PHYSICS OF PROTEIN FOLDING

by Hue Sun Chan

As a physico-chemical phenomenon, protein folding is one of the most basic self-assembly processes in biology. Thousands of different types of proteins are responsible for the functioning of life. Understanding and predicting the three-dimensional structure and dynamic properties of proteins from one-dimensional amino acid sequences requires a basic knowledge of the molecular forces involved. This article provides a brief summary of how a physicists' approach to statistical mechanical and polymer modeling has helped in this endeavor. Results from exact enumeration and Monte Carlo simulations of simplified heteropolymer models are highlighted to elucidate the sequence-structure mapping, molecular evolution, thermodynamic properties of proteins and their folding kinetics. As well, the consequences of protein misfolding and rudimentary physical scenarios for the propagation of disease-causing misfolded protein structures are also discussed.

Understanding and predicting the three-dimensional structure and dynamic properties of proteins from one-dimensional amino acid sequences requires a basic knowledge of the molecular forces involved.

WHY IS THE PROTEIN FOLDING PROBLEM INTERESTING AND IMPORTANT?

Proteins are biomolecules essential for life as we know it on earth. They come in thousands of types and shapes, performing many different functions. Some are responsible for the transport of ions across cell membranes, some provide structural integrity for organisms, and many proteins are enzymes that regulate and catalyze various biochemical reactions. Indeed, the ability of enzymes to speed up reactions by many orders of magnitude is remarkable. Most biological reactions would practically not occur in the absence of enzymes because their uncatalyzed rates are exceedingly slow in comparison with typical biological time scale [1].

Proteins are linear chain molecules. They carry vital information consisting of specific sequences of amino acids synthesized according to the dictate of a given organism's genetic make-up [1-3]. The basic covalent structure of a protein is quite simple. It is given in almost every biochemistry textbook, and can also be found in a recent *Physics in Canada* article on proteins [3]. When placed in appropriate aqueous environments, many proteins - at least the small "globular" proteins - can spontaneously assemble, each sequence folding up into a single complex shape capable of carrying out certain biological functions. The folded form of a protein is often referred to as its native structure [4]. Many of these structures have been determined experimentally by X-ray crystallography and NMR. Their atomic coordinates are available from the Protein Data Bank (http://www.rcsb.org/pdb/). Thus, the process of protein folding transforms the linear, one-dimensional information of a protein's sequence into a three-dimensional functional entity. Figure 1 illustrates this phenomenon by showing the sequence and native structure of a small protein called chymotrypsin inhibitor 2 (CI2). The function of CI2 is to inhibit the action of chymotrypsin, one of the protein-digestive enzymes secreted by the pancreas. The space-filling depiction in Fig. 1 conveys that protein native structures are often quite compact. In fact, it has long been known [5] that the average packing density of the interior of a folded protein resembles that of solids [6].

Protein folding is often characterized as a "problem" [2] because we don't yet have a sufficient understanding of the physics involved to predict the native structure of a protein solely from the knowedge of its amino acid sequence, though Nature can do this with ease. In other words, we don't yet have a very good grasp of what is contributing to the schematic arrow in Fig. 1. The protein folding problem is of tremendous intellectual and practical interest. As physicists, we would like to understand how such a remarkable self-assembly feat arises from the fundamental laws of physics [7]. It is also obvious that better knowledge of this basic biological process would advance all areas of biomedical research. In this view, the main goal in tackling the protein folding problem is to gain insight into the underlying energetics. For this reason, attention has often been focused on smaller, soluble globular proteins such as CI2 (Fig. 1), in the hope that their energetics would be easier to decipher. Nonetheless, one should not lose sight of the bigger picture that proteins can be much larger and more complex, and that not only the compact folded form but disordered forms of certain proteins can have biological functions as well (for a recent computational study of such phenomena, see Ref. [8]).

STATISTICAL MECHANICS OF FOLDING AND MISFOLDING

A protein molecule is a polymer chain with rotatable chemical bonds along its backbone. As a result, it can take up many different conformations. The number of possible conformations increases approximately exponentially with the number of amino acids that make up the protein. The many degrees of conformational freedom is the main reason why the protein folding problem is difficult. The study of protein folding entails deciphering how the interplay of physical driving forces conspires to favor the native conformation (or a small ensemble of native conformations) among an exponentially large number of possible conformations. The basic forces responsible for protein folding are easily identifiable. They include van der Waals

H.S. Chan (chan@arrhenius.med.utoronto.ca) Departments of Biochemistry and of Medical Genetics and Microbiology, University of Toronto, Medical Sciences Building -5th Fl., 1 King's College Circle, Toronto, Ontario M5S 1A8

