14 Monte Carlo studies of biological molecules

14.1 INTRODUCTION

The combination of improved experimental capability, great advances in computer performance, and the development of new algorithms from computer science have led to quite sophisticated methods for the study of certain biomolecules, in particular of folded protein structures. One such technique, called 'threading', picks out small pieces of the primary structure of a protein whose structure is unknown and examines extensive databases of known protein structures to find similar pieces of primary structure. One then guesses that this piece will have the same folded structure as that in the known structure. Since pieces do not all fit together perfectly, an effective force field is used to 'optimize' the resultant structure, and Monte Carlo methods have already begun to play a role in this approach. (There are substantial similarities to 'homology modeling' approaches to the same, or similar, problems.) Of course, the certainty that the structure is correct comes primarily from comparison with experimental structure determination of crystallized proteins. One limitation is thus that not all proteins can be crystallized, and, even if they can, there is no assurance that the structure will be the same *in vivo*. Threading algorithms have, in some cases, been extraordinarily successful, but since they do not make use of the interactions between atoms it would be useful to complement this approach by atomistic simulations. (For an introductory overview of protein structure prediction, see Wooley and Ye (2007).) Biological molecules are extremely large and complex; moreover, they are usually surrounded by a large number of water molecules. Thus, realistic simulations that include water explicitly and take into account polarization effects are inordinately difficult. There have also been many attempts to handle this task by means of molecular dynamics simulations, but the necessity of performing very long runs of very large systems makes it extremely difficult (if not impossible) to reach equilibrium. We are thus in the quite early stages in the study of biological molecules via Monte Carlo methods, but, as we shall see, results are quite promising for the future. The field is developing rapidly, and in some ways we are passing through the same kind of maturation period as Monte Carlo enthusiasts in physics did 30–40 years ago in which simulations were, at first, not taken too seriously by experimentalists. This will surely change.

Generally speaking, there are two classes of problems associated with the simulation of biological molecules. First, the interactions are complex and difficult to describe in terms of simple, classical, phenomenological potentials, or 'force fields,' that can be used for a simulation. Second, because the free energy landscapes for many of the molecules of interest are complex, long time scales may tend to obscure the correct behavior of the system and inventive sampling methods are often needed. Molecular dynamics methods are quite useful for describing dynamical behavior over quite short time scales, but the maximum times for which the integration of the equations of motion can be performed are often orders of magnitude too short to describe the physical range of interest. Thus, the use of innovative Monte Carlo algorithms may offer the only hope of producing understanding of the behavior of many of these molecules as observed in the laboratory (or in living beings).

14.2 PROTEIN FOLDING

14.2.1 Introduction

One exceedingly important set of problems in modern biological science centers around obtaining an understanding of how proteins obtain their folded structures and how to develop a predictive capability to determine what the folded structure will be for an arbitrary protein. Proteins may be viewed (somewhat simplistically) as linear polymers with the naturally occurring amino acids as monomers. For a given sequence of amino acids we would then like to know what structure will result after the protein has folded. This is an exceedingly difficult problem that has two distinct aspects that must be examined. First of all, the nature of the model to be used must be considered. The physical characteristics of proteins are complex, and, in principle, covalent forces between atoms on the 'backbone', van der Waals forces and hydrogen bonds between atoms on different parts of the protein, and long range, shielded electrostatic forces describing the effects of solvent, all need inclusion. Consequently, the corresponding range of independent 'coordinates' that need to be varied is huge. To date it has simply not been possible to examine these problems using realistic Hamiltonians that include all degrees of freedom, and some degree of simplification has been needed. A reasonable compromise is then to use a somewhat simplified Hamiltonian to describe the system in which a combination of bonded and non-bonded forces is used. For simplicity the bond lengths and bond angles are kept constant and the degrees of freedom are constrained to the rotations about the fixed bonds, expressed in terms of dihedral angles (see, e.g., Hansmann and Okamoto, 1999). Once this is done, the behavior of the system is given by the usual formulae of statistical mechanics, e.g. the partition function

