Eur. Phys. J. D **51**, 33–40 (2009)
DOI: 10.1140/epjd/e2008-00086-2

THE EUROPEAN PHYSICAL JOURNAL D

Regular article

Computer simulation of proteins: thermodynamics and structure prediction

J.H. Meinke[1], S. Mohanty[1], W. Nadler[1], O. Zimmermann[1], and U.H.E. Hansmann[2,a]

[1] John-von-Neumann Institute for Computing, Forschungszentrum Jülich, 52425 Jülich, Germany
[2] Department of Physics, Michigan Technological University, Houghton, MI 49931, U.S.A.

Received 24 January 2008 / Received in final form 10 March 2008 / Published online 7 May 2008

Abstract. Over the last decade, computer simulations have become an increasingly important tool to study proteins. They now regularly complement experimental investigations and often are the only instrument to probe processes in the cell. Here, we summarize some of the algorithmic advances and review recent results that exemplify the progress over the last years. Our focus is on the thermodynamics and structure prediction of proteins, with information on the kinetics and dynamics inferred only indirectly.

PACS. 87.15.Cc Folding and sequence analysis – 87.15.Ak Monte Carlo simulations – 87.15.-v Biomolecules: structure and physical properties

1 Introduction

Proteins are molecular nanomachines transporting molecules, catalyzing biochemical reactions, and fighting infections. A detailed knowledge of protein structure and function is, therefore, of critical importance for understanding the molecular machinery of cells. Despite decades of research, both experimental and theoretical, the mechanisms of protein folding and interaction are still only poorly understood. Reliable tools that allow studies of these phenomena in computer experiments would open the way to understand the various diseases that are caused by misfolding of proteins, and enable the design of novel drugs with customized properties.

From a computational point of view, the difficulties arise from the complex form of the forces within and between molecules. Containing both repulsive and attractive terms these forces lead to a rough energy landscape with a huge number of local minima. A typical thermal energy of the order of k_BT is much smaller than the energy barriers a protein has to overcome. As a consequence, simple canonical Monte Carlo or molecular dynamics simulations get trapped in local minima and normally do not thermalize within the available CPU time. Even simulations of the few existing "mini-proteins" (less than 50 residues) are computationally hard tasks, and the effective computational cost increases exponentially with the number of residues for canonical methods.

The search for techniques that can overcome this sampling problem in computational protein studies is an active area of research. In the following, we describe methods that sample the high dimensional conformational space of proteins efficiently. Our focus is on the thermodynamics and structure prediction of proteins. Together with increasing computer power, which slowly approaches the petaflop range, these techniques now put first principle simulations of the thermodynamics of small proteins within reach. We review some recent results in order to demonstrate that these improved algorithms allow us to analyze the folding and aggregation of small proteins in full detail. Note that by construction our techniques only allow us to obtain indirect information on the kinetics and dynamics. With up to 50 residues, these proteins are small compared to the average size of a cellular protein (~ 250 residues). For this reason, we give an outlook how these methods can be combined with local structure constraints to predict the structure of much larger proteins and summarize some machine learning based algorithms for calculation of such constraints.

2 Algorithms for protein simulations

Efficient sampling of the conformational space of proteins requires sampling of low-energy configurations and avoiding to get trapped in local minima. The latter problem is what renders conventional Monte Carlo as well as molecular dynamics simulations at room temperature so inefficient: the crossing of an energy barrier of height ΔE is

[a] e-mail: hansmann@mtu.edu

suppressed by a factor $\propto \exp(-\Delta E/k_BT)$ (k_B being the Boltzmann constant and T the temperature of the system), and typical barriers are much larger than thermal energies. One way to achieve faster sampling is to introduce artificial weights that lead to a uniform distribution of one or more selected physical quantities in a Monte Carlo or molecular dynamics simulation. For instance, in multicanonical sampling [1] the weight $w(E)$ is set so that the distribution of energies, $P(E)$, is given by:

$$P(E) \propto n(E)w(E) \approx \text{const}, \tag{1}$$

where $n(E)$ is the spectral density. In this way, a free random walk in the energy space is performed that allows the simulation to escape from any local minimum. The thermodynamic average of a physical quantity A can now be calculated by re-weighting [2–4]:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx \mathcal{A}(x) w^{-1}(E(x))\, e^{-E(x)/k_BT}}{\int dx w^{-1}(E(x)) e^{-E(x)/k_BT}}. \tag{2}$$

Here, x stands for configurations. Note that the weights $w(E)$ are not known a priori, and estimators have to be determined. Commonly used are the iterative procedures described in references [1,5], an overview of them and other attempts is given in reference [6].

In a variant of this idea, "energy landscape paving" (ELP) [7], the search is dynamically steered away from those parts of the energy landscape that have already been explored. For this purpose, the energy is modified by a function of the time-dependent histogram. This function increases over time while the system stays in a particular minimum until the weight of the minimum has decreased sufficiently to escape from it. The system will continue the search until the next minimum is found.

$$w(\tilde{E}) = e^{-\tilde{E}/k_BT} \quad \text{with} \quad \tilde{E} = E + f(H(q,t)). \tag{3}$$

T is a (low) temperature, $\tilde{E}$ the generalized time-dependent 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., the fraction of native contacts. Eliminating the time dependence reduces ELP to other generalized-ensemble methods, for instance to multicanonical sampling for $f(H(q,t)) = \ln H(E)$.