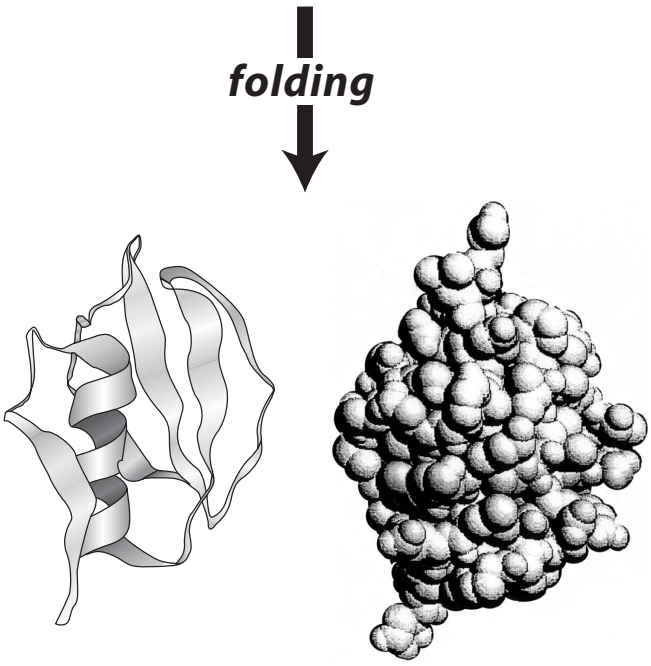

Fig. 1 **The protein folding phenomenon is exemplified here by CI2 (Protein Data Bank accession code 2ci2), whose sequence of 83 amino acids is given at the top of this figure by the standard one-letter symbols for the amino acids (see, *e.g.*, pages 6-7 of Ref. [4]). This particular sequence specifically encodes for the native structure of this protein and is presented schematically in two different representations to emphasize different structural features: The ribbon representation (left) shows how the backbone of the protein chain molecule is folded, highlighting regular structural elements such as helices, strands (essentially linear zigzags), and sheets formed from more than one strand. On the other hand, the space-filling representation (right) indicates the volume occupied by the atoms that make up the protein, underscoring the compactness of the native structure. Preparation of this figure was aided by the software RasMol.**

interactions, hydrogen bonding, and electrostatic interactions [1,2,4]. But for systems as big and as complex as a protein and its surrounding solvent molecules, a complete quantum mechanical description is currently out of the question. Even approximate classical molecular dynamics simulations using all-atom representations are computationally extremely costly; and no such simulation has yet successfully generated a trajectory from an open conformation of a protein with 50 or more amino acids to a global energy-minimum conformation that coincides with its known native structure. In this context, simplified modeling techniques have proven to be very useful in addressing general physical principles of protein folding [9-14]. We focus on such approaches in this article.

A main component of the organizational forces in protein folding is the hydrophobic effect. This effect arises from the interaction between nonpolar chemical groups and water. Among the 20 amino acids, some are more nonpolar. Like oil, these amino acids tend to avoid water, and are referred to as being hydrophobic. The physical reason for this behavior is that nonpolar groups tend to restrict or break up hydrogen bonds among water molecules without replacing them with equal or more favorable interactions. Hence, placing them in water is energetically unfavorable. On the other hand, some amino acids are polar, which implies that they favor being solvated by water. A typical protein sequence has both hydrophobic and polar amino acids. The hydrophobic effect drives protein folding because it forces the protein chain to organize itself to sequester most of its hydrophobic amino acids in a core (so they can avoid contact with water) and to leave most of its polar amino acids on the surface.

This mechanism is illustrated in Fig. 2 using one of the simplest caricatures of protein folding known as the HP (hydrophobic-polar) model, a minimalist lattice construct designed to capture the essential physics of hydrophobicity [10,11]. Instead of 20 amino acid types (a 20-letter alphabet), sequences in this model consist of only two types of units (often referred to as residues), namely H and P, from a reduced 2-letter alphabet. Only non-bonded nearest-neighbor contacts between H residues are favorable. Each of such contacts is assigned an energy ε (< 0). Contacts between H and P and that between P and P residues are taken to be neutral (*i.e.*, have zero energy). As such, this model neglects hydrogen bonding, which has to be taken into account when more detailed considerations are desired. In the HP model, the many conformations available to a protein are modeled by the set of all possible lattice self-avoiding walks accessible to a model HP sequence. Using a computer, these walks can be exactly enumerated for short chains. Figure 2 shows the thermodynamic equilibrium between the native conformation with a hydrophobic core containing the maximum number of HH contacts achievable by the given HP sequence and the other (denatured) conformations with fewer HH contacts. At equilibrium, chain populations in different conformations follow the Boltzmann distribution. Thus, at a given temperature, the native folded conformation (N state) is favored when the HH contact energy is strong

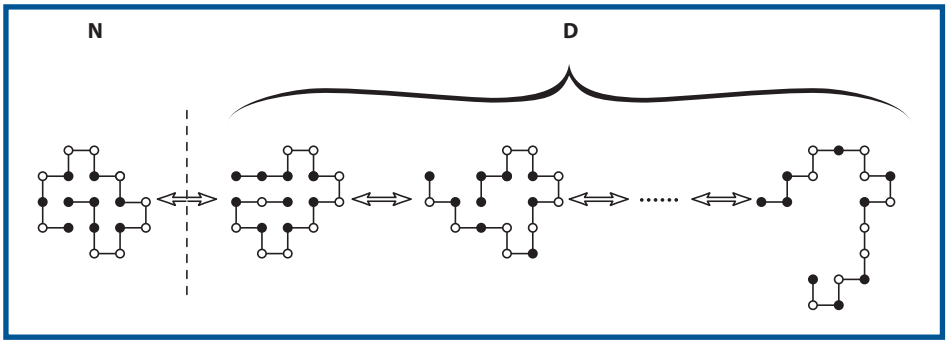

Fig. 2 **Equilibrium between protein native (N) and denatured (D) states: A two-dimensional (2D) HP model illustration. In this article, H and P amino acids are represented by filled and open circles, respectively. For the HP sequence in this figure, the unique (single) ground-state conformation (N) has 9 hydrophobic-hydrophobic (HH) contacts. All other conformations (5,808,334 total) have fewer than 9 HH contacts; and they are classified as belonging to the denatured (D) ensemble. Three examples of the large ensemble of D conformations are shown. The arrows in this figure indicate that conformations can thermally inter-convert; the vertical dashed line demarcates conformations of the N and D states.**

(ε more negative). On the other hand, when the HH contact energy is weak, the ensemble of denatured conformations (D state) is favored because its conformational entropy is immensely larger than that of the N state. It is noteworthy that only a small fraction of all possible model HP sequences have a unique ground-state conformation. The one in Fig. 2 does, but most HP sequences have degenerate ground states [15]. This feature echoes the experimental observation that most random amino acid sequences don't fold to a unique native structure like naturally occurring proteins [16,17] , which are the products of eons of natural selection. Besides the HP model, other lattice models have also been used extensively to study protein folding. A review of their strengths and weaknesses can be found in Ref. [14].