$$Z = \sum_{\text{configurations } i} e^{-E_i/k_\text{B}T} \equiv \sum_E g(E) e^{-E/k_\text{B}T} \tag{14.1}$$

where the first sum is over all configurations of the system, and the second sum is over all energies with $g(E)$ being the density of states. As for spin glass models discussed in Chapters 4, 5, and 7, the resultant energy landscape is quite rough and standard Monte Carlo methods tend to be trapped in metastable states. For the case of proteins this often means that the polymer folds, but into a state that does not have the lowest free energy and that is widely separated in phase space from the correct groundstate. This, then, is a challenging problem but one where the sophisticated methods described in earlier chapters may be brought to bear.

Several nice reviews of algorithmic advances and recent results in the computer simulation of protein folding are available (Shaknovich, 2006; Meinke *et al.*, 2009), and the latter even includes a discussion on the parallel implementation of a force field. With the number of processors in a machine increasing far more rapidly than individual processor performance, parallelization of both the force field and the sampling process will be needed to enhance the overall effectiveness of a simulation.

14.2.2 How to best simulate proteins: Monte Carlo or molecular dynamics?

An early comparison of Monte Carlo and molecular dynamics simulations for the protein bovine pancreatic trypsin inhibitor in vacuum provided a very pessimistic view of the suitability of Monte Carlo for such studies (Northrup and McCammon, 1980). Recently, however, Hu *et al.* (2005) demonstrated that, with an appropriately chosen move set, Monte Carlo can be competitive, or even superior, for certain biomolecular systems. (This conclusion was similar to that found by Jorgensen and Tirado-Rives (1996) in their comparison of the two methods for the simulation of liquid hexane.) They probed the efficiency of different trial moves for several peptides and both implicit and explicit solvent. Consequently, they provided an implementation of a Monte Carlo module for the commonly used computational biochemistry program CHARMM. Since new Monte Carlo algorithms, as well as more efficient implementation of existing methods, appear with almost frightening regularity, we believe that Monte Carlo methods will play an increased role in the future for problems for which the explicit time dependence is not the ultimate goal.

14.2.3 Generalized ensemble methods

The umbrella sampling, multicanonical sampling, parallel tempering method, and Wang–Landau sampling discussed earlier are all suitable for the study of protein folding. A ‘standard model’ for the testing of simulational methods is Met-enkephalin which has the amino acid sequence Tyr–Gly–Gly–Phe–Met. For this system the probability weight

$$w(E) = \left(1 + \frac{\beta(E - E_0)}{n_F}\right)^{-n_F} \tag{14.2}$$

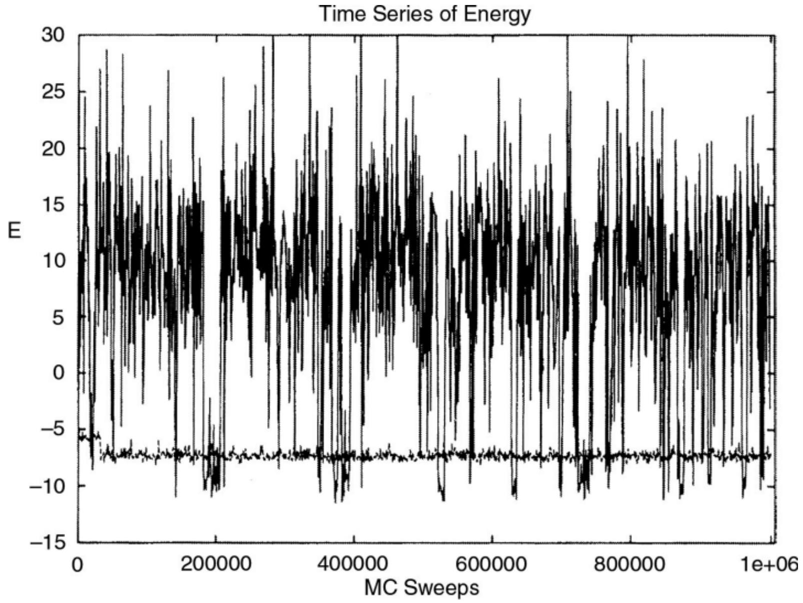

Fig. 14.1 Time sequence of the energy of simulations for Met-enkephalin from a regular canonical simulation at $T=$ 50 K (dotted curve) and from a generalized ensemble simulation (solid curve). After Hansmann and Okamoto (1999).

