Published in final edited form as:

Biochim Biophys Acta. 2008 January ; 1784(1): 252–258.

Understanding protein folding: Small proteins in silico

Olav Zimmermann^a and Ulrich H.E. Hansmann^{a,b,*}

^a John von Neumann Institut für Computing, Research Centre Jülich, 52425 Jülich, Germany

^b Department of Physics, Michigan Technological University, Houghton, MI 49931, U.S.A

Abstract

Recent improvements in methodology and increased computer power now allow atomistic computer simulations of protein folding. We briefly review several advanced Monte Carlo algorithms that have contributed to this development. Details of folding simulations of three designed mini proteins are shown. Adding global translations and rotations has allowed us to handle multiple chains and to simulate the aggregation of six beta-amyloid fragments. In a different line of research we have developed several algorithms to predict local features from sequence. In an outlook we sketch how such biasing could extend the application spectrum of Monte Carlo simulations to structure prediction of larger proteins.

Keywords

protein folding; aggregation; generalized-ensemble Sampling; structure prediction

1 Introduction

Understanding protein folding in detail is a cornerstone for the successful design of protein inhibitors and other therapeutic proteins. While knowledge based approaches are very successful in the prediction of protein structures [1], they are less helpful in unraveling the folding process. This is because they are derived from native structures and thus strongly biased against any non-native intermediates. Such intermediates may play a vital role in accelerating folding and preventing aggregation. While in principal atomistic models and carefully calibrated physical force fields allow to simulate the folding of a protein from a random conformation to the native state, straightforward approaches are computationally demanding. The computational complexity originates from the large conformational search space and the rugged free energy landscape, whose many local minima may trap the search.

In recent years many algorithmic improvements have been devised to overcome these problems. Together with increasing computer power, they now put atomistic folding simulations of small proteins within reach. We review a range of Monte Carlo techniques to sample efficiently the high dimensional conformational search space. We show results for three designed proteins of 20–25 residues forming stable monomers. These results demonstrate that improved algorithms allow us to analyze the folding process of small proteins in full detail. Note that while this size range appears small compared to the average size of a cellular protein (≈ 250 residues) it is comparable to many protein kinase inhibitor peptides (see [2] and

* Corresponding author. Tel.: +1 906 487 4552; fax: +1 906 487 2933. Email address: hansmann@mtu.edu (Ulrich H.E. Hansmann).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

references therein). In an outlook, we sketch how these methods can be used together with local structure constraints to predict the structure of much larger proteins and show some machine learning based algorithms that allow to obtain such constraints.

2 Generalized-ensemble Monte Carlo

The basic requirements for a simulation are a realistic force field, a move set, and a strategy which ensures efficient sampling of the conformational space. The following discussion focuses on the latter point. Although all atom folding simulations of small proteins have been recently accomplished by both Monte Carlo (MC) and Molecular Dynamics (MD) methods [3], we concentrate on MC approaches. Apart from the simpler setup of MC simulations, which do neither require coupling to a heat bath nor a differentiable potential function, it has been found that protein folding simulations using MC algorithms are several times more efficient than those with MD [4]. All simulations reported here use the ECEPP/3 force field [5] as implemented in our simulation package SMMP [6,7]. ECEPP/3 represents a typical physics based force fields which can be summarized as

$$\begin{aligned}
 E_{\text{ECEPP/3}} &= E_C + E_{LJ} + E_{HB} + E_{Tor} \\
 &= \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}} \\
 &\quad + \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\
 &\quad + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\
 &\quad + \sum_l U_l (1 \pm \cos(n_l \xi_l)),
 \end{aligned} \tag{1}$$

where r_{ij} is the distance between atoms i and j , q_i is the charge of atom i , ϵ is the dielectric constant, ξ_l and n_l are the torsional angle of chemical bond l and its respective multiplicity, and A_{ij} , B_{ij} , C_{ij} , D_{ij} , U_l are parameters that have been derived from experimental structures.

In canonical simulations crossing of an energy barrier of height ΔE is suppressed by a factor $\propto \exp(-\Delta E/k_B T)$ (k_B is the Boltzmann constant and T the temperature of the system). Hence, at biologically relevant temperatures the protein likely will get trapped in a local minimum. Work in our group focuses on the development and use of techniques that allow location of local minima, but also enable escape out of the minima to continue the search process. An example are generalized ensemble methods [8]. In most of these methods, one introduces artificial weights that will lead to a uniform distribution of one or more preselected physical quantities in a MC or MD simulation. For instance, in *multicanonical sampling* [9] the weight $w(E)$ is chosen such that the distribution of energies $P(E)$ is given by:

$$P(E) \propto n(E) w(E) = \text{const}, \tag{2}$$

where the spectral density $n(E)$ is the number of different configurations with energy E (see Fig. 1(a)).

