landscapes² uses the CSA method to generate decoys and thus ignores the entropy factor, which consequently makes it very hard for methods such as MD or REM to reach low-energy regions with UNRES. However, the advantage of introducing MCM to REM is that the CSA global minimum can be attained. The convergence appears to be faster for REMCM than REM in Fig 4.5., showing an improvement over the traditional Replica Exchange method. Finally, the traditional Replica Exchange method not only converged more slowly but also failed to locate the global energy basin as seen in figure 4.5; the lowest-RMSD structures differed by 8 Å from the native and were high in energy (not shown here). The low-energy structures obtained by REM had RMSD's of 14 Å from the native.

4.4.3 Comparison of REMCM to other methods

To evaluate the effectiveness of REMCM, a comparison was made to other global optimization methods, specifically to Monte Carlo-with-Minimization (MCM),^{7, 8} Conformational Space Annealing (CSA),¹⁶⁻¹⁸ and Conformational Family Monte

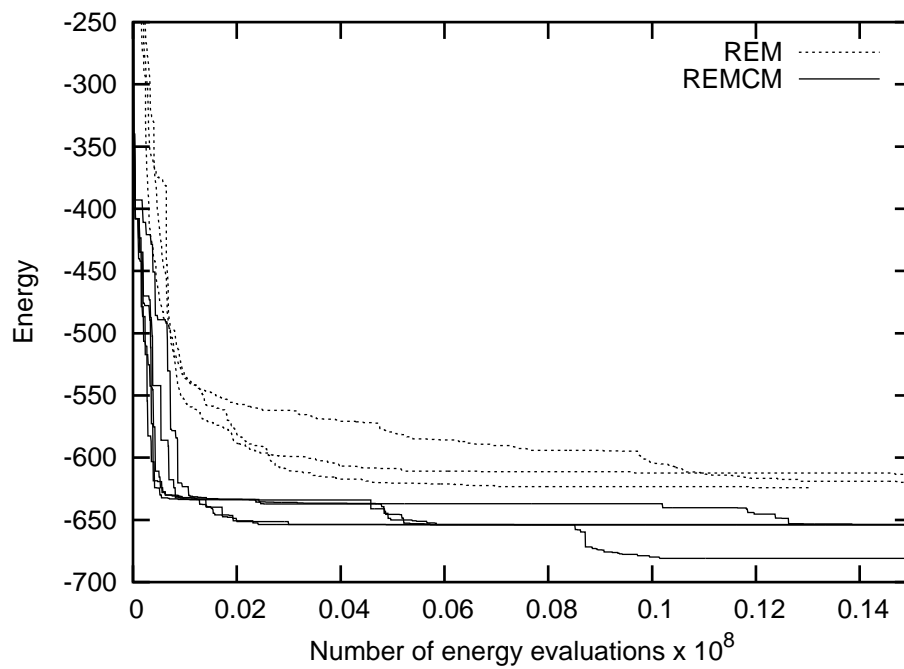


Figure 4.5: The graph shows the lowest energy obtained at a given energy-evaluation step for the given number of energy evaluations. Different curves correspond to different simulations with the same starting parameters for REMCM (solid line) and REM (dotted line).

Carlo (CFMC).¹¹ The comparison was carried out for all five proteins considered in section 4.3.1 with the 4P force field.³

The performance comparison between MCM and REMCM is shown in Figure 4.6. The plots illustrate the lowest-energy structure obtained at a given energy-evaluation step for the given number of energy evaluations. It can be seen that MCM has problems consistently locating the global minimum. For 1E0L, only two out of five MCM runs converged, while the other three were trapped. For 1GAB and 1CLB, the statistics for MCM was even worse, showing that only one out of five runs was successful in locating the low-energy basins, although the absolute energies were not quite as low as for REMCM. 1BDD showed good results with four out of five MCM runs, converging slightly slower than the REMCM runs.

Table 4.2 contains a summary of the simulation results with all proteins for REMCM, CSA, CFMC, MCM and one comparison for 1GAB. The Table entries show both the energy and C^α RMSD with respect to the native structure for a best run with each method. The row marked as l corresponds to the lowest energy structure found in that simulation run. The row marked as 10 (20) corresponds to the best RMSD structure found up to 10 (20) kcal/mol higher than the lowest energy observed. The empty fields indicate that the value was not improved by including structures of higher energy. In comparison to CSA, REMCM obtains comparable energies for all proteins except for 1IGD. For 1GAB, REMCM obtains a structure lower than CSA by 12 kcal/mol. CFMC and MCM on the other hand appear to become trapped in higher-energy conformations.

The comparison results for REMCM with CSA and CFMC are shown in Fig 4.7. The plots show the lowest energy structure obtained at a given energy-evaluation step for the given number of energy evaluations. In comparing the results, it is

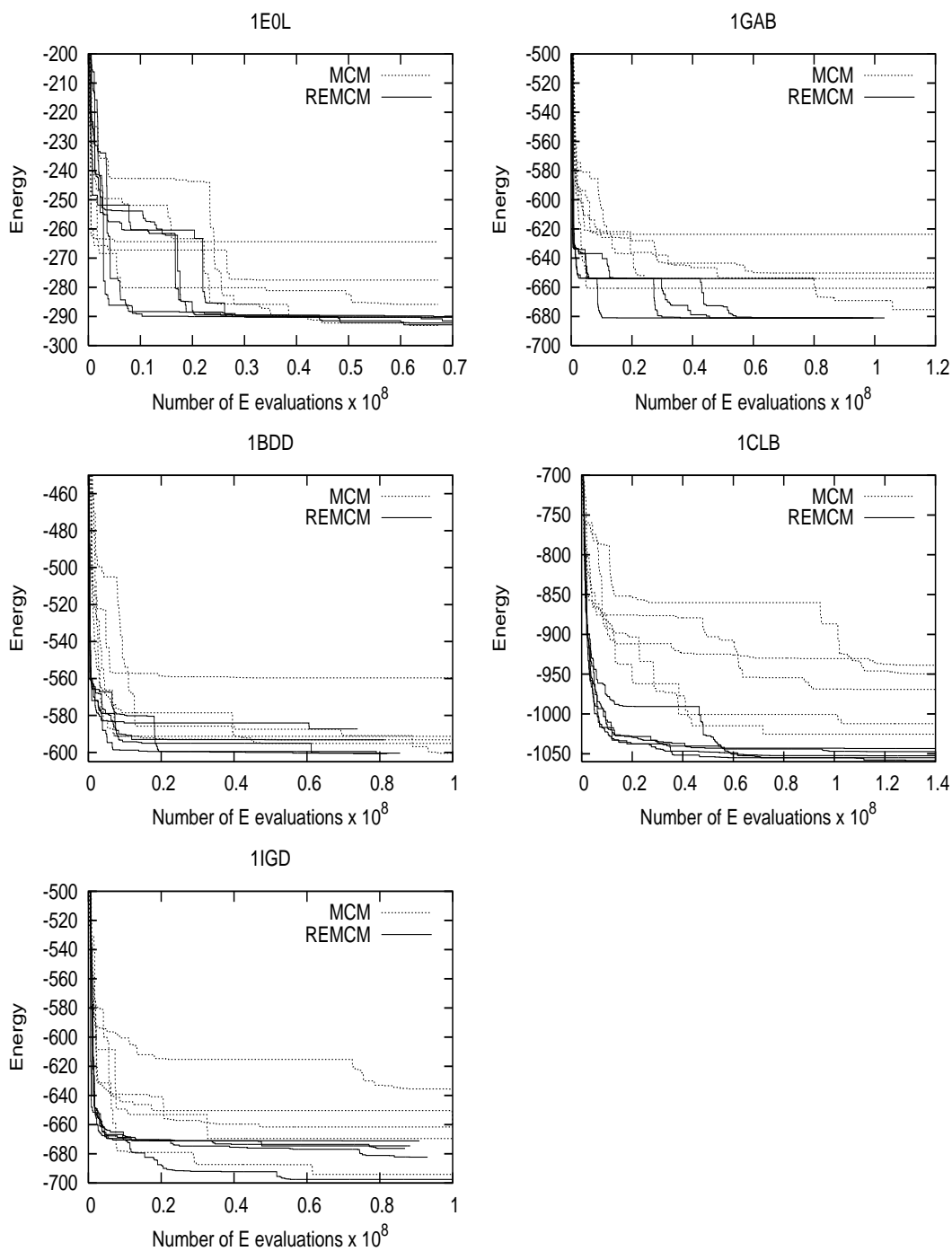


Figure 4.6: Performance comparison of MCM and REMCM. The comparison was carried out for five proteins, and the plots denote the lowest energy obtained at a given energy-evaluation step for the given number of energy evaluations. Different Curves correspond to different simulations with the same starting parameters for REMCM (solid line) and MCM (dotted line). It can be seen that REMCM converges faster and is also more consistent in locating the global energy minima.

Table 4.2: Best simulation runs for REMCM, CSA, CFMC, and MCM for each protein. The results show both the energy (first row) and the corresponding C $^{\alpha}$ RMSD (second row) with respect to the native. 1 corresponds to the lowest energy structure found in a given simulation. 10 and 20 correspond to the best RMSD structures found 10 (20) kcal/mol higher than the lowest energy observed. Empty fields indicate that the value above is not improved in the higher energy structures.

			REMCM	CSA	CFMC	MCM	REM
1E0L	1	kcal/mol	-293	-296	-274	-293	
		Å	(4.79)	(4.73)	(5.10)	(4.76)	
	10	kcal/mol	-286	-288	-274	-290	
		Å	(4.232)	(3.780)	(4.867)	(4.14)	
	20	kcal/mol	-277	-282			
		Å	(3.616)	(3.633)			
1GAB	1	kcal/mol	-681	-669	-672	-675	-627
		Å	(2.64)	(2.93)	(2.75)	(2.59)	(14.63)
	10	kcal/mol	-673	-668	-666	-672	-619
		Å	(2.31)	(2.89)	(2.32)	(2.54)	(14.41)
	20	kcal/mol	-667				
		Å	(2.27)				
1BDD	1	kcal/mol	-601	-605	-592	-601	
		Å	(5.73)	(4.77)	(10.07)	(5.74)	
	10	kcal/mol	-590	-598	-583	-591	
		Å	(4.34)	(4.79)	(3.74)	(3.98)	
	20	kcal/mol	-586	-591	-581	-583	
		Å	(3.85)	(3.51)	(3.40)	(3.51)	
1CLB	1	kcal/mol	-1059	-1057	-1039	-1025	
		Å	(5.61)	(4.84)	(5.28)	(6.45)	
	10	kcal/mol	-1050	-1050	-1030	-1021	
		Å	(5.38)	(4.69)	(4.63)	(5.92)	
	20	kcal/mol	-1041	-1043			
		Å	(4.70)	(4.31)			
1IGD	1	kcal/mol	-698	-747	-694	-675	
		Å	(13.23)	(5.61)	(13.08)	(10.88)	
	10	kcal/mol	-691	-738	-684	-671	
		Å	(12.86)	(4.70)	(10.73)	(10.71)	
	20	kcal/mol	-680	-733			
		Å	(12.21)	(4.40)			

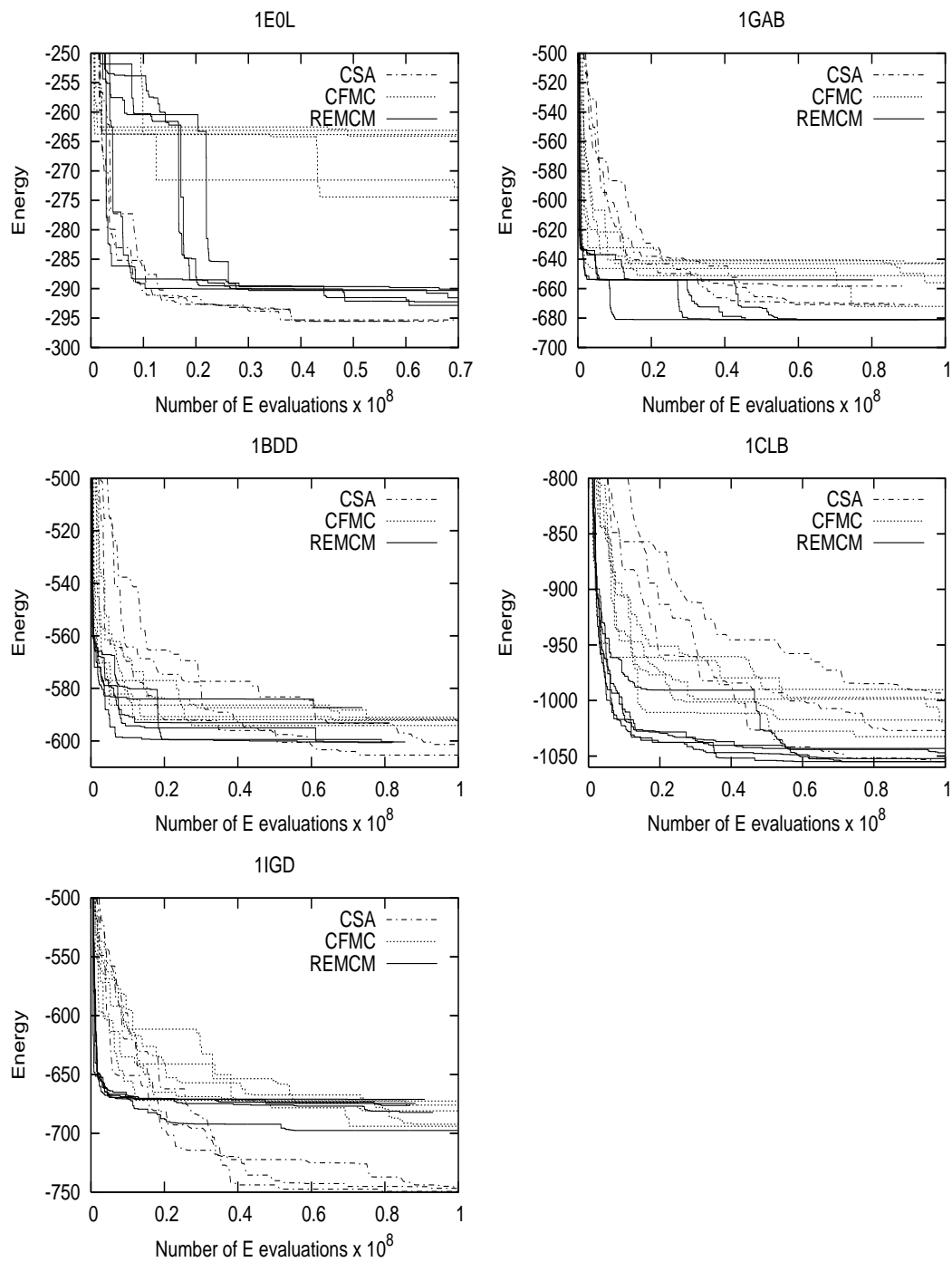
worthwhile to point out the difference in the overall shape of the simulation progress in the different methods. CSA seems to descend in a smooth way, whereas CFMC and REMCM seem to have a very sharp drop at the beginning. After this drop, they either find the native basin fairly quickly or become trapped in a different part of the surface, in which case it seems to take some time to locate the native basin. For 1E0L, the most dependable procedure appears to be CSA, which consistently converged to the global minimum, whereas CFMC became trapped at around -260 kcal/mol. REMCM obtained the global minimum, although in three simulations it took a considerable number of energy evaluations before it reached the native basin. When it did reach the native basin right away, however, it converged as fast as CSA. For 1GAB and 1BDD, both REMCM and CFMC converged faster and, overall, reached lower energies than CSA. Both 1E0L, and 1GAB were training proteins for the 4P force field, and, in contrast to 1E0L, 1GAB posed no challenge to REMCM and CFMC. From the graph for 1CLB in Fig 4.7, it is evident that REMCM converges much faster than the other two methods. However, as mentioned in the comparison of REMCM with MCM, only three of five simulations actually located the native basin. The other simulations reached very low energies, but these were populated mainly by non-native structures (as shown in Figure 4). It appears that current UNRES parameters make 1CLB a hard target, especially since the difference in energy between the native and non-native ensembles is very small. Overall, however, REMCM was very efficient in locating low-energy structures. Finally, the plot for 1IGD shows a complete failure of REMCM and CFMC to locate the global minimum. This minimum was observed by CSA at around -747 kcal/mol, but no variation of parameters in REMCM succeeded in obtaining structures even below the -700 kcal/mol level. This protein was also used in the

training set for the 4P force field, and like 1E0L (and unlike 1GAB) it posed difficulty for REMCM and CFMC.

