Article No. jmbi.1999.3006 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 293, 283–293

JMB

Protein Folding: From the Levinthal Paradox to Structure Prediction

Barry Honig

Department of Biochemistry and Molecular Biophysics Columbia University 630 West 168 St. New York, NY 10032, USA

This article is a personal perspective on the developments in the field of protein folding over approximately the last 40 years. In addition to its historical aspects, the article presents a view of the principles of protein folding with particular emphasis on the relationship of these principles to the problem of protein structure prediction. It is argued that despite much that is new, the essential elements of our current understanding of protein folding were anticipated by researchers many years ago. These elements include the recognition of the central importance of the polypeptide backbone as a determinant of protein conformation, hierarchical protein folding, and multiple folding pathways. Important areas of progress include a detailed characterization of the folding pathways of a number of proteins and a fundamental understanding of the physical chemical forces that determine protein stability. Despite these developments, fold prediction algorithms still encounter difficulties in identifying the correct fold for a given sequence. This may be due to the possibility that the free energy differences between at least a few alternate conformations of many proteins are not large. Significant progress in protein structure prediction has been due primarily to the explosive growth of sequence and structural databases. However, further progress is likely to depend in part on the ability to combine information available from databases with principles and algorithms derived from physical chemical studies of protein folding. An approach to the integration of the two areas is outlined with specific reference to the PrISM program that is a fully integrated sequence/structural-analysis/fold-recognition/homology model building software system.

Keywords: protein folding; folding pathways; protein stability; protein structure prediction

Introduction

This article represents a personal account of the developments in the field of protein folding over the past 40 years. Although I hope that the historical perspective offered in the article will be of interest, it is my intention as well to present a view of the principles of protein folding with particular emphasis on the relationship of these principles to the problem of protein structure prediction. The historical aspects of the article are entirely subjective and reflect the evolution of my own thinking in the area and the influence of others in this process.

I first became aware of the field of protein folding when, in 1970, I became a postdoc, in the laboratory, of Cyrus Levinthal in the Department of Biological Sciences of Columbia University. The field has changed and grown dramatically since that time and has, in fact, split into two fairly distinct research areas, one involving the study of the physical chemical principles that underlie protein stability and folding pathways, and the other involving pure protein structure prediction. There is much overlap between the two, and they have begun to merge again, but the goals and methods used in each field are still quite different. In the 1970s we believed that protein structure prediction required first an understanding of folding energetics and folding pathways. This has clearly

Abbreviations used: MM, molecular mechanics; PB, Poisson-Boltzmann.

E-mail address of the corresponding author: bh6@columbia.edu

0022-2836/99/420283–11 $30.00/0

proved not to be the case as is evident from the fact that so much of structure prediction today is based on the identification and exploitation of related sequences and structures, as in homology modeling and fold recognition. Nevertheless, it is likely that a complete solution to all aspects of the protein folding problem will require that the information available from the analysis of large structural and sequence databases be combined with the principles and methods of physical chemistry and computational chemistry. One of the goals of this article is to point to ways in which the two research areas can be integrated.

I have included the ''Levinthal Paradox'' in the title of this article because it has become so central to current discussions of ''new'' and ''old'' views of protein folding. After a discussion of this and other issues that relate primarily to folding pathways, the article proceeds to a discussion of protein energetics and protein stability. These sections rely heavily on work from my own lab with reference to early work upon which many of my own opinions are based. These sections are followed by a section on structure prediction based primarily on the analysis of sequence and structural databases. Finally, a brief attempt will be made to integrate the different sections with an eye to the future.

The Levinthal paradox and folding pathways

Much has been written about the Levinthal paradox and its various resolutions. The paradox involves the observation that there is insufficient time to randomly search the entire conformational space available to a polypeptide chain as an unfolded protein (Levinthal, 1968). The obvious resolution, and this was Levinthal's point, is that proteins have to fold through some directed process. The challenge then is not to resolve the paradox itself, but rather to address scientific questions such as the nature of the directed process, the identity of folding nuclei, and how the primary sequence codes for kinetics as well as for thermodynamic stability.