was chosen with $n_F = 19$ and $E_0 = E_{\mathrm{GS}} = -12.2$ kcal/mol (E_{GS} being the known groundstate energy). As shown in Fig. 14.1, the canonical simulation at $T = 50$ K is trapped in a low-lying metastable state whereas the generalized ensemble simulation explores both higher-lying and lower-lying states. A multicanonical ensemble simulation had found earlier that the mean energy at this temperature, i.e. in the canonical ensemble, should be –11.1 kcal/mol.

Similar studies were carried out using parallel tempering, and this approach proved to be effective in overcoming the problem of multiple minima. (For a comparison of parallel tempering with canonical Monte Carlo and molecular dynamics, see Hansmann, 1997.)

As an example of the ability of simulations to describe folded structures, in Fig. 14.2 we show a comparison of the low energy conformation found from simulation with the structure determined from X-ray data. The superposition of the two structures shows that the simulation reproduces the tertiary structure quite well. This is quite gratifying since it shows that the Monte Carlo simulation is well on its way to becoming a predictive tool.

It is now known that in nature there are proteins that tend to fold into more than one state, and the mis-folding of some proteins is believed to be responsible for some neurological diseases. A simple example is a peptide with the amino acid sequence EKAYLRT (glutamine–lysine–alanine–tyrosine–leucine–arginine–threonine) which appears at positions of both α-helices and β-sheets in naturally occurring proteins. (The use of the alphabet to denote amino acid sequences is standard in the biochemical/biological community. See, e.g., Guo and Guo (2007).) EKAYLRT is an excellent system for the study of whether the folding process depends upon the intrinsic properties of the protein or upon the interaction of the protein with its environment. Peng

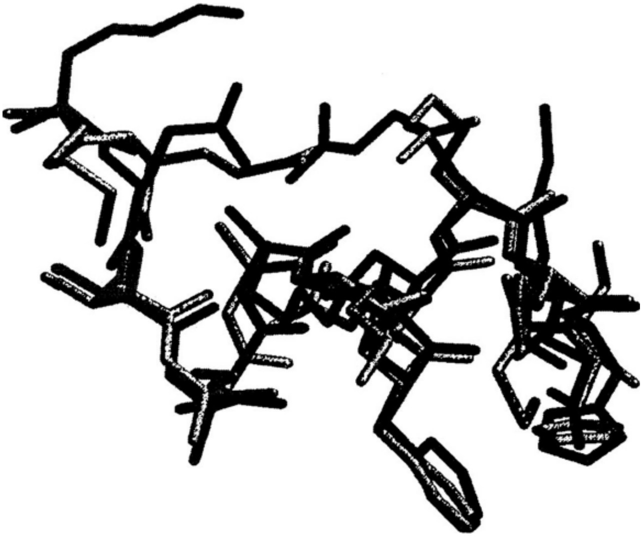

Fig. 14.2 Structure of the C-peptide of ribonuclease A: (black sticks) lowest energy state obtained from a multicanonical Monte Carlo study; (gray sticks) the X-ray structure. After Hansmann and Okamoto (1999).

and Hansmann (2003) used multicanonical simulations to study the behavior of the peptide as both an isolated molecule as well as when interacting with another peptide, using an all-atom representation with interactions between all atoms in a standard force field. They concluded that EKAYLRT by itself has the tendency to form an α-helix, but when it is close to another strand it forms a β-sheet. While not conclusive, this multicanonical study offers a very promising view of the utility of this system for increasing our understanding of various neurodegenerative illnesses.