In this way, a free random walk in the energy space is realized that allows the simulation to escape from any local minimum. The thermodynamic average of a physical quantity A e.g. radius of gyration can now be calculated by re-weighting: [10,11]

$$\langle A \rangle_T = \frac{\int dx A(x) w^{-1}(E(x)) e^{-E(x)/k_B T}}{\int dx w^{-1}(E(x)) e^{-E(x)/k_B T}}. \tag{3}$$

Here, x stands for configurations. Note that the weights $w(E)$ are not a priori known, and estimators have to be determined by an iterative procedure described in Refs. [9,12].

A variant of this idea is the global optimization technique *energy landscape paving* (ELP) [13]. Here, low-temperature Monte Carlo simulations are performed with a modified "energy" that steers the search away from regions already explored:

$$w(E) = e^{-\tilde{E}/k_B T} \quad \text{with} \quad \tilde{E} = E + f(H(q, t)). \quad (4)$$

T is a (low) temperature, $\tilde{E}$ serves as a replacement of the energy E , and $f(H(q, t))$ is a function of the histogram $H(q, t)$ in a pre-chosen "order parameter" q e.g. fraction of native contacts. Within ELP the weight of a local minimum state decreases with the time the system stays in that minimum till it is no longer favored. The system will then explore higher energies till it falls into a new local minimum (see Fig. 1(c)). Obviously, for $f(H(q, t)) = f(H(q))$ the method reduces to the various generalized-ensemble methods [8] (for instance for $f(H(q, t)) = \ln H(E)$ to multicanonical sampling).

Another way of enhancing the sampling of low-energy protein configurations is *parallel tempering* (also known as replica exchange) [14–16] which was first introduced to protein science in Ref. [17]. In its most common form, one considers N *non-interacting* copies of the molecule, each at a different temperature T_i . In addition to standard Monte Carlo [17] or molecular dynamics moves [17] that affect only one copy, parallel tempering introduces a new *global* update [14]: the exchange of conformations between two copies i and $j = i + 1$ with probability

$$P(C_i \rightarrow C_j) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))), \quad (5)$$

where $\beta = 1/k_B T$. This exchange of conformations leads to a faster convergence than is observed in regular low temperature canonical simulations (see Fig. 1(b)). Note that parallel tempering does not require Boltzmann weights but can be combined easily with other generalized-ensemble techniques [17].

Variants that rely on random walks in other control parameters than the temperatures have been developed by us. In "*model hopping*" [18] we enhance sampling of low energy configurations by performing a random walk through an ensemble of systems with slightly altered energy functions. In that way, information is exchanged between varying stages of coarse graining or different local environments. We assume that the energy function can be separated in two terms: $E = E_A + aE_B$. As in parallel tempering (see next chapter), "*model hopping*" considers N non-interacting copies of the molecule, whose configurations evolve through standard Monte Carlo or Molecular Dynamics moves, but also are exchanged between two adjacent copies with probability

$$\begin{aligned} P(C_i \rightarrow C_j) &= \min(1, \exp(-\beta(E_A(C_j) + a_i E_B(C_j) + E_A(C_i) + a_j E_B(C_i) \\ &\quad - E_A(C_i) - a_i E_B(C_i) - E_A(C_j) - a_j E_B(C_j)))) \\ &= \min(1, \exp(\beta \Delta a \Delta E_B)). \end{aligned} \quad (6)$$

Here, $\Delta a = a_j - a_i$ and $\Delta E_B = E_B(C_j) - E_B(C_i)$. Configurations perform a random walk on a ladder of models with $a_1 = 1 > a_2 > a_3 > \dots > a_N$ that differ by the relative contributions of E_B to the total energy E of the molecule. For instance, barriers in the energy landscape of proteins often arise from van der Waals repulsion between atoms that come too close. Hence, we have considered an implementation of "*model hopping*" with successively smaller contributions from the van der Waals energy. While the "physical" system is on one side of the ladder (at $a_1 = 1$), the (non-physical) model on the other end of the ladder (at $a_N \ll 1$) may

allow atoms to share the same position in space. As the protein "tunnels" through energy barriers, sampling of low-energy configurations is enhanced in the "physical" model (at $a_1 = 1$).