In parallel tempering [8,9] — also known as replica exchange method — first introduced to protein science in reference [10], standard Monte Carlo or molecular dynamics moves are performed in parallel at different values of a control parameter, most often the temperature. At certain times the current conformations of replicas at neighboring temperatures T_i and $T_{j=i+1}$ are exchanged with probability

$$P(C_i \to C_j) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))), \tag{4}$$

with $\beta = 1/k_bT$. For a given replica the swap moves induce a random walk from low temperatures, where barriers lead to long relaxation times, to high temperatures, where equilibration is rapid, and back. This random walk results in a faster convergence at low temperatures.

Even with generalized-ensemble and parallel tempering techniques bottlenecks and barriers can lead to slow relaxation. In parallel tempering, convergence can be gauged by the frequency of statistically independent visits to the lowest temperature. A lower bound for this number is the rate of round-trips n_{rt} between the lowest and highest temperatures, T_0 and T_N. We define $n_{up}(i)$ ($n_{dn}(i)$) as the number of replicas at temperature T_i that came from T_0 (T_N). The fraction of replicas moving up

$$f_{up}(i) = \frac{n_{up}(i)}{n_{up}(i) + n_{dn}(i)} \tag{5}$$

describes a stationary distribution of probability flow between temperatures T_0 and T_N. Trebst et al. [11–13] have investigated this *flow* and have provided an iterative scheme for adjusting the discretization to optimize the flow distribution. Maximizing the number of round-trips n_{rt} results in a constant transition probability between neighboring nodes, and a linear flow distribution among the nodes [14]:

$$f_{up}^{opt}(i) = i/N. \tag{6}$$

Based on this result an iteration scheme can be devised for assigning control parameter values to nodes [12,13] illustrated in Figure 2 of reference [13]:

(i) an initial set of control parameters $\beta_0 > \beta_1 > ... > \beta_{N-1} > \beta_N$, e.g., inverse temperatures, gives rise to a flow distribution, e.g., $f_{up}(0) = 1 \geq f_{up}(1) \geq ... \geq f_{up}(N-1) \geq f_{up}(N) = 0$;
(ii) their values define a function $g[f]$, with $g[f(i)] = \beta_i$, in particular $g[1] = \beta_0$ and $g[0] = \beta_N$, and piecewise linear interpolation to calculate $g(x)$ for intermediate values x;
(iii) the new control parameter values are determined from this function by the relation $\beta_i' = g[1 - i/N], i = 1, ..., N-1$, keeping β_0 and β_N fixed.

If the relaxation at a particular temperature is slower than hopping in temperature, the state space partitions itself into disjoint free energy basins connected only via neighboring nodes, forming a tree-like hierarchical network (see Fig. 1). Optimizing the temperature distribution [13,14] in this case of broken ergodicity leads again to a linear flow distribution, but the acceptance probabilities are no longer constant. Similarly, one can show that weights optimizing the flow through order parameter space (for instance, energy) do not lead to a flat distribution in the case of broken ergodicity [13,14]. These results are in contrast to previous approaches based on an analysis of replica exchange acceptance rates [15–18]. They allow in addition an analytical expression of the optimal number of replicas in parallel tempering [19] and optimized replica exchange move sets for molecular dynamics [20] that go beyond the techniques previously employed [21].

The random walk of replicas is not restricted to one in temperatures. For instance, in "model hopping" [22] the

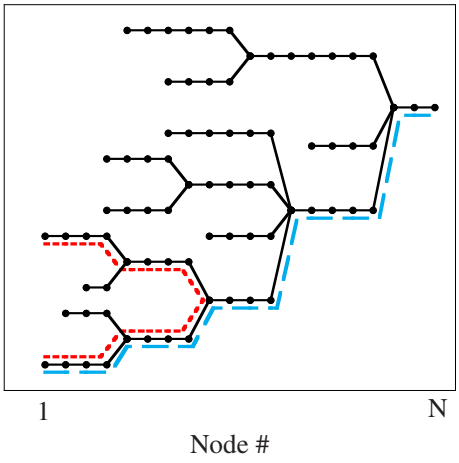

Fig. 1. (Color online) Sketch of broken ergodicity in a parallel tempering; for certain nodes (temperatures) the system partitions into several disjoint free energy wells. Flow among temperatures is necessary (blue) to obtain relaxation between the various free energy wells (e.g. red).

system performs a random work through an ensemble of energy functions. This allows to exchange information between different levels of coarse graining or various local environments. As a variant of parallel tempering (see above), "model hopping" employs N non-interacting copies of the molecule. Standard Monte Carlo or molecular dynamics moves are used for the evolution of the configurations, but the probability for the exchange of two neighboring copies is calculated as

$$\begin{aligned} P(C_i \to C_j) = \min(1, \exp\{-\beta\,[&E_A(C_j) + a_i E_B(C_j) \\ &+ E_A(C_i) + a_j E_B(C_i) - E_A(C_i) \\ &- a_i E_B(C_i) - E_A(C_j) - a_j E_B(C_j)]\}) \\ = \min(1, \exp(&\beta \Delta a \Delta E_B)). \end{aligned} \tag{7}$$