14.2.4 Globular proteins: a case study

The understanding of globular protein crystallization is important for conquering many pathological diseases, and Monte Carlo simulations are beginning to play a role for these systems. Pagan *et al.* (2004) simulated the ten Wolde–Frenkel model which uses a modified Lennard–Jones pair-wise potential whose range of attractive interaction is small compared to the protein diameter. In this model, for particles a distance r apart, the interaction potential is

$$V(r) = \begin{cases} \infty, & r < \sigma \\ \dfrac{4\varepsilon}{\alpha^2}\left(\dfrac{1}{\left[(r/\sigma)^2 - 1\right]^6} - \dfrac{\alpha}{\left[(r/\sigma)^2 - 1\right]^3}\right), & r \geq \sigma \end{cases} \tag{14.3}$$

where σ is the hard-core radius and ε is the depth of the potential well. In chemical potential-temperature space this model shows fluid–fluid coexistence up to a critical point. One reason for examining this study so closely is that they took advantage of multiple methods that we have described earlier in this text. They employed Metropolis sampling (see Section 4.2) in the grand-canonical ensemble and analyzed the data using histogram reweighting (see Section 7.2),

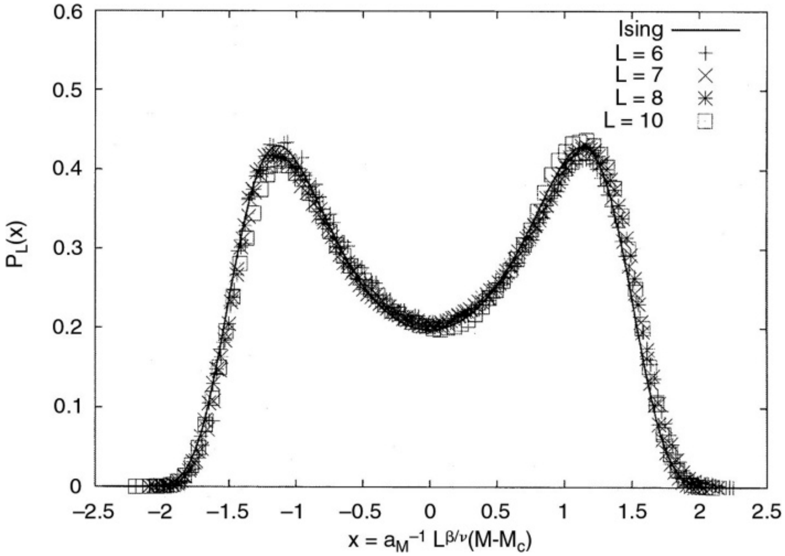

Fig. 14.3 Scaling for the order parameter (density) of the ten Wolde-Frenkel model. Shown for comparison is the universal fixed-point ordering operator function (solid curve). From Pagan *et al.* (2004).

finite size scaling (see Section 4.2.3), and field mixing (see Section 4.2.3.5). Data were obtained for $L \times L \times L$ simulation cells with L varying from 6–10 σ and with periodic boundary conditions. Long runs, extending from 10^8 to 10^9 MCS, were used to take data, and distributions of both the density and energy were constructed. A finite size scaling plot of the distribution of the order parameter (the density) is shown in Fig. 14.3. A field mixing analysis was used to determine the location of the phase coexistence region and the critical point. From the variation of the coexistence densities as the critical density is approached they extracted an estimate for the critical exponent β that is consistent with the three-dimensional Ising value.

