



PROTEIN FOLDING *IN SILICO*: AN OVERVIEW

By Ulrich H.E. Hansmann

PROTEINS ARE ONE OF THE MOST COMMON YET IMPORTANT CLASSES OF MOLECULES IN LIVING SYSTEMS—COMMON BECAUSE THE HUMAN BODY MAKES AT LEAST 50,000 DIFFERENT PROTEINS, AND VITAL BECAUSE MUSCLES AND CONNECTIVE TISSUES

are built from them. As enzymes, proteins catalyze and regulate a cell's biochemical reactions; as antibodies, they are an important part of the immune system. Proteins differ substantially in size and structure,¹ but chemically, all are long, linear chain molecules, with the 20 naturally occurring amino acids acting as monomers. Locally, the chain of amino acids forms regular structures such as helices, sheets, and turns—a protein's so-called *secondary* structures. A major class of proteins, globular proteins, folds into compact configurations (the *tertiary* structure) in which they are biologically active. Considerable evidence suggests that proteins fold spontaneously into this structure, which is unique and determined solely by the protein's amino acid sequence (the *primary* structure).

A specific protein's sequence of amino acids is specified as DNA code in the human genome. With the human genome project's successful completion, we now know in principle the chemical composition of all the proteins in our bodies, but that information does not tell us what these proteins do or how they work. Because proteins are functional only if they fold into their tertiary structure's specific shape, and misfolded proteins can

cause a variety of diseases, understanding how the structure and function of proteins emerge from the amino acid sequence takes on increased importance. For instance, detailed knowledge of this relationship could help us understand the function of enzymes and how the immune system works. Design of new enzymes, hormones, antibodies, and biosensors are just a few of the possible applications.

The attempt to understand the mechanism that drives a protein into its unique, biologically active structure, and to predict this structure and the protein's corresponding function from knowledge about its amino acid sequence, is called the *protein-folding problem*. The inherent difficulties in solving experimentally a protein's tertiary structure only amplify the problem. Whereas it takes only hours to days to determine an amino acid sequence, for example, it would take months to years to discover its corresponding 3D shape by X-ray crystallography or nuclear magnetic resonance experiments. Equally challenging are experiments that explore the folding process's kinetics and dynamics. In short, efficient computational methods could help us tackle the protein-folding problem.

Computer Simulations of Proteins

Given its importance, it's not surprising that the protein-folding problem has raised considerable interest over the past 20 years. Although no one has found a complete solution yet, several promising research avenues have emerged over the years. One possible approach is to analyze the data sets of known sequences and structures for correlations and then use neural networks and other pattern-recognition techniques to find further relations.² Taxonomic lists exist for single, pairs, and triplets of amino acids that appear in helices, sheets, or turns, and they can help predict new structures. Such methods successfully predict the 3D structure of sequences with strong homology to proteins for which the structure is known, but they aren't suited for exploring directly the mechanisms underlying the folding.

In principle, it is possible both to investigate the thermodynamics and kinetics of folding and to predict the folded conformation through computer simulations.³ When successful, such an approach helps us understand a protein's folding solely from the underlying physical laws. Given a suitable description of the relevant forces, for instance, we can solve numerically the equations of motion for each atom in a protein and follow the trajectory in time via a molecular dynamics (MD) simulation, integrating the fundamental equation $F = m(a)$ (force = mass \times acceleration). This lets us

study explicitly the folding process, identify the folded state, and calculate equilibrium properties by computing averages over the sampled set of conformations.