The old view has been defined as describing folding as akin to chemical reactions involving distinct intermediates and transition states. Although this language was sometimes used, I do not believe that many people thought in these simplistic terms. As an example, in a 1976 paper (Honig *et al*., 1976) Levinthal and I wrote that ''We assume that proteins fold by following a multiply branched pathway in which the first stage is the formation of local secondary structure governed by interactions that are near each other in the peptide chain. Subsequently, these structures, such as α-helices and antiparallel β-strands, would interact, perhaps being modified in the process, to produce larger structural fragments which then undergo further assembly to yield the native conformation.'' In retrospect, a multiply branched pathway is somewhat vague concept, not surprising given the paucity of experimental information at that time about folding pathways. The phrase was intended to imply that even though there is no fixed sequence of events in folding, there is a general order of events that can be described in a phenomenological sense in terms of the progressive accumulation of tertiary structure from smaller fragments. We were well aware, of course, that these fragments were not stable in isolation and I do not recall any thought that a ''pathway'' required the existence of an observable intermediate. Rather, our thinking was influenced by the existing literature on the helix coil transition of polyamino acids in which secondary structure formation was described in terms of ensembles of states (Lifson & Roig, 1961; Zimm & Bragg, 1959).

We were not alone in this hierarchical view of protein folding. Ptitsyn (1973) had expressed similar views a few years earlier while Karplus & Weaver (1976) introduced ''the diffusion collision model'' which described the coalescence of secondary structure units that were in themselves unstable. Hierarchic folding as implied here is lucidly discussed in two recent articles by Baldwin & Rose (1999a,b). As they point out, the concept allows for alternate pathways of self-assembly with many possible folding routes.

The new view describes protein folding in terms of statistical ensembles of states (Dill & Chan, 1997; Dobson & Karplus, 1999; Wolynes *et al*., 1995) and focuses on the general features of folding on a complex multidimensional potential energy functional. This has been described in terms of a folding funnel that embodies the idea that there is a large ensemble of states available to the unfolded protein and far fewer to the folded protein. The width of the funnel is related to configurational entropy of the polypeptide chain, while the depth of the funnel depicts a free energy function that does not include the protein's internal degrees of freedom. A funnel implies that for protein folding there is a decrease in energy and concomitant loss of entropy with increasing structure. This of course must be the case, since any increase in structure requires a loss of configurational entropy which must be balanced by a decrease in energy if the state is to be populated. As an example, Mirankar & Dobson (1996) depicted a hypothetical folding pathway of a helical protein in terms of a funnel but, interestingly, the pathway appears closely related to the old view of hierarchical folding.

Funnel diagrams provide a framework for a statistical mechanical treatment of folding in terms of a few critical parameters describing the funnel shape and the flatness of its side. These, in turn, can be conveniently related to computer simulations of protein folding that have been made possible by simple lattice models (Dill & Chan, 1997; Dobson & Karplus, 1999; Wolynes *et al*., 1995). Pande *et al*. (1998) have coined the term neoclassical view to emphasize the point that the use of statistical mechanical ensembles and the results of lattice models are not necessarily inconsistent with the classical pathway picture. What appears

new to me is that many concepts are now defined much more precisely in a theoretical sense, while experimental progress has made it possible both to test specific hypotheses and to provide the data required for more precise formulations of the mechanism of protein folding than were previously possible.

In parallel with the increasingly rigorous theoretical descriptions of possible folding pathways, there have been major experimental advances in recent years. Detailed and compelling descriptions of the folding of a number of proteins have been provided (see, for example, Englander *et al*., 1998; Fersht *et al*., 1992; Hughson *et al*., 1991; Matthews, 1993; Radford *et al*., 1992)) and this has enabled a far more focused and meaningful discussion of the mechanism of protein folding than was possible in the past. Many controversies still exist, but in my view a fairly coherent picture of protein folding is emerging. In many ways, this picture is remarkably consistent with older ideas in that much of the discussion involves secondary structure formation as an early event in folding. The alternative that has been proposed is that the earliest event corresponds to a hydrophobic collapse into a state from which secondary structure can then form.