14.2.5 Simulations of membrane proteins

Membrane proteins form a particularly interesting and complex sub-branch of protein folding research because the protein–membrane interaction produces another degree of complexity into the problem. Several recent Monte Carlo studies of transmembrane helix behavior in glycophorin A and Bacteriorhodopsin have used ‘state of the art’ techniques from statistical physics. First, Chen and Xu (2005) invoked parallel tempering (see Section 5.4.2) to simulate both systems. The helices were modeled in the united atom representation using the CHARMM19 force field, and an explicit knowledge-based potential was developed to describe the residue level interaction between the helices and the membrane. (To speed up the simulations they kept the internal structure of the helix backbone fixed, but allowed the dihedral angles to change. Global moves of the helices were allowed as well.) Monte Carlo simulations were performed both with and without the helix–membrane coupling, and results suggested that the contributions from the helix–membrane interaction play an extremely important role in determining the packing of the helices in the membrane. Thus, the work by Chen and Xu (2005) addressed both challenges mentioned in the introduction to this chapter: improving the description of the interactions and improving the simulational methods. This study was followed by a further examination of these two transmembrane proteins using Wang–Landau sampling (see Section 7.8) at both the residue and the atomic level (Chen and Xu, 2006). In their implementation they reduced the modification factor to $\ln f = 10^{-7}$ and used a fairly standard flatness criterion ($p = 0.8$). Individual runs at the atomistic level took about one month of CPU time on a single processor workstation, and from the resultant data Chen and Xu examined energy landscapes and structural properties. Wang–Landau sampling at the residue level took only a few hours because of the coarse graining of the system. They concluded that a hierarchical approach to membrane protein structure prediction via simulation was promising: candidate structures can first be selected at the residue level and then refined with atomistic detail.

A two-step Monte Carlo procedure was then developed by Gervais *et al.* (2009) to obtain the free energy landscape for membrane proteins. They considered the dimerization process in glycophorin A, including both helix–helix interactions and a helix–membrane coupling. (The system under consideration is composed of two identical α-helices, A and B, of 22 residues each, EITLIIFGVMAGVIGTILLISY.) The helix backbone is a perfect α-helix and was kept fixed. A unified atom representation was employed where, in addition to all heavy atoms, only polar hydrogen atoms susceptible to being involved in hydrogen bonding were explicitly included (the total number of atoms was 378). The energy density of states of the system was first estimated with Wang–Landau sampling, and then a production run, with fixed density of states, was performed, during which various observables were sampled to provide insight into the folding thermodynamics of the protein in question. The procedure was used to study glycophorin A, and the dimerization process of this homo-dimer was found to be highly hierarchical. All seven residues of the motif LIxxGVxxGVxxT play a dominating role in the dimerization, manifesting in two distinct transitions: (i) contact formation between the two helices at a temperature of 800 K followed by (ii) collapse of the system to the native state at 300 K. The advantages of this procedure are its flexibility and its broad range of applicability. The specific heat thus calculated for glycophorin A, shown in Fig. 14.4, shows the two-step acquisition of the native state is remarkably similar to what is found in the HP lattice protein model (see Fig. 14.2). (Note that in a lattice system with discrete variables, the specific heat goes to zero as T approaches zero, whereas the model for glycophorin A has continuous variables so the specific heat does not go to zero, within the framework of classical statistical mechanics.) Of course the peak at ~800 K has no physical meaning since the membrane/protein system would not exist at such a high temperature; however, the model is defined at the intersection of biology and statistical mechanics, and the specific heat curve is meaningful within this context. For a different membrane protein study using replica exchange see Ulmschneider *et al.* (2007).

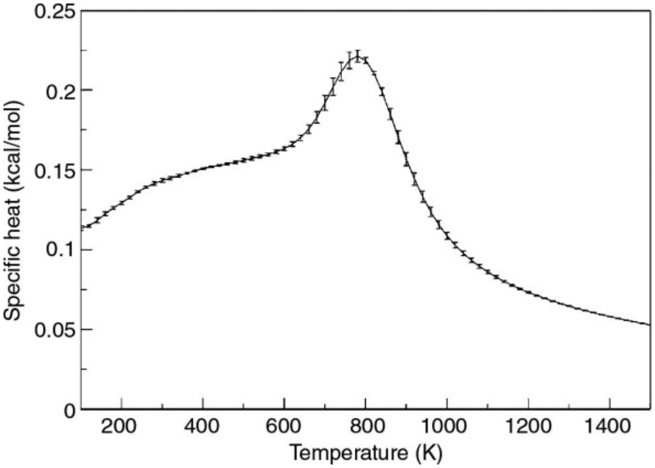

Fig. 14.4 Specific heat for glycophorin A as determined by Wang–Landau sampling. Errors were estimated by a jackknife analysis for 10 independent runs. Note that only the temperature range for which results are reliable (i.e. $T >$ 100 K) is shown. From Gervais *et al.* (2009).