with $\Delta a = a_j - a_i$ and $\Delta E_B = E_B(C_j) - E_B(C_i)$. The random walk of the configurations is performed on a ladder of models with $a_1 = 1 > a_2 > a_3 > \ldots > a_N$ that differ by the relative contribution of E_B to the total energy E of the molecule. The van der Waals repulsion between close or even overlapping atoms frequently leads to high barriers in the energy landscape of proteins [23]. Hence, we have considered an implementation of "model hopping" with decreasing contributions from the van der Waals energy. With "model hopping" the (non-physical) model at one end of the ladder (at $a_N << 1$) may allow atoms to share the same position in space thereby "tunneling" through an energy barrier. Hence, at the "physical" end of the ladder (at $a_1 = 1$) the sampling of low-energy configurations is increased. Using this technique we were able to "predict" the structure of protein A by an all-atom simulation with an accuracy of 3.2 Å rmsd [22]. For a comparison with other methods see also, for instance, references [24–28]. Model hopping also allows guiding a simulation by information obtained from homologous structures [29]. Such spatial constraints can introduce an additional roughness into the energy landscape and often leads to extremely slow convergence of the simulation. This problem is circumvented in model hopping through a random walk in an ensemble of replicas that differ by the strength with that the constraints are coupled to the system.

3 Parallel implementation of a force field

With generalized-ensemble sampling, replica exchange techniques and related methods the numerical effort in simulations of small proteins is expected to increase no longer exponentially with number of residues, but only with a power law. Under optimal circumstances, the computational effort in generalized-ensemble algorithms scales $\propto \hat{X}^2$ where $\hat{X}$ is the range in the ensemble coordinate X. For instance, in the multicanonical algorithm, this coordinate is the potential energy $X = E$. Since $E \propto N^2$, the computational effort increases in multicanonical simulations with the number of residues at least as $\approx N^4$ [30]. This scaling clearly limits the size of proteins and protein complexes that can be studied. Hence, the above described simulation techniques need to be implemented in software that utilizes efficiently the computational power of a few thousand processors commonly available in today's supercomputers. An example is the protein simulation package SMMP [31–33], which the authors now run regularly on 4096 processors of the IBM BlueGene/L JUBL installed at the Jülich Supercomputing Centre (JSC).

Our investigation into strategies for parallelization of force fields have focused on ECEPP/3 [34] defined by

$$\begin{aligned} E_{ECEPP/3} &= E_C + E_{LJ} + E_{HB} + E_{Tor} \\ &= \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}} \\ &\quad + \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ &\quad + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ &\quad + \sum_{l} U_l (1 \pm \cos(n_l \xi_l)) \;, \end{aligned} \tag{8}$$

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. The protein-water interaction is approximated by a solvent accessible surface term

$$E_{\text{solv}} = \sum_i \sigma_i A_i \;. \tag{9}$$

The sum goes over the solvent accessible areas A_i of all atoms i weighted by solvation parameters σ_i as determined in [35], a common choice when the ECEPP/3 force field is utilized. Note that E_{solv} is a rather crude approximation

of the interaction between the polypeptide and the surrounding water motivated by the low computational costs compared to simulations with explicit water molecules. The energy $E_{\mathrm{ECEPP/3}}$ from equation (8) involves sums over all atom pairs since the definition of the ECEPP/3 force field does not include any cutoffs plus a sum over the dihedral angles in the protein. Each term in these sums can be calculated independently. In SMMP, each atom is associated with a dihedral angle. This makes it natural to use a loop over the dihedral angles as outermost loop in the implementation of the energy calculation and distribute the work across processors based on the number of interactions associated with each dihedral angle. The solvent energy term E_{solv} in equation (9), on the other hand, is defined by adjacency relations in space and calculated by dividing up space into boxes and distributing the work based on the spatial grid.

We ran our benchmark on 4 different platforms: JUMP, an IBM p690 cluster with 32 processors and 112 GB of shared memory per node; JUBL, an IBM BlueGene/L with 8 racks and a total of 16384 Power 4 processor at 700 MHz; JULI a PC cluster using dual-core PowerPC 970 MP processors at 2.5 GHz with an InfiniPath network, and NICOLE, an Opteron based PC cluster with a clock speed of 2.4 GHz using Infiniband networking. Except for the setup of the communicators used for the energy calculation on BG/L, we used the same source code for all measurements. We performed 50 sweeps of a Monte Carlo simulation of the designed protein TOP7 [36] starting from a stretched chain. Data was written to disk every 10 sweeps. On JUBL, we used multiple replicas in parallel with the indicated number of processors per replica to fill a half plane (512 processors).

Using a single processor, the fastest system in the test (JULI, 18 min) finished the benchmark more than 6 times faster than JUBL (2 h). Using the maximum number of processors that still lead to an increase in speed, JUMP becomes the fastest system finishing the calculation in 80 s using 64 processors almost 3.5 times faster than JUBL using 64 processors per replica as well. For the PC clusters scaling breaks down before reaching 64 processors as the overhead due to communication becomes too large. On all of our test systems, the speedup is much better for the calculation of $E_{\mathrm{ECEPP/3}}$ than the speedup of the calculation of the solvent energy.

To run multiple replicas in parallel, we divide the processors into groups with each group responsible for the energy calculation of a single replica. The number of replicas is only limited by the number of available processors. Because of that, the low cost per processor makes BlueGene/L an attractive platform for protein simulations. With the large number of processors available on JUBL, we can run simulations with 64 replicas at a quarter of the cost and at the same speed as on JUMP.