Isolated helices solve the Levinthal paradox through a nucleation propagation mechanism (Lifson & Roig, 1961; Zimm & Bragg, 1959; Zwanzig, 1995) and this can be an extremely fast process. Most experimental evidence clearly points to the conclusion that secondary structure formation occurs at the earliest stages in protein folding. Indeed global hydrophobic collapse in the absence of secondary structure formation has not to my knowledge been observed. Early stages in protein folding almost certainly involve the transient formation or ''flickering'' of units of secondary structure (Anfinsen, 1973) which are stabilized, or perhaps nucleated, by tertiary interactions with either other units of secondary structure or with hydrophobic residues in still relatively unstructured parts of the chain (Honig & Cohen, 1996). This seems to be the case in the folding of chymotrypsin inhibitor 2 (CI2) where there is no detectable intermediate and where the folding nucleus has been described in terms of local secondary structure stabilized by tertiary interactions (Fersht, 1997). Fersht has used the term nucleation-condensation to describe this type of process and distinguishes it from classical nucleation in which structure grows from a strong localized nucleus (Fersht, 1997). The multiphasic folding kinetics observed for a number of larger proteins might, in principle, point to a different mechanism, but it has been pointed out (Baldwin & Rose, 1999a,b; Miranker & Dobson, 1996) that the difference may not be fundamental but, rather, may simply reflect differences in stability of intermediates along a folding pathway.

How does the coherent picture about folding kinetics that is now emerging help us in the problem of sequence analysis and structure prediction? Numerous studies indicate that local conformational preferences are coded for in the sequence (see, for example, Baldwin & Rose, 1999a; Wright *et al*., 1988) and that, for example, regions of the sequence that are observed to be helical in a folded protein or in folding intermediates have large helical propensities in isolation (Dyson *et al*., 1992). This is not true for all secondary structure elements in a given protein, some of which may be able to change conformation in different environments (Minor & Kim, 1996) but it is true for some and presumably these are the ones that play a crucial role in driving the folding process. The approximately 70 % success rate of secondary structure prediction algorithms (Rost & Sander, 1993) is of course entirely consistent with the important role of local sequence patterns in determining secondary structure. The fact that the success rate is not 100 % is certainly due in part to the role of tertiary interactions in fixing some secondary structural elements. It is likely to be important in fold prediction to find a way to discriminate weak from strong secondary structure signals in analyzing a particular sequence or sequence family.

Do protein folding studies also provide us with information as to how sequence might code for topology? One conclusion is that despite the apparently important role for close packing in determining protein stability (see below) packing is unlikely to be an important topological determinant. This is suggested from the fact that many features of the native topology are determined at the stage of folding intermediates even though a tightly packed core has not yet formed (see, for example, Alm & Baker, 1999; Jennings & Wright, 1993). It is not clear however what sequence elements code for topology. One factor is certainly the existence of hydrophobic faces on the surface of secondary structure elements, but the problem, in general, may turn out to be more complicated. This conclusion is suggested from the fact that it is often difficult in fold recognition challenges to identify the correct topology of secondary structure elements even when the protein class (all α, α/β, all β) is correctly identified. The problem may well be due to an incomplete understanding of the energetic determinants of protein folding and to the possibility that the free energy difference between a subset of alternative protein folds for a particular sequence is smaller than generally expected. This issue is discussed in further detail in the next section.

What stabilizes folded proteins and folding intermediates?

Calculations at the atomic level

Molecular mechanics (MM) calculations have made important contributions to our understanding of protein stability. Harold Scheraga's group (Scheraga, 1968) and, in parallel, Shneior Lifson's group (Levitt & Lifson, 1969) made seminal advances in MM methodology and in the develop-

ment of force fields that were fit to experimental measurements on small molecules. The Levinthal group developed its own program that combined MM calculations with interactive computer graphics (Katz & Levinthal, 1972). A natural evolution of the molecular mechanics model of proteins led to the field of the molecular dynamics of proteins that has had such major impact. Much of the seminal work in this area originated in Martin Karplus' laboratory (see, for example, McCammon *et al*., 1979).

