STATEMENT OF RESEARCH INTERESTS

Ulrich H.E. Hansmann

A) Overview

Computational science has extended the range of phenomena that can be investigated within the framework of physics. Complex systems such as spin glasses or neural networks, are hampered by common problems and can be studied with similar techniques. My research focuses on a biological example, the physics of proteins. These macromolecules are key components in the molecular machinery of cells, for instance catalyzing biochemical reactions. A detailed understanding of folding and interaction of proteins will lead to new insights into the molecular working of cells, as needed in many medical and biotechnological applications.

While simulations can complement experiments in probing folding, aggregation, binding and other fundamental processes in cells, they are computationally challenging for realistic protein models. This is because all-atom models lead to a rough energy landscape with a vast number of local minima separated by high barriers [1]. For a typical single-domain protein such as the 153 amino-acid myoglobin the task becomes impractical. On a supercomputer capable of trillions of floating point operations per second a single folding trajectory of $\approx 10^{-3}$s would take years with molecular dynamics simulations [2].

A significant part of my research is concerned with overcoming this bottleneck. Its center piece is the development and advancement of numerical techniques such as the *generalized-ensemble* approach [3] with the final goal of folding domains in proteins (usually on the order of 50-200 residues). Related is the development and dissemination of new software. Our programs are collected in the free program package SMMP (**S**imple **M**olecular **M**echanics for **P**roteins) [4]. This method oriented research is supported by the National Science Foundation (NSF) under contract CHE-0809002.

Current applications of our techniques focus on carefully chosen proteins ranging from the 28-residue Fsd-Ey up to the 93-residue TOP7 probing the mechanism of folding in small proteins and the conditions under which proteins misfold and aggregate (implicated in the outbreak of neurological diseases). Protein-ligand binding and protein interaction networks belong to the same research direction and provide an interface for collaborations with experimental groups. This application oriented research is supported by the National Institutes of Health (NIH) under contract GM62838.

B) Background: Generalized-Ensemble Sampling and Related Techniques

The key idea behind generalized-ensemble based techniques is to replace the canonical simulations, where the 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), with schemes that both ensure sampling of low-energy configurations *and* avoid trapping in local minima. For instance, in multicanonical sampling [5] the weight $w(E)$ leads to a distribution

$$P(E) \propto n(E)\, w_{mu}(E) = \text{const}, \tag{1}$$

with $n(E)$ the density of states. A free random walk in the energy space is performed that allows the simulation to escape from any local minimum. From this simulation one can calculate the thermodynamic average of any physical quantity A by re-weighting: [6]

$$< \mathcal{A} >_T \;\; = \;\; \frac{\int dx \; \mathcal{A}(x) \; w^{-1}(E(x)) \; e^{-E(x)/k_B T}}{\int dx \; w^{-1}(E(x)) \; e^{-E(x)/k_B T}} \; , \tag{2}$$

where x labels the configurations. Note that the weight $w(E)$ is not *a priori* known and estimators have to be determined by an iterative procedure described in Refs [5, 7].

We have introduced an optimization technique called Energy Landscape Paving (ELP) [8] that relies on a modified energy expression steering the search away from regions already explored:

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

Here, $\tilde{E}$ is an "effective" energy, and $f(H(q,t))$ is a function of the histogram $H(q,t)$ in a pre-chosen "order parameter" q. The weight of a local minimum decreases with the time the system stays in that minimum till it is no longer favored, and the system continues it search. For $f(H(q,t)) = f(H(q))$ the method reduces to the various generalized-ensemble methods [3] (for instance for $f(H(q,t)) = \ln H(E)$ to multicanonical sampling).

In parallel tempering (also known as replica exchange method) [9], first introduced to protein folding by me in Ref. [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 a probability

$$w(\; \mathbf{C}^{old} \rightarrow \; \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) \; . \tag{4}$$

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 results in a faster convergence at low temperatures.

For both Monte Carlo [11] and molecular dynamics [12] we have demonstrated that generalized-ensemble based techniques are superior in locating low-energy conformers [7,13]. For a critical evaluation of these now widely used methods, see Ref. [14].

C) Current work

In the past, my coworkers and I have demonstrated that the generalized-ensemble approach allows de novo folding simulations of peptides and proteins with up to $30 - 50$ residues [15–18]. My group continues to further advance these techniques, with the goal of enabling folding domains in proteins (usually consisting of 50-200 amino acids). One avenue is to improve the computational efficiency of these techniques [19,20] which is often below the theoretical optimum. A second research direction focuses on identifying "order parameters" or "reaction coordinates" [21] that allow tailoring of generalized ensembles which are most suitable for the simulation of proteins or classes of proteins. Many of our algorithms [8, 10, 16] are implemented in the freeware program package SMMP [4] available from either the program library of *Computer Physics Communications* or directly from the authors (*www.hansmann-lab.com/cbpc/smmp/smmp.php*).

Our algorithms are tested on carefully chosen proteins ranging from the 28-residue FsdEy to the 93-residue TOP7. The data obtained in these simulations are used further to create new analysis techniques, and to examine the limitations set by energy functions and solvent models. For instance, my co-workers and I have introduced partition function zeros analysis and the measurement of the fractal dimension of energy landscapes as tools for characterizing transitions in biomolecules [22, 23]. We also investigate how the distribution of low-energy states depends on the solvent model. By separating the effects of intramolecular interactions and solvation, such research allows one to study the extent that folding is determined by intrinsic properties of the protein [24].