4 Folding of small proteins

Modern simulation techniques together with experimental advances have led to a unified picture of protein folding as a stochastic process in a high dimensional energy landscape. The native 3D structure of a given protein is the global minimum of free energy at physiological temperatures. Each molecule finds a path to fold into the native state, sampling only a small part of the astronomically large conformation space [37–41]. Recent evaluations of protein energy landscapes are reported, for instance, in references [42–45].

Following the experimental observation of a correlation between the topological complexity of the native state and the folding rates of a wide variety of proteins [46], it is widely believed that the final structure determines the rate and mechanism of the folding transition. The more complex the native structure of a protein is the slower it folds. In reference [46], this complexity was captured by the "relative contact order" of a protein, which is the average sequence separation between residues in contact in the native state, normalized by the chain length. Proteins with mostly helical structures have small contact orders whereas complex β-sheet structures can have relatively high contact orders. The observed correlation between relative contact order and folding rates persists over 6 orders of magnitude in the folding rates. A further indication of the importance of the native state topology comes from the observation that diverse proteins with similar structures, but little sequence similarity, have comparable folding rates [47].

Another experimental measurable used to characterise the process of protein folding is the distribution of structures in the so called transition state [48]. This is the state through which the protein has to pass from the unfolded state with high entropy to the native state with low entropy and energy. The distribution of structures in transition states can be deduced from the effect of site specific mutations on the folding rate (see also [49], and references therein). Mutations to residues which contribute to crucial stabilizing structures in the transition state have large effects on the folding rate, whereas mutations at sites which are disordered in the transition state have little effect. Structures in the transition state ensemble are insensitive to large differences in the sequences, if the native topologies are similar [49]. Note however that there are many examples of proteins with similar structures, sometimes even of similar sequences, that fold using different pathways [50].

Consistent with the expectation that small contact orders mean simple folding pathways and faster folding, most successful atomistic folding simulations have been for helical proteins. Using completely unrestrained all-atom molecular dynamics simulations with the AMBER force field, folding of the 20 residue helical peptide 1L2Y (trp cage) was reproduced in detail [51]. In our own computational studies, we have found that small helical proteins (such as the 23 residue 1RIJ) indeed exhibit simple funnel like folding free-energy landscapes [52]. The helix hydrogen bonds form in no particular order, although the two ends of a helix show greater tendency to dissolve and reform.

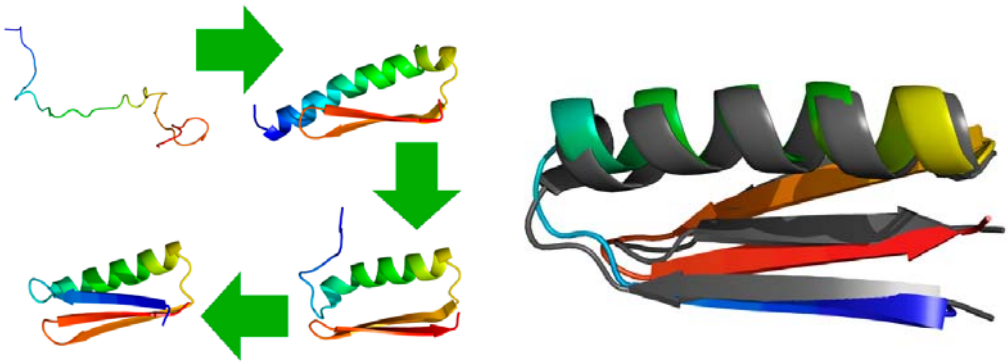

Fig. 2. (Color online) Left: The non-trivial unexpected pathway for the folding of the molecule CFr revealed in the simulations. Right: The free-energy minimum structure seen in all-atom simulations (coloured) superimposed on the experimentally measured structure (gray).

On average, β-sheet structures have larger contact order, and correspondingly fold slower. Among these, β-hairpins are the simplest, as the hydrogen bonded residues are close in sequence. Many simulations of small peptides exist that fold into β-hairpins, for instance (not a comprehensive list!), references [53–56]. These studies have shown two main folding mechanisms. For the 9 residue β-hairpin peptide YQNPDGSQA, explicit water molecular dynamics simulations with the AMBER force field showed folding is initiated by interactions between the two arms of the hairpin before the backbone hydrogen bonds stabilize the structure [56]. For the 3-stranded β-sheet beta3s, folding proceeds in a zipper mechanism from the turns towards the ends of the hairpins [52]. The folding of the two β-hairpins is more cooperative than the folding of helices, and once formed, the β-sheets show a greater resilience towards unfolding. While simulations based on Monte Carlo methods do not allow a direct temporal interpretation of the observed folding trajectories, qualitative inferences can be drawn from the data. For the same length of the peptides, it is our experience that a much greater number of independent folding events are observed for helical peptides compared to β-sheets of the same size in the same total Monte Carlo time. This indicates that structures with small contact order are more easily available, consistent with the experiments.