Early calculations on proteins tended to ignore solvent or to account for solvent effects implicitly in the form of parameter adjustment or distant-dependent dielectric constants. In time, solvent molecules were explicitly included in the simulations, but this approach added significant computer time and was of uncertain accuracy in the treatment of electrostatic interactions. In addition, ionic strength effects were extremely difficult if not impossible to treat. This was in part the motivation of my group in returning to classical electrostatics in the form of the Poisson-Boltzmann (PB) equation as a means of treating solvent effects. Warwicker & Watson (1982) used finite difference methods to solve the PB equation in 1982 (the FDPB method), but had used their method primarily to obtain graphical depictions of electrostatic potentials. My group, particularly Michael Gilson, Alex Rashin, Kim Sharp and Anthony Nicholls, in collaboration with Rick Fine, was able to demonstrate that classical electrostatics could yield numerical results of near experimental accuracy for phenomena such as charge-charge interaction energies, solvation free energies, a variety of salt effects on binding, and electrostatically enhanced diffusion (reviewed by Honig & Nicholls, 1995). This suggested that it should be possible to apply the methodology to the study of protein stability.

The free energy balance

In 1962 Tanford attempted to describe the energetics of protein folding primarily in terms of a balance between hydrophobic interactions and configurational entropy (Tanford, 1962). Tanford's conclusions were that the hydrophobic effect and configurational entropy made approximately equal in magnitude but opposite in sign contributions to protein stability. He discussed the requirement that all polar groups form hydrogen bonds either with the solvent or with other polar groups in the protein, but implicitly assumed that hydrogen bonds made little or no contribution to the stability of proteins. Rather he viewed the necessity of forming hydrogen bonds as a constraint that limits the total number of conformations available to a folded polypeptide chain. Tanford also assumed that ionizable amino acids also made little or no contribution to stability. An updated discussion of these and many other related issues is provided in an excellent review by Dill (1990).

In parallel with advances in our understanding of protein folding kinetics, the last decade has witnessed major advances in our understanding of protein stability. A seminal paper in these developments was published by Murphy, Privalov and Gill (MPG) (Murphy *et al*., 1990). The paper reported plots of entropy, enthalpy and free energy as a function of heat capacity for a number of proteins and provided similar plots for the dissolution of gaseous, liquid and solid hydrocarbons in water. The data were surprising in a number ways, and consequently galvanized much theoretical and experimental work as well as considerable discussion (Lazaridis *et al*., 1995; Makhatadze & Privalov, 1995; Yang *et al*., 1992). Privalov, Gill and coworkers had already shown that heat capacity changes upon unfolding were correlated with expectations based on an analysis of the entropies of dissolution of hydrocarbons (Privalov & Gill, 1988). Moreover, observed heat capacity changes were consistent with expectations based on the relative hydrophobicity of different proteins (Makhatadze & Privalov, 1993, 1995). Thus myoglobin which buries a large amount of non-polar suface area (about 60 % nonpolar) has a large heat capacity increment upon unfolding (on a per residue basis), while ribonuclease which has a much smaller heat capacity change buries less than 50 % non-polar area. Yet a plot of free energy *versus* heat capacity change (i.e. relative hydrophobicity) is essentially flat (Murphy *et al*., 1990), indicating that despite the apparently dominant contribution of hydrophobicity to protein stability, relative hydrophobicity is not correlated with relative stability. This result is, on the face of it, quite surprising.

Since entropies of unfolding were found to be correlated in the expected way with heat capacity changes and relative hydrophobicities, the source of the surprise was in the enthalpies of unfolding which are inversely correlated with heat capacity changes. Simply stated, as proteins bury more nonpolar area per amino acid upon folding their enthalpies of unfolding become less positive. What is the source of this correlation? The most straightforward answer is that more polar proteins have more internal hydrogen bonds and that these stabilize proteins enthalpically. However, there are a number of problems with this explanation. First, many proteins such as myoglobin have enthalpies of unfolding that are close to zero on a per residue basis (Makhatadze & Privalov, 1993, 1995). It is not clear how this is possible if hydrogen bonds provide an enthalpic driving force for folding. A second problem with assuming that hydrogen bonds stabilize folded proteins is that theoretical calculations suggest that polar groups prefer to be fully solvated in water rather than hydrogen bonded in the interior of a protein (Honig & Yang, 1995). The major counter argument to the theoretical predictions is the quite general experimental observation that removing one member of a hydrogen bonded pair destabilizes proteins (Shirley *et al*.,

1992). However, as we have discussed, this type of experiment does not address the question of the contribution of a hydrogen bond to protein stability as it simply leaves an unsatisfied hydrogen bond in the protein interior. It appears then that hydrogen bonds make little or no direct contribution to protein stability but, as Tanford pointed out many years ago (Tanford, 1962), they provide a crucial constraint on allowable folds of the polypeptide chain.