4.5 Conclusions

In this work, Replica Exchange coupled with Monte Carlo-with-Minimization was applied to search the conformational space of five test proteins with the United Residue force field. Adding minimization altered the behavior of a typical Replica Exchange simulation, so that the high temperature replicas in REMCM also sample some of the low-energy subspace of the energy landscape. The test of this procedure led to results which were compared to other optimization methods used with UNRES, namely MCM, CSA and CFMC. Overall, REMCM was successful on four out of five proteins tested; furthermore, it performed much more consistently than MCM and CFMC in locating the global minima, and converged to these low-energy regions much more efficiently than CSA. It failed to locate the global minimum of 1IGD, for which only CSA was able to reach this native region. Since 1IGD was also in the training set for the force field used in the computations, this raises an interesting question as to why REMCM and CFMC (methods not used in the energy parameterization process) were unable to obtain the global minimum. A possible answer might lie in the fact that both REMCM and CFMC use similar kinds of perturbation moves, whereas CSA uses genetic operators to evolve the population of conformations. In particular, our recent implementation of CSA¹⁸ exchanges β -hairpins and non-local strand pairs between conformations, thus enhancing the probability of forming β -structures. Without these moves even CSA could not locate the global minimum of 1IGD because the β -structures dis-

Figure 4.7: Performance comparison of different global optimization methods. Solid line represents runs with REMCM, dotted line CFMC, while dash-dot lines are CSA runs. The comparison was carried out for five proteins, and the plots denote the lowest energy obtained at a given energy-evaluation step for the given number of energy evaluations. It is evident that CFMC and REMCM seem to converge faster than CSA, although CFMC has a tendency to become stuck (e.g., 1E0L). CSA outperforms the other two methods for 1IGD where it systematically reaches a lower energy basin. REMCM appears to perform the best on 1CLB where it both converges faster and reaches lower energy structures, although three of the runs (similar to the right-hand panel in Fig 4.3) produced non-native structures as the lowest energy. Thus, the best results for REMCM are for 1GAB and for 1BDD.



appeared during the course of CSA simulations in favor of α -helices.¹⁸ No similar operators or genetic algorithm are present in REMCM, which could explain the failure of this method to locate the global minimum of IIGD. Because CSA was used in the parameterization of the force field, this might also suggest that the landscape is more biased towards this method, making it easier for CSA to explore energy basins.

Nevertheless, REMCM seems to have some interesting features and has potential as a stand-alone global optimization method applied to biological macromolecules. Moreover, because it is easy to implement and has few parameters to adjust, it is very suitable for implementation in the future revision of our hierarchical optimization procedure.² This optimization procedure is based on a hierarchical design of the potential-energy landscape such that the energy decrease follows the increase of native-likeness.³⁹ REMCM would add the aspect of updating the conformations on the fly, thereby reducing the number of full CSA runs, which are typically required after every iteration of the optimization procedure. This could speed up the entire optimization process considerably, thus allowing us to include more and larger proteins in the training set, resulting in a force field with much greater predicting power. Finally, subjecting the parameterization process to more than one optimization method might improve its performance with other optimization methods.

BIBLIOGRAPHY FOR CHAPTER 4

- [1] Scheraga, H.A.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Ripoll, D.R.; Vila, J.A.; Kazmierkiewicz, R.; Saunders, J.A.; Arnautova, Y.A.; Jagielska, A.; Chinchio, M.; Nancias, M., *Front. Biosci.* 2005, 9, 3296.
- [2] Oldziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16934.
- [3] Oldziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nancias, M.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16950.
- [4] Anfinsen, C. B., *Science* 1973, 181, 223.
- [5] Metropolis, N.; Ulam, S., *J. Am. Stat. Assoc.* 1949, 44, 335.
- [6] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., *J. Chem. Phys.* 1953, 21, 1087.
- [7] Li, Z.; Scheraga, H. A., *Proc. Natl. Acad. Sci., U. S. A.* 1987, 84, 6611.
- [8] Li, Z.; Scheraga, H. A., *J. Molec. Str. (Theochem)* 1988, 179, 333.
- [9] Ripoll, D. R.; Scheraga, H. A., *Biopolymers* 1988, 27, 1283.
- [10] Ripoll, D. R.; Scheraga, H. A., *J. Protein Chem.* 1989, 8, 263.
- [11] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.
- [12] Piela, L.; Kostrowicki, J.; Scheraga, H. A., *J. Phys. Chem.* 1989, 93, 3339.
- [13] Pillardy, J.; Olszewski, K. A.; Piela, L., *J. Phys. Chem.* 1992, 96, 4337.
- [14] Pillardy, J.; Liwo, A.; Groth, M.; Scheraga, H. A., *J. Phys. Chem. B* 1999, 103, 7353.
- [15] Pillardy, J.; Liwo, A.; Scheraga, H.A., *J. Phys. Chem. A* 1999, 103, 9370.
- [16] Lee, J.; Scheraga, H. A.; Rackovsky, S., *J. Comput. Chem.* 1997, 18, 1222.
- [17] Lee, J.; Scheraga, H. A., *Int. J. Quant. Chem.* 1999, 75, 255.
- [18] Czaplewski, C.; Liwo, A.; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [19] Hukushima, K.; Nemoto, K., *J. Phys. Soc. Jpn.* 1996, 65, 1604.
- [20] Hansmann, U. H. E., *Chem. Phys. Lett.* 1997, 281, 140.

- [21] Swendsen, R. H.; Wang, J. S., *Phys. Rev. Lett.* 1986, 57, 2607.
- [22] Gront, D.; Kolinski, A.; Skolnick, J., *J. Chem. Phys.* 2001, 115, 1569.
- [23] Kolinski, A.; Gront, D.; Pokarowski, P.; Skolnick, J., *Biopolymers* 2003, 69, 399.
- [24] Fenwick, M. K.; Escobedo, F. A., *J. Chem. Phys.* 2003, 119, 11998.
- [25] Romiszowski, P.; Sikorski, A., *Physica A* 2004, 336, 187.
- [26] Fukunishi, H.; Watanabe, O.; Takada, S., *J. Chem. Phys.* 2002, 116, 9058.
- [27] Jang, S.; Shin, S.; Pak, Y., *Phys. Rev. Lett.* 2002, 91, 058305.
- [28] Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids*, Oxford University Press; New York 1987.
- [29] Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H., *Nat. Struct. Biol.* 2000, 7, 375.
- [30] Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I., *Biochemistry* 1992, 31, 9665.
- [31] Skelton, N. J.; Kordel, J.; Chazin, W. J., *J. Mol. Biol.* 1995, 249, 441.
- [32] Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drankenberg, T.; Bjorck, L., *J. Mol. Biol.* 1997, 266, 859.
- [33] Derrick, J. P.; Wigley, D. B., *J. Mol. Biol.* 1994, 243, 906.
- [34] Kertész, J.; Kondor, I., *Advances in Computer Simulation*, Springer-Verlag; Berlin Heidelberg 1998.
- [35] Gay, D. M., *ACM Trans. Math. Software* 1983, 9, 503.
- [36] Chinchio, M.; Scheraga, H. A., *J. Comp. Chem.* 2005, To be submitted for publication.
- [37] Gront, D.; Kolinski, A.; Skolnick, J., *J. Chem. Phys.* 2000, 113, 5065.
- [38] Liwo, A.; Khalili, M.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 2362.
- [39] Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Oldziej, S.; Pillardy, J.; Scheraga, H.A., *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 1937.

Chapter 5

Replica Exchange and Multicanonical Algorithms with the coarse-grained UNRES force field *

5.1 Introduction

Efficient sampling algorithms have been an essential component of methods for studying protein structure and dynamics in structural biology and theoretical chemistry. A variety of sampling algorithms have been used in our laboratory and, depending on whether the goal is global optimization or folding simulations, they can be categorized in the following way.

For successful prediction of the three-dimensional structure of a protein (based solely on its amino acid sequence), several classes of algorithms have been used. The first class includes modifications of the Metropolis Monte Carlo procedure,^{1, 2} such as Monte Carlo-with-Minimization (MCM),^{3, 4} electrostatically-driven Monte Carlo (EDMC),^{5, 6} Conformational Family Monte Carlo (CFMC),⁷ and Replica Exchange Monte Carlo-with-Minimization (REMCM).⁸ The second class includes deformation-based methods, such as the diffusion-equation method (DEM),⁹ the distance-scaling method (DSM),¹⁰ and the self-consistent basin-to-deformed-basin

*Published as Nancias, M.; Czaplewski, C.; Scheraga, H.A., *J. Chem. Theo. Comp.* 2005, in press. Copyright (2006) American Chemical Society.

method (SCBDBM).^{11, 12} The third class includes genetic algorithms such as the Conformational Space Annealing (CSA) method.¹³⁻¹⁵ For the study of protein-folding pathways, recently-applied Molecular Dynamics with the united-residue (UNRES) force field¹⁶⁻¹⁹ has been shown to be particularly effective. To evaluate thermodynamic properties, another class of sampling methods is necessary. This is because minimization-based methods violate the condition of microscopic reversibility required for producing Boltzmann statistics and, although methods such as Molecular Dynamics or Metropolis Monte Carlo can be used for estimating thermodynamic properties as well as for a global search, they easily become trapped for complex systems, and thus are not the most effective methods for studying large systems.

The origins of one of the most popular advanced sampling methods, the Replica Exchange method (also known as Exchange Monte Carlo,²⁰ or Parallel Tempering²¹), can be traced back to the work carried out by Swendsen and Wang²² for spin-glass systems, and the more familiar form of the algorithm was developed by Geyer²³ with his use of Metropolis-coupled Markov chain Monte Carlo. In the Replica Exchange method, several copies (replicas) of the system are simulated with standard Metropolis Monte Carlo^{1, 2} or Molecular Dynamics procedures (each replica differing from the others in a particular way, usually in temperature), while permitting an exchange among the replicas, and thus surmounting barriers in the rugged conformational energy landscapes. This method has been applied extensively in protein-folding simulations using both lattice²⁴⁻²⁷ and off-lattice models.²⁸⁻³²

Recently, much attention has been paid to generalized ensemble algorithms whose advantage is efficient sampling of the conformational energy landscape. In this approach, efficient sampling does not mean locating the global minimum as

quickly as possible, but rather covering the landscape in such a way as to provide accurate statistics. Two well-known methods are the multicanonical algorithm^{33, 34} (also known as entropy sampling^{35, 36}), and simulated tempering³⁷ (also referred to as the method of expanded ensembles³⁸). The multicanonical algorithm performs a one-dimensional random walk in energy space, while simulated tempering follows a random walk in temperature space, thereby inducing a random walk in the space of potential energy. Although these algorithms are generally too expensive for locating global minima,³⁹ they are useful for producing accurate statistics for thermodynamic averages of observed variables. However the application of these algorithms is nontrivial and very tedious; in particular, the need to obtain the proper sampling weights often limits the use of generalized ensemble techniques.⁴⁰

Due to the fact that the Replica Exchange method alleviates the problem of the tedious estimation of weight factors in the multicanonical algorithms, combinations of replica exchange with generalized ensemble methods have been developed, e.g., REMUCAREM⁴¹ i.e., Replica Exchange Multicanonical Algorithm with Replica Exchange; others include Replica Exchange Simulated Tempering, or Simulated Tempering Replica Exchange (REST, STREM, respectively).⁴² Other modifications of Replica Exchange include Replica Exchange with Solute Tempering,⁴³ Model Hopping,⁴⁴ Hamiltonian Replica Exchange,⁴⁵ and the Replica-Exchange Method Using a Generalized Effective Potential.⁴⁶

Having demonstrated that the coarse-grained united-residue (UNRES) protein model is helpful in surmounting problems with all-atom models,^{18, 47} we apply the Replica Exchange method (REM), the Replica Exchange Multicanonical Method (REMUCA), and the Replica Exchange Multicanonical Method with Replica Exchange (REMUCAREM), in both Monte Carlo and Molecular Dynamics versions, to the UNRES model in the present work. The advantage of Replica Exchange lies

in its simplicity and, in contrast to other methods, it is not very sensitive to the few parameters involved therein (such as the cooling schedule in simulated tempering, or the successful estimation of weight factors in multicanonical algorithms). The power of REMUCA lies in the effective estimate of the multicanonical weight factors from Replica Exchange simulations. REMUCAREM further exploits the idea of running several replicas of multicanonical simulations with different set of multicanonical weights. The motivation behind the present work is to test the applicability of these algorithms to determine the thermodynamic properties of large systems. The ability to compute thermodynamic properties will thereby enable us to improve our UNRES model, and consequently improve protein folding simulations, i.e., bring our simulated results closer to experimental ones.

5.2 Methods

5.2.1 The UNRES force field

All the above-mentioned algorithms were implemented with the United Residue force field, hence in this section, the UNRES model of polypeptide chains and the corresponding force field is described briefly. In chapter 3 the UNRES model used with Monte Carlo procedures was described, this section extends the description of UNRES force field for Molecular Dynamics.

Molecular Dynamics with UNRES requires an extra degree of freedom, namely the vibrations of the virtual-bond lengths, which are treated with an additional harmonic potential. The complete UNRES potential-energy function for Molecular Dynamics is then expressed by the following equation:¹⁸

$$U_{MD} = U_{MC} + w_{vib} \sum_i U_{vib}(d_i) \quad (5.1)$$

where U_{MC} is the Monte Carlo UNRES potential energy described in chapter 2

(eq. 3.2) and $U_{vib}(d_i)$, d_i being the length of the i th virtual bond, are the simple harmonic potentials defined as $U_{vib}(d_i) = (1/2) k_{d_i} (d_i - d_i^o)^2$, where k_{d_i} is the force constant of the i th virtual bond, currently set at 500 kcal/(mol \times \AA^2) and d_i^o is the average length (corresponding to that used in the fixed-bond UNRES potential) of the i th virtual bond; e.g., $d_i^o = 3.8 \text{ \AA}$ for a $C^\alpha \cdots C^\alpha$ virtual bond corresponding to a *trans* peptide group. As in previous work,¹⁸ the weight w_{vib} was arbitrarily set at 1.

5.2.2 Replica Exchange Method (REM)

The Replica Exchange method is an extension of the Metropolis Monte Carlo, or Molecular Dynamics, methods. The underlying idea is to run different copies (replicas) of the system at different levels of a certain property (such as temperature). To summarize the method, a Monte Carlo (MC) or Molecular Dynamics (MD) simulation is carried out on each selected conformation at its assigned temperature for a determined number of MC or MD steps, after which the neighboring replicas undergo an exchange with the acceptance criterion described below (in eq. 5.3). Let

$$\Delta \equiv \left[(\beta_m - \beta_n) \{ E(Y) - E(X) \} \right] \quad (5.2)$$

where β_m is the inverse temperature defined as $1/(k_B T_m)$, $E(X)$ is the energy of conformation X. If one adopts the Metropolis method, the replica-exchange transition probability can be expressed as

$$W(X, \beta_m | Y, \beta_n) = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases} \quad (5.3)$$

i.e., if Δ is less than or equal to 0, the exchange is performed (since the probability is 1); otherwise a random number between 0 and 1 is generated and compared to

the factor $\exp(-\Delta)$. If the value of this factor is smaller, the exchange is performed; otherwise the exchange is rejected.

To evaluate thermodynamic quantities at any temperature, it is essential to extract maximum information from all replicas. For this purpose a multi-histogram reweighting technique^{48, 49} can be used. For a replica exchange simulation with M replicas at M distinct temperatures, a set of M energy histograms $N_m(E)$ is obtained. The densities of states $[n(E)]$ are then obtained self-consistently from the following WHAM^{48, 49} equations:

$$n(E) = \frac{\sum_{m=1}^M g_m^{-1} N_m(E)}{\sum_{m=1}^M g_m^{-1} n_m \exp(f_m - \beta_m E)} \quad (5.4)$$

and

$$\exp(-f_m) \equiv \sum_E n(E) \exp(-\beta_m E) \quad (5.5)$$

where $N_m(E)$ is the histogram at temperature T_m , $\beta_m = 1/(k_b T_m)$ is the inverse temperature, n_m is the total number of samples in the m th replica, $g_m = 1 + 2\tau_m$, and τ_m is the integrated autocorrelation time at temperature T_m . In biomolecular systems, g_m is approximately constant⁴⁹ and, therefore, can be canceled in eq. 5.4. The WHAM equations 5.4 and 5.5 are evaluated self-consistently and the resulting densities of states are used to evaluate the expectation value of any observable A in equation 5.6:

$$\langle A \rangle_T = \frac{\sum_E A(E) n(E) \exp(-\beta E)}{\sum_E n(E) \exp(-\beta E)} \quad (5.6)$$

5.2.3 Multicanonical Algorithm (MUCA)

A single canonical simulation (MC or MD) by definition samples a very restricted energy region. Furthermore, when sampling the conformations of the protein in

low-energy regions, the multiple-minima problem is usually encountered and the simulation can be trapped in a particular local energy minimum, making it difficult to obtain a reliable estimate of the density of states of proteins. In determining the density of states of a large system by simulation procedures, a clear criterion is needed about the stage of simulations at which all of the conformational space of the protein has been sampled sufficiently. Traditional MC or MD procedures do not provide such a convergence criterion. For these reasons, a multicanonical algorithm^{33, 34} (also known as Entropy Sampling^{35, 36}) has been used for protein studies. In section 5.2.4, we show why MUCA is combined with REM to produce REMUCA, whose efficiency is explored in the present work. For this purpose, we first outline MUCA. In the next paragraph, we present the background of Entropy Sampling and tie it together with the Multicanonical Algorithm notation.