As protein folding is essential for life's many functions, it is not too surprising that protein misfolding can lead to various diseases. With recent advances, an increasing number of ailments are found to be related to protein misfolding and aggregation. These include prion diseases such as bovine spongiform encephalopathy (mad cow disease) and related Creutzfeldt-Jakob disease in humans, and Alzheimer's disease (see introductory discussions in Ref. [18-20] and references therein).

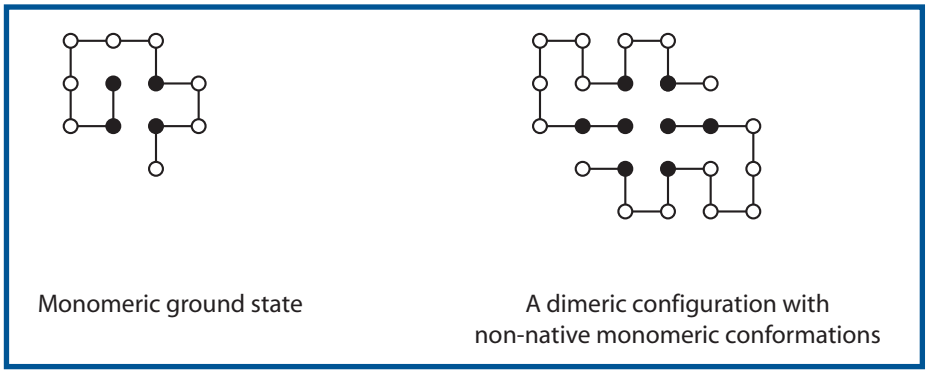

Fig. 3 Protein misfolding and aggregation. This simple 2D HP model example demonstrates that dimerization can readily lead to stabilization of non-ground-state single-chain conformations.

Figure 3 provides a simple physical illustration of how protein aggregation can lead to misfolding [19]. On the left side of this figure is an HP sequence shown in its unique monomeric (single-chain) ground-state conformation with three HH contacts. This may be taken as the "healthy" form of the protein. However, when there is more than one copy of this protein in close proximity (as in certain physiological settings), there are energetically more favorable ways to pack the chains together than to let each individual chain retain its monomeric ground-state conformation. An example is shown on the right side of Fig. 3. This dimeric (two-chain) configuration has a total of seven HH contacts, whereas any configuration with the two chains adopting their individual monomeric ground state has a total of only six HH contacts. It follows then, that if individual chains of this protein being not in the "healthy" conformational form would lead to the protein malfunctioning, the aggregated configuration of the protein on the right may be interpreted as the "disease-causing" form. In this scenario, a collection of protein chains in the healthy conformation is prevented from adopting the thermodynamically more favorable disease-causing form only by a high kinetic barrier. The result is extremely slow conversion dynamics similar to the glassy behavior observed in some physical systems. However, in such situations, the introduction of a dimeric disease-causing form can catalyze the conversion and lead to propagation of the disease-causing form at an accelerated rate. This physical picture thus provides a rationalization for the infectivity of prion diseases [14,20].

KNOWLEDGE-BASED VS PHYSICS-BASED APPROACHES

Aside from efforts to understand protein folding in terms of basic physical forces, a commonly used complementary approach is to utilize the vast amount of available protein sequence and structural data to gain insight into how protein sequences relate to their folded structures. This "knowledge-based" methodology relies on knowing the protein's native structure that was obtained from experiment. Such efforts are a part of a fast-expanding and productive discipline known as bioinformatics. A key element in this approach is the usage of "statistical potentials" (Fig. 4) [14,17,21-25]. These are energy-like parameters derived from the statistics of spatial contacts between various chemical entities among large collections of protein native structures. For example, assuming that such contact frequencies follow a Boltzmann-like distribution, an effective interaction energy e_{ij} between two amino acid types i and j may be estimated [21,22] using the formula $\exp(-e_{ij}/k_BT) = \tilde{n}_{ij}\,\tilde{n}_{00}/\,\tilde{n}_{i0}\,\tilde{n}_{j0}$. Here k_BT is the Boltzmann constant times the absolute temperature, "0" denotes solvent; and $\tilde{n}_{ij}$ is an appropriately normalized average number of contacts between amino acid types i and j, a number directly obtainable from the frequency of its occurrences among known protein native structures. Details of this formulation can be found in Ref. [21,22].

It is reassuring that the energy-like parameters emerged from the above and other similar knowledge-based analyses exhibit trends consistent with physical expectations. This suggests that the assumptions used in deriving them enjoy at least some degree of validity. For example, statistical potentials between hydrophobic amino acids are largely favourable [14,17,21,22]. However, at a fundamental level, it should be recognized that the precise relationship between

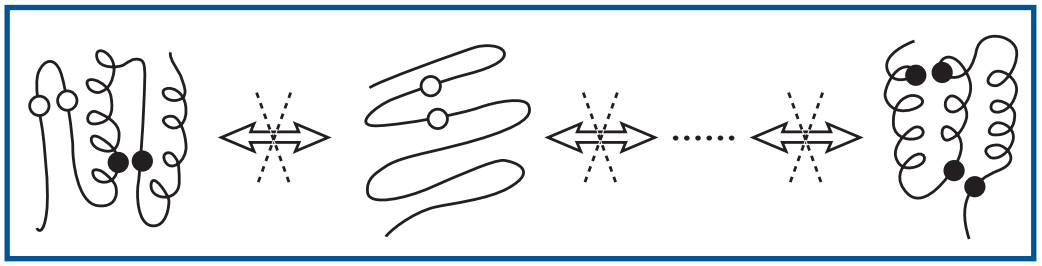

Fig. 4 Knowledge-based statistical potentials for protein folding studies. Energy-like parameters are often deduced by the relative abundance of a given interaction from protein native structures found in the Protein Data Bank. The schematics here correspond to the situation when interactions between two "black" residues occur more frequently than those between two "white" residues. As a result, the interaction between two black residues is deemed more favorable. The fact that different native structures cannot interconvert is emphasized by the crosses on the arrows. See text for further details.