If we accept the evidence that the solvation of polar groups enthalpically favors the unfolded state, the small net enthalpies of unfolding imply that there must be a compensating enthalpic term that favors the folded state (Yang *et al*., 1992). This term almost certainly corresponds to close packing, that is, enhanced van der Waals interactions in the tightly packed protein interior relative to those in the aqueous environment of the unfolded state (Nicholls *et al*., 1991). Given the evidence that the interiors of proteins have densities that correspond to those of organic crystals, there must be enhanced packing interactions in the protein interior (Richards & Lim, 1993). The nature of these interactions and the ability of proteins to accommodate alternate packing arrangements have been clarified by the studies of Sauer and co-workers (Lim & Sauer, 1989). Based on an analysis of the melting of hydrocarbons, we have shown that the enthalpic contribution of close packing is quite large (Nicholls *et al*., 1991). However, the free energy contribution is less significant since freezing of side chains must always be accompanied by an entropic penalty.

If enthalpies of unfolding represent a balance between solvation effects and close packing, how does this help us understand the negative slope of the MPG enthalpy plot. That is, why are proteins enthalpically stabilized as they become more polar. One possibility is suggested by the important observation that entropies of unfolding are positively correlated with compressibility (Phelps *et al*., 1998). Since entropies of unfolding are in turn correlated with hydrophobicity (Murphy *et al*., 1990), the implication is that proteins with more non-polar interiors are more compressible, and hence less well packed than more polar proteins. Although the source of the effect is uncertain, it appears likely that the close approach of atoms made possible by hydrogen bonds and the energetic barriers preventing them from breaking lead to a more closely packed interior than a region that consists predominantly of non-polar groups. In such a region, the uniform size of the atoms and the lack of a strong restoring force would be expected to result in a less solid-like local environment. Thus, hydrogen bonds may provide an indirect non-local driving force stabilizing proteins even though the direct local contribution of hydrogen bonding groups appears to be neutral or slightly destabilizing (see also below).

The accumulated evidence then suggests that protein folding is driven by the hydrophobic effect and the enthalpic stabilization afforded by close packing in the solid-like protein interior (Honig & Yang, 1995). The contribution of hydrogen bonds to protein stability is small, although, given the argument above, it appears difficult to quantify. Electostatic interactions involving ionizable groups have been predicted, based on FDPB calculations, to make only a small relative contribution to protein stability. Overall they appear to be slightly destabilizing due to the partial desolvation that accompanies any folding process (Yang & Honig, 1993, 1994). Of course, since pH changes can denature proteins electrostatic effects are clearly important but their relative effect is small within the total free energy balance that governs folding. It is important to recognize in this regard that given that the free energies of folding for most proteins are on the order of only 10 kcal/mol, interactions of this magnitude can have a significant influence on observed stability while being relatively insignificant within the total free energy balance.

Due to desolvation effects, individual ion pairs buried in a low dielectric environment have been predicted theoretically to be less stable than the isolated ion pairs free in aqueous solution (Honig & Hubbell, 1984). Using the FDPB method, Hensch & Tidor (1994) found that most buried salt bridges are destabilizing, thus predicting that replacing ion pairs with non-polar groups of the same size will stabilize proteins. This prediction has been verified by the experiments of Waldberger *et al*. (1995) on the Arc repressor. In certain cases ion pairs can, however, be stabilizing. They appear to be so when located on the protein surface or if they are organized in the protein interior in networks in which they undergo stabilizing interactions with one another. These features appear to be exploited by many hyperthermophilic proteins so as to gain added stability relative to their mesophilic homologs (Xiao & Honig, 1999).