In the present work, we use the term "conformation" to indicate a particular structure and the term "state" to denote all the conformations that either have a given energy or are within a small energy interval. The probability of occurrence of a conformation x with energy E , denoted as $P(x)$, and the probability of occurrence of a state with energy E , denoted as $P(E)$, are related to each other in a canonical ensemble by the following relations, with E being written for $E(x)$:

$$P(x) \propto \exp(-\beta E) \quad (5.7)$$

$$P(E) \propto n(E)\exp(-\beta E) = \exp[S(E)/k_B - \beta E] \quad (5.8)$$

where k_B is the Boltzmann constant, $\beta = 1/k_B T$ with T being the temperature, $n(E)$ is the number of conformations with energy E (i.e., density of states), and $S(E) = k_B \ln[n(E)]$ is the entropy of the state with energy E .

The Entropy Sampling method is based on an artificial distribution of states, in which the probability of occurrence of a state with energy E is scaled by the exponential of the *negative* of the entropy of the state, $S(E)$. In Entropy Sampling,

the probabilities of occurrence of a conformation x and a state with energy E , respectively, are defined as

$$P(x) \propto \exp\{-S[E(x)]/k_B\} \quad (5.9)$$

$$P(E) \propto n(E)\exp[-S(E)/k_B] \quad (5.10)$$

where $n(E)$ and $S(E)$ have similar meanings as described above. Equations 5.9 and 5.10 can be related to equations 5.7 and 5.8 by first setting $\beta = 0$ (i.e., temperature to infinity) in equations 5.7 and 5.8 and then multiplying the resulting probabilities by the weight factor $\exp[-S(E)/k_B]$. The physical meaning of this modification is that the larger the conformational entropy of a state, the smaller is the weight given to the state. In this way, the probabilities of occurrence of all states with different energies are constant in the new distribution, i.e., $P(E)$ of equation 5.10 is a constant, taken as 1.

To connect the Entropy Sampling formalism to the commonly-used Multicanonical Algorithm, we can define a new variable, the multicanonical energy E_{mu} , in the following way

$$E_{mu}(E; T_0) = T_0 S(E) = k_B T_0 \ln[n(E)] \quad (5.11)$$

where T_0 is the reference temperature, and $S(E)$ is the microcanonical entropy as above. The reference temperature is the temperature at which the MC or MD multicanonical simulation is carried out. It should be noted that the reference temperature theoretically plays no role in calculating thermodynamics, because the formula for obtaining thermodynamic quantities (eq. 5.6) is independent of T_0 ; however, in practice, the value chosen for T_0 affects the sampling efficiency of numerical simulations. Equations 5.9 and 5.10 then become

$$P(x) \propto \exp\{-E_{mu}[E(x); T_0]/T_0 k_B\} \quad (5.12)$$

$$P(E) \propto n(E) \exp[-E_{mu}(E; T_0)/T_0 k_B] \quad (5.13)$$

Consequently, the multicanonical Monte Carlo simulation is carried out with the following modified Metropolis acceptance criterion:

$$W(X|Y) = \begin{cases} 1 & \text{for } \Delta E_{mu} \leq 0 \\ \exp(-\beta_0 \Delta E_{mu}) & \text{for } \Delta E_{mu} > 0 \end{cases} \quad (5.14)$$

where $\beta_0 = 1/k_B T_0$ and $\Delta E_{mu} \equiv E_{mu}[E(Y); T_0] - E_{mu}[E(X); T_0]$.

The multicanonical molecular dynamics simulation is carried out by integrating the following modified Newton equation;⁵⁰⁻⁵² see eq. 21 of reference 50:

$$\dot{\mathbf{p}}_k = -\frac{\partial E_{mu}(E; T_0)}{\partial \mathbf{q}_k} = \frac{\partial E_{mu}(E; T_0)}{\partial E} \mathbf{f}_k \quad (5.15)$$

where \mathbf{p}_k is the momentum, \mathbf{q}_k is the generalized coordinate of the kth atom, and \mathbf{f}_k is the force on the kth atom. Specifically the UNRES MD equation of motion (eq. 32 of reference 16) is modified as

$$\ddot{\mathbf{q}}(t) = -G^{-1} \frac{\partial E_{mu}(U; T_0)}{\partial U} \nabla_{\mathbf{q}} U[\mathbf{q}(t)] \quad (5.16)$$

where U [being $U(x)$] is the UNRES potential energy (U_{MD} of eq. 5.1), $\mathbf{q}(t)$ are the generalized coordinates at time t , and G is the mass matrix (eq. 26 of reference 16). In practice, one can use cubic splines to approximate $\partial E_{mu}(U; T_0)/\partial U$.

Because the density of states is usually not known *a priori*, the multicanonical weights are usually obtained by iterating short runs;^{36, 53-55} i.e. E_{mu} is obtained such that equation 5.13 is constant for all energies E . For this purpose, one uses the single histogram reweighting technique to obtain a new estimate of the densities of states after each iteration:

$$n(E) = \frac{N_{mu}(E)}{\exp[-\beta_0 E_{mu}(E; T_0)]} \quad (5.17)$$

where N_{mu} is the histogram obtained from the multicanonical simulation (either MC or MD), and $\exp[-\beta_0 E_{mu}(E; T_0)] = 1/n(E)$ are the input multicanonical

weights. The new estimates of the density of states are then used in equation 5.11 to obtain new values of E_{mu} and hence new input weights. This procedure is repeated until the histogram N_{mu} obtained from the multicanonical simulation is sufficiently flat (i.e. the probability of visiting any part of the energy space is constant). The resulting weights are then used for a long multicanonical simulation, from which thermodynamic quantities can be calculated.

To obtain expected averages from a multicanonical simulation, the single histogram reweighting technique (eq. 5.17) is first used to obtain a new estimate of the densities of states. The new estimates of densities of states are then used in equation 5.6 to obtain the thermodynamic averages.

5.2.4 Replica Exchange Multicanonical Algorithm (REMUCA)

MUCA without REM converges very slowly and consequently is inefficient.^{56–58} Therefore, we have explored the use of REMUCA, which differs from MUCA in how the starting weights for the simulation are obtained. While MUCA requires short iterative multicanonical simulations, REMUCA obtains the starting weights from a short Replica Exchange simulation, by first obtaining the densities of states from REM, which are then used to estimate the multicanonical weights $\{exp[-E_{mu}(E; T_0)/k_B T_0]\}$ with equation 5.11. In practice, the values for the multicanonical potential energy, $E_{mu}(E; T_0)$, obtained from replica exchange, are reliable only in the range of $\langle E \rangle_{T_{min}} \leq E \leq \langle E \rangle_{T_{max}}$, where T_{min} and T_{max} are the lowest and highest temperatures in REM, and $E_{min} = \langle E \rangle_{T_{min}}$ and $E_{max} = \langle E \rangle_{T_{max}}$ are the canonical expectation values at those temperatures; i.e., we use multicanonical sampling only in the region between E_{min} and E_{max} , and canonical sampling outside of this region. The reason why the weights are reliable only between E_{min}

and E_{max} is because T_{min} and T_{max} (which determine E_{min} and E_{max}) are chosen arbitrarily for the REM simulation, such that the region sampled by overlapping replicas between E_{min} and E_{max} contains both the native structure and the most probable non-native structures. Therefore, the best region sampled by REM is the one between E_{min} and E_{max} , which determines that the multicanonical input weights should be reliable only between E_{min} and E_{max} . In principle, any sampling can be used below E_{min} and above E_{max} as long as the simulation returns back to the multicanonical region which should contain both the native structure and the most probable non-native structures; in practice this calculation has been carried out with canonical sampling.

The only reason to explore the canonical region is to force a random walk from the multicanonical region, which may have wandered out of the multicanonical region, to return to the multicanonical region. In essence, by sampling for thermodynamic data only in the multicanonical region, it is being assumed that the multicanonical region is large enough to encompass both the native structure and the more probable (i.e., lower-energy) parts of the ensemble of non-native structures. In addition, at the upper (E_{max}) and lower energy (E_{min}) boundaries between the multicanonical and canonical regions, the constant probability in the multicanonical region decreases in the canonical region.

The canonical sampling is carried out by extrapolating the multicanonical energies $[E_{mu}(E, T_0)]$ linearly.⁵⁶ It should be noted that only data from the multicanonical region (between E_{min} and E_{max}) are used for calculating thermodynamic properties. Hence, the energy space in REMUCA is divided into three regions as

follows:

$$\epsilon_{mu}^0(E) \equiv \begin{cases} E_{mu}(E_{min}; T_0) + \left. \frac{\partial E_{mu}(E; T_0)}{\partial E} \right|_{E_{min}} (E - E_{min}) & \text{for } E \leq E_{min} \\ E_{mu}(E; T_0) & \text{for } E_{min} \leq E \leq E_{max} \\ E_{mu}(E_{max}; T_0) + \left. \frac{\partial E_{mu}(E; T_0)}{\partial E} \right|_{E_{max}} (E - E_{max}) & \text{for } E \geq E_{max} \end{cases}$$

where $\epsilon_{mu}^0(E)$ is substituted for $E_{mu}(E; T_0)$ in eq. 5.14 (for MC) and 5.16 (for MD), and T_0 is the reference temperature for the Monte Carlo and Molecular Dynamics simulation (the temperature at which the MC or MD simulation is carried out). Again the reference temperature bears no significance in the results of the thermodynamic quantities (because eq. 5.6 is independent of T_0). The rest of the simulation for both MC and MD proceeds as in traditional MUCA simulation (eq 5.14 for MC, and eq. 5.16 for MD) with ϵ_{mu}^0 replacing E_{mu} .

5.2.5 Multicanonical Replica-Exchange Method (MUCAREM)

We also explore the use of the REMUCAREM algorithm, whose core is the same as that of the MUCAREM algorithm. Therefore, we first present the theoretical background of MUCAREM, and later extend the discussion to REMUCAREM. Just as REM consists of several replicas of canonical MC or MD simulations, MUCAREM consists of several replicas of multicanonical simulations. The difference between REM and MUCAREM is that the replicas in REM are associated with different temperatures whereas, in MUCAREM, the replicas are associated with different energy ranges over which multicanonical simulations are carried out. The advantage of the MUCAREM approach over the traditional REM is that the probability distributions of energies of different replicas are broader in MUCAREM than in REM; therefore, a smaller number of replicas is required to cover the entire energy range.

The starting weights are obtained by short iterations of MUCA simulations, as described earlier in section 5.2.3. The following procedures are carried out in *each* cycle:

1. Select an energy range for each replica, for which the replica will carry out the MUCA simulation. This energy range of a given replica should overlap the energy ranges of the neighboring replicas, and the combined energy range from all replicas should cover the whole energy space (i.e., the combined energy range should contain the native structure and the most probable non-native structures). Assign a different random protein conformation to each energy range.
2. A MUCA simulation with MC or MD is carried out on each selected conformation within its energy range for a determined number of MC or MD steps. The MC or MD simulations are carried out with equations 5.14 or 5.16, respectively, where E_{mu} is replaced by ϵ_{mu}^m defined as follows:

$$\epsilon_{mu}^m(E) \equiv \begin{cases} E_{mu}(E_{min}^m; T_m) + \frac{\partial E_{mu}(E; T_m)}{\partial E} \Big|_{E_{min}^m} (E - E_{min}^m) & \text{for } E \leq E_{min}^m \\ E_{mu}(E; T_m) & \text{for } E_{min}^m \leq E \leq E_{max}^m \\ E_{mu}(E_{max}^m; T_m) + \frac{\partial E_{mu}(E; T_m)}{\partial E} \Big|_{E_{max}^m} (E - E_{max}^m) & \text{for } E \geq E_{max}^m \end{cases}$$

where m is the replica index ($m = min \dots max$), and min and max are the lowest and highest temperature replicas. E_{min}^m is then the canonical expectation value of the energy of the m th replica at temperature T_{min}^m [$E_{min}^m = \langle E \rangle_{T_{min}^m}$], and similarly E_{max}^m is the canonical expectation value of the energy of the m th replica at temperature T_{max}^m [$E_{max}^m = \langle E \rangle_{T_{max}^m}$] for the m th multicanonical replica. It should be noted that T_{min}^m and T_{max}^m are different for different replicas (for different m 's) and thus determine a different multicanonical energy range E_{min}^m and E_{max}^m for different replicas.

Therefore, the multicanonical simulation with each replica is carried out in a different energy range (E_{min}^m and E_{max}^m).

3. After carrying out a selected number of MC or MD steps, stop the simulation of each replica and attempt an exchange of the whole conformations between neighboring replicas with the following transition probability:

$$W(Y|X) = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases} \quad (5.18)$$

where $\Delta \equiv \beta_{m+1} \left\{ \epsilon_{mu}^{m+1}[E(Y)] - \epsilon_{mu}^{m+1}[E(X)] \right\} - \beta_m \left\{ \epsilon_{mu}^m[E(Y)] - \epsilon_{mu}^m[E(X)] \right\}$

4. Continue the simulation with each newly formed conformation at each new energy range as in step 2.
5. Iterate points 3 and 4 until the system sufficiently covers the entire energy range.

As in REM, the densities of states are obtained from self consistent evaluation of the following modified WHAM equations:

$$n(E) = \frac{\sum_{m=1}^M g_m^{-1} N_m(E)}{\sum_{m=1}^M g_m^{-1} n_m \exp(f_m - \beta_m \epsilon_{mu}^m(E))} \quad (5.19)$$

and

$$\exp(-f_m) \equiv \sum_E n(E) \exp(-\beta_m \epsilon_{mu}^m(E)) \quad (5.20)$$

where $N_m(E)$ is the histogram at temperature T_m , $\beta_m = 1/(k_b T_m)$ is the inverse temperature, n_m is the total number of samples in the m th replica, g_m is defined as in section 5.2.2. The resulting densities of states are then used to evaluate the

expectation value of any observable in equation 5.6, with g_m cancelling out, as in eq. 5.4.

5.2.6 Replica Exchange Multicanonical with Replica Exchange Method (REMUCAREM)

MUCAREM without input weights from REM converges very slowly and consequently is inefficient.⁵⁶⁻⁵⁸ Therefore we have explored the use of REMUCAREM, which, as in REMUCA, obtains the starting weights from Replica Exchange simulations as opposed to iterative short MUCA simulations. Everything else proceeds in the same manner as in MUCAREM.

5.3 Implementation Details

All the simulations were carried out on one peptide (20 residues of Alanine with free ends; ala₂₀) and two small proteins, namely the B-domain of staphylococcal protein A (an α -protein; 46 residues; 1BDD),⁵⁹ and the E. Coli MltD LysM Domain (an $\alpha+\beta$ -protein; 48 residues; 1E0G).⁶⁰ The ala₂₀ peptide was used to check whether the algorithms perform correctly, and the proteins were chosen so that basic α and $\alpha+\beta$ topologies were tested, and their size was reasonable with respect to the computational time. As in our previous work,⁶¹ the length of protein 1BDD was shortened from the original 60 residues in the PDB to 46 residues. The set of UNRES energy parameters, designated as the 4P force field⁶¹ and used in the present work, was derived by optimizing the parameters for four proteins simultaneously: 1E0L⁶² (a β -protein ; 37 residues), 1E0G⁶⁰ (an $\alpha + \beta$ protein; 48 residues), 1IGD⁶³ (an $\alpha + \beta$ protein; 61 residues) and 1GAB⁶⁴ (an α -protein; 53 residues).

The Monte Carlo (MC) simulations with REM, REMUCA and REMUCAREM were carried out as follows. All four UNRES angles in every residue of the protein were subjected to a perturbation. One MC sweep consisted of updating all of these angles for each residue in the sequence, with a Metropolis evaluation after each perturbation. The Molecular Dynamics (MD) simulations with these same algorithms were carried out with the Berendsen thermostat,⁶⁵ using the velocity Verlet algorithm⁶⁶ with variable time step to integrate the equations of motion. The variable time step was accomplished by scaling the time step δt by powers of 2.¹⁶ The cutoff change of acceleration δa_{cut} for the scaling procedure was increased to $\delta a_{cut} = 4 \text{ \AA}/\text{mtu}$,¹⁶ to allow for the multiplication of the forces in the modified Newton equation (in eq. 5.16, MUCA MD utilizes a factor that multiplies the forces, i.e., accelerations, which would cause the maximum change of acceleration δa_{max} to exceed the cutoff value δa_{cut} , and thus the time step would be unnecessarily reduced). The time step was set at 4.89 fs to yield stable trajectories.¹⁶ However, this is only a formal time step and, because of the reduction of the number of degrees of freedom in UNRES, the time step is several times larger compared with all-atom MD (see reference 16 for details). The coupling constant to the thermal bath was increased to 0.2445 ps to overcome the limitation of the Berendsen thermostat and produce a more Boltzmann-like distribution.¹⁷ Replica Exchange MD was carried out using multiplexing,⁶⁷ in which several replicas were simulated at each temperature. Since MC lacks the gradient and is consequently much less efficient at exploring the energy space than MD, the temperature range in the MC version of REM was lower than that of the REM MD simulations (so that the low-temperature replicas in REM MC would involve a sufficient number of moves to explore the low energy basins), and the number of replicas and the frequency of exchange in REM was much higher in MC. In all the simulations (both MC and

MD), the system was equilibrated for 20% of the simulation length, and the last 80% of the simulation was used for the calculations. All Monte Carlo simulations were started from random conformations, and the starting point for all molecular dynamics simulations was an extended chain; because the system was equilibrated and, because REM uses high-temperature replicas, and both REMUCA and REMUCAREM perform a random walk in the energy space, the simulations were independent of the starting conditions.