MD is the method of choice for investigating the kinetics of folding, but for calculating proteins' equilibrium properties, a Monte Carlo (MC) simulation at relevant temperatures is an alternative approach. Here, trial moves are generated randomly and accepted or rejected according to the Boltzmann weight (canonical thermal probability $\propto \exp(-E/k_B T)$). In the Metropolis algorithm, for instance, a trial configuration with energy E^{new} replaces the current configuration (that has energy E^{old}) if

$$\min\left(1, e^{(E^{new} - E^{old})/k_B T}\right) \geq R, \quad (1)$$

where R is a random number and takes values between 0 and 1, and e is the Euler number. The Metropolis algorithm satisfies detailed balance. Thus, if each configuration can be reached in a finite number of steps (ergodicity), the resulting Markov process will converge to the canonical distribution, which will thus contain the protein's folded conformation. We again calculate thermodynamic quantities by computing averages over the sampled conformations.

Minimal Protein Models

The basic ingredient in computer simulations of proteins is a model that describes the protein's relevant physics. It is tempting for physicists to use minimal models that capture only a few, presumably important, interactions in real proteins.⁴ These include chain connectivity, excluded volume, hydrophobicity as the driving force, and

sequence heterogeneity.

In one often-used minimal model, a protein is regarded as a self-avoiding walk of monomers on a simple cubic lattice. The model considers only two kinds of amino acids: H (ydrophobic) and P (olar), which is hydrophilic. The analog to the amino acid sequence in a protein is thus the set $\{\sigma_j\}$ of monomers along the chain. For this so-called HP-model, we now define a Hamiltonian as

$$H = \sum_{i < j} E_{\sigma_i, \sigma_j} \Delta_{ij}, \quad (2)$$

where $E_{H,H}$, $E_{H,P}$, and $E_{P,P}$ are the energies of $H-H$, $H-P$, and $P-P$ contacts, and whose values define the model's specifics. The contact matrix $\Delta_{i,j}$ is given by $\Delta_{i,j} = 1$, if the residues i and j are not adjacent along the chain and simultaneously occupy nearest-neighbor sites on the lattice. In all other cases, $\Delta_{i,j} = 0$.

The advantage of using such minimal protein models lies in their simplicity and the resulting ease in computer simulation. These characteristics let us explore fully the system's properties as a function of the system's parameters. Consequently, minimal protein model simulations have led to significant new insight over the past few years into the dynamics of folding. The resulting new view of folding assumes that the energy landscape of proteins resembles a funnel, with a free-energy gradient toward the biologically active structure; such a funnel does not exist for random heteropolymers. The funnel itself is partially rough and riddled with traps in which the protein can transiently reside. There is no unique pathway, but a multiplicity of folding routes point toward the native state (that is, the structure in which the protein is biologically active).

Competition between the tendency toward the folded state and trapping due to the landscape's ruggedness thus characterizes the folding process. A protein folds faster the smoother its landscape is. Although computer simulations of minimalist models suggest a diversity of possible scenarios, a common one might be that the polypeptide chain first collapses from a random coil to a compact state. In the second stage, the protein explores a set of compact structures. The final stage involves a transition from one of the many local minima in the set of compact structures to the native conformation.

Detailed Protein Models

Although minimal protein models prove successful in exploring the general characteristics of possible folding mechanisms, they have their limitations. Researchers commonly assume that a protein's native structure corresponds to the free energy's global minimum. However, calorimetric measurements show that the protein in its native state is only marginally more stable (approximately 10 to 20 kcal/mol, compared to a total energy of roughly 10^7 kcal/mol) than the ensemble of the denatured conformations. Because idealized and simplified models lack the necessary precision to describe such small differences, they do not allow further probing for the details of structural transitions and folding in specific proteins. For this purpose, detailed representations of proteins that take into account the interactions among all atoms are more appropriate.

We can divide such interactions into two groups:² the interactions between all atoms within the protein (leading to an energy $E_{protein}$) and a term E_{solv} describing the protein's interaction with the surrounding solvent,

$$E_{tot} = E_{protein} + E_{solv}. \quad (3)$$

The latter term is an especially serious hurdle because explicit inclusions of solvent molecules are computationally demanding. Hence, we often must rely on implicit solvent models that approximate the protein–solvent interaction—for instance, a solvent-accessible surface term that accounts in an empirical way for the hydrophobic forces on the protein:⁵

$$E_{solv} = \sum_i \sigma_i A_i. \quad (4)$$

Here, A_i is the i th atom's solvent-accessible surface area and depends on the configuration, and σ_i is atom i 's empirically determined solvation parameter.