Secondary structure

It seems clear that many of the essential elements of protein energetics can be derived from understanding secondary structure formation and secondary structure propensities. Much of my own thinking in this area has been influenced by the early experimental work by Scheraga, Katchalsky and others on the physical chemical properties of amino acid homo- and heteropolymers (Ingwall *et al*., 1968), by the work of Baldwin and co-workers who succeeded in stabilizing shorter helices (Marqusee *et al*., 1989), and by the work of Dyson, Wright and colleagues (Dyson & Wright, 1991) who demonstrated that even short peptides in solution can be partially structured. The work of Serrano and co-workers should also be highlighted in this regard (see, for example, Munoz & Serrano, 1994).

In our own laboratory, An-Suei Yang embarked on a series of studies aimed at understanding the

energetic basis of secondary structure formation. We used the ability to reproduce experimentally derived observations as a test of the accuracy of our calculations. Using methods and parameters derived either from small molecule data or quantum mechanics, we were able to account for the magnitude of the σ and s parameters that arise in the Zimm-Bragg theory of the helix coil transition (Yang & Honig, 1995a), the enthalpy change associated with helix formation, the handedness and range of observed twists of β-sheets in globular proteins (Yang & Honig, 1995b) and the sequence dependence of the stability of various β-turns in proteins (Yang *et al*., 1996). The success of the calculations suggests that the overall picture of protein energetics they convey has many correct elements. This conclusion is supported by the fact that MD simulations on a number of related problems have yielded overall similar results as well as a similar physical picture (see, for example, Tobias & Brooks, 1991; Yan *et al*., 1993).

The calculations support a description of the determinants of polypeptide and protein stability that was briefly summarized in the second part of this article. Polyalanine α-helices are found to be stabilized by hydrophobic interactions and close packing, primarily involving partial burial of the β carbon atom, while hydrogen-bonding groups, make little or no contribution to helix stability and may even be marginally destabilizing due to the desolvation of polar groups upon helix formation (Yang & Honig, 1995a). In a sense, helix formation can be viewed as a form of constrained hydrophobic collapse in that hydrogen bond formation provides a necessary constraint on the allowable conformations of the collapsed state (i.e. helical polypeptide). Polyalanine β-sheets are found to be less stable than α-helices due in large part to the absence of stabilizing electrostatic interactions between aligned dipoles in the helix (Yang & Honig, 1995b). This accounts for the fact that alanine has a greater helix than sheet propensity. The β-sheets are stabilized primarily by non-polar interactions between residues on adjacent strands. There must be enough of these interactions to overcome the intrinsic instability of β-sheets and this accounts for the fact that peptide fragments that form β-sheets in proteins tend to be hydrophobic.

Overall, our calculations support the notion that there are strong local sequence propensities to form specific secondary structures including β-turns. This is consistent with the success of secondary structure prediction algorithms and with the fact that some peptide fragments that are helical in a folded protein have a significant tendency to form helices when isolated in solution (Dyson *et al*., 1992). Of course in many cases secondary structure is determined by tertiary interactions, and in fact there are a number of striking examples where a segment of a protein changed its secondary structure in response to environmental changes. Given the subtle and delicate balance between the forces that determine protein folding, perhaps this is not too surprising.

Implications for folding pathways

The preceding discussion is couched entirely in terms of free energy changes in a hypothetical folding process and in itself says nothing about folding pathways. However the discussion has what I believe to be a number of important implications for pathways. First, the existence of marginally stable secondary structure elements is an intrinsic property of any polypeptide chain; indeed the formation of any secondary structure element is a downhill process energetically for essentially any sequence. (It is not necessarily downhill in free energy due to the configurational entropy price). This implies that isolated secondary structure elements can be stabilized by specific sequences, consistent with the experiments mentioned above. However, it also implies that secondary structure elements can easily change conformation in the presence of a relatively small number of tertiary interactions. That is, the free energy difference between an α-helical, β-hairpin and coil conformations for most sequences is small enough that their relative populations can be easily shifted. For example, individual helices can be transformed into β-sheets by changing just a few amino acids (Cordes *et al*., 1999) which can constitute a change in driving force of no more than a few kcal/mol. This demonstrates that proteins have a structural plasticity which allows them to change conformation readily.