5.4 Results and Discussion

5.4.1 Poly-L-alanine

First, to test the algorithms, a very simple poly-L-alanine system (20 residues) was chosen, and REM, REMUCA, and REMUCAREM simulations were carried out with both MC and MD. The parameters used in all simulations for ala₂₀ are shown in Table 5.1. REM simulations were carried out first, from which the densities of states were obtained. It was found that the densities of states obtained from REM simulations were not precise enough for REMUCA, because REMUCA simulations did not perform a random walk (i.e. did not have flat energy histograms). Therefore, after the first iteration of REMUCA simulations, the densities of states were reweighted with eq. 5.17 and, with these weights, a second iteration of REMUCA simulations was carried out. The second set of weights used for REMUCA were also used for REMUCAREM simulations. The simulation weights for alanine are shown as a solid or dashed curve in Figure 5.1. The dashed line shows an example of the multicanonical energy function (eq. 5.11), used in the modified Metropolis criterion in MC simulations (eq. 5.14), while the solid line shows its derivative, a

Table 5.1: Parameters used in ala₂₀ simulations. Replicas column shows the number of replicas used for each simulation. Temp shows the reference temperature (K) or range of temperatures for simulations (for REMUCA MC and REMUCAREM MC, the reference temperature cancels out in the equations; therefore, the corresponding fields are empty). Step is the number of UNRES MD time steps, where the maximum time step was set to 4.9 fs in all MD simulations. A sweep is defined as perturbing all four angles at all the positions along the peptide sequence (for ala₂₀, sweep is equal to 80 energy evaluations).

Simulation	Replicas	Temp	Steps/Sweeps
REM MD	16	400-2000	16,000,000
REMUCA MD	1	100	10,000,000
REMUCAREM MD	2	100,101	20,000,000
REM MC	30	100-2000	2,000,000
REMUCA MC	1		1,000,000
REMUCAREM MC	2		1,000,000

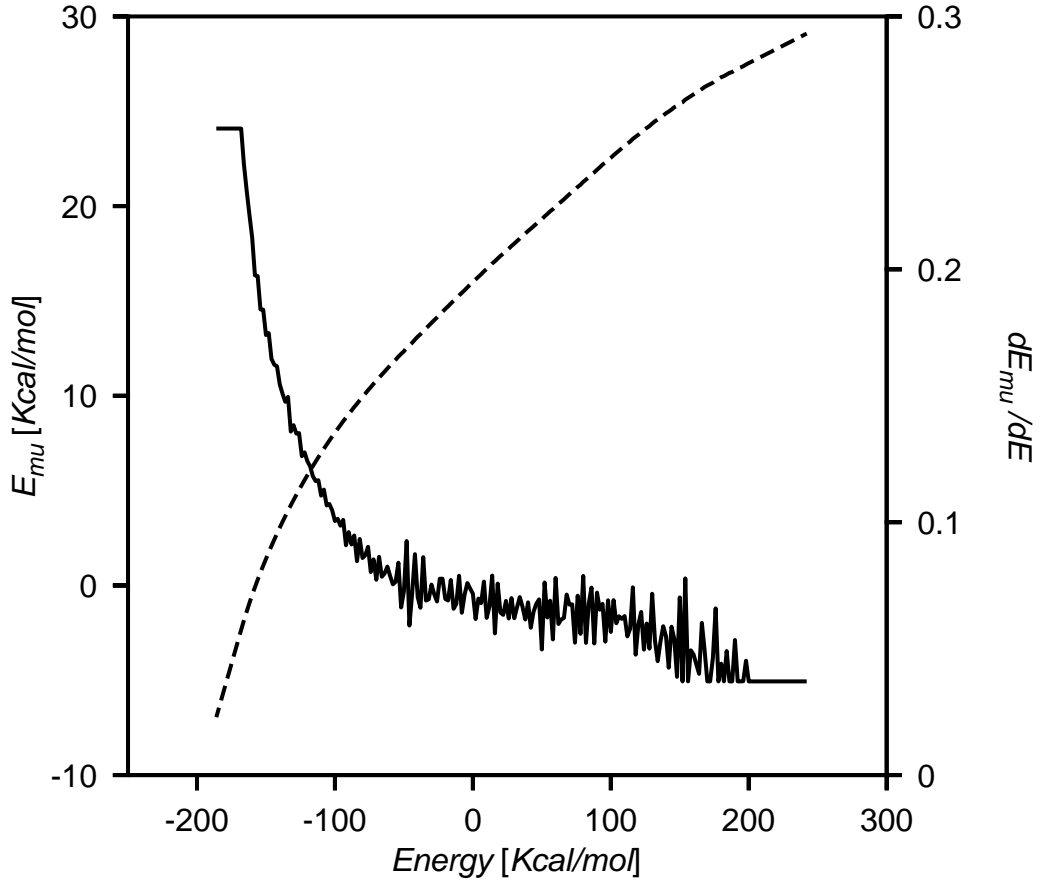


Figure 5.1: The parameters used for multicanonical simulations. The dashed line denotes the multicanonical energy function (eq. 5.11), while the solid line denotes the derivative of this function fitted with cubic splines. The derivatives are used as a multiplicative factor $[\partial E_{mu}(E; T_0)/\partial E]$ in the modified Newton equation (eq. 5.15) in molecular dynamics. The flat regions of the derivative curve show where the multicanonical simulation changes to the canonical simulation.

factor multiplying the force in the modified Newton equation (eq. 5.16).

The results are summarized in Figures 5.2 and 5.3. Figure 5.2 consists of six plots. Three plots on the top correspond to MC simulations, whereas the three plots on the bottom correspond to MD simulations. The two plots in each column are for REM, REMUCA, and REMUCAREM simulations, respectively. Each plot depicts the logarithm of the probabilities $\ln[P(E)]$ as a function of energy (E) for the given simulation. By comparing the top row to the bottom row, it can be seen

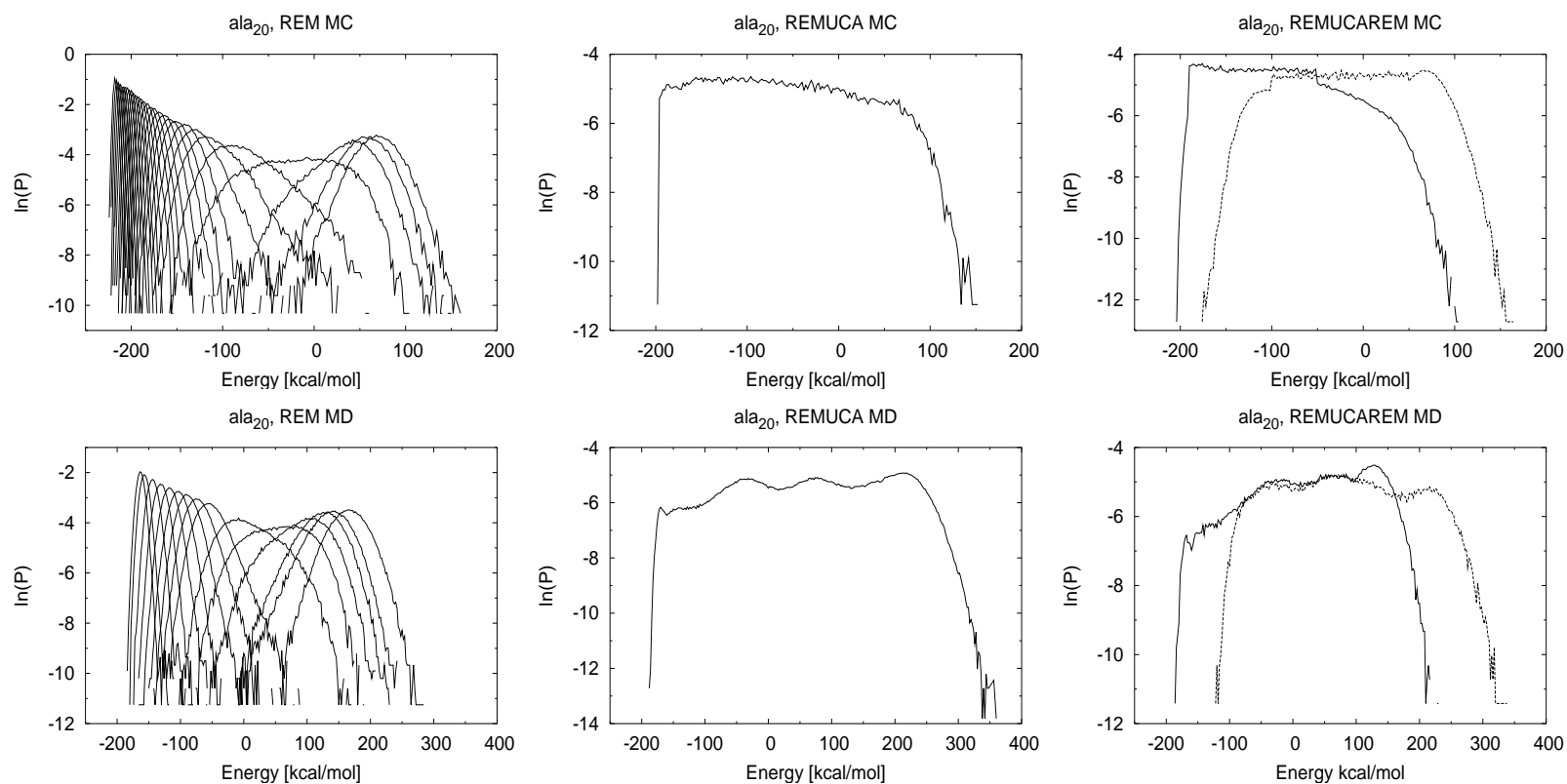


Figure 5.2: Histogram curves for simulations with alanine. The plots depict the logarithm of the probabilities as a function of energy. The top-row plots are from MC simulations (REM, REMUCA, REMUCAREM, from left to right respectively). The bottom-row plots are from MD simulations. For REM, and REMUCAREM (left, and right columns) each curve corresponds to an individual replica at a different temperature (for REM) or different energy range (for REMUCAREM); see Table 5.1 for the number of such replicas.

that MC simulations cover a smaller energy range than their MD counterparts. This is due to the fact that the MD energy function contains the extra vibration term (eq. 5.1) adding to the energy range for MD simulations. It is evident from the plots that REMUCA MC and REMUCAREM MC are flatter $\{\text{constant } \ln[P(E)]\}$ than REMUCA MD and REMUCAREM MD. This discrepancy probably arises from the fact that the MD versions of multicanonical simulations utilize the derivative of the multicanonical energy function (eq. 5.16), whereas the MC simulations use only the multicanonical energy function itself (eq. 5.14, Fig. 5.1). As mentioned in the Methods section, the derivatives are fitted using cubic splines, which can cause problems if the entropy function is not smooth (the derivative will be rough, which will cause numerical instabilities in the integration of eq. 5.16).

By comparing the plots for REM MC and REMUCA MC, it can be seen that REMUCA MC does not cover the entire low-energy region, but rather stops before -200 kcal/mol. This is because we shifted the low-energy boundary for multicanonical sampling up from the canonical average evaluated by the lowest temperature replica. The reason for doing this is that, when the boundary was lower in energy, the MC multicanonical simulations would walk in the entire energy range until they encountered the low-energy region, at which point the simulations would become trapped in deep local minima out of which they did not escape for the remainder of the simulation (data not shown). This issue was easily resolved for ala₂₀ MC simulations by simply raising the low-energy boundary, but the issue reappears during both MC and MD simulations with 1BDD and 1E0G, and is discussed further when describing the results for 1BDD and 1E0G.

Figure 5.3 also shows two rows of plots, one for MC and one for MD simulations. The first column corresponds to simulations with poly-L-alanine. Each plot con-

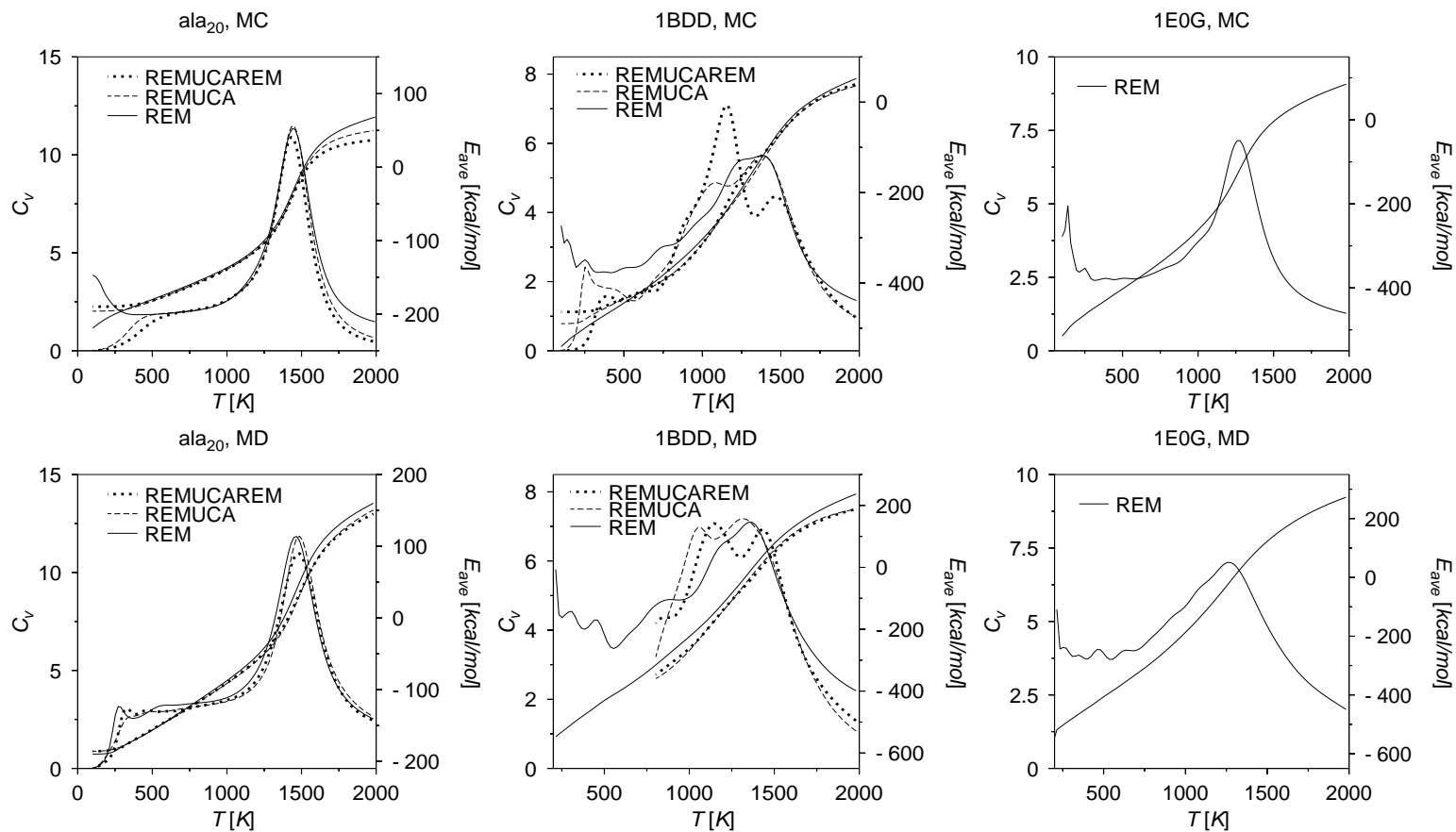


Figure 5.3: Thermodynamic quantities calculated by various methods for ala₂₀, 1BDD and 1E0G Heat capacity as well as average energy as a function of temperature for REM (solid line), REMUCA (dashed line) and REMUCAREM (dotted line) simulations with MC (top row) and MD (bottom row). The columns correspond to ala₂₀, 1BDD, and 1E0G, from left to right, respectively. Good agreement for all three simulations for both MC and MD versions can be observed for ala₂₀; some overlap is observed for 1BDD, and only REM results (see text) are shown for 1E0G.

sists of two graphs, one is the heat capacity, and the other is the average energy as a function of temperature. Each graph contains three curves, each corresponding to REM, REMUCA and REMUCAREM simulations, respectively. The average energy was calculated with eq. 5.6, and the heat capacity was evaluated according to the following formula:

$$C_V = \beta^2 \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{N} \quad (5.21)$$

For both MC and MD simulations with ala₂₀, all the curves overlap, suggesting that the simulations converged to the same distribution. The main peak of the specific heat curve indicates the temperature of the peptide collapse. For a simple system such as ala₂₀, the collapse occurs simultaneously with folding to the native α -helical state. This temperature appears to be 1400 K for MC and 1500 K for MD. It is important to note that the UNRES temperature has no relevance to the experimental temperature because UNRES is a coarse-grained potential in which the non-essential degrees of freedom have been averaged out, and energy parameter optimization was carried out with a hierarchical procedure⁶⁸ to provide the steepest decrease of energy with increasing native likeness⁶⁹ while ignoring the correspondence between the simulated and experimental thermodynamic characteristics of folding. Moreover, the decoy sets were generated using the CSA method which walks only in the space of local minima, thus violating the detailed balance condition. As mentioned further in the Conclusions section, we are currently revising our hierarchical force field optimization procedure,⁶⁹ to introduce entropy using methods applied in the present work, and consequently to capture as much physics as possible.