On the other hand, $E_{protein}$ is the sum of all intramolecular interactions and modeled by various atomic force fields. One example is the ECEPP energy function,⁶ which is given by the sum of the electrostatic term E_{es} , the van der Waals energy E_{vdW} , and the hydrogen-bond term E_{hb} for all pairs of atoms in the peptide together with the torsion term E_{tors} for all torsion angles:

$$E_{ECEPP} = E_{es} + E_{vdW} + E_{hb} + E_{tors}, \quad (5)$$

$$E_{es} = \sum_{(i,j)} \frac{332q_i q_j}{\epsilon r_{ij}}, \quad (6)$$

$$E_{vdW} = \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right), \quad (7)$$

$$E_{hb} = \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \quad (8)$$

$$E_{tors} = \sum_l U_l (1 \pm \cos(n_l \alpha_l)). \quad (9)$$

Here, r_{ij} is the distance between the

atoms i and j , and α_l is the torsion angle for the chemical bond l . The parameters q_i , A_{ij} , B_{ij} , C_{ij} , D_{ij} , U_l , and n_l are calculated from crystal structures of amino acids using a semiempirical molecular orbital method. Note that the ECEPP energies are in kcal/mol, which leads to the factor 332 in Equation 6.

The Need for Sophisticated Techniques

Unfortunately, computer simulations of folding are notoriously difficult for such detailed protein models. The complex form of the intramolecular forces and of the interaction with the solvent, which contains both repulsive and attractive terms, leads to a rough energy landscape with a huge number of local minima. Hence, a typical thermal energy of the order $k_B T$ is much less than the energy barriers that the protein has to overcome in the low-temperature region. Simple canonical MC or MD simulations will get trapped in a local minimum and often won't thermalize within a finite amount of available CPU time. This makes accurately calculating physical quantities quite difficult.

With the development and availability of ever-faster computers, new attempts to solve the protein-folding problem with brute force have emerged—for example, by performing long-time MD simulations of all-atom models of protein–water systems at suitable temperatures. Probably the best-known example is the $1\mu s$ MD trajectory of the 36-residue villin headpiece subdomain (called HP-36; discussed in more detail later in this article). Yong Duan and Peter Kollman found configurations with a root-mean-square deviation of only 5.7 Å to the experimentally determined structure.⁷

Another interesting approach in this direction is *folding@home*, a project in

which interested participants can download a screensaver or daemon and, after installation, donate unused CPU cycles for protein-folding simulations.⁸ In this way, *folding@home* acts as a distributed, inhomogenous but massively parallel computer that can speed up folding simulations.

Although such computer-intensive calculations (when they are possible) set a gold standard for simulations, they also corroborate the need for more sophisticated techniques that allow a faster sampling of the relevant protein configurations.⁹ The choice of method depends in part on the problem under investigation. For many years, the emphasis in protein studies was on protein-structure prediction. Assuming that the native structure is thermodynamically stable, we can identify the global-minimum conformation in the free energy at $T \approx 300$ K (with the lowest potential energy conformation) and search for this conformation with powerful optimization techniques such as genetic algorithms and simulated annealing. Only recently, with the recognition of energy landscape theory and funnel concepts, has interest increased in the thermodynamics of folding. Investigating such questions requires going beyond global optimization techniques; we must sample a set of configurations from a canonical ensemble and take an average of the chosen quantity over this ensemble.