physical interactions and knowledge-based energy-like parameters is far from clear. In fact, an insightful test using the HP model showed that certain common procedures for extracting statistical potentials do not always recover the correct physical interactions [23]. Procedures for obtaining statistical potentials often assume that the probabilities of contacts in a collection of Protein Data Bank structures depend solely on the contacts' interaction energy. As such, these assumptions neglect the severe constraints imposed by chain connectivity, and the fact that a collection of folded structures is not a Boltzmann ensemble because native structures of different proteins cannot interconvert into one another (Fig. 4). In view of these difficulties, efforts have been made to provide better justifications [24] and to design more sophisticated extraction procedures to enhance the predictive power of statistical potentials [25].

ENERGETIC INGREDIENTS: SOLVENT-MEDIATED INTERACTIONS

One limitation of statistical potentials, because they are derived from folded structures, is that they are intrinsically inadequate for providing spatial dependencies for the interactions [23]. Moreover, to extract knowledge-based energy parameters, functional forms for various interactions have to be presupposed. However, some basic properties of intra-protein interactions simply cannot be deduced from the mere knowledge of a collection of protein native structures, no matter how extensive the collection.

For protein folding, the physics can be rather complicated, in large measure because folding takes place in an aqueous environment. Consequently, all interactions among chemical groups along the protein chain are mediated by the aqueous solvent, adding many degrees of freedom to the problem: Electrostatic interactions can be shielded by ions in the solvent, folding kinetics can be affected by solvent viscosity, etc.

A classic water-mediated phenomenon is hydrophobicity. To illustrate the complexity of solvent effects, here we give a brief account of an atomistic consideration of hydrophobic interaction, the driving force that the HP model described above seeks to capture in a highly coarse-grained manner. As a prototypical hydrophobic interaction [26], Fig. 5 shows the free energy function (potential of mean force) of bringing two methane molecules together in water, computed by averaging over the water degrees of freedom using an explicit, geometrically accurate model of water [27-29]. Simulations of this sort have had a long history [27,28]. They provide atomic details that are often neglected in the HP and other "implicit-solvent" models that do not treat water explicitly [30].

Consistent with intuitive expectations, Fig. 5 indicates that the free energy is lowest (most favorable) when the two methanes are in contact ($\xi \approx 4$ Å). The important role of the aqueous solvent is underscored by the fact that the contact interaction between two methanes is much more favorable in water under ambient conditions (298 K and 368 K curves) than in vacuum (LJ curve). The spatial variation of the two-methane interaction in water is more complex than that in vacuum. In particular, for the water-mediated interaction, a desolvation free energy barrier develops at $\xi \approx 6$ Å, corresponding to the point at which water molecules are being squeezed out between the two methanes as they approach each other. Although the basic atomic interactions in the model in Fig. 5 are temperature-independent, the water-mediated two-methane interaction is temperature-dependent as a result of averaging over water configurations. In particular, Fig. 5 shows that as temperature is increased from 298 K to 368 K, the contact between two methanes becomes more favorable, whereas the desolvation barrier is reduced at higher temperatures [14,29]. Similar simulations were applied to address other aspects of hydrophobic interactions as well. These investigations include ascertaining the dependence of two-methane hydrophobic interactions on denaturants (chemicals that unfold proteins) [31], and the extent to which water-mediated hydrophobic interactions deviate from the assumption (as by the HP model) of pairwise additivity [32]. Results from these recent atomistic considerations have provided novel rationalizations for several protein thermodynamic phenomena that were rather puzzling in the context of more simplistic views of hydrophobicity [31,32].

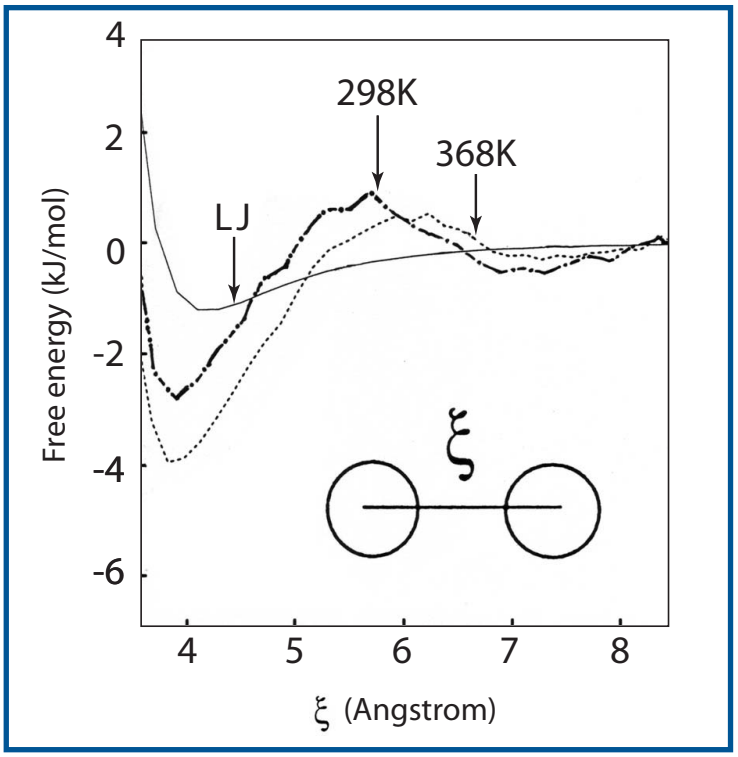

Fig. 5 **Solvent-mediated interactions. The free energy of association of two methanes in water is shown as an elementary example. The potential of mean force under atmospheric pressure between two methanes as a function of their separation ξ is obtained by Boltzmann-averaging over the water degrees of freedom at two temperatures (as indicated). The Lennard-Jones interaction potential between two methanes in vacuum (LJ curve) is included for comparison (adapted from Ref. [14]). Computational details are provided in Ref. [29].**