5.4.2 1BDD

We repeated the same procedure for 1BDD as for ala₂₀. The parameters used for the simulations with 1BDD are described in Table 5.2. Similarly, as for ala₂₀, the results for 1BDD are shown in Figure 5.4. First, since 1BDD has more degrees of freedom than ala₂₀, we used a larger number of replicas in both REM MC and REM MD algorithms and, in REM MD, we additionally multiplexed each replica to have more trajectories from which to sample. Although it might appear that, by using more replicas, REM would perform much better than both REMUCA and REMUCAREM, the advantage of REMUCAREM (as mentioned in section 5.2.5) is that a smaller number of replicas is required to cover the entire energy range. To provide a fair comparison, we used the same number of steps for both REM and REMUCAREM (see Table 5.2); although many more steps were used in REMUCAREM than in REMUCA, the results with REMUCAREM are not substantially improved over those with REMUCA, as discussed later in this section. As for poly-L-alanine, the density of states from the Replica Exchange simulations was insufficient to carry out a random walk with REMUCA and REMUCAREM; therefore, the densities of states were reweighted. The multicanonical histogram curves in Figure 5.4 correspond to one iteration of reweighting. Additionally, we encountered a trapping problem in the low-energy region for both MC and MD simulations. As for ala₂₀, we increased the low multicanonical energy boundary to escape the trapping regions (Fig 5.4 shows that REMUCA and REMUCAREM MC and MD do not sample all the way to the lowest energy; i.e. not beyond -500 kcal/mol). To verify whether moving the multicanonical energy boundary is acceptable, we show the RMSD results in Figure 5.5. The left column shows the energy versus RMSD profile for Replica Exchange simulations. As can be seen

Table 5.2: Parameters used in 1BDD and 1E0G simulations. Replicas column shows the number of replicas used for each simulation. Temp shows the reference temperature (K) or range of temperatures for simulations (for REMUCA MC and REMUCAREM MC, the reference temperature cancels out in the equations; therefore, the corresponding fields are empty). Step is the number of UNRES MD time steps, where the maximum time step was set to 4.9 fs in all MD simulations. A sweep is defined as 192 and 184 energy evaluations (4 angles for each residue in the chain) for 1BDD and 1E0G, respectively. (b) Multiplexed replicas. 30(x4) means that 4 replicas for each temperature (with 30 temperatures) were simulated.

Protein	Simulation	Replicas	Temp	Steps/Sweeps
1BDD	REM MD	30(x4) ^b	200-1800	240,000,000
	REMUCA MD	1	50	20,000,000
	REMUCAREM MD	8	50- 400	240,000,000
	REM MC	50	50-1800	10,000,000
	REMUCA MC	1		1,000,000
	REMUCAREM MC	2		1,000,000
1E0G	REM MD	30(x4) ^b	200-1800	240,000,000
	REM MC	50	50-2000	10,000,000

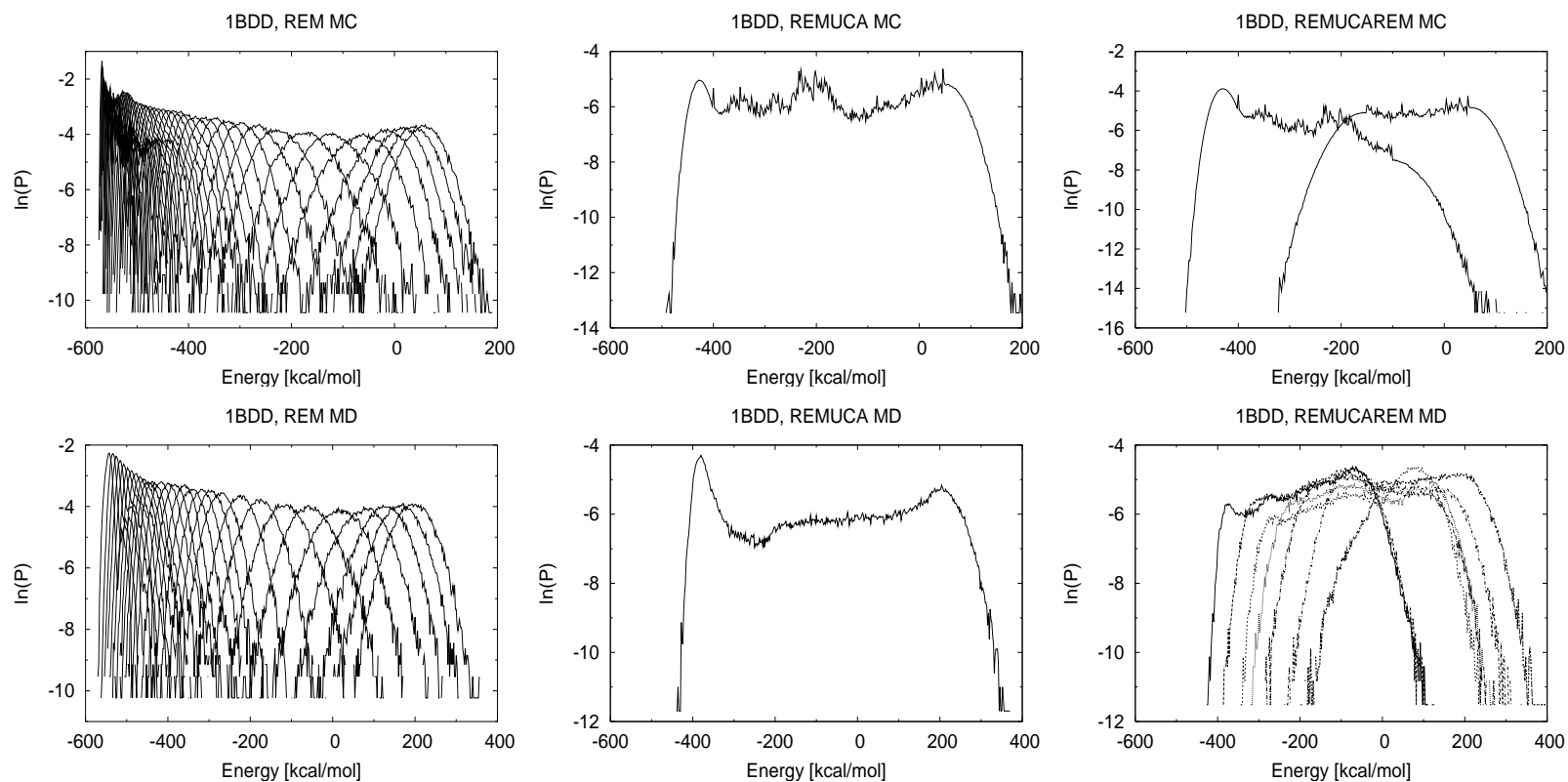


Figure 5.4: Histogram curves for simulations with 1BDD. The plots depict the logarithm of the probabilities as a function of energy. The top-row plots are from MC simulations (REM, REMUCA, REMUCAREM, from left to right respectively). The bottom-row plots are from MD simulations. For REM, and REMUCAREM (left, and right columns) each curve corresponds to an individual replica at a different temperature (for REM) or different energy range (for REMUCAREM); see Table 5.2 for the number of such replicas.

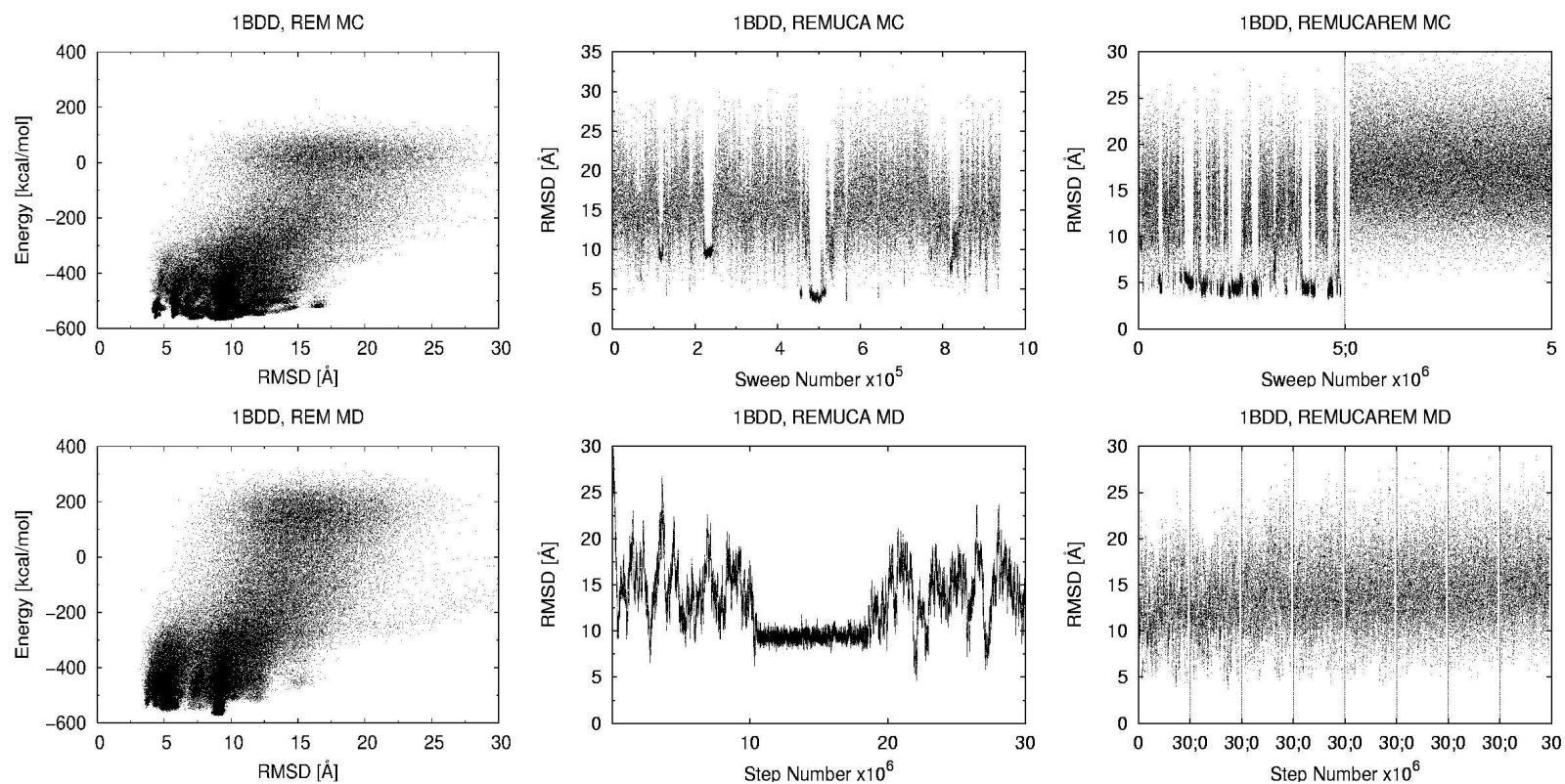


Figure 5.5: Simulation results for 1BDD. The top-row plots are from MC simulations (REM, REMUCA, REMUCAREM, from left to right, respectively). The bottom-row plots are from MD simulations. The left-column shows Energy versus RMSD coverage of the energy space. The middle-column shows the random walk of the REMUCA simulations, and the right-column shows the random walk for all REMUCAREM replicas (one after another).

from this column, both REM MC and REM MD cover a wide conformational space, which includes the native structure (centered ~ 4.5 Å for REM MC, and ~ 4.0 Å for REM MD). The middle and the right columns show an RMSD trajectory for REMUCA and REMUCAREM simulations, respectively. It can be seen that the system folds and unfolds several times over the course of the run, i.e., attains the low-RMSD region. Even though the multicanonical simulation should perform a random walk in the energy space, it is more important that the simulation fully samples the conformational space, which can be observed in both the REMUCA and REMUCAREM RMSD trajectories.

The middle column of Figure 5.3 shows the calculated heat capacities and average energies for both MC and MD REM simulations with 1BDD. By contrast to the simulations with poly-L-alanine, 1BDD heat capacities have broad irregular peaks. The irregular peak is an overlap of two peaks, one corresponding to a collapse to a more compact state but without the final folding, and one corresponding to a transition to the native state, as will be shown later in Figure 5.7. For 1BDD, REM, REMUCA and REMUCAREM peaks do not coincide as they do for poly-L-alanine. The fact that all simulations differ in the shape of their heat capacity curve suggests that all simulations have not converged to the same distribution. The reason why the REMUCA and REMUCAREM curves do not cover the whole temperature range is that the multicanonical region was restricted to avoid trapping (i.e., the low multicanonical energy boundary was increased).

5.4.3 1E0G

Finally, for 1E0G, Replica Exchange successfully sampled the energy space, and produced reasonable statistics for thermodynamic quantities (Fig. 5.6). The left

column of Figure 5.6 shows the histograms for Replica Exchange simulations with both MC (top) and MD (bottom). The middle column depicts plots of energy as a function of RMSD from the experimental structure, showing that the simulations cover an extended portion of the energy space. It can be seen that the REM MD simulation reaches the native state within an RMSD of around 4.5 Å and has low energy, whereas the REM MC simulation barely touches 5 Å RMSD, without reaching the low-energy region, which suggests incomplete N- and C-terminal β -strand contacts (correct β -strand packing provides a large contribution to decreasing the energy of the native structure, and is necessary for the RMSD to be below 5 Å). For multicanonical simulations (REMUCA and REMUCAREM), we were unable to obtain proper multicanonical weights, which would enable the system to carry out a random walk in the energy space. Even after several iterations of reweighting, the system would walk towards the low energy states, where it would stay for the remainder of the simulation. This behavior is shown in the right column of Figure 5.6, where a REMUCAREM simulation is shown for MC and a REMUCA simulation for MD. For REMUCAREM MC, it is evident that the lower energy replica (replica 1) reaches low energies and remains trapped in a low-energy region, whereas the high energy replica (replica 2) carries out a random walk. A similar behavior is observed for MD simulations (trapping of REMUCA MD is shown in Figure 5.6). This observation is similar to that from a study carried out by Bhattacharya and Sethna, who showed that, in the case of glassy systems, even multicanonical simulations have problems carrying out a random walk, and instead become trapped in metastable states.⁷⁰ They implemented the Entropy Sampling version of the algorithm with Lennard-Jones glasses, and observed that simulations that have dynamic updating of the microcanonical entropy