New Simulation Techniques

Researchers have developed several novel simulation techniques over the past few years that promise improved sampling. One of the more successful methods is the so-called *generalized ensemble* approach,¹⁰ which is characterized by the condition that an MC–MD simulation leads to a uniform distribu-

tion of a prechosen physical quantity. In multicanonical sampling, for example, the weights $w(E)$ are chosen such that the distribution of energies $P(E)$ is given by

$$P(E) \propto n(E)w(E) = \text{const}, \quad (10)$$

where $n(E)$ is the spectral density—that is, the number of configurations with a given energy E . The simulation performs a random walk in energy space that allows it to escape from any local minimum. The large range of energies sampled lets us calculate the thermodynamic average of any physical quantity \mathcal{A} via the reweighting technique:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx \mathcal{A}(x) w^{-1}(x) e^{-\beta E(x)}}{\int dx w^{-1}(x) e^{-\beta E(x)}}. \quad (11)$$

Here, x stands for configurations, $\beta = 1/k_B T$, and k_B is the Boltzmann constant. Unlike the canonical ensemble, however, the weights are not a priori known for simulations in generalized ensembles. Instead, an iterative procedure (which can require up to 50 percent of the available computer time) must determine estimators.

Another effective method for protein simulations is parallel tempering, also known as replica exchange or the multiple Markov chain method.^{11,12} In its most common form, parallel tempering assumes an artificial system of N noninteracting copies of the molecule, each at a different temperature T_i . Let C_i be the configuration of the molecule in the i th copy, and $C = \{C_i\}$. Now, in addition to standard MC or MD moves that affect only one copy, parallel tempering introduces a new global update: the exchange of conformations between two copies i and $j = i + 1$ with

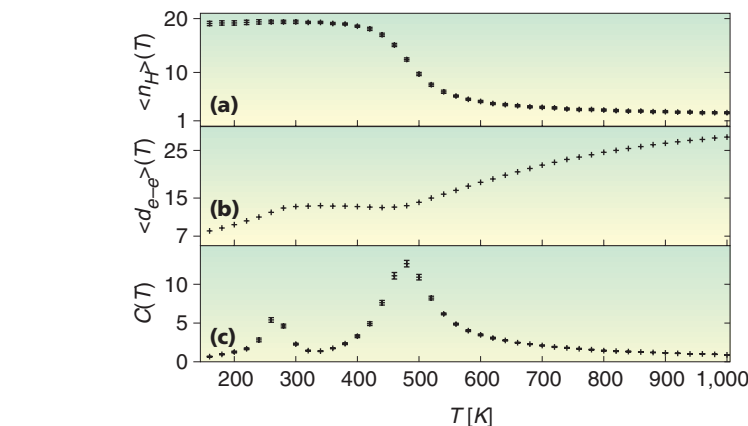


Figure 1. The thermodynamics of folding for the artificial peptide Ala₁₀-Gly₅-Ala₁₀. This illustration shows the (a) average number $\langle n_H \rangle$ of helical residues, (b) end-to-end distance $\langle d_{e-e} \rangle$, and (c) specific heat $C(T)$ as function of temperature T . The data are calculated from a multicanonical simulation of Ala₁₀-Gly₅-Ala₁₀.

probability

$$w(C^{old} \rightarrow C^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))). \quad (12)$$

The exchange of conformations will lead—especially at low temperatures—to a faster convergence of the Markov chain toward the stationary distribution than we observe in regular canonical simulations with only local moves.

These sophisticated, newer, simulation techniques let us obtain accurate estimates for physical quantities in simulations of a specific protein, but all-atom simulations that rely on these techniques are still very expensive in computer time. Hence, the emphasis is less on predicting protein structures than on exploring the folding mechanism and on determining the limitations of the protein models employed.