EVOLUTIONARY LANDSCAPES

It is not surprising from the above atomistic consideration that although the HP model has been instrumental for many conceptual advances, the simple pairwise additive hydrophobic interaction it embodies is by itself insufficient for a quantitative account of certain key generic thermodynamic and kinetic properties of protein folding [33,34]. However, notwithstanding these limitations, the simple HP model remains particularly useful in providing a tractable physical mapping between sequences and native structures. This is because for naturally occurring proteins, contributions from various energetic com-

ponents tend to consistently favor the same native structure [9]. Therefore, even though the HP potential does not account for all energetic contributions, the correspondence between a sequence's H/P pattern and its ground-state conformation(s) should still essentially mirror that of real proteins [14,35]. Notably, this stipulation is supported by the fact that the H/P patterns observed among short 2D HP sequences with unique ground-state conformations (*i.e.*, sequences considered to be model proteins, see above) are similar to that of proteins found in the Protein Data Bank [36].

Exact enumerations in the HP model have led to exhaustive mappings between sequences and conformational structures [37], making the HP model an instructive testbed for evolutionary ideas that require considering a broad coverage of both sequence and conformational spaces [14,15,35-39]. Other simplifed lattice models (see, *e.g.*, Ref. [40,41]) have also been applied to the study of molecular evolution. One important finding from HP model studies is that not all compact conformations are encodable [37]. An encodable structure is one for which there exists at least one sequence that has the given structure as the sequence's unique ground-state conformation. An encodable structure is said to be more "designable" if it has a larger set of sequences that encode for it [38]. A set of sequences that encode for the same structure and are interconnected by single-point substitutional mutations is known as a neutral net [39]. As an example of simple exact model approaches to evolution [15], the topology of mutational connections of an HP model neutral net is shown in Fig. 6.

Of particular interest here is our finding that the centrally located prototype sequence tends to have the maximum native thermodynamic stability among sequences in the neutral net. At the same time, it also tends to be mutationally most stable. (The prototype sequence in Fig. 6 corresponds to the sequence shown in Fig. 2.) Moreover, the sequences in the neutral net in Fig. 6 (as in many other HP neutral nets) conform to the paradigm of a "superfunnel" sequence-space landscape in that native thermodynamic stability tend to increase as a sequence's Hamming distance from the prototype sequence is decreased [see arrows in Fig. 6(b)]. Sequence-space topology thus emerges as an important determining factor in evolutionary dynamics. A general analysis has shown that steady-state populations of prototype sequences can be enhanced relative to other sequences solely because of their higher mutational stabilities. If model prototype sequences are identified with naturally occurring proteins, the superfunnel paradigm rationalizes why most but not all mutations on natural proteins destabilize their native structures [15,39]. More recently, we have extended HP model evolutionary studies to incorporate effects of crossovers and recombinations. We found that certain sequentially local H/P patterns (segments of the full chain) are significantly more conducive to the folding of the full chain to a unique native state than other local H/P patterns. Intriguingly, some of these favorable sequentially local H/P patterns are reminiscent of the autonomous folding units observed in real proteins [15,35].

FOLDING COOPERATIVITY: A CHALLENGE TO PHYSICAL MODELING

Despite numerous advances, current knowledge about the various physical interactions contributing to protein folding is still rather limited. An obvious avenue to evaluate our understanding of protein energetics is to compare model predictions with experiment. In this regard, since certain protein properties are