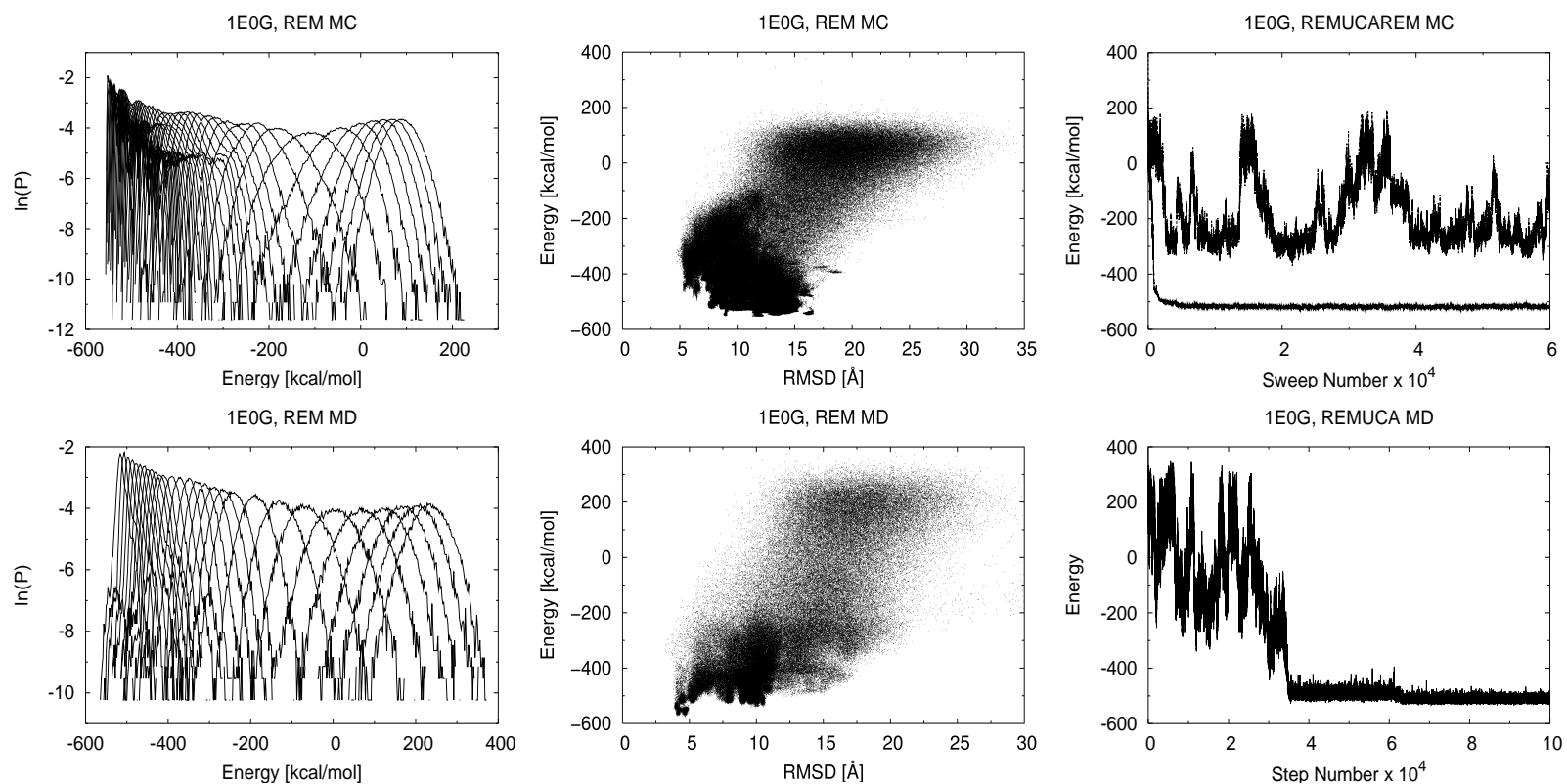


Figure 5.6: Simulation results for 1E0G. The top-row plots are from MC simulations, whereas the bottom-row plots are from MD simulations. The left-column shows the histogram curves for REM. Each curve corresponds to an individual replica at a different temperature. The middle column shows energy versus RMSD coverage of the energy space. The right-column shows energies at a series of steps of REMUCAREM for MC (top, with two replicas), and REMUCA for MD (bottom).

function perform a random walk in the energy space, while the simulations with fixed weights (precomputed by iterative procedures) became trapped in metastable states. The dynamic updating of the weights (i.e., eq. 5 of ref. 36) is essentially a single histogram reweighting on the fly with the difference that not all regions might be visited, and typically the time between updates is much shorter. Dynamic updating ensures that the system does not remain in the same conformation for a long time. However, it also introduces discontinuities, and negative gradients into the E_{mu} function, which poses problems for the MD version of the REMUCA algorithm, with MD being more sensitive to the input weights because of its use of derivatives. The dynamic updating procedure pushes the system out of trapped states, but this violates the detailed balance condition, and thus no longer guarantees convergence to the proper distribution or correct estimates of thermodynamic quantities. Because of the trapping problem, we did not calculate average energies and heat capacities from both REMUCA and REMUCAREM simulations for 1E0G (see Fig. 5.3).

The third column of Figure 5.3 shows the calculated heat capacities and average energies for both MC and MD REM simulations with 1E0G. A sharp single peak for the heat capacity is observed for REM MC whereas a broader peak is observed for REM MD simulations, and in both cases it is centered at around 1270 K. As mentioned above (energy vs. RMSD plot in Fig. 5.6), the REM MC simulation does not quite sample the native region. This observation, and the fact that the heat capacity for REM MC has a sharper peak, suggests that REM MC predicts a collapse to a more compact state but without the final folding (i.e., there is no low-energy structure below 5 Å RMSD as shown in the energy vs. RMSD plot in Fig. 5.6). On the other hand, the statistics from REM MD contains the native

region (shown in the energy vs. RMSD plot in Fig. 5.6) and thus incorporates the contribution of the native region to the thermodynamic quantities. The collapse to a more compact structure and final folding do not seem to coincide (see the upcoming discussion about Fig. 5.7), which broadens the heat capacity curve. For MC, the sharp peak is centered at 1270 K (Fig. 5.3) which corresponds roughly to -130 kcal/mol of average energy. From the energy vs. RMSD plot in Figure 5.6, it can be seen that the highest allowed energy for the collapsed structure (RMSD ~ 5 Å) is also around -130 kcal/mol. Folding to the native state for MD occurs at lower energies, which broadens its heat capacity peak (see the discussion about Fig. 5.7 in section 5.4.4)

5.4.4 Free energy diagrams

From our tests on ala₂₀, 1BDD, and 1E0G, we conclude that Replica Exchange Molecular dynamics is the most efficient method for sampling and calculating thermodynamic quantities with a rugged energy landscape such as the 4P force field, applied to larger systems. Since the free energy is the most important quantity for the description of equilibrium properties of proteins, we used REM MD to calculate free energy profiles for ala₂₀, 1BDD, and 1E0G. For this purpose, we used the densities of states obtained from the multi-histogram analysis (eq. 5.4). From the densities of states, we calculated the microcanonical entropy, $S(E_i) = k_B \ln [n(E_i)]$, for all conformations collected from the simulations, and used it to compute the microcanonical free energies with the following expression: $F(E_i, T) = E_i - TS(E_i)$. To plot the restricted canonical free energy as a function of RMSD (r) and radius of gyration (ρ), we calculated the restricted canonical free energy by evaluating

the following expression for each grid point:

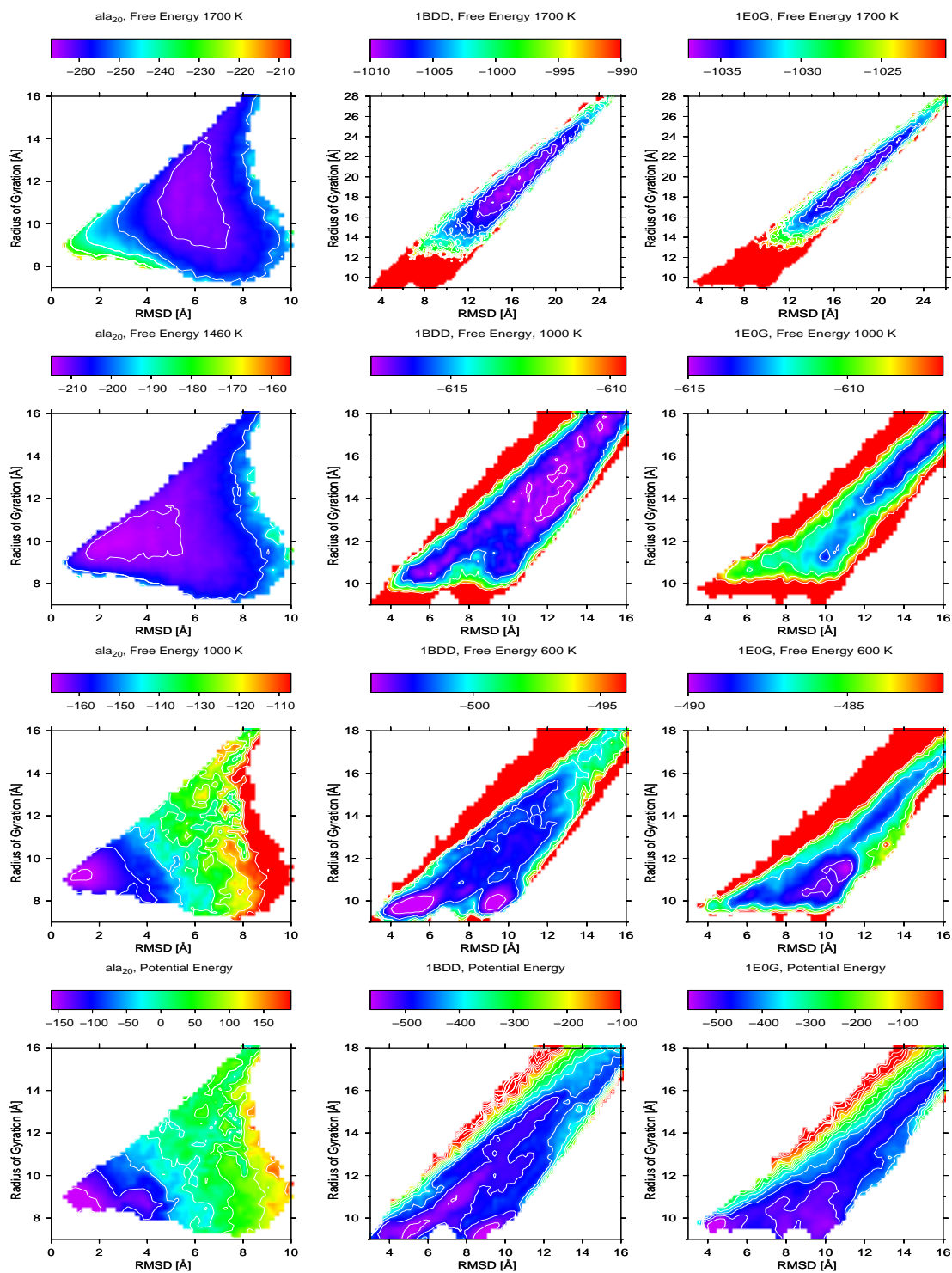
$$F(r, \rho, T) = -k_B T \ln \sum_{E_i \in N(r, \rho)} \exp\left(\frac{-F(E_i, T)}{k_B T}\right) \quad (5.22)$$

where the index i enumerates conformations within the histogram bins, $N(r, \rho)$, for given ranges of RMSD and radius of gyration.

Figure 5.7 shows the restricted canonical free energy plots as a function of RMSD and radius of gyration for various temperatures. Each column corresponds to simulations with ala₂₀, 1BDD and 1E0G, from left to right, respectively. The temperatures are chosen so that the highest temperature is higher than that of the heat capacity peak (first row), within the peak (second row), below the peak (third row), and at zero K (fourth row) from top to bottom, respectively.

The highest temperature free energy plot for ala₂₀ shows that, at this temperature, the peptide is preferentially completely unfolded, as indicated by the high RMSD (greater than 5 Å) and the high radius of gyration (greater than 9 Å), whereas at the heat capacity peak temperature (1460 K) the lowest free energy region connects both the native and the non-native basins (RMSD between 2 and 5 Å). For 1000 K, the free energy surface already appears very similar to the free energy surface at 0K, which represents the potential energy surface. The native state (RMSD lower than 2 Å) is the lowest free energy at this temperature, confirming our observation from the heat capacity curve. It should be noted that the range of energies observed in the potential energy plot is much larger than the range observed with non-zero temperatures, showing that the search for the native state is very much facilitated in the restricted canonical free energy surface. In other words, the restricted canonical free energy differences do not need to be very large in order to pass from the unfolded to the folded state, whereas large potential energy barriers must be crossed to pass from the unfolded to the folded state in

Figure 5.7: Free energy (in kcal/mol, indicated by the colored bars at the top of each graph) as a function of RMSD and radius of gyration for various temperatures. The free energy surfaces were calculated from the REMD simulations (see text). The columns correspond to simulations with ala₂₀, 1BDD, 1E0G, from left to right, respectively. The temperatures are chosen so that the highest temperature is higher than that of the heat capacity peak (first row), within the peak (second row), below the peak (third row), and at zero K (fourth row) for comparison.



the potential energy surface. For ala₂₀, we conclude that, even though the force field was optimized without any thermodynamics, we still observe a correct folding behavior.

For protein A (1BDD), the restricted canonical free energy plots look similar to the plots for ala₂₀. At high temperature, unpacked, open structures with high RMSD and radius of gyration are observed. At 1000 K, the low free energy region connects unfolded non-native states with compact states (both native and non-native). At much lower temperature (600 K), the lowest free energy regions belong to the native basin (centered around 5 Å RMSD) and to the mirror image (centered around 9 Å RMSD). It should be noted that, for ala₂₀, the native region had the lowest free energy at 1000 K whereas, for 1BDD, the temperature had to be lowered to 600 K for this to occur. Finally, the potential energy plot is again similar to the low-temperature free energy plot, but has a much larger energy range. It should be noted that, at 600 K, the free energy has well defined regions of low free energy whereas, for the potential energy, the native state is more evenly connected with compact but non-native states, which has been observed previously in MD studies with protein A in our laboratory (all 10 simulations successfully folded protein A with the 4P force field at 800 K).¹⁸

For 1E0G, the high-temperature plot again shows a preference for unfolded structures. For 1000 K, the compact structures are not quite preferential in free energy. From previous MD work with 1E0G in our laboratory,¹⁸ it was found that the successful folding trajectory starts with formation of non-interacting helical structures, which then collapse to a native HTH motif (15 Å RMSD) and finally to one with 3.9 Å RMSD from the experimental structure. The HTH motif structures appear to be preferable in terms of free energy at 1000 K, which is still within the

broad peak of the heat capacity for 1E0G. For low temperature, such as 600 K, the low free energy region connects the HTH motif to compact native-like structures without β -strand contacts (around 6 Å RMSD). However, the fully formed native structure (centered at 4.5 Å RMSD) is at higher free energy, and it appears at the lowest free energy region only at very low temperatures (where the free energy plot is similar to the potential energy plot). Liwo et al observed that only 6 out of 10 canonical MD simulations at 800 K yielded native-like structures.¹⁸ Our free energy calculations show that the lowest free energy corresponds to non-native compact structures (i.e., with low radius of gyration, but high RMSD); however the native structures (with RMSD less than 5 Å) have slightly higher free energy. Therefore the non-native conformations are more probable, but the native structures still have a finite probability to occur. Thus, our free energy calculations agree with the results obtained by Liwo et al.

Since the temperature must be extremely low in order for the native state to be the global minimum of the free energy, the entropy contribution is much larger than that for the same temperature in protein A and ala₂₀. A larger contribution from entropy means more accessible conformations for a given temperature. Therefore, the multicanonical simulations have to sample a larger number of accessible conformations, which becomes difficult for 1E0G.

From Fig. 5.7, it can be seen that, for a simple system such as ala₂₀, the collapse occurs simultaneously (at 1460 K) with folding to the native α -helical state (RMSD values and radii of gyration for low free energy regions decrease simultaneously with temperature from 1700 K to 1460 K to 1000 K). For protein A and 1E0G, the low free energy region at 1000 K extends all the way to the low radius of gyration and high RMSD values. For protein A, two low free energy

regions remain as the temperature is decreased to 600 K, one being the native, and one being the mirror image. For 1E0G, the low free energy region at 600 K with low radius of gyration but high RMSD appears first and, as the temperature is lowered (not shown here), the native region becomes the lowest free energy basin. However, this occurs at very low temperatures, as described above. This explains why the heat capacity peaks for both protein A, and 1E0G are broad and irregular. The two main events, collapse, and folding to the native state, occur at different temperatures.

5.5 Conclusions

In the present work, we implemented REM, REMUCA and REMUCAREM algorithms with the UNRES force field, utilizing Monte Carlo and Molecular Dynamics techniques. First, we tested all the algorithms on a simple poly-L-alanine system. For both the MC and MD algorithms, we obtained good agreement for heat capacity and average energy curves, which shows that all the simulations converged to the same distribution, and that our implementation works as expected.

Next, we applied the simulations to two proteins, namely to 1BDD and 1E0G. First, the 1BDD simulations performed reasonably well. The best performance was observed for the Replica Exchange algorithm in both the MC and MD simulations, since REM appeared to be much less sensitive to the input parameters (the only important parameter is the distribution of temperatures). In order to carry out a random walk, REMUCA and REMUCAREM depend on a proper estimation of the input weights and, as for ala₂₀, both REMUCA and REMUCAREM simulations had to be reweighted in order to obtain reasonably flat histograms. A trapping problem occurred at low energies, which was alleviated by raising the lower energy

boundary for multicanonical simulations. However, by excluding a certain energy region from being sampled, the agreement among the heat capacity curves for all simulations was not so good.

Since 1E0G has a more complicated fold than 1BDD, multicanonical simulations broke down, and only Replica Exchange simulations were capable of exploring the energy region and computing the thermodynamic averages. This observation agrees with that from the study by Aleksenko et al⁷¹ who concluded that the generalized ensemble approach is a useful study tool for proteins up to 30-40 residues with simple topology such as the α -helix. Furthermore, since the MD version of REMUCA and REMUCAREM use the derivative of the entropy function, MD multicanonical simulations are even more sensitive than their MC counterparts; therefore, they are more difficult to implement. Conversely, MD is much more capable of exploring the energy landscape than MC; hence, MD simulations are much more useful for larger systems.

Finally, we analyzed data from our REM MD simulations for all three test systems, and calculated free energy maps as a function of RMSD and radius of gyration. The free energy calculations show the correct folding behavior for poly-L-alanine and protein A while, for 1E0G, the native structure had the lowest free energy only at very low temperatures; hence, the entropy contribution is much larger than that for the same temperature in protein A and ala₂₀. The larger contribution from entropy means more accessible conformations for a given temperature. For the same temperature, ala₂₀ has the smallest entropy contribution, followed by protein A, and then by 1E0G.