Exploring the Folding Mechanism

As an example for the first line of research, I present here results from simulations¹³ of a simple artificial protein, Ala₁₀-Gly₅-Ala₁₀, that was obtained in collaboration with Nelson Alves at the University of Sao Paulo, Brazil. The molecule is built up from two chains, each of which has 10 alanine residues

connected by five glycine residues. Three quantities are measured in multicanonical simulations (see Figure 1). The first quantity (Figure 1a) is the average number $\langle n_H \rangle$ of residues that are part of an α -helix drawn as a function of temperature. At high temperatures, few residues are part of a helix, so this number is small. However, at low temperatures, helices form, and almost all the alanine residues are part of an α -helix.

The transition between the two temperature regions at $T = 483 \pm 8$ K is sharp and corresponds to a pronounced peak in the specific heat $C(T)$. However, Figure 1c shows a second, smaller peak at the lower temperature $T_f = 265 \pm 7$ K, indicating yet another transition. To understand this second peak, look at Figure 1b: it shows the average end-to-end distance $\langle d_{e-e} \rangle(T)$ as a function of temperature. This quantity is a measure for a protein conformation's compactness. Observe that $\langle d_{e-e} \rangle(T)$ decreases when the temperature is lowered. Below the helix-coil transition T_{bc} , the curve becomes almost flat at a value of $\langle d_{e-e} \rangle \approx 10$ Å, with little further change in the molecule's compactness. At the temperature T_f , however, the end-to-end distance decreases again sharply toward a new value $\langle d_{e-e} \rangle = 6.1$ Å. Hence,

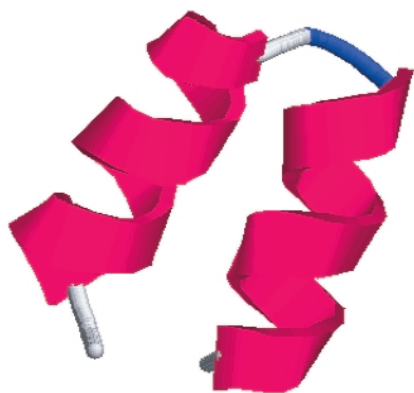


Figure 2. The global minimum conformation of Ala₁₀-Gly₅-Ala₁₀.

T_f marks the folding of the molecule into a defined compact structure with the two terminal ends of the peptide close together.

Figure 2 supports this scenario, displaying the lowest-energy configuration. It consists of two helices (made up of the alanine residues) connected by a turn (built out of flexible glycine residues) to a hairpin-like structure that is consistent with the small value of the end-to-end distance $\langle d_{e-e} \rangle$ at temperatures below T_f in Figure 1b.

The analysis of our peptide's thermodynamics suggests that in the gas phase, Ala₁₀-Gly₅-Ala₁₀ folds in a two-step process. The first step is the formation of α -helices and can be characterized by a helix-coil transition temperature $T_{hc} = 483 \pm 8$ K. The formation of α -helices then restricts the possible configuration space. Energetically most favorable is the folding of two α -helices (made out of the alanine residues) into a hairpin. This second step can be characterized by a lower folding temperature $T_f = 265 \pm 7$ K.

This scenario is reminiscent of the well-known framework and collision-diffusion model of folding, which proposes that local elements of native local secondary structure form independently of tertiary structure. These elements diffuse until they collide and coalesce to give such a structure. In our case, the molecule's thermal energy in the

temperature region between T_f and T_{hc} does not allow coalescing of the helix fragments, which thus form and decay. Some stabilization happens when these fragments form a U-turn-like bundle of two (antiparallel) α -helices connected by a turn of glycine residues. This is the most stable structure, and below T_f thermal energies can no longer overcome the energy gap that separates this configuration from its competitors.

Determining the Protein Model's Limitations

To explore the limitations of current energy functions in protein simulations, Luc Wille (at Florida Atlantic University) and I have simulated the villin headpiece subdomain, which is a 36-residue peptide (HP-36).¹⁴ HP-36 is one of the smallest peptides that can fold autonomously; Duan and Kollman chose it recently for a 1-microsecond MD simulation of protein folding.⁷ Figure 3 shows the structure of HP-36 (PDB code 1vii) as obtained from the Protein Data Bank (www.rcsb.org/pdb). It consists of three helices between residues 4 to 8, 15 to 18, and 23 to 32, respectively, which are connected by a loop and a turn. We obtain as energy (ECEPP/2 + solvation term) of the native structure $E_{nat} = -276$ kcal/mol.