A second implication is that early in the folding process there must be many different combinations of secondary structure elements with very similar stability. Indeed, as we have argued, in the absence of tertiary interactions almost any secondary structure fragment can appear at almost any location along the sequence (Yang *et al*., 1996). Of course, only the proper combination of secondary structure elements that can form favorable tertiary contacts will condense to the native structure. Secondary structure propensities appear to provide a first tier of ''screening'' and result in small but significant populations of ''correct'' secondary structure elements for a given sequence. (Correct here does not mean identical with native but rather similar to native.) However, it is the tertiary interactions that make the final selection as to the actual native topology.

This is not meant to imply that proteins always begin to fold by forming isolated secondary structure elements. The picture that emerges from purely energetic considerations is that early stages in folding involve the transient formation of secondary structure elements which are stabilized by a combination of long-range and local interactions, both of which are primarily hydrophobic (Honig & Cohen, 1996). In this sense, secondary and tertiary structure are expected to appear simultaneously in some cooperative process, although there are

clearly a number of ways for this too happen. For example, Brooks and co-workers find that secondary and tertiary structure form simultaneously in a helical protein (Boczko & Brooks, 1995); in a protein containing β-sheets, tertiary structure is also formed concomitantly with the appearance of secondary structure elements although the correct hydrogen bonds are formed relatively late in the folding process (Sheinerman & Brooks, 1998).

A third implication of polypeptide chain energetics is that there must be many folding pathways, although many of these are likely to be related in that they involve elements of the same secondary structure fragments that are seen in the native structure. In the early stages of folding the free energy differences between alternate conformations of a partially unfolded polypeptide chain are too small to limit folding to a single pathway. At a later stage in folding the free energy differences between alternate conformations may also not be large. This follows in part from the argument that the driving force for tertiary structure formation from secondary structure elements is relatively small (Yang *et al.*, 1992). Thus, although the number of pathways is clearly restricted at later stages in folding, there is always likely to be an ensemble of related but non-identical structures. The possibility that the free energy difference between at least a few alternate chain topologies is relatively small may account for some of the difficulties encountered by fold recognition scoring functions (see below).

Protein structure prediction

Over approximately the last ten years the protein prediction problem has become increasingly distinct from the protein folding problem. The former is concerned with predicting the final structure, while the latter has focused on folding pathways. The two communities have become quite diverse as well; protein prediction is primarily the domain of researchers with expertise in sequence and structure analysis, while protein folding has increasingly attracted physical chemists and polymer physicists. The latter group has fold prediction from first principles as one of its goals, but the increasing success of homology modeling and fold recognition techniques suggests that *ab initio* folding, despite the enormous fascination with this problem, is unlikely to be competitive with methods that rely primarily on the rapidly growing database of existing structures. Given the importance of structural information in modern molecular biology, there is a strong biological imperative to predict structure independent of the means used for that prediction. It becomes relevant then to ask how useful the enormous efforts that have been devoted to understanding folding pathways and folding energetics will be in structure prediction.

In my view they will be extremely useful, but it will be necessary to find a way to integrate the two fields. The flow of information can be in both ways in the sense that what we learn about the relationship between sequence and structure will help us understand folding pathways and folding energetics. In parallel, improved understanding about kinetics and stability should help us in the design of prediction algorithms. The following paragraphs describe possible modes of interplay between the two areas that are based on our own work. A key element in this work has been the development by An-Suei Yang of the PrISM (protein informatics system for modeling) program (Yang & Honig, 1999). The program, which can be run either interactively or in automatic mode, consists of a variety of linked modules which include the facility to carry out sequence analysis, structure-based sequence alignment, fast structure-structure superposition using a unique structural domain database, multiple structure alignment, fold recognition and homology model building. Many of the individual algorithms are similar to those developed by other workers, but we believe the nature of their integration as well as a number of novel features have produced a particularly effective program. PrISM was used, with a considerable level of success (Yang & Honig, 1999), to make predictions for all 43 targets at the recent CASP3 conference.