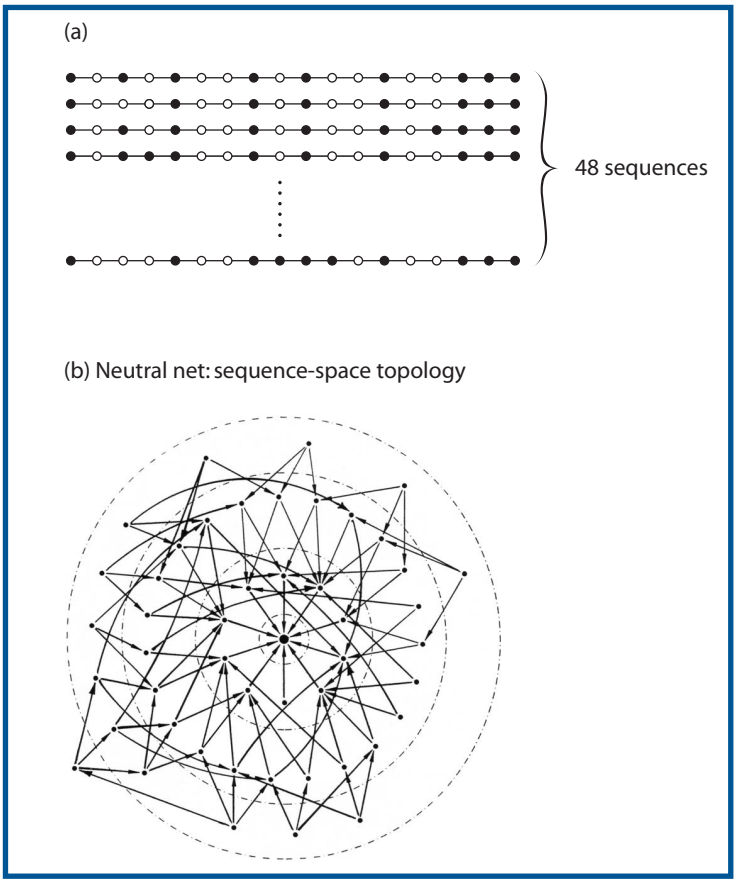

Fig. 6 **Modeling evolutionary landscapes. (a) Convergence in the 2D HP model. A total of 48 different HP sequences of length 18 have the same native structure shown in Fig. 2 as their unique ground-state conformation. A complete list of the sequences can be found in Fig. 16.7 of Ref. [14]. (b) The neutral net formed by the 48 sequences (adapted from Ref. [39], each sequence represented by a dot) is depicted by the single-residue H to P or P to H mutations connecting pairs of sequences (represented by arrows pointing towards the sequence with higher native thermodynamic stability). The concentric circles indicate the number of single-residue mutations (Hamming distance) separating a given sequence in the neutral net with the centrally located prototype sequence. See text and Ref. [39] for details.**

so special and remarkable *vis-à-vis* that of other polymers, we found that even generic folding behavior - features common to many proteins and not particularly sequence-specific - can provide stringent constraints on modeling [33]. One such generic property is calorimetric cooperativity, which means that for many real, small proteins, the distribution of enthalpy at the folding/unfolding transition mid-point temperature (which is identifed by a prominent peak in the heat capacity) is sharply bimodal with little conformational population having enthalpies intermediate between the native and denatured values [14,34,42,43]. Somewhat surprisingly, we found that many protein chain models (including the HP model) do not satisfy this strict constraint [34,42].

The apparent two-state folding kinetics of many small proteins provide an even more stringent cooperativity criterion [44-46]. Figure 7 compares an experimental apparent two-state "chevron plot" (left panel) [46] with one generated from a simplified implicit-solvent continuum model [45,47-49] of the same protein (right panel). The hallmarks of apparent two-state cooperative

kinetics [44] are that (i) the folding and unfolding relaxations are essentially single-exponentials; (ii) the logarithmic folding and unfolding rates (ln k_f and ln k_u) at constant temperature are essentially linear in chemical denaturant concentration, *i.e.*, both arms of the "chevron plot" are linear; and (iii) the equilibrium ratio of native to denatured conformational population is well-predicted by k_f/k_u. In contrast to the experimental data in the left panel of Fig. 7, the Langevin dynamics [47] results in the right panel show that even a model with explicit energetic biases favoring the known native structure [48] do not satisfy all the criteria for apparent two-state folding/unfolding kinetics [45]. The dashed V-shape in this panel represents a hypothetical chevron plot that would satisfy criterion (iii) above, which is needed for a consistent two-state description of thermodynamics and kinetics. But both the folding and unfolding arms of the theoretical chevron plots are significantly nonlinear, exhibiting so-called "chevron rollovers" and deviating substantially from the hypothetical two-state chevron plot (dashed V-shape). More detailed analyses indicate that chevron rollovers in these continuum models are caused by mild kinetic trapping. Similar chevron rollovers have been observed in three-dimensional lattice models as well [44]. These findings strongly suggest that physical interactions in real, small proteins are much more specific than we otherwise imagined. Mechanisms must exist to allow some proteins to avoid kinetic traps and to create a high free energy barrier between the folded and denatured conformations. While the precise nature of these mechanisms are yet to be discovered, we have little doubt that the elucidation of their physical origins would contribute immensely to the deciphering of protein energetics in general.

OUTLOOK

This article has focused primarily on simplified modeling approaches to protein folding. Given the current state of our knowledge, simplified models are extremely useful for identifying problems, posing questions and testing new concepts. For instance, the basic issue of protein folding cooperativity (see above) has not been adequately addressed by detailed molecular dynamics simulations. However, using simplified three-dimensional lattice models, we have recently shown that apparent two-state protein folding kinetics can arise from a cooperative interplay between conformational preferences that affect sequentially local parts of the protein chain and favorable contact interactions among amino acid residues far apart along the chain sequence [50]. We view this mechanism of local-nonlocal coupling as a "mesoscopic" organizing principle. Such higher organizing principles [7] deduced from simple modeling are expected to be useful in guiding further investigation into the atomistic origins of protein folding. In this endeavor - as in any area of science - it is important to acknowledge failures as well as recognize successes. Failures of current models are often starting points of deeper understanding, as in the case of protein folding cooperativity [51]. The many approaches to protein folding should be viewed as being complementary. These include simplified modeling as well as atomistic simulation, knowledge-based inductive constructs as well as physics-based deductive methods [14]. Protein folding is a complex phenomenon. Simply, we need all the information we can get if we ever hope to get a firm grasp of the physics necessary to solve this "second genetic code" [2].