In our simulations, we find for the lowest-energy conformation that $E_{min} = -277$ kcal/mol. Its radius of gyration is $R_g = 10.1$ Å, indicating that the numerically obtained structure is slightly less compact than the regularized experimental structure ($R_g = 9.6$ Å). Still,

our structure looks very similar to the PDB structure. It consists of three helices, with the first stretching from residue 2 to residue 11 and looking more elongated than the corresponding helix in the native structure (residues 4 to 8). The second helix consists of residues 13 to 17 (compared to residue 15 to 18 in the native structure), and the third helix stretches from residue 23 to 33 (residues 23 to 32 in the PDB structure). Our numerically obtained structure has 95 percent of the native helical content and 65 percent of the native contacts formed. The root-mean-square deviation to the native structure is 5.8 Å. All three values are comparable with Duan and Kollman's results of the 1-microsecond MD simulation (which required orders of magnitude less computer time). In their simulation, the optimal structure showed 80 percent of native helical content and 62 percent of native contacts and had a root-mean-square deviation of 5.7 Å to the native structure.

However, if solvation effects are neglected and we explore the ECEPP/2 conformations of HP-36, we find that the lowest-energy structure has an energy of $E_{GP} = -192$ kcal/mol and differs significantly from the regularized PDB-structure, which now has a higher energy ($E_{nat} = -176$ kcal/mol). Only by adding the solvation term can we obtain an essentially correct structure as a global minimum. Our results point out the need to include solvent effects in protein simulations.

Even in the case where we have included a solvent term in our simulation, the quality of our structure predictions for this molecule are still limited to an RMSD of ≈ 6 Å. This demonstrates that our folding simulations are limited by the energy function's accuracy. An important part of all-atom protein simulations is thus to

explore the limitations of these energy functions, which can lead to the development of better protein models.

Recent years have seen an increase in interest and activity in the protein-folding problem. Although researchers have made remarkable progress over the past decade, there is still a need for new ideas and better algorithms. New techniques together with a refinement of the force fields (including improved approximations of the solvent effects) will lead eventually to an increased understanding of the protein-folding problem. As of today, computer simulations of small proteins (with approximately 50 amino acids) seem to be feasible and are started now by an increasing number of groups, including ones I'm affiliated with. For (a partial) listing of research groups see www.fccc.edu/research/labs/roder/folding_groups.html.

Acknowledgments

The results on Ala₁₀-Gly₅-Ala₁₀ and HP-36 are published elsewhere^{12,13} and originate from collaborations with Nelson A. Alves (University of Sao Paulo) and Luc T. Wille (Florida Atlantic University). I gratefully acknowledge financial support from the US National Science Foundation's research grant CHE-9981874.

References

1. C. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd ed., Garland Publishing, 1998.
2. F. Eisenhaber, B. Persson, and P. Argos, "Protein Structure Prediction: Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acids Sequence," *Critical Rev. Biochemistry and Molecular Biology*, vol. 30, no. 1, 1995, pp. 1-94.
3. D. Frenkel and B. Smit, *Understanding Molecular Simulations*, Academic Press, 1996.
4. K.A. Dill and H.S. Chan, "From Levinthal to