Proteins with both helix and β-sheet elements pose a special challenge to all-atom sequence based models. The simplest examples have a structure with a helix and a β-hairpin. We have examined two such systems and found distinct folding mechanisms. The 23 residue BBA5 molecule has a small β-hairpin where the turn region is stabilized by a synthetic amino acid D-proline. This hairpin and the helix of BBA5 form on their own, and only later make hydrophobic contacts [52]. The protein FSD-EY has a similar hairpin structure, however, the hydrophobic residues of the helix line up on one side, providing a template around which the hairpin forms [57]. The hairpin of FSD-EY never forms independently of the helix in our simulations, nor is there any clear evidence of a zipper like mechanism.

The formation of structures with non-local β-sheet contacts are highly non-trivial. While one part of the chain is synthesized it can not find its binding partner before that part is also synthesized. In the meantime there is a danger that the first-formed β-strand interacts with nearby molecules leading to potentially harmful aggregates of incompletely folded proteins. We have done extensive folding simulations of one such molecule, the C-terminal fragment CFr (PDB id: 2 GJH) of the designed 93 residue protein Top7 (PDB id: 1 QYS). The protein folds to about 1.8 Å backbone rmsd from the native state in all-atom parallel tempering Monte Carlo simulations starting from random initial conformations. The contact order of CFr puts it near the region of intermediate contact orders in the contact order vs. fold rate plot [46,49]. Along the sequence from N- to C-terminus, the secondary structure profile of the molecule CFr is: strand – helix – strand – strand (see Fig. 2). The two strands at the C-terminus make a β-hairpin. The strands at the N- and C-termini are also adjacent in the 3 stranded β-sheet. None of the simple folding mechanisms discussed above could give rise to this arrangement.

Our simulations revealed an unanticipated mechanism (see Fig. 2) for folding of this structure [58]. The N-terminal β-strand first folds into a non-native extension of the native helix. The β-hairpin at the C-terminus forms independently. When the helix and the C-terminal hairpin make the correct tertiary contacts, the non-native part of the helix unfolds to release the N-terminal residues. These subsequently form β-sheet contacts with the hairpin and complete the native structure. By “caching” the residues of the N-terminal β-strand as a non-native extension of a helix, the molecule protects them from premature contacts with other regions with strong β-strand propensities that would lead to misfolding or very slow folding. The caching of the N-terminal strand, accelerates folding of CFr by avoiding many misfolded states. We speculate that this is a common mechanism in molecules where adjacent strands in a β-sheet have large sequence separation. It can also protect a nascent N-terminal β-strand which is synthesized early, from intermolecular interactions that

could lead to aggregation, until the rest of the molecule is synthesized and properly folded.

The folding of CFr brings the possible role of non-native interactions in protein folding into focus. Experimental [59–61] as well as computational studies with simplified models [62,63] have indicated the presence of non-native α-helical structures early in the folding process of predominantly β-sheet proteins. The simulations of CFr provide a detailed picture of how such non-native interactions [64] might arise spontaneously and channel the folding pathway. They indicate that some proteins with large contact order fold faster than other with similar complexity by utilizing an accelerating mechanism such as caching mechanism. This might be a cause of the large fluctuations observed in the contact order vs. fold rate plot in reference [46], in the mid-contact order proteins.

An interesting question is whether there are different distinct transitions in the folding process, and what their thermal order and relation is. An important example is the role of side-chain ordering. Generalized ensemble and parallel tempering methods are well-suited to overcome the "slowness" of this process observed in previous canonical simulations [65,66]. In recent studies on homopolymers [67,68], a de-coupling of backbone and side-chain ordering was found for certain amino acids. Important characteristics of the side chain ordering process do not depend on the details of the environment, i.e., whether the molecules are in gas phase or solvent, but solely on the particular side groups. Side-chain ordering exhibits a transition-like character, marked by an accompanying peak in the specific heat. In a related investigation, the role of charged end groups in stabilizing and de-stabilizing secondary structures in gas phase was established [69]. These latter results are important for a comparison of simulations with molecular beam experiments of biological peptides [70–74].

For the villin headpiece subdomain HP-36 (PDB ID: 1VII), one of the smallest proteins (596 atoms) with well-defined secondary and tertiary structure [75], our results indicate a thermal hierarchy of ordering events with side-chain ordering appearing at temperatures below the helix-coil transition, i.e., secondary structure formation, but above the final folding transition to the native state [76]. We conjecture that side-chain ordering facilitates the search for the correct backbone topology. This assumption is consistent with various computational [77–79] and experimental [80] studies that also identify the formation of helical segments as the time limiting factor in the folding of HP-36.

Simulations of FSD-EY and HP-36 described in this section were carried out using the SMMP package [33] and the ECEPP force field [34] described in Section 2 and aimed also at testing our parallel implementation of the force field. Simulations of 1RIJ, beta3s, and Top7 CFr were done with the PROFASI package [81] and the Lund force field [82]. As in ECEPP, the bond lengths and angles are fixed in the Lund force field, but simplified interaction terms are assumed that allow a faster energy calculation (on single processors): a purely repulsive excluded volume term, a local backbone electrostatic term, a directional hydrogen bond term, and a pairwise additive hydrophobic term to approximate the effects of the solvent.

5 Constraint generation for structure prediction

Structure prediction is concerned with the search for the global minimum of a protein folding energy landscape. Atomistic folding simulations with physics based force fields for average size proteins ($\approx$ 250 residues) are computationally far beyond the capabilities of even the largest supercomputers of the day.