This can be seen in Fig. 2 where we compare the times series of a model hopping run with a regular canonical simulation. The latter got trapped in a local minimum and never reached the energies found in model hopping. Equilibration with model hopping was about 8 times faster than in the canonical simulation and about 3 times faster than in a parallel tempering simulation. With this realization of "model hopping" we could "predict" the structure of a 46-residue protein A in an all-atom simulation within a root mean square deviation (rmsd) of 3.2 Å [18].

Model Hopping also allows guiding a simulation by information obtained from homologous structures [19]. Usually, such spatial constraints introduce an additional roughness into the energy landscape and therefore often leads to extremely slow convergence of the simulation. This problem is circumvented in our approach through a random walk in an ensemble of replicas that differ by the strength with which the constraints are coupled to the system. We have demonstrated the usefulness of this approach on some examples of the CASP6 competition [19].

3 In silico folding of small proteins

In order to point out the differences in the role of α -helix or β -sheet formation in the folding process we have investigated three proteins with distinct native folds: the all-helical 1RIJ (22 residues) [20], the all-sheet beta3s (20 residues) [21], and BBA5 (23 residues) [22] which has a mixed helix-sheet fold [23]. In all three cases we find the native structure as global minimum in free energy at room temperature (Fig. 3). However, the folding mechanism differs and is correlated with the specific fold.

The free energy landscape of 1RIJ [20] is characterized by a funnel-like topology around a dominating minimum indicating that collapse of the protein chain and helix formation are synchronous. The corresponding lowest energy conformer has a backbone rmsd of $r_{rmsd} \approx 2.7$ Å and appears at $T = 274$ K with a frequency of about 50% which is smaller than the experimentally observed probability of 90% [20]. For the all-sheet beta3s [21] we find a global minimum that has an rmsd of 2 Å to the native state. About 10% of the configurations have a rmsd smaller than 3 Å and can be considered to be similar to the native state. The experimentally observed propensity for this structure is 13–31% at 284 K [21]. We observe that collapse precedes the formation of the native hydrogen bonds. Formation of one of the two hairpins in a zipper-like fashion catalyzes the formation of the second, by acting as a template for the remaining part of the three-sheet structure. But there is no particular order in that hairpins form. Finally, for BBA5 [22] we find that the two secondary structure elements form independently. The N-terminal hairpin has a D-proline which facilitates the formation of the type II' turn. But the hydrogen bonds corresponding to a good β -hairpin structure have a slightly lower probability than the C-terminal helix. Contacts between the helix and the hairpin are less stable than the secondary structures themselves, which indicates a folding mechanism in which local structures form first and then assemble into their native-like arrangement. Data of the above simulations are also used to optimize the Lund force field [24].

Another application is our investigation of the oligomerization of a β -amyloid fragment. Together with an enhanced move set including global translation and rotation of one molecule we have studied the A β (16–22) peptide which is a key element in the aggregation of the Alzheimer's β -amyloid. In a system of 6 independent molecules we could observe that the aggregates form predominantly antiparallel β -sheets while monomers are mainly helical (Fig. 4) [25]. The antiparallel β -sheets are more stable than parallel ones due to better side chain packing.

4 Biased Monte Carlo for Structure Prediction

While we have focused so far on the dynamic simulation of protein folding, Monte Carlo algorithms can be used to search for the conformation at the global energy minimum which is assumed to be the native state. Natural proteins have an average size of 250 residues. Hence the search space has to be reduced to make the search feasible, even at the expense of losing the correct thermodynamics. Biased Monte Carlo implements prior knowledge of the native structure as constraints, usually modeled as Gaussian energy terms that have their minimum at a particular distance of the constraint. In an extreme case, the so-called G ϕ -models [26], the force field can be substituted entirely by these constraint terms. For smaller proteins it has recently been shown that a biased Monte Carlo move which updates a dihedral angle pair to the cluster centers of a database derived dihedral angle statistics can significantly accelerate the conformational search [27]. While constraints may help to steer the search when many energetically similar conformations are available i.e. on flat energy plateaus, they often lead to additional high energy barriers and local minima. Some proteins e.g. form end to end contacts. If the distance constraints force the ends to join too early, the elements in between may not be able to reach their native conformation. Only crossing the high energy barrier of dissolving the end to end binding will then lead to the native structure.