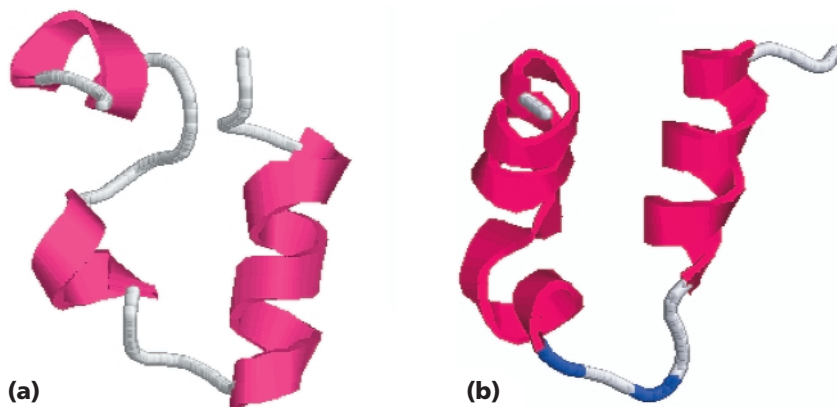


Figure 3. The HP-36 peptide's structure. The (a) experimentally determined structure of HP-36 and (b) lowest energy conformation of that peptide as obtained from a generalized-ensemble simulation.

- Pathways to Funnels," *Nature Structural Biology*, vol. 4, no. 1, 1997, pp. 10-19.
5. T. Ooi et al., "Accessible Surface Areas as a Measure of the Thermodynamic Parameters of Hydration of Peptides," *Proc. Nat'l Acad. Science*, vol. 84, 1987, pp. 3086-3090.
6. M.J. Sippl, G. Némethy, and H.A. Scheraga, "Intermolecular Potentials from Crystal Data 6: Determination of Empirical Potentials for O-HLO = C Hydrogen Bonds from Packing Configurations," *J. Physical Chemistry*, vol. 88, 1994, pp. 6231-6233.
7. Y. Duan and P.A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science*, vol. 282, 1998, pp. 740-744.
8. M. Shirts and V. Pande, "Screen Savers of the World, Unite!" *Science*, vol. 290, 2000, pp. 1903-1904.
9. U.H.E. Hansmann and Y. Okamoto, "New Monte Carlo Algorithms for Protein Folding," *Current Opinion in Structural Biology*, vol. 9, 1999, pp. 177-184.
10. U.H.E. Hansmann and Y. Okamoto, "The Generalized-Ensemble Approach for Protein Folding Simulations," *Ann. Reviews in Computational Physics*, vol. 6, D. Stauffer, ed., World Scientific, 1998, pp. 129-157.
11. K. Hukushima and K. Nemoto, "Exchange Monte Carlo Method and Applications to Spin Glass Simulations," *J. Physical Soc. (Japan)*, vol. 65, 1996, pp. 1604-1608.
12. G.J. Geyer and E.A. Thompson, "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference," *J. Am. Statistical Assoc.*, vol. 90, no. 431, 1995, pp. 909-920.
13. N.A. Alves and U.H.E. Hansmann, "Helix Formation and Folding in an Artificial Peptide," *J. Chemical Physics*, vol. 117, no. 5, 2002, pp. 2337-2343.
14. U.H.E. Hansmann and L. Wille, "Global Optimization by Energy Landscape Paving," *Physical Rev. Letters*, vol. 88, no. 6, 2002, p. 068105 (1:4).

Ulrich H.E. Hansmann is an associate professor in the Department of Physics at Michigan Technological University. His research interests include simulations of the protein-folding problem, simulation of complex systems, and global optimization techniques. He has a PhD in physics from the Freie Universität, Berlin. Contact him at the Dept. of Physics, Michigan Technological Univ., Houghton, MI 49931; hansmann@mtu.edu.

The IEEE Computer & Communications Societies present

This new quarterly magazine aims to advance mobile and ubiquitous computing by bringing together its various disciplines, including peer-reviewed articles on:

IEEE
pervasive
COMPUTING
MOBILE AND UBIQUITOUS SYSTEMS

- Hardware and Software Technologies
- Human-Computer Interaction
- Security, Scalability, and Privacy
- Real-World Sensing

<http://computer.org/pervasive/>