14.3 MONTE CARLO SIMULATIONS OF RNA STRUCTURES

Determination of the structure of RNA is of great importance in molecular biology. Studies of RNA are complicated by the presence of pseudoknots, a feature that renders the problem NP-complete. Recently, Lou and Clote (2010) used Wang–Landau sampling (see Section 7.8) to study the thermodynamic properties, and thus the melting temperature, of RNA without restricting the model to exclude pseudoknots. They considered single RNA molecules as well as hybridized complexes of two RNA molecules. As can be seen in Fig. 14.5, exept for very short RNA sequences the execution time is dramatically reduced by using Wang–Landau sampling instead of one of the 'standard' programs *RNAfold* (Lou and Clote, 2010).

Simulations were performed for the toy sequence of 5′-AGCGA-3′ hybridized to its reverse complement 3′-UCGCU-5′ using both a 'standard' program (UNAFold) that restricts the allowed molecular structures and Wang–Landau sampling. Although both give similar melting temperatures, UNAFold shows a single-step process whereas the specific heat from Wang–Landau sampling indicates a two-step process. Clearly, then, the restriction of pseudo-knots has a significant effect on the resultant thermodynamic behavior.

14.4 MONTE CARLO SIMULATIONS OF CARBOHYDRATES

Monte Carlo simulations have not been used very often for the study of carbohydrates. One significant physical difference, as compared to proteins, is that carbohydrate molecules tend to stay rather 'floppy' and do not always form a well defined 'native state'. Nonetheless, there have been a number of Monte Carlo studies of carbohydrates, and it is likely that more studies

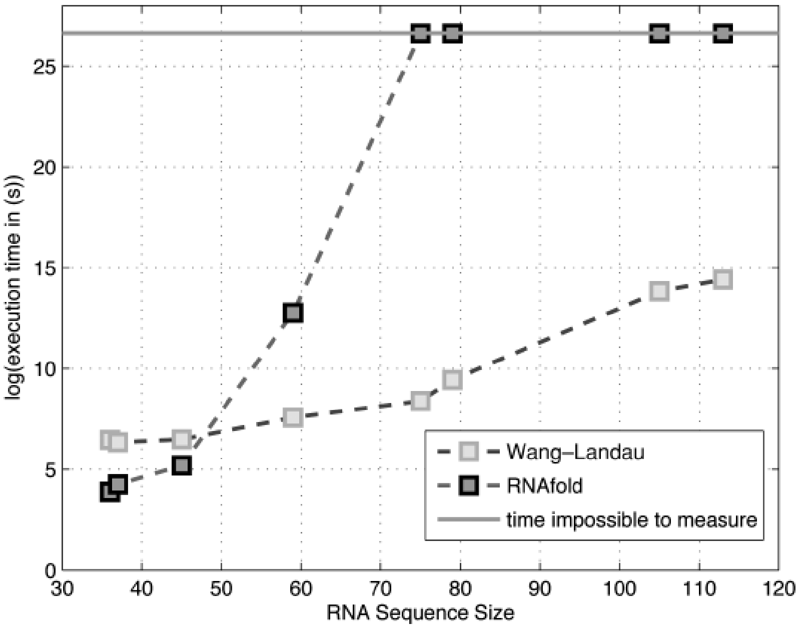

Fig. 14.5 Comparison of execution times between the program RNAsubopt-D and Wang–Landau sampling for RNA sequences of varying size. After Lou and Clote (2010).

will be forthcoming in the future. In an early study, Stuike-Prill and Pinto (1995) performed Metropolis Monte Carlo simulations of four sub-structures (increasingly complex oligosaccharides) from the cell-wall polysaccharide antigen of *Streptococcus* group A. The authors chose a modified HSFA potential energy in the GEGOP force field for simulations *in vacuo* and adjusted the maximum amount by which the dihedral angles and glycosidic bond angles were allowed to change within a trial move to obtain an acceptance rate between 30 and 60%. (The different force fields used in this area are quite different than those associated with more traditional problems in statistical physics, but there are a number of references given by Stuike-Prill and Pinto (1995) that will enlighten the interested reader.) Elevated temperatures were used to insure a broad sampling of the conformational state, but overall the structures were conformationally surprisingly restricted. In Fig. 14.6 we show an overlay of 50 different structures that were produced by the simulation and which depict a relatively well formed structure. Interestingly, the authors note that they find much higher flexibility about the glycosidic linkages than was found with earlier MD simulations. Overall, they found quite good agreement with experimental averaged proton–proton distances obtained from NMR spectroscopy.