Although both REMUCA and REMUCAREM seem to have potential as sampling methods applied to smaller systems, Replica Exchange utilizing MD, coupled

with multiplexing, appears to offer more insight into the behavior of protein folding for more complicated systems with a rough energy landscape. Moreover, since Replica Exchange is easy to implement and has few parameters to adjust, it is very suitable for implementation in the future revision of our hierarchical optimization procedure,⁶⁹ which is currently under development in our laboratory. The new optimization procedure is based on a hierarchical design of the potential-energy landscape such that the energy decrease follows the increase of native-likeness⁶⁸ and utilizes MD as a sampling method to capture as much physics as possible. Preliminary tests (unpublished data) show that Replica Exchange together with Umbrella Sampling⁷² (introduced when the native region is not sufficiently covered with the initial parameter set) covers a broader region of conformational space, and thus produces better statistics for hierarchical optimization. Consequently, this will allow us to produce a coarse-grained force field suitable for Molecular Dynamics simulations, which will be capable of more accurate evaluation of thermodynamic quantities.

BIBLIOGRAPHY FOR CHAPTER 5

- [1] Metropolis, N.; Ulam, S., *J. Am. Stat. Assoc.* 1949, 44, 335.
- [2] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., *J. Chem. Phys.* 1953, 21, 1087.
- [3] Li, Z.; Scheraga, H. A., *Proc. Natl. Acad. Sci., U. S. A.* 1987, 84, 6611.
- [4] Li, Z.; Scheraga, H. A., *J. Molec. Str. (Theochem)* 1988, 179, 333.
- [5] Ripoll, D. R.; Scheraga, H. A., *Biopolymers* 1988, 27, 1283.
- [6] Ripoll, D. R.; Scheraga, H. A., *J. Protein Chem.* 1989, 8, 263.
- [7] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.
- [8] Nancias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H.A., *J. Comp. Chem.* 2005, 26, 1472.
- [9] Piela, L.; Kostrowicki, J.; Scheraga, H. A., *J. Phys. Chem.* 1989, 93, 3339.
- [10] Pillardy, J.; Olszewski, K. A.; Piela, L., *J. Phys. Chem.* 1992, 96, 4337.
- [11] Pillardy, J.; Liwo, A.; Groth, M.; Scheraga, H. A., *J. Phys. Chem. B* 1999, 103, 7353.
- [12] Pillardy, J.; Liwo, A.; Scheraga, H.A., *J. Phys. Chem. A* 1999, 103, 9370.
- [13] Lee, J.; Scheraga, H. A.; Rackovsky, S., *J. Comput. Chem.* 1997, 18, 1222.
- [14] Lee, J.; Scheraga, H. A., *Int. J. Quant. Chem.* 1999, 75, 255.
- [15] Czaplewski, C.; Liwo, A.; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [16] Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H.A., *J. Phys. Chem. B* 2005, 109, 13785.
- [17] Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H.A., *J. Phys. Chem. B* 2005, 109, 13798.
- [18] Liwo, A.; Khalili, M.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 2362.
- [19] Khalili, M.; Liwo, A.; Scheraga, H.A., *J. Mol. Biol.*, in press.
- [20] Hukushima, K.; Nemoto, K., *J. Phys. Soc. Jpn.* 1996, 65, 1604.

- [21] Hansmann, U. H. E., *Chem. Phys. Lett.* 1997, 281, 140.
- [22] Swendsen, R. H.; Wang, J. S., *Phys. Rev. Lett.* 1986, 57, 2607.
- [23] Geyer, C.J., *Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface*, American Statistical Association, New York 1991.
- [24] Gront, D.; Kolinski, A.; Skolnick, J., *J. Chem. Phys.* 2001, 115, 1569.
- [25] Kolinski, A.; Gront, D.; Pokarowski, P.; Skolnick, J., *Biopolymers* 2003, 69, 399.
- [26] Fenwick, M. K.; Escobedo, F. A., *J. Chem. Phys.* 2003, 119, 11998.
- [27] Romiszowski, P.; Sikorski, A., *Physica A* 2004, 336, 187.
- [28] Sugita, Y.; Okamoto, Y., *Chem. Phys. Lett.* 1999, 314, 141.
- [29] Zhou, R.H.; Berne, B.J.; Germain, R., *Proc. Natl. Acad. Sci. USA* 2001, 98, 14931.
- [30] Sanbonmatsu, K.Y.; Garcia, A.E., *Proteins* 2002, 46, 225.
- [31] Garcia, A.E.; Onuchic, J.N., *Proc. Natl. Acad. Sci. USA* 2003, 100, 13898.
- [32] Lin, C.-Y.; Hu, C.-K.; Hansmann, U. H. E., *Proteins: Struct., Funct., Genet.* 2003, 52, 436.
- [33] Berg, B. A.; Neuhaus, T., *Phys. Lett. B* 1991, 267, 249.
- [34] Berg, B. A.; Neuhaus, T., *Phys. Rev. Lett.* 1992, 68, 9.
- [35] Lee, J., *Phys. Rev. Lett.* 1993, 71, 211.
- [36] Hao, M.A.; Scheraga, H.A., *J. Phys. Chem.* 1994, 98, 4940.
- [37] Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N., *J. Chem. Phys.* 1992, 96, 1776.
- [38] Marinari, E.; Parisi, G., *Europhys. Lett.* 1992, 19, 451.
- [39] Gront, D.; Kolinski, A.; Skolnick, J., *J. Chem. Phys.* 2000, 113, 5065.
- [40] Hansmann, U.H.E., *Phys. Rev. E* 1997, 56, 6200.
- [41] Mitsutake, A.; Sugita, Y.; Okamoto, Y., *J. Chem. Phys.* 2003, 118, 6664.
- [42] Mitsutake, A.; Okamoto, Y., *J. Chem. Phys.* 2004, 121, 2491.
- [43] Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J., *Proc. Natl. Acad. Sci. USA* 2005, 102, 13749.

- [44] Kwak, W.; Hansmann, U.H.E., *Phys. Rev. Lett.* 2005, 95, 138102.
- [45] Fukunishi, H.; Watanabe, O.; Takada, S., *J. Chem. Phys.* 2002, 116, 9058.
- [46] Jang, S.; Shin, S.; Pak, Y., *Phys. Rev. Lett.* 2002, 91, 058305.
- [47] Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nancias, M.; Vila, J.A.; Khalili, M.; Arnautova, Y.A.; Jagielska, A.; Makowski, M.; Schafroth, H.D.; Kazmierkiewicz, R.; Ripoll, D.R.; Pillardy, J.; Saunders, J.A.; Kang, Y.K.; Gibson, K.D.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 7547.
- [48] Ferrenberg, A.M.; Swendsen, R.H., *Phys. Rev. Lett.* 1989, 63, 1195.
- [49] Kumar, S.; Bouzida, D.; Swendsen, R.H.; Kollman, P.A.; Rosenberg, J.M., *J. Comp. Chem.* 1992, 13, 1011.
- [50] Hansmann, U.H.E.; Okamoto, Y.; Eisenmenger, F., *Chem. Phys. Lett.* 1996, 259, 321.
- [51] Nakajima, N.; Nakamura, H.; Kidera, A., *J. Phys. Chem. B* 1997, 101, 817.
- [52] Bartels, C.; Karplus, M., *J. Phys. Chem. B* 1998, 102, 865.
- [53] Berg, B.A., *Int. J. Mod. Phys. C* 1992, 3, 1083.
- [54] Hansmann, U.H.E.; Okamoto, Y., *J. Phys. Soc. Jpn* 1994, 63, 3945.
- [55] Hansmann, U.H.E.; Okamoto, Y., *Physica A* 1994, 212, 415.
- [56] Sugita, Y.; Okamoto, Y., *Chem. Phys. Lett.* 2000, 329, 261.
- [57] Mitsutake, A.; Sugita, Y.; Okamoto, Y., *J. Chem. Phys.* 2003, 118, 6664.
- [58] Mitsutake, A.; Sugita, Y.; Okamoto, Y., *J. Chem. Phys.* 2003, 118, 6676.
- [59] Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I., *Biochemistry* 1992, 31, 9665.
- [60] Bateman, A.; Bycroft, M., *J. Mol. Biol.* 2000, 299, 1113.
- [61] Oldziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nancias, M.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16950.
- [62] Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H., *Nat. Struct. Biol.* 2000, 7, 375.
- [63] Derrick, J. P.; Wigley, D. B., *J. Mol. Biol.* 1994, 243, 906.
- [64] Johansson, M. U.; de Chateau, M.; Wikstrom, M.; Forsen, S.; Drankenberg, T.; Bjorck, L., *J. Mol. Biol.* 1997, 266, 859.

- [65] Berendsen, H.J.C.; Postma, J.P.M.; van Gunsteren, W.F.; DiNola, A.; Haak, J.R., *J. Chem. Phys.* 1984, 81, 3684.
- [66] Swope, W.C.; Andersen, H.C.; Berens, P.H.; Wilson, K.R., *J. Chem. Phys.* 1982, 76, 637.
- [67] Rhee, Y.M.; Pande, V.S., *Biophys. J.* 2003, 84, 775.
- [68] Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H.A., *Proc. Natl. Acad. Sci. U.S.A.* 2002, 99, 1937.
- [69] Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A., *J. Phys. Chem. B* 2004, 108, 16934.
- [70] Bhattacharya, K.K.; Sethna, J.P., *Phys. Rev. E* 1998, 57, 2553.
- [71] Aleksenko, V.; Kwak, W.; Hansmann, U.H.E, *Physica A* 2005, 350, 28.
- [72] Frenkel, D.; Smit, B., *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, San Diego 2002.

Chapter 6

Conclusion

In the present work, we studied protein folding with a coarse-grained representation of the polypeptide chain. For this purpose, three projects were reported in present work.

First, to extend the scope of our method of energy-based protein-structure prediction to large proteins, we developed an efficient method for searching for optimal packing of α -helices.¹ It treats α -helices as rigid bodies and uses a simplified Lennard-Jones potential with Miyazawa-Jernigan contact-energy parameters² to describe the interactions between the α -helical elements in this coarse-grained system. Global conformational searches to generate packing arrangements rapidly are carried out with a CFMC type of approach. The results for 42 proteins show that the approach reproduces native-like folds of α -helical proteins as low-energy local minima of this highly simplified potential function. The method was applied with very good results in the CASP6 exercise; we correctly predicted the topology of target T0198 (a 235-residue protein). Currently, only α -helices are treated by this simple procedure, but inclusion of β -strands and sheets in the model would extend the applicability of the procedure to many proteins. For this, it will be necessary to address the issue of hydrogen bonds which is currently not treated.

Next, we combined the Replica Exchange Monte Carlo (REMC)³ method with our Monte Carlo-Minimization (MCM)^{4, 5} method into the Replica Exchange Monte Carlo with Minimization (REMCM) method⁶ and applied it to global conforma-

tional searches of UNRES chains. Like MCM, the REMCM method is based on perturbation of the current conformation and subsequent local energy minimization; then the acceptance/rejection of the new conformation is based on the Metropolis test. However, as in REMC, trajectories are run at various temperatures and conformations can change their assignment to particular temperatures based on a modified Metropolis test. Application of this method to test proteins: protein A (α), 1CLB (α), 1E0L ($\alpha + \beta$), and 1IGD ($\alpha + \beta$) showed that REMCM performs better than MCM, REMC, and CFMC⁷ and comparably to CSA.⁸ Although REMCM is a promising global optimization technique, the focus in our laboratory has been slowly shifting to predict not only the final folded structure but also the kinetics and mechanism of folding. Because the global optimization methods using minimization violate the detailed balance condition, methods which provide canonical sampling such as MC or MD are much more useful for this purpose.

Finally, we implemented efficient methods for calculating thermodynamic averages with UNRES,⁹ namely a Replica Exchange method (REM),³ a Replica Exchange Multicanonical method (REMUCA),¹⁰ and Replica Exchange Multicanonical with Replica Exchange (REMUCAREM),¹⁰ in both Monte Carlo and Molecular Dynamics versions. Application to a small peptide (ala₂₀) and two small proteins (1BDD, 1E0G) showed that calculated thermodynamic averages, such as canonical average energy and heat capacity, were in good agreement among all simulations for poly-L-alanine, showing that the algorithms were implemented correctly, and that all three algorithms are equally effective for small systems. For larger systems, such as 1BDD and 1E0G, Replica Exchange appeared as the most capable technique for sampling rugged energy surfaces such as UNRES. Especially Replica

Exchange Molecular Dynamics coupled with multiplexing appears to be a powerful and scalable method for calculating thermodynamic quantities. For these reasons we used REM MD to calculate free energy surfaces for all systems, which enabled us to visualize the deficiencies of the UNRES force field with current energy parameters.

The last project in the present work describes a first attempt to calculate thermodynamic averages with the UNRES force field. In order to bring the calculated results closer to experimental ones, the UNRES energy function must be reparameterized for canonical simulations such as MC or MD. For this to occur, the CSA method has to be replaced by MD as the component of the hierarchical optimization of the UNRES energy function, responsible for providing decoy sets for different levels. This would effectively replace a database representing local minima of the conformational space, with a database of MD decoys, which would provide configurational entropy for the system. Decoy sets would be generated by running REM MD simulations at temperatures corresponding to complete unfolding, partial folding, and complete folding. These temperatures can be chosen based on experimental data of folding for the training proteins. The MD runs would be carried out with restraints imposed on the quantitative measures of native-likeness of parts of the molecule and/or the entire molecule. Different restraints would correspond to different extent of folding according to the pre-defined hierarchy. The Weighted Histogram Analysis Method (WHAM)^{11, 12} can be used to remove the restraining potentials from the calculated free energies and averages. This would provide much better coverage of the conformational space compared to the procedure described in section 3.2 where only global and local (in the neighborhood of the experimental structure) CSA searches are carried out. Once a more physical

UNRES force field is obtained, kinetic and thermodynamic studies can be carried out on large systems.

Although REMD is a powerful method for exploring free energy landscapes, it does not provide direct information about kinetics. To circumvent this problem, the algorithm developed by Andrec et al.¹³ could be implemented. In this algorithm the power of REMD sampling is combined with a kinetic network model to provide kinetics. REMD simulations are used to generate a lattice of states which are then constructed into a network, and kinetic transitions between states that have sufficient structural similarity are allowed. The qualitative features of the kinetics and corresponding pathways between macrostates can be understood by analyzing the overall network structure or constructing kinetic Monte Carlo "trajectories" that consist of Markovian random walks on the lattice.

Because the secondary degrees of freedom are removed in the UNRES representation of the polypeptide chain, UNRES provides both a decrease in the cost of computation and extension of the time scale. Recently, kinetic studies with the UNRES force field on Staphylococcal protein A were carried out, using 400 Langevin dynamics trajectories.¹⁴ The results suggest that the UNRES force field is well suited for studying the kinetics of folding. It is evident that Replica Exchange MD provides an improvement in sampling of the conformational space over traditional MD, and therefore if kinetic information about the system could be retrieved from such simulations, this should allow for kinetic studies with systems for which traditional molecular dynamics is ineffective. By implementing the kinetic network model with REMD using UNRES and applying it to larger systems, a powerful tool would be created especially in helping to clarify issues such as the nature of folding funnels, intermediates, and kinetic bottlenecks.

BIBLIOGRAPHY FOR CHAPTER 6

- [1] Nancias, M.; Chinchio, M.; Pillardy, J.; Ripoll, D.R.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2003, 100, 1706.
- [2] Miyazawa, S.; Jernigan, R. L., *J. Mol. Biol.* 1996, 256, 623.
- [3] Swendsen, R. H.; Wang, J. S., *Phys. Rev. Lett.* 1986, 57, 2607.
- [4] Li, Z.; Scheraga, H. A., *Proc. Natl. Acad. Sci., U. S. A.* 1987, 84, 6611.
- [5] Li, Z.; Scheraga, H. A., *J. Molec. Str. (Theochem)* 1988, 179, 333.
- [6] Nancias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H.A., *J. Comp. Chem.* 2005, 26, 1472.
- [7] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.
- [8] Czaplewski, C.; Liwo, A; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [9] Nancias, M.; Czaplewski, C.; Scheraga, H.A., *J. Chem. Theo. Comp.* 2005, submitted.
- [10] Mitsutake, A.; Sugita, Y.; Okamoto, Y., *J. Chem. Phys.* 2003, 118, 6664.
- [11] Ferrenberg, A.M.; Swendsen, R.H., *Phys. Rev. Lett.* 1989, 63, 1195.
- [12] Kumar, S.; Bouzida, D.; Swendsen, R.H.; Kollman, P.A.; Rosenberg, J.M., *J. Comp. Chem.* 1992, 13, 1011.
- [13] Andrec, M.; Felts, A.K.; Gallicchio, E.; Levy, R.M., *Proc. Natl. Acad. Sci. USA* 2005, 19, 6801.
- [14] Khalili, M.; Liwo, A.; Scheraga, H.A., *J. Mol. Biol.*, in press.