For this reason, most structure prediction methods successful in blind tests such as CASP [83] are based on non-physical approaches. Several methods use fragments of native structures to assemble native like trial structures. These are either selected from large libraries of short representative fragments [84] or from those parts of template structures that can be aligned to the target sequence [85]. Often, knowledge-based potentials are used where propensities, derived from statistics of native structures, are translated into pseudo energies by means of Boltzmann's law [86,87]. These pseudo energy landscapes can be explored with similar search techniques as those described above for physical energies [88]. The so sampled structures can be evaluated in a second step by physics-based potentials [89].

A common strategy to reduce the conformational search space is the use of constraints derived from the experimental structures of related proteins. The low and decreasing rate of novel folds among newly resolved structures suggests that the protein fold space explored by nature is limited to a few thousand classes [90] and there is evidence that the 45 000 entries of the protein data bank (PDB) contain a structural representative for almost every fold on the single domain level [91]. It is thus likely that for any given target sequence a suitable structure template is available. Successful fold recognition, i.e., identifying the structure templates that best fit the target sequence and sequence-structure alignment are thus essential prerequisites to derive useful template based constraints. 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 [92].

Long range distance constraints, i.e., prediction of positions close in structure but far apart in sequence, can be derived from fold recognition or correlated mutation analysis. As these constraints have a rather low accuracy, and errors are more likely to lead to frustrated conformations, local constraints are preferred. The most-widely used prediction algorithms generate local constraints from secondary structure. Usually they predict the membership of a protein residue to one of 3 classes: α-helix, β-sheet or coil, where coil is simply defined as absence of both α-helix and β-sheet. Although algorithms like

PSIPRED [93] achieve an accuracy beyond 75%, mapping of these classes to dihedral angle constraints is not straightforward as most programs use the secondary structure definition of the program DSSP [94] as a reference. DSSP however defines secondary structure classes in terms of H-bond patterns and thus largely ignores dihedral angles. As a consequence only for about half of the residues, i.e., those located in the non-terminal parts of helices and beta strands, a direct mapping to a narrow region in the dihedral space is possible.

An alternative strategy is the direct prediction of dihedral angle values for each residue [95–97]. For instance, DHPRED predicts for each residue its dihedral angle region [97]. Although this prediction is again a 3-class classification it provides dihedral information independent of an H-bond centered definition. As most coil residues are members of the alpha or beta dihedral angle region (only 7% of all residues lie outside), the number of informative constraints is much higher. The algorithm has a three layer structure and is based on classification by Support-Vector-Machines (SVM) [98], a supervised machine learning algorithm. The first layer uses the sequence profile information from a PSI-BLAST alignment to obtain an initial dihedral angle region prediction for each residue. In the second layer this initial classification is added to the sequence profile information for an iterative update of the dihedral angle region predictions. After convergence a nearest neighbor heuristics is applied in the last layer to resolve remaining ambiguities. The performance of DHPRED is comparable to PSIPRED and provides a direct mapping to dihedral angles. In addition the dihedral angle regions for many coil residues can be identified thus providing local dihedral constraints for the entire protein chain. An extension of these ideas is the program BETTY [99] that classifies 88% of all β-residues correctly as parts of a parallel anti-parallel β-sheet. Combined with PSIPRED, 79.3% of all residues can be correctly classified by BETTY into parallel-β, anti-parallel-β, α-helix, and coil.

6 Conclusions

Progress in hardware and algorithm development now allow the physics based simulation of biological macromolecules. For small proteins atomistic simulation of the entire folding process has become possible. The thermodynamics of the protein folding process can be revealed and used for explanation of experimental observations from first principles. The increasing accuracy of sequence based prediction methods to obtain local structural constraints suggests a growing range of applications for biased simulations in structure prediction.

This work was supported in part by the National Institutes of Health (USA) Grant GM62838 and National Science Foundation (USA) Grant CHE-0313618. Most calculations were done on computers of the John von Neumann Institute for Computing, Research Centre Jülich, Germany.