A rather different approach was taken by Bernardi *et al.* (2004), who used a hybrid Monte Carlo/stochastic dynamics method with the AMBER* force field to simulate the behavior of the conformation and the dynamics of several mannobiodides. A two-step process was used, beginning with a Monte Carlo/energy minimization search followed by the Monte Carlo/stochastic dynamics simulation. Rather extended cutoffs were used for the van der Waals

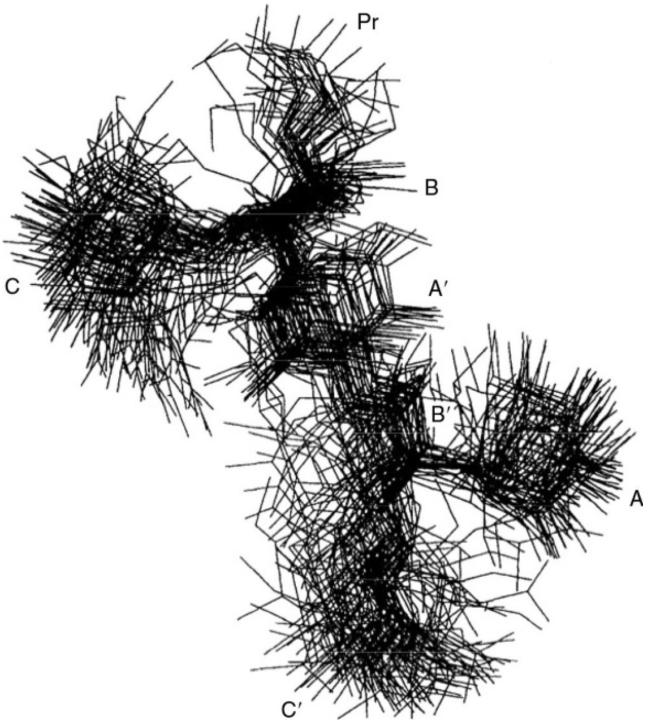

Fig. 14.6 Overlay of 50 configurations generated by a Monte Carlo simulation of the hexasaccharide 4 at 700 K. From Stuike-Prill and Pinto (1995).

and electrostatic couplings (25 Å) and hydrogen bonds (15 Å) with water solvation modeled by a continuum solvent model. Their results were in agreement with available experimental data.

14.5 DETERMINING MACROMOLECULAR STRUCTURES

In earlier sections we described how Monte Carlo simulations based upon atomic potentials could be used to simulate biological molecules. Other methods that we have outlined in previous sections have also been used to help determine or understand conformations of biological molecules in different ways. One such example was the use of inverse Monte Carlo simulations (see Section 5.9.4) to compute inter-residue couplings from radial distribution functions (Bathe and Rutledge, 2003) that could come from, e.g., X-ray scattering data. They tested the approach on a simple homopolymer made up of freely jointed beads and then for a heteropolymer model with interactions chosen to mimic the three-helix bundle fragment of *Staphyloccus aureus* protein A. From Monte Carlo simulations for these two models, radial distribution functions (total non-bonded radial distribution functions for the homopolymer and individual residue specific radial distribution functions for the protein model) were extracted to be used as input for the inverse Monte Carlo procedure. The effectiveness of the method was evaluated for random coil, random globule, and ordered globule states.