It is hence important to choose suitable constraints. An obvious source are local properties of protein structure. Secondary structure can be predicted with high accuracy and thus has become a frequent choice. However, this choice cannot provide any constraints for those 45% of residues that are in coil regions i.e. neither belonging to an α -helix, nor to a β -sheet. For this reason, our group has developed various algorithms to predict local structure properties based on the amino acid sequence of a protein. Our program DHPRED predicts the dihedral angle region (helical, extended or outlier) for each residue [28]. DHPRED uses a multistep procedure based on Support Vector Machines (SVM) [29], a class of supervised machine learning algorithm which uses positive and negative training examples for learning a nonlinear classifier. The SVM classifiers for DHPRED have been trained on overlapping 15-residue fragments from 500 non-redundant protein structures. To harness implicit information on the structural requirements at a sequence position we use the position specific scoring matrices of PSIBLAST to encode the sequence information. A second SVM layer uses the prediction results of the first layer to account for correlations between dihedral preferences of neighboring residues. The prediction accuracy of DHPRED for residues in α -helices and β -sheets is comparable to PSIPRED, a state-of-the-art secondary structure prediction program [30]. In addition, it also predicts the dihedral region for two thirds of the coil residues correctly (Fig. 5). The performance demonstrates that large parts of the coil regions in proteins contain a local ordering although their dihedral signatures are less trivial than those in α -helices and β -sheets. Among these orderings β -turns [31] represent the most abundant class in coil regions. We have therefore developed a classifier to distinguish between the different subtypes of β -turns. Each turn type corresponds to a different dihedral signature i.e. a different set of correlated dihedral constraints [32]. In order to derive constraints for the topology of a protein chain we have trained the SVM-based algorithm BETTY [33] to distinguish between parallel and anti-parallel β -sheets. On 7-fold cross validated data it correctly classifies 88% of all β -residues. Complemented by PSIPRED for secondary structure prediction the four-class accuracy Q4 (parallel- β , anti-parallel- β , α -helix, coil) is 79.3%.

A range of other constraints could be derived from the experimental structures of related proteins or from predictions of local structure. These predictions can be either translated into biased distributions which substitute the standard distributions (uniform or Gaussian) from which the Monte Carlo algorithm randomly selects the conformational updates or they can be used as a control parameter in a model hopping scheme. We are currently working on a Monte

Carlo based structure prediction algorithm which implements such a biased conformational search.

5 Conclusion

Recent progress in Monte Carlo algorithms and increased computer power has led to unprecedented performance in the physics based simulation of biological macromolecules. We have shown that for small proteins atomistic simulation of the entire folding process is possible. The computer simulation reveals details of protein folding on various time scales that can be used for both validation and physical explanation of experimental observations. With increasing accuracy of sequence based prediction methods to obtain local structural constraints, we envisage an increasing range of applications for biased Monte Carlo simulations in structure prediction.

Acknowledgements

This work was supported in part by the National Institutes of Health (USA) Grant GM62838 and National Science Foundation (USA) Grant CHE-0313618.

References

1. Kolinski A, Bujnicki JM. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* 2005;61(Suppl 7):84–90. [PubMed: 16187348]
2. Bogoyevitch MA, Barr RK, Ketterman AJ. Peptide inhibitors of protein kinases—discovery, characterisation and use. *Biochim Biophys Acta* 2005;1754:79–99. [PubMed: 16182621]
3. Lin CY, Hu CK, Hansmann UHE. Parallel Tempering Simulations of HP-36, Proteins: Structure. Function and Genetics 2003;52:436–445.
4. Ulmschneider JP, Ulmschneider MB, Di Nola A. Monte Carlo vs molecular dynamics for all-atom polypeptide folding simulations. *J Phys Chem B* 2006;110:16733–16742. [PubMed: 16913813]
5. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J Phys Chem* 1992;96:6472–6484.
6. Eisenmenger F, Hansmann UHE, Hayryan S, Hu CK. [SMMP] A modern package for simulation of proteins. *Comp Phys Comm* 2001;138:192–212.
7. Eisenmenger F, Hansmann UHE, Hayryan S, Hu CK. An Enhanced Version of SMMP - Open-Source Software Package for Simulation of Proteins. *Comp Phys Comm* 2006;174:422–429.
8. Hansmann, UHE.; Okamoto, Y. The Generalized-Ensemble Approach for Protein Folding Simulations. In: Stauffer, D., editor. *Annual Reviews of Computational Physics. VI.* World Scientific; Singapore: 1999. p. 129-157.
9. Berg B, Neuhaus T. Multicanonical algorithms for first order phase transitions. *Phys Lett B* 1991;267:249–253.
10. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. *Phys Rev Lett* 1988;61:2635–2638. [PubMed: 10039183]
11. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo Data Analysis. *Phys Rev Lett* 1989;63:1195–1198. [PubMed: 10040500]
12. Hansmann UHE, Okamoto Y. Comparative Study of Multicanonical and Simulated Annealing Algorithms in the Protein Folding Problem. *Physica A* 1994;212:415–437.
13. Hansmann UHE, Wille Luc. Global Optimization by Energy Landscape Paving. *Phys Rev Lett* 2002;88:068105. [PubMed: 11863858]
14. Hukushima K, Nemoto K. Exchange Monte Carlo Method and Applications to Spin Glass Simulations. *J Phys Soc (Japan)* 1996;65:1604–1608.