References

1. B.A. Berg, T. Neuhaus, Phys. Lett. B **267**, 249 (1991)
2. Z.W. Salsburg, J.D. Jacobson, W. Fickett, W.W. Wood, J. Chem. Phys. **30**, 65 (1959)
3. G.M. Torrie, J.P. Valleau, Chem. Phys. Lett. **28**, 578 (1974)
4. A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. **63**, 1603(E) (1089), and references given in that erratum
5. U.H.E. Hansmann, Y. Okamoto, Physica A **212**, 415 (1994)
6. B.A. Berg, Comput. Phys. Commun. **153**, 397 (2003)
7. U.H.E. Hansmann, L.T. Wille, Phys. Rev. Lett. **88**, 068105 (2002)
8. C.J. Geyer, A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995)
9. K. Hukushima, K. Nemoto, J. Phys. Soc. (Jpn) **65**, 1604 (1996)
10. U.H.E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997)
11. S. Trebst, D.A. Huse, M. Troyer, Phys. Rev. E **70**, 046701 (2004)
12. H.G. Katzgraber, S. Trebst, D.A. Huse, M. Troyer, J. Stat. Mech. Theor. Exp. P03018 (2006)
13. S. Trebst, M. Troyer, U.H.E. Hansmann, J. Chem. Phys. **124**, 174903 (2006)
14. W. Nadler, U.H.E. Hansmann, Phys. Rev. E **75**, 026109 (2007)
15. D.A. Kofke, J. Chem. Phys. **117**, 6911 (2002); Erratum: J. Chem. Phys. **120**, 10852 (2004)
16. C. Predescu, M. Predescu, C. Ciabanu, J. Chem. Phys. **120**, 4119 (2004)
17. D.A. Kofke, J. Chem. Phys. **121**, 1167 (2004)
18. A. Kone, D.A. Kofke, J. Chem. Phys. **122**, 206101 (2005)
19. W. Nadler, U.H.E. Hansmann, Phys. Rev. E **76**, 065701R (2007)
20. W. Nadler, U.H.E. Hansmann, Phys. Rev. E **76**, 057102 (2007)
21. Y. Sugita, Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999)
22. W. Kwak, U.H.E. Hansmann, Phys. Rev. Lett. **95**, 138102 (2005)
23. J. Chen, W. Im, C.L. Brooks III, J. Comp. Chem. **26**, 1565 (2005)
24. J.M. Carr, D.J. Wales, J. Chem. Phys. **123**, 234901 (2005)
25. M. Nanias, M. Chinchio, J. Pillardy, D.R. Ripoll, H.A. Scheraga, Proc. Natl. Acad. Sci. USA **100**, 1706 (2003)
26. J. Pillardy, C. Czaplewski, A. Liwo, W.J. Wedemeyer, J. Lee, D.R. Ripoll, P. Arlukowicz, S. Oldziej, Y.A. Arnautova, H.A. Scheraga, J. Phys. Chem. B **105**, 7299 (2001)
27. A. Verma, A. Schug, K.H. Lee, W. Wenzel, J. Chem. Phys. **124**, 044515 (2006)
28. A. Verma, W. Wenzel, J. Phys. Cond. Mat. **19**, 285213 (2007)
29. D. Gront, A. Kolinski, U.H.E. Hansmann, Int. J. Quant. Chem. **105**, 826 (2005)
30. U.H.E. Hansmann, Y. Okamoto, J. Chem. Phys. **110**, 1267 (1999)
31. F. Eisenmenger, U.H.E. Hansmann, S. Hayryan, C.-K. Hu, Comput. Phys. Commun. **138**, 192 (2001)
32. F. Eisenmenger, U.H.E. Hansmann, S. Hayryan, C.-K. Hu, Comput. Phys. Commun. **174**, 422 (2006)