A different type of conformational problem arises because the binding of transcription-factor proteins to specific DNA sequences plays an important role in gene expression. DNA binding sites are often identified using weight matrices, and the identification of low energy binding sites can, in turn, allow the construction of accurate weight matrices. For this reason, Endres and Wingreen (2006) used a Wang–Landau algorithm (see Section 7.7) to sample high affinity binding sites to extract weight matrices. They found that this procedure matched well with a slow but exact 'dead-end elimination method' and offered significant computational improvement over more standard Monte Carlo methods. They used this approach to demonstrate homology modeling by changing the amino-acid sequence in a co-crystal X-ray structure of a native protein–DNA complex, Zif268, and recovered a weight matrix typical of Zif268 when the protein is allowed to be flexible.

14.6 OUTLOOK

In this brief chapter we have only attempted to give the reader a mild taste of the use of Monte Carlo methods to study biological molecules. Simulations of proteins and carbohydrates have progressed in recent years so that many studies now use some of the most sophisticated methods developed within the statistical physics community. While the argument about whether Monte Carlo or molecular dynamics is superior for such systems is likely to continue, it is clear that Monte Carlo has become an important, mature alternative for the study of biological molecules. Currently, many problems, e.g. translocation of DNA through pores in biological membranes (see Chapter 10), are only accessible by studies within the framework of highly coarse-grained models lacking any chemical detail; in coming years, progress with algorithms and hardware will allow the investigation of such problems with more realistic models.

REFERENCES

Bathe, M. and Rutledge, G. C. (2003), *J. Comput. Chem.* **34**, 876.

Bernardi, A., Colombo, A., and Sanchez-Medina, I. (2004), *Carbohydr. Res.* **339**, 967.

Chen, Z. and Xu, Y. (2005), *Proteins: Structure, Function, and Bioinform.* **62**, 539.

Chen, Z. and Xu, Y. (2006), *J. Bioinform. and Comput. Biol.* **4**, 317.

Endres, R. G. and Wingreen, N. S. (2006), *Phys. Rev. E* **68**, 061921.

Gervais, C., Wuest, T., Landau, D. P., and Xu, Y. (2009), *J. Chem. Phys.* **13**, 215106.

Guo, H. and Guo, H. (2007), in *Computational Methods for Protein Structure Prediction and Modeling*, eds. Y. Xu, D. Xu, and J. Liang (Springer Science Business Media, New York), vol. 2, p. 29.

Hansmann, U. H. E. (1997), *Chem. Phys. Lett.* **281**, 140.

Hansmann, U. H. E. and Okamoto, Y. (1999) in *Annual Reviews of Computational Physics VI*, ed. D. Stauffer (World Scientific, Singapore).

Hu, J., Ma, A. and Dinner, A. R. (2005), *J. Comput. Chem.* **27**, 203.

Jorgensen, W. L. and Tirado-Rives, J. (1996), *J. Phys. Chem.* **100**, 14, 508.

Lou, F. and Clote, P. (2010), *Bioinformatics* **26**, 1278.

Meinke, J. H., Mohanty, S., Nadler, W., Zimmermann, O., and Hansmann, U. H. E. (2009), *Eur. Phys. J. D* **51**, 33.

Northrup, S. H. and McCammon, J. A. (1980), *Biopol.* **19**, 1001.

Pagan, D. L., Gracheva, M. E., and Gunton, J. D. (2004), *J. Chem. Phys.* **120**, 8292.

Peng, Y. and Hansmann, U. H. E. (2003), *Phys. Rev. E* **68**, 041911.

Shaknovich, E. (2006) in *Computer Simulations in Condensed Matter: From Materials to Chemical Biology*, eds. M. Ferrario, G. Ciccotti, and K. Binder (Springer, Heidelberg), vol. 2, p. 563.

Stuike-Prill, R. and Pinto, B. M. (1995), *Carbohydr. Res.* **279**, 59.

Ulmschneider, J. P., Ulmschneider, M. B., and Di Nola, A. (2007), *Proteins: Structure, Function, and Bioinformatics* **69**, 297.

Wooley, J. C. and Ye, Y. (2007), in *Computational Methods for Protein Structure Prediction and Modeling*, eds. Y. Xu, D. Xu, and J. Liang (Springer Science Business Media, New York), vol. 1, p. P.