An example of our investigations into the folding mechanism of proteins is our study of the 49-residue protein CFr which is characterizes by an end-to-end β-sheet. As the N-terminal β-strand is synthesized early on, but cannot bind to the C-terminus before the chain is fully synthesized, there must be a mechanism by which the N-terminal β-strand avoids interaction with other parts of the chain or nearby molecules that may lead to misfolding and aggregation. Our simulations [17, 25] indicate that this risk is avoided by a mechanism that relies on the chameleon behavior of one of the terminal β-strands to facilitate folding. The N-terminal residues are cached transiently in a helix, preventing the premature formation of contacts with other molecules. Only after formation and proper arrangement of the remaining protein, these residues refold into the third strand of a β-sheet, completing the native structure. We are probing now whether this mechanism exists also in proteins with similar fold such as ferredoxins. Proteins with ferredoxin-like fold are not only involved in the production of hydrogen gas and heavy metal detoxification in bacteria, but also are connected with copper-deficiency and related diseases in humans.

For many proteins, the biologically active structure is not only determined by the sequence of amino acids but also by the interaction with other proteins. Such environment-modulated structural changes include ligand-binding and chaperone-assisted folding of proteins, but also the autocatalysis and aggregation of mis-folded proteins. The latter case is particullarly interesting as protein misfolding and aggregation are implicated in a number of illnesses such as Huntington's, spongiform encephalopathies (prion-mediated), or most common, Alzheimer's disease [26]. Simulations of the conformational transitions in these proteins and their subsequent aggregation may contribute to understanding of the disease mechanisms at a molecular level. Results of our preliminary investigations can be found in Ref. [27–29].

Self-assembly of protein complexes and protein-ligand binding are other examples for the interactions between proteins and other molecules that are studied in my group. For instance, the CFr monomer leaves a β-strand with several hydrophobic residues exposed. This suggests immediate dimerization explaining why only dimers are observed in experiments. We have tested this hypothesis and found the energy difference between the isolated monomers and a typical dimer to be of the order of 30 kcal/mol [17]. We are now extending this research to the self-assembly of protein complexes such as the 84-residue homotetrameric BBAT2. The complexity of systems of interacting proteins will require further development of methods and algorithms, directed toward the long-term goal of simulating cells on all time and length scales relevant for medical and technological applications. The algorithmic advances will be also relevant in other areas of nanophysics. As an example, we are studying now the use of proteins for sorting carbon-nanotubes.

References

[1] U.H.E. Hansmann, *Comp. Sci. Eng.* **5** (2003) 64.

[2] F. Allen et al., *IBM Systems Journal* **40** (2001) 310.

[3] O. Zimmermann and U.H.E. Hansmann, *BBA - Proteins and Proteomics*, **1784** (2008) 252.

[4] F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **138** (2001) 192; *Comp. Phys. Comm.* **174** (2006) 422; J. H. Meinke, S. Mohanty, F. Eisenmenger, U. H. E. Hansmann, *Comp. Phys. Comm.*, **178** (2008) 459.

[5] B. Berg and T. Neuhaus, *Phys. Lett.* **B267** (1991) 249.

[6] A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61** (1988) 2635.

[7] U.H.E. Hansmann and Y. Okamoto, *Physica A* **212** (1994) 415.

[8] U.H.E. Hansmann and L.T. Wille, *Phys. Rev. Let.*, **88** (2002) 068105.

[9] K. Hukushima and K. Nemoto, *J. Phys. Soc. (Japan)*, **65** (1996) 1604; G.J. Geyer, *Stat. Sci.* **7** (1992) 437.

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

[11] U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem* **14** (1993) 1333.

[12] U.H.E. Hansmann, Y. Okamoto, F. Eisenmenger, *Chem. Phys. Lett.* **259** (1996) 321.

[13] U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem.* **18** (1997) 920.

[14] D. Gront, A. Kolinski, J. Skolnick, *J. Chem. Phys.* **113** (2000) 5065.

[15] C.-Y. Lin, C.-K. Hu and U.H.E. Hansmann, *Proteins*, **52** (2003) 436.

[16] W. Kwak and U.H.E. Hansmann, *Phys. Rev. Lett.* **95** (2005) 138102.

[17] S. Mohanty, J.H. Meinke, O. Zimmermann and U.H.E. Hansmann, *Proc. Nat. Acad. Sci. (USA)*, **105** (2008) 8004.

[18] J.H. Meinke and U.H.E. Hansmann, *J. Comp. Chem.* **30** (2009) 1642.

[19] W. Nadler and U.H.E. Hansmann, *Phys. Rev. E* **75** (2007) 026109; **76** (2007) 057102.

[20] W. Nadler and U.H.E. Hansmann, *J. Phys. Chem.B* **112** (2008) 10386.

[21] U.H.E. Hansmann, *J. Chem. Phys.* **120** (2004) 417.

[22] N.A. Alves and U.H.E. Hansmann, *Int. J. Mod. Phys. C*, **11** (2000) 301.

[23] N.A. Alves and U.H.E. Hansmann, *Phys. Rev. Lett.* **84** (2000) 1836.

[24] Y. Peng and U.H.E. Hansmann, *Biophysical J.*, **82** (2002) 3269.

[25] S. Mohanty and U.H.E. Hansmann, *J. Phys. Chem. B* **112** (2008) 15134.

[26] J-C. Rochet and P.T. Lansbury, *Curr Op Struc Biol* **10** (2000) 60.

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

[28] J. Meinke and U.H.E. Hansmann, *J. Chem. Phys.* **126** (2006) 014706.

[29] P. Anand, F.S. Nandel and U.H.E. Hansmann, *J. Chem. Phys.* **128** (2008) 165102; **129** (2008) 195102.