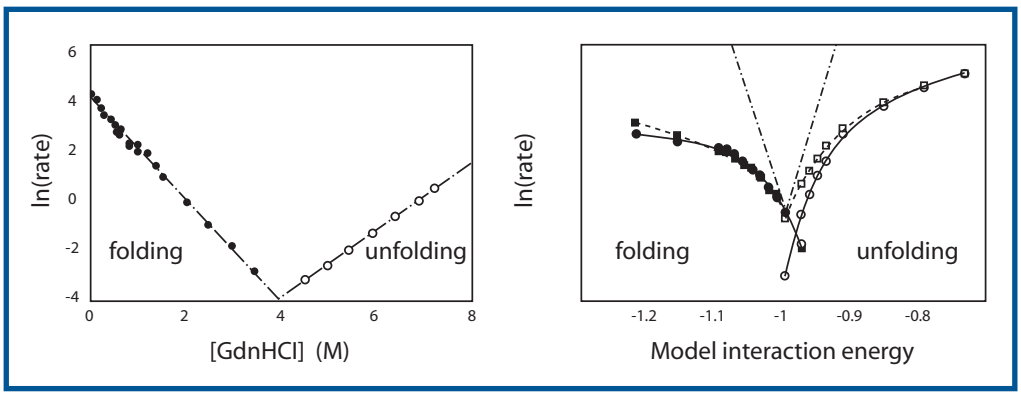

Fig. 7 **Protein folding cooperativity. Left: Experimentally measured logarithmic folding and unfolding rates of a 64-residue truncated form of the protein CI2 (cf. Fig. 1) as functions of guanidinium chloride (GdnHCl, a denaturant) concentration (data from Fig. 4 of Jackson and Fersht [46]). Because of their V-shapes, such plots are known as "chevron plots" in the protein folding literature. Right: Corresponding dependence of folding and unfolding rates on interaction energy in two native-centric (Go-like) three-dimensional continuum models of truncated CI2 with pairwise additive contact energies (adapted from Fig. 7 of Kaya and Chan [45]). The qualitative differences between experiment (left) and theory (right) underscore the deficiencies of current understanding of the high degree of cooperativity exhibited in the folding of many small proteins. See text and Ref. [45] for details.**

ACKNOWLEDGEMENTS

My research effort at the University of Toronto has been supported by the Canadian Institutes of Health Research (CIHR grant no. MOP-15323), a Premier's Research Excellence Award from the Province of Ontario, the Canadian Protein Engineering Network of Centres of Excellence (PENCE), the Ontario Centre for Genomic Computing at the Hospital for Sick Children in Toronto, and the Canada Research Chair Program.

REFERENCES:

1. T.E. Creighton, "Proteins: Structures and Molecular Properties", Freeman, New York, 1993.
2. H.S. Chan and K.A. Dill, "The Protein Folding Problem", *Physics Today*, **46**(2):24-32, (1993).
3. J.Z.Y. Chen, "Understanding Protein Folding from Polymer Models", *Physics in Canada*, **59**(2):93-102, (2003).
4. C. Branden and J. Tooze, "Introduction to Protein Structure", Garland, New York and London, 1991.
5. F.M. Richards, "Areas, Volumes, Packing and Protein Structure", *Annu. Rev. Biophys. Bioeng.*, **6**:151-176, (1977).
6. M. Levitt, M. Gerstein, E. Huang, S. Subbiah and J. Tsai, "Protein Folding: The Endgame", *Annu. Rev. Biochem.*, **66**:549-579, (1997).
7. R.B. Laughlin and D. Pines, "The Theory of Everything", *Proc. Natl. Acad. Sci. USA*, **97**:28-31, (2000).
8. G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S.T. Freer and P.W. Rose, "Simulating Disorder-Order Transitions in Molecular Recognition of Unstructured Proteins: Where Folding Meets Binding", *Proc. Natl. Acad. Sci. USA*, **100**:5148-5153, (2003).
9. J.D. Bryngelson, J.N. Onuchic, N.D. Socci and P.G. Wolynes,"Funnels, Pathways, and the Energy Landscape of Protein Folding - A Synthesis", *Proteins*, **21**:167-195, (1995).
10. K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas and H.S. Chan, "Principles of Protein