Appendix A

Chapter 2: CFMC; Chapter 4: CSA, CFMC

The following sections describe the global optimization methods used with the UNRES force field.

A.1 Conformational Space Annealing (CSA)

Conformational Space Annealing (CSA)¹⁻³ is a powerful global optimization method that has been used successfully with UNRES in the CASP3 through CASP6 blind structure prediction exercises.⁴⁻⁶ CSA employs a genetic algorithm and maintains a population of parent structures which evolve using genetic operators. It differs from other genetic optimization methods by carrying out local energy minimization for all conformations, using the Secant Unconstrained Minimization Solver (SUMSL),⁷ and by employing a similarity measure to maintain a database of conformations.

The algorithm anneals in the conformational space by decreasing the similarity measure over the course of the run, which enables CSA not only to search the entire conformational space globally for low-energy fold families (at the start of the search), but also to search the candidate fold families more locally for the lowest-energy representatives (at the end of the search). The purpose of the similarity

cutoff is to maintain a diverse population of structures, i.e. to make sure that the saved conformations are suitably different from one another. At the beginning of a search, the similarity measure cutoff is large, which forces the diverse population to be scattered sparsely in conformation space. As the search progresses, the cutoff decreases and the conformations in the population are allowed to have a higher degree of similarity. By the end of the search, the cutoff is small, and structures in the population can be close, permitting a better local search of the low-energy regions discovered earlier in the search.

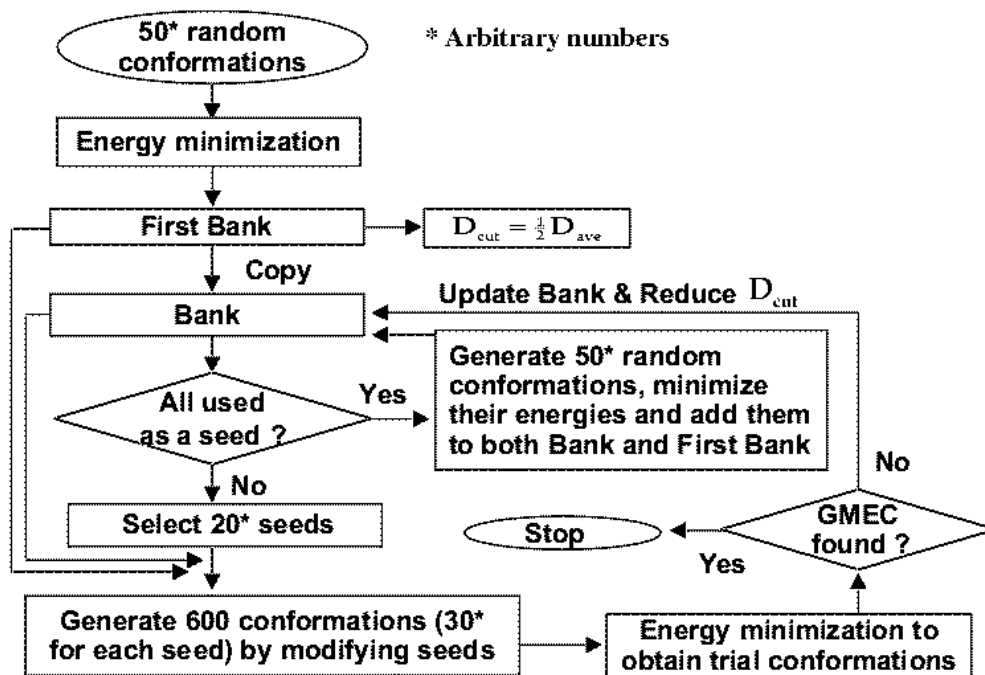


Figure A.1: CSA algorithm^{1, 2, 8}

Figure A.1 shows the basic CSA algorithm. First, a set of random conformations is produced by generating random, non-overlapping conformations and minimizing them locally. This random set is the first CSA bank and its conforma-

tions are unchanged through the entire search. The CSA bank is a changing set of conformations that represent the best structures at the current stage of the search. At the beginning of the search, the CSA bank is just a copy of the first CSA bank. The initial cutoff distance is taken as one-half the average distance between all the structures in the first CSA bank.

Most of CSA work takes place in a loop in which the parent conformations generate new trial conformations which are subsequently minimized and used to update the CSA bank. Each such iteration is called a CSA step. Seed conformations used to generate new conformations are chosen from the CSA bank. To ensure that all the conformations in the CSA bank are eventually used as seeds, and that particular conformations are not overused, CSA keeps track of which seeds have already been used, and preferentially selects unused seeds. Within a single Step, CSA also tries to select seeds that are conformationally diverse by picking additional seeds that are not close to the seeds already selected during the Step.

A variety of methods are used to generate the trial conformations. A seed conformation is taken from the CSA bank and then perturbed by copying portions of local structure from a different conformation in the CSA bank or first CSA bank. The combinations of perturbing variables that can be taken from the other conformation are:

1. The side-chain α and β angles for a single residue.
2. The backbone γ and θ angles for a single residue.
3. All four angles for a single residue.
4. All four angles for a window of consecutive residues.

5. All four angles for a window of consecutive residues comprising a β -hairpin.
6. Pair of remote interacting β -strands.
7. Several non-genetic moves have been added³ to improve the sampling efficiency of the CSA method:
 - (a) A seed can be perturbed by shifting the position of one of its β -hairpins by one or two residues.
 - (b) Carefully designed local move (where only a portion of a backbone can be perturbed, while keeping the rest of the protein frozen).⁹

Each trial conformation is checked for overlapping side chains, which are removed and carefully designed side-chain moves are applied to relax the conformations.⁹ During each minimization stage not only trial conformations, but also the seed structures, are minimized, in case the seed structures had not been fully minimized previously due to a cutoff in the allowed number of energy or gradient evaluations per minimization.

To update the bank, a new structure is either added, rejected or it replaces an old structure. If the new structure has higher energy than every conformation in the Bank, it is discarded. If the new structure is a reminimization of a seed conformation, it replaces the seed conformation if it is lower in energy, otherwise it is discarded. If the new structure is not a reminimized seed structure, then its distances from all conformations in the Bank are computed. If no Bank conformations are within the cutoff distance, the new conformation replaces the highest-energy structure in the Bank. If the new structure within the cutoff distance of one or more structures in the Bank, their energies are compared and the lower energy conformation is kept.

Once the Bank has been updated, the distance cutoff is decreased (until it reaches a minimum cutoff value), and CSA begins a new CSA step by selecting a new set of seed conformations. If all the CSA bank structures have been used as seeds, then all the conformations are once again considered unused so that all can be used as seeds again. This recycling of previous seed structures is usually limited to two times. If the maximum number of recyclings has already been reached and the search is not yet complete, the sizes of the First Bank and Bank are increased by an amount equal to their initial sizes. CSA then continues from the the beginning, generating new random conformations that are then added to the First Bank and Bank. The search continues with a larger pool of random and current structures from which to choose.

The conformational “distance” between two conformations in the UNRES geometry is defined as the absolute value of the difference in backbone dihedral angles between the two conformations, averaged over all such angles in the structure.

A.2 Conformational Family Monte Carlo (CFMC)

Conformational Family Monte Carlo incorporates some of the features of CSA, but uses a Monte Carlo approach rather than a genetic algorithm, and explicitly clusters the population of structures into conformational families. A database of conformations is maintained and updated throughout the search, with the conformations divided into families based upon their C^α coordinate rms distances from one another by means of a minimal-tree clustering method.¹⁰ A maximum of N_{fam} families are allowed at any time, and the families are separated by a distance cutoff d_{fam} ; i.e., no conformation in one family will be within d_{fam} of a conformation in another family. Within a family, a maximum of N_{conf} conformations are allowed,

and no conformations are permitted to be within a distance cutoff d_{conf} of one another. These distance cutoffs may be decreased over the course of a search; such an annealing scheme serves a similar purpose as in CSA—it focuses on smaller, low-energy regions later in the search. By using the metropolis criterion CFMC also anneals in temperature over the course of the run, thus the annealing scheme is expanded into two dimensions.

Figure A.2 shows the basic CFMC algorithm. The first step of CFMC is to generate a starting database of distinct conformations. N_{fam} random structures are generated and locally minimized, with each one starting as a representative of a different family. Of course, any conformations within d_{fam} of others are eliminated and replaced with new ones to ensure that all the starting conformations in the database truly represent different families.

After the starting database is constructed, CFMC enters its main loop, in which new conformations are generated, minimized, and then incorporated into the database. Initially, the lowest-energy family (judging by the lowest-energy structure in each) in the database is chosen as the generating family. A random conformation from the generating family is chosen based on a Boltzmann criterion of the energies of the conformations in the family relative to the energy spread of the family. This structure is then perturbed in one of ten possible ways, falling into two broad “global” and “local” categories. In each broad category there are five move types:

1. Backbone perturbations in a single residue.
2. Side-chain perturbations in a single residue.
3. Backbone perturbations in a window of consecutive residues.

4. Side-chain perturbations in a window of consecutive residues.
5. Interpolation of two conformations, in which the variables of two structures are averaged to generate a new conformation.

Each of the first four types of perturbations have two subclasses; the angles affected by the moves can be perturbed by adding random values to them, or by taking the values from another conformation in the database. For small, local moves, the range of the random perturbations is small, and values can be taken only from other conformations in the same family. For larger, global moves, the range of the random perturbations is larger, and values can be taken only from conformations in other families. The last type of move averages all the variables of the current structure with another structure in the database, with a random relative weighting of the two conformations. Again, the other conformation used in the interpolation is another conformation in the same family, or a conformation in a different family, depending on whether the move is a local or global one. Once the new conformation is generated, it is locally minimized and then evaluated for inclusion in the database.

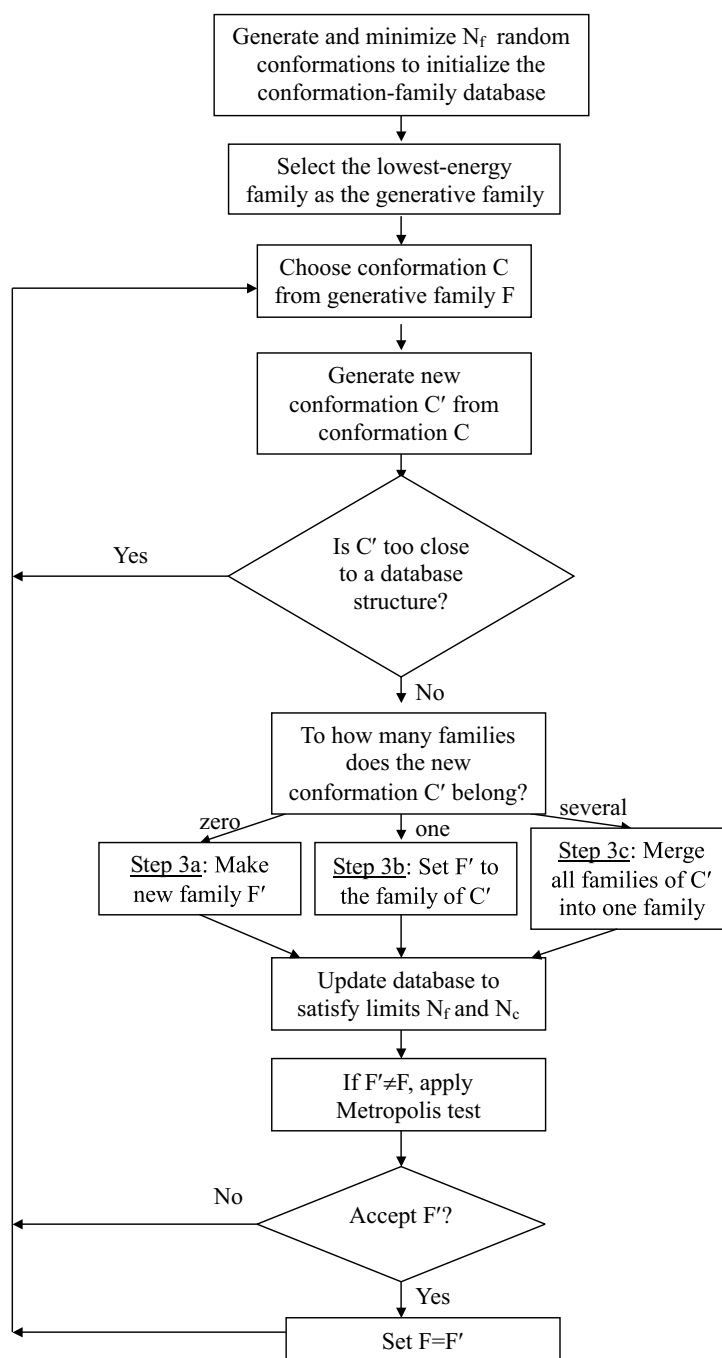
When evaluating a new conformation, its distance from all the existing conformations in the database must be calculated. With these distances, the families to which the conformation could belong (i.e., those containing a conformation closer than d_{fam}), and the conformations to which it is very close (i.e., closer than d_{conf}), can be determined. If there is a very close conformation in the database that has lower energy than the current structure being evaluated, then the current structure is rejected. If there are no existing families to which the new conformation could belong, it represents a new family. If there is room for another family in the database, it is added; otherwise, it replaces the highest-energy family if it has

a lower energy than that family. The second case to consider is when the new conformation clearly belongs in only one existing family. If the new conformation is not very close to one of the existing ones, and there is room in the family for another conformation, it is added to the family. If it is very close to existing conformations, the very close conformations are eliminated, since a check has already been made that the new conformation has lower energy than all those very close to it. If there aren't any very close conformations, but there isn't room for a new one in the family, the new conformation replaces the highest-energy conformation in the family, if that highest-energy conformation is also higher in energy than the new conformation. Since this process may have eliminated conformations that connected the family together, the family must be checked to see if it has split into more than one family. If it has split, and there isn't room for more families in the database, the highest-energy families are eliminated. The last case is when the new conformation potentially belongs to several different families. In this case, all those families are merged into one, any very close conformations are eliminated, the high-energy conformations are eliminated if the family is too large, and the family is reclustered to check if it split as the result of any eliminations. If it has split, and there isn't room for more families in the database, the highest-energy families are eliminated.

Once the database is updated, the family to which the new conformation belongs is compared to the generating family. If the family to which the new family belongs is different than the generating family, the Metropolis criterion is applied to determine whether the generating family will switch to the new one. If the generating family was eliminated or merged during the updating procedure, then the generating family is changed to the new family. If both the generating family and

the new conformation were eliminated, however, then a new generating family is chosen based on a Boltzmann criterion of the family energies relative to the family energy spread. The energy of a family is defined as the energy of its lowest-energy member.

In order to parallelize CFMC, several threads with different generating families are started concurrently, and each thread can produce new conformations for minimization and testing on several processors at once. Since all the threads are simultaneously modifying the same database and the generating families for the threads are continually changing, all the threads are periodically reset and the generating families set to distinct families. Also, since several different conformations for a single thread are generated from the same generating family, special care is taken to maintain an appropriate sequence of current conformations and generating families.¹¹

Figure A.2: CFMC algorithm¹¹

BIBLIOGRAPHY FOR APPENDIX A

- [1] Lee, J.; Scheraga, H. A.; Rackovsky, S., *J. Comput. Chem.* 1997, 18, 1222.
- [2] Lee, J.; Scheraga, H. A., *Intl. J. Quantum Chem.* 1999, 75, 255.
- [3] Czaplewski, C.; Liwo, A.; Pillardy, J.; Oldziej, S.; Scheraga, H. A., *Polymer* 2004, 45, 677.
- [4] Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Gibson, K. D.; Scheraga, H. A., *Intl. J. Quantum Chem.* 2000, 77, 90.
- [5] Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kaźmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y.-J.; Scheraga, H. A., *Proc. Natl. Acad. Sci., USA* 2001, 98, 2329.
- [6] Oldziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nancias, M.; Vila, J.A.; Khalili, M.; Arnautova, Y.A.; Jagielska, A.; Makowski, M.; Schafroth, H.D.; Kazmierkiewicz, R.; Ripoll, D.R.; Pillardy, J.; Saunders, J.A.; Kang, Y.K.; Gibson, K.D.; Scheraga, H.A., *Proc. Natl. Acad. Sci. USA* 2005, 102, 7547.
- [7] Gay, D. M., *ACM Trans. Math. Software* 1983, 9, 503.
- [8] Lee, J.; Pillardy, J.; Czaplewski, C.; Arnautova, Y.; Ripoll, D. R.; Liwo, A.; Gibson, K. D.; Wawak, R. J.; Scheraga, H. A., *Comp. Phys. Comm.* 2000, 128, 399.
- [9] Chinchio, M.; Scheraga, H. A., *J. Comp. Chem.* 2005, To be submitted for publication.
- [10] Späth, H., *Cluster analysis algorithms for data reduction and classificati of objects*, Halsted Press, New York 1980.
- [11] Pillardy, J.; Czaplewski, C.; Wedemeyer, W. J.; Scheraga, H. A., *Helvetica Chimica Acta* 2000, 83, 2214.