15. Geyer GJ, Thompson EA. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *J Am Stat Assn* 1995;90:909–920.
16. Tesi MC, van Rensburg EJJ, Orlandini E, Whittington SG. Monte Carlo Study of the Interacting Self-Avoiding Walk Model in Three Dimensions. *J Stat Phys* 1996;82:155–181.
17. Hansmann UHE. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem Phys Lett* 1997;281:140–150.
18. Kwak W, Hansmann UHE. Efficient Sampling of Protein Structures by Model Hopping. *Phys Rev Lett* 2005;95:138102. [PubMed: 16197184]
19. Gront D, Kolinski A, Hansmann UHE. Exploring protein energy landscape with hierarchical clustering. *Int J Quant Chem* 2005;105:826.
20. Liu Y, Liu Z, Androphy E, Chen J, Baleja JD. Design and characterisation of helical peptides that inhibit the e6 protein of papillomavirus. *Biochemistry* 2004;43:7421–7431. [PubMed: 15182185]
21. de Alba E, Santoro J, Rico M, Jiménez MA. De novo design of a monomeric three-stranded antiparallel β -sheet. *Protein Sci* 1999;8:854–865. [PubMed: 10211831]
22. Struthers M, Ottesen J, Imperiali B. Design and nmr analyses of compact, independently folded bba motifs. *Fold Des* 1998;3:95–103. [PubMed: 9565754]
23. Mohanty S, Hansmann UHE. Folding of proteins with diverse folds. *Biophys J* 2006;91:3573–3578. [PubMed: 16950845]
24. Mohanty S, Hansmann UHE. Improving of an all-atom force field. *Phys Rev E* 2007;76:012901.
25. Meinke JH, Hansmann UHE. Aggregation of β -amyloid fragments. *J Chem Phys* 2007;126:014706. [PubMed: 17212510]
26. Ueda Y, Taketomi H, Go N. Studies on Protein Folding, Unfolding, and Fluctuations by Computer Simulation. II. A Three-Dimensional Lattice Model of Lysozyme. *Biopolymers* 1978;17:1531–1548.
27. Chen WW, Yang JS, Shakhnovich EI. A Knowledge-based Move Set for Protein Folding. *Proteins: Structure, Function, and Bioinformatics* 2007;66:682–688.
28. Zimmermann O, Hansmann UHE. Support vector machines for prediction of dihedral angle regions. *Bioinformatics* 2006;22:3009–3015. [PubMed: 17005536]
29. Schölkopf, B.; Smola, AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press; Cambridge, MA: 2002.
30. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202. [PubMed: 10493868]
31. Lewis PN, Monany FA, Scheraga HA. Chain reversals in proteins. *Biochim Biophys Acta* 1973;303:211–229. [PubMed: 4351002]
32. Zimmermann, O.; Hansmann, UHE. Dihedral Angle Patterns in Coil Regions of Protein Structures. In: Ulrich, HE.; Hansmann, JH.; Meinke, S.; Mohanty, O.; Zimmermann, editors. *From Computational Biophysics to Systems Biology (CBSB07)*, NIC-Series. 36. Jülich; Germany: 2007. p. 301-304.
33. Zimmermann O, Wang L, Hansmann UHE. BETTY: Prediction of β -strand type from sequence. In *Silico Biol* 2007;7:0037.

Fig. 1. Scheme of different MC techniques: (a) multicanonical simulation, (b) parallel tempering, (c) energy landscape paving.

Fig. 2.

Time series of total energy as obtained in a MH run (for $a = 1$) at $T = 300K$ and in a regular canonical Monte Carlo simulation at the same temperature. The figure is taken from Ref. [18].

Fig. 3.
Superposition of minimum free energy structures with their native counterparts
a) beta3s, rmsd=2.22 Å
b) BBA5, rmsd=2.36 Å
c) 1RIJ, rmsd=2.36 Å

Fig. 4. Three different conformations of local minima. Values for free energy are given in kcal/mol. The figure is taken from Ref. [25]

Fig. 5.

Prediction of dihedral angle regions for 3 CASP6 targets. Color code for predictions: black = correct, white = wrong, gray = not evaluated (ambiguous or chain end). The figures are taken from Ref. [28]

a) T238, PDB-id 2BLK:A

b) T242, PDB-id 1W33:A

c) T273, PDB-id 1WDJ:A