Folding -A Perspective from Simple Exact Models", *Protein Sci.*, **4**:561-602, (1995).
11. H.S. Chan and K.A. Dill, "Polymer Principles in Protein Structure and Stability", *Annu Rev. Biophys. Biophys. Chem.*, **20**:447-490, (1991).
12. S.S. Plotkin and J.N. Onuchic, "Understanding Protein Folding with Energy Landscape Theory - Part I: Basic Concepts", *Q. Rev. Biophys.*, **35**:111-167, (2002).
13. S.S. Plotkin and J.N. Onuchic, "Understanding Protein Folding with Energy Landscape Theory - Part II: Quantitative Aspects", *Q. Rev. Biophys.*, **35**:205-286, (2002).
14. H.S. Chan, H. Kaya and S. Shimizu, "Computational Methods for Protein Folding: Scaling a Hierarchy of Complexities", in *Current Topics in Computational Molecular Biology*, Eds. T. Jiang, Y. Xu and M.Q. Zhang, The MIT Press, Cambridge, Massachusetts, 403-447, 2002.
15. H.S. Chan and E. Bornberg-Bauer, "Perspectives on Protein Evolution from Simple Exact Models", *Appl. Bioinformatics,* **1**:121-144, (2002).
16. A.R. Davidson, K.J. Lumb and R.T. Sauer, "Cooperatively Folded Proteins in Random Sequence Libraries", *Nature Struct. Biol.*, **2**:856-864, (1995).
17. H.S. Chan, "Folding Alphabets", *Nature Struct. Biol.* **6**:994-996, (1999).
18. H. Wille, M.D. Michelitsch, V. Guenebaut, S. Supattapone, A. Serban, F.E. Cohen, D.A. Agard and S.B. Prusiner, "Structural Studies of the Scrapie Prion Protein by Electron Crystallography", *Proc. Natl. Acad. Sci. USA,* **99**:3563-3568, (2002).
19. P.M. Harrison, H.S. Chan, S..Prusiner and F. E. Cohen, "Thermodynamics of Model Prions and its Implications for the Problem of Prion Protein Folding", *J. Mol. Biol.*, **286**:593-606, (1999).
20. P.M. Harrison, H.S. Chan, S. Prusiner and F.E. Cohen, "Conformational Propagation with Prion-Like Characteristics in a Simple Model of Protein Folding", *Protein Sci.*, **10**:819-835, (2001).
21. S. Miyazawa and R.L. Jernigan, "Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation", *Macromolecules,* **18**:534-552, (1985).
22. S. Miyazawa and R.L. Jernigan, "Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term for Simulation and Threading", *J. Mol. Biol.*, **256**:623-644, (1996).
23. P.D. Thomas and K.A. Dill, "Statistical Potentials Extracted from Protein Structures: How Accurate are They?", *J. Mol. Biol.*, **257**:457-469, (1996).
24. J. Skolnick, L. Jaroszewski, A. Kolinski and A. Godzik, "Derivation and Testing of Pair Potentials for Protein Folding. When is the Quasichemical Approximation Correct?", *Protein Sci.*, **6**:676-688, (1997).
25. H.Y. Zhou and Y.Q. Zhou, "Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction", *Protein Sci.*, **11**:2714-2726, (2002).
26. L.R. Pratt and D. Chandler, "Theory of the Hydrophobic Effect", *J. Chem. Phys.*, **67**:3683-3704, (1977).
27. A. Geiger, A. Rahman and F.H. Stillinger, "Molecular Dynamics Study of the Hydration of Lennard-Jones Solutes", *J. Chem. Phys.*, **70**:263-276, (1979).
28. C. Pangali, M. Rao and B.J. Berne, "A Monte Carlo Simulation of the Hydrophobic Interaction", *J. Chem. Phys.*, **71**:2975-2981, (1979).
29. S. Shimizu and H.S. Chan, "Temperature Dependence of Hydrophobic Interactions: A Mean Force Perspective, Effects of Water Density, and Nonadditivity of Thermodynamic Signatures", *J. Chem. Phys.*, **113**:4683-4700, (2000).
30. B. Roux and T. Simonson, "Implicit Solvent Models", *Biophys. Chem.*, **78**:1-20, (1999).
31. S. Shimizu and H.S. Chan, "Origins of Protein Denatured States Compactness and Hydrophobic Clustering in Aqueous Urea: Inferences from Nonpolar Potentials of Mean Force", *Proteins,* **49**:560-566, (2002).
32. S. Shimizu and H.S. Chan, "Anti-Cooperativity and Cooperativity in Hydrophobic Interactions: Three-Body Free Energy Landscapes and Comparison with Implicit-Solvent Potential Functions for Proteins", *Proteins,* **48**:15-30, (2002).
33. H.S. Chan, "Protein Folding: Matching Speed and Locality", *Nature,* **392**:761-763, (1998).
34. H.S. Chan, "Modeling Protein Density of States: Additive Hydrophobic Effects are Insufficient for Calorimetric Two-State Cooperativity" *Proteins,* **40**:543-571, (2000).
35. Y. Cui, W.H. Wong, E. Bornberg-Bauer and H.S. Chan, "Recombinatoric Exploration of Novel Folded Structures: A Heteropolymer-Based Model of Protein Evolutionary Landscapes", *Proc. Natl. Acad. Sci. USA,* **99**:809-814, (2002).
36. A. Irbäck and E. Sandelin, "On Hydrophobicity Correlations in Protein Chains", *Biophys. J.*, **79**:2252-2258, (2000).
37. H.S. Chan and K.A. Dill, "Comparing Folding Codes for Proteins and Polymers", *Proteins,* **24**:335-344, (1996).
38. H. Li, R. Helling, C. Tang and N. Wingreen, "Emergence of Preferred Structures in a Simple Model of Protein Folding", *Science,* **273**:666-669, (1996).
39. E. Bornberg-Bauer and H.S. Chan, "Modeling Evolutionary Landscapes: Mutational Stability, Topology and Superfunnels in Sequence Space", *Proc. Natl. Acad. Sci. USA,* **96**:10689-10694, (1999).
40. S. Govindarajan and R.A. Goldstein, "Evolution of Model Proteins on a Foldability Landscape", *Proteins,* **29**:461-466, (1997).
41. G. Tiana, R.A. Broglia and E.I. Shakhnovich,"Hiking in the Energy Landscape in Sequence Space: A Bumpy Road to Good Folders", *Proteins,* **39**:244-251, (2000).
42. H. Kaya and H.S. Chan, "Polymer Principles of Protein Calorimetric Two-State Cooperativity", *Proteins,* **40**:637-661, (2000).
43. H. Kaya and H.S. Chan, "Energetic Components of Cooperative Protein Folding", *Phys. Rev. Lett.*, **85**:4823-4826, (2000).
44. H. Kaya and H.S. Chan, "Origins of Chevron Rollovers in Non-Two-State Protein Folding Kinetics", *Phys. Rev. Lett.*, **90**:258104, (2003).
45. H. Kaya and H.S. Chan, "Solvation Effects and Driving Forces for Protein Thermodynamic and Kinetic Cooperativity: How Adequate is Native-Centric Topological Modeling?", *J. Mol. Biol.*, **326**:911-931, (2003).
46. S.E. Jackson and A.R. Fersht, "Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition", *Biochemistry,* **30**:10428-10435, (1991).
47. T. Veitshans, D. Klimov and D. Thirumalai, "Protein Folding Kinetics: Timescales, Pathways and Energy Landscapes in Terms of Sequence-Dependent Properties", *Fold. Des.*, **2**:1-22, (1997).
48. C. Clementi, H. Nymeyer and J.N. Onuchic, "Topological and Energetic Factors: What Determines the Structural Details of the Transition state Ensemble and 'En-Route' Intermediates for Protein Folding? An Investigation for Small Globular Proteins", *J. Mol. Biol.*, **298**:937-953, (2000).
49. T. Head-Gordon and S. Brown, "Minimalist Models for Protein Folding and Design", *Curr. Opin. Struct. Biol.*, **13**:160-167, (2003).
50. H. Kaya and H.S. Chan, "Contact Order Dependent Protein Folding Rates: Kinetic Consequences of a Cooperative Interplay Between Favorable Nonlocal Interactions and Local Conformational Preferences", *Proteins,* **52**:524-533(2003).
51. H.S. Chan, S. Shimizu and H. Kaya, "Cooperativity Principles in Protein Folding", *Methods Enzymol.*, **380**:350-379, (2004).