33. J.H. Meinke, S. Mohanty, F. Eisenmenger, U.H.E. Hansmann, Comput. Phys. Commun. DOI:10.1016/j.cpc.2007.11.004.
34. G. Nemethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H.A. Scheraga, J. Phys. Chem. **96**, 6472 (1992)
35. T. Ooi, M. Oobatake, G. Nemethy, H.A. Scheraga, Proc. Natl. Acad. Sci. USA **84** 3086 (1987)
36. B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Science **302**, 1364 (2003)
37. J.D. Bryngelson, P.G. Wolynes, Proc. Natl Acad. Sci. USA **84**, 7524 (1987)
38. J.N. Onuchic, Z. Luhey-Schulten, P.G. Wolynes, Ann. Rev. Phys. Chem. **48**, 545 (1997)
39. K.A. Dill, H.S. Chan, Nat. Struct. Biol. **4**, 10 (1997)
40. E.I. Shakhnovitch, Curr. Opin. Struc. Biol. **7**, 29 (1997)
41. T. Veitshans, D. Klimov, D. Thirumalai, Fold. Des. **2**, 1 (1997)
42. T. Herges, W. Wenzel, Structure **13**, 661 (2005)
43. P. Pokarowski, A. Kolinski, J. Skolnick, Biophys. J. **84**, 1518 (2003)
44. A. Liwo, P. Arukowicz, C. Czaplewski, S. Oldziej, J. Pillardi, H.A. Scheraga, Proc. Natl. Acad. Sci. USA **99**, 1937 (2002)
45. D.A. Evans, D.J. Wales, J. Chem. Phys. **121**, 1080 (2004)
46. K.W. Plaxco, K.T. Simons, D. Baker, J. Mol. Biol. **277**, 985 (1998)
47. D.E. Kim, H. Gu, D. Baker, Proc. Natl. Acad. Sci. USA **95**, 4982 (1998)
48. A. Matouschek, J. Kellis Jr, L. Serrano, A.R. Fersht, Nature **340**, 122 (1989)
49. D. Baker, Nature **405**, 39 (2000)
50. A. Zarrine-Afsar, S.M. Larson, A.R. Davidson, Curr. Opin. Struct. Biol. **15**, 42 (2005)
51. C. Simmerling, B. Strockbine, A.E. Roitberg, J. Am. Chem. Soc. **124**, 11258 (2002)
52. S. Mohanty, U.H.E. Hansmann, Biophys. J. **92**, 3573 (2006)
53. R. Zhou, B.J. Berne, R. Germain, Proc. Natl. Acad. Sci. USA **98**, 14931 (2001)
54. W.Y. Yang, J.W. Pitera, W.C. Swope, M. Gruebele, J. Mol. Biol. **336**, 241 (2004)
55. W. Wenzel, Europhys. Lett. **76**, 156 (2006)
56. X. Wu, B.R. Brooks, Biophys. J. **86**, 1946 (2004)
57. S. Mohanty, U.H.E. Hansmann, Phys. Rev. E **76**, 1539 (2007)
58. S. Mohanty, J. Meinke, O. Zimmermann, U.H.E. Hansmann, Proc. Natl. Acad. Sci. USA, DOI: 10.1073/pnas.0708411105
59. D. Hamada, S. Segawa, Y. Goto, Nat. Struct. Biol. **3**, 868 (1996)
60. K. Kuwata, M. Hoshino, S. Era, C.A. Batt, Y. Goto, J. Mol. Biol. **283**, 731 (1998)
61. K. Kuwata, R. Shastry, H. Cheng, M. Hoshino, C.A. Batt, Y. Goto, H. Roder, Nat. Struct. Biol. **8**, 151 (2001)
62. G. Chikenji, M. Kikuchi, Proc. Natl. Acad. Sci. USA **97**, 14273 (2000)
63. G. Chikenji, Y. Fujitsukab, S. Takada, Chem. Phys. **307**, 157 (2004)
64. S.S. Plotkin, Protein Struct. Func. Genet. **45**, 337 (2001)
65. J. Shimada, E. Kussell, E.I. Shakhnovich, J. Mol. Biol. **308**, 79 (2001)
66. E. Kussell, E.I. Shakhnovich, Phys. Rev. Lett. **89**, 168101 (2002)
67. Y. Wei, W. Nadler, U.H.E. Hansmann, J. Chem. Phys. **125**, 164902 (2006)
68. Y. Wei, W. Nadler, U.H.E. Hansmann, J. Phys. Chem. B **111**, 4244 (2007)
69. Y. Wei, W. Nadler, U.H.E. Hansmann, J. Chem. Phys. **126**, 204307 (2007)
70. P. Dugourd, R.R. Hudgins, D.E. Clemmer, M.F. Jarrold, Rev. Sci. Instrum. **68**, 1122 (1997)
71. R.R. Hudgins, J. Woenckhaus, M.F. Jarrold, Int. J. Mass. Spectrom. Ion. Proc. **165/166**, 497 (1997)
72. M. Kohtani, J.E. Schneider, T.C. Jones, M.F. Jarrold, J. Am. Chem. Soc. **126**, 16981 (2004)
73. M. Kohtani, M.F. Jarrold, J. Am. Chem. Soc. **126**, 8454 (2004)
74. B.S. Kinnear, D.T. Kaleta, M. Kohtani, R.R. Hudgins, M.F. Jarrold, J. Am. Chem. Soc. **122**, 9243 (2000)
75. C.J. McKnight, P.T. Matsudaira, P.S. Kim, Nat. Struct. Biol. **4**, 180 (1997)
76. Y. Wei, W. Nadler, U.H.E. Hansmann, J. Chem. Phys. **128**, 025105 (2008)
77. H. Lei, C. Wei, H. Liu, Y. Duan, Proc. Natl. Acad. Sci. USA **104**, 4930 (2007)
78. G.M.S. De Mori, G. Colombo, M. Micheletti, Proteins **58**, 459 (2005)
79. G. Jayachandran, V. Vishal, V.S. Pande, J. Chem. Phys. **124**, 164902 (2006)
80. J. Kubelka, W.A. Eaton, J. Hofrichter, J. Mol. Biol. **329**, 625 (2003)
81. A. Irbäck, S. Mohanty, J. Comput. Chem. **27**, 1548 (2006)
82. A. Irbäck, S. Mohanty, Biophys. J. **88**, 1560 (2005)
83. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, A. Tramontano, Proteins **69**, 3 (2007)
84. K.T. Simons, C. Kooperberg, E. Huang, D. Baker, J. Mol. Biol. **268**, 209225 (1997)
85. Y. Zhang, Proteins **S8**, 108 (2007)
86. A. Kolinski, Acta Biochim. Pol. **51**, 349 (2004)
87. Y. Zhang, A. Kolinski, J. Skolnick, Biophys. J. **85**, 1145 (2003)
88. Y. Zhang, D. Kihara, J. Skolnick, Proteins **48**, 192 (2002)
89. A. Verma, W. Wenzel, *From Computational Biophysics to Systems Biology (CBSB07)*, edited by H.E. Hansmann, J. Meinke, S. Mohanty, O. Zimmermann, NIC Series, Vol. 36 (2007) pp. 283-286
90. Y.I. Wolf, N.V. Grishin, E.V. Koonin, J. Mol. Biol. **299**, 897 (2000)
91. Y. Zhang, I.A. Hubner, A.K. Arakaki, E. Shakhnovich, J. Skolnick, Proc. Natl. Acad. Sci. USA **103**, 2605 (2006)
92. W.W. Chen, J.S. Yang, E.I. Shakhnovich, Proteins **66**, 682 (2007)
93. D.T. Jones, J. Mol. Biol. **292**, 195 (1999)
94. W. Kabsch, C. Sander, Biopolymers **22**, 2577 (1983)
95. M.J. Wood, J.D. Hirst, Proteins **59**, 476 (2005)
96. O. Dor, Y. Zhou, Proteins **68**, 76 (2007)
97. O. Zimmermann, U.H.E. Hansmann, Bioinformatics **22**, 3009 (2006)
98. B. Schökopf, A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA, 2002)
99. O. Zimmermann, L. Wang, U.H.E. Hansmann, In Silico Biol. **7**, 0037 (2007)