We have attempted, in retrospect, to consider the extent to which our prior work on protein stability influenced the design and applications that have been built into PrISM. One connection has been that secondary structure prediction and secondary structure topology are central elements in our various prediction algorithms. This is reflected in our domain database in which loops and the specific connectivity between secondary structure elements are important elements (A.-S. Yang & B.H., unpublished results). That is, we insist that residues in a single domain must be contiguous in sequence. The rationale in part is to obtain a continuous sequence profile for each domain. However, an additional factor results from our view that domains with similar topologies are likely to fold in similar ways and hence there should be sequence profiles characteristic of the final domain structure but also of the folding pathways.

PrISM also has a structure superposition module and structure similarity score (SSS) that rely heavily on topology as well as the number of secondary structure elements that can be superimposed (A.-S. Yang & B.H., unpublished results). This should be contrasted with similarity measures that rely on rmsd alone or on contact maps between amino acids. The energetic interplay between secondary and tertiary structure is reflected in a scoring scheme for sequence to structure alignment (used in threading and in homology model building) where secondary structure propensities are included but where they can be ''overruled'' by tertiary interactions. In this way PrISM has built into it the facility of a sequence to change its

secondary structure in the presence of specific tertiary interactions.

We can also use PrISM to extract information that will feed back to our attempt to understand the energetic and kinetic basis of protein folding. Specifically, PrISM has a multiple structure superposition module that makes it possible to extract multiple sequence alignments for protein families based solely on geometric similarities. We have obtained such alignments for groups of proteins for which a meaningful sequence alignment would be difficult if not impossible to achieve based on sequence alone. These alignments were extremely useful in building homology models for CASP3 targets with low sequence identity to the template structure (Yang & Honig, 1999). Moreover, multiple structure alignment reveals residues that are likely to play a crucial role in the protein folding process or in stabilizing the folded structure, since they are conserved among most or all structural homologs, even in cases of low sequence identity. These residues are obvious targets for mutation experiments or for theoretical attempts to understand their role in folding.

Structure-based sequence alignments offer the possibility of discovering fundamentally new insights as to the rules that determine the relationship between sequence and structure. However the optimal utilization of this type of data will require a proper theoretical framework for its interpretation. This then provides a bridge between database analysis and physical chemical studies of proteins. An area where physical chemical studies can be of enormous help is in the refinement of homology models. These suffer from alignment problems, from the problem of generating accurate loop conformations, and from the fact that the accuracy of homology models, even when the alignment is perfect, depends on the rmsd between the template and actual structure. That is, there still does not appear to be a reliable procedure where one begins with a homology model based on some template, and then relaxes the structure, using MD for example, to yield a conformation that is close to native. This is an important problem where database analysis cannot help.

A related problem is to understand the frequent failures of fold recognition algorithms to predict the proper chain topology even when the protein architecture is correctly predicted (for example a β-sandwich is correctly predicted but some of the strands are predicted to be in the wrong sheet). These difficulties may be due in part to the possibility, mentioned in the previous section, that the free energy differences between alternate chain topologies may not be large. If two alternate conformations are not in fact that different in free energy, an energetic ''scoring function'' would have to be extremely accurate to distinguish between them. Failures in fold recognition may reflect fundamental shortcomings in energy functions based on statistics alone. It remains to be seen whether physical chemical studies can solve the problem.

For the most part, existing scoring functions have been based on ''knowledge-based'' potentials that are derived from the distribution of inter-residue distances in a database of known structures (Sippl, 1995). These have proved particularly effective in fold recognition applications, but statistics should become increasingly less effective when one wishes, for example, to construct accurate ''high-resolution'' homology models. Recently, a number of labs including our own have shown that free energies calculated from force fields and solvation models are quite successful in discriminating the native conformation from well-constructed but misfolded ''decoys'' (Laziridis & Karplus, 1999; Vorobjev *et al*., 1998; D. Petrey & B.H., unpublished results). In a number of cases the energy gap between the native and misfolded structure is relatively small suggesting, in agreement with the discussion above, that there may be a number of alternate conformations that are similar in stability. Given the fact that folding free energies are so small, the incorrect conformations may be unstable relative to the unfolded manifold.

If the calculated free energy differences are at all meaningful, there are clear implications for the problem of fold prediction. Specifically, the discrimination of perhaps similar but alternate chain topologies may only be possible at high resolution. As discussed by Sauer and co-workers (Cordes *et al*., 1996), there may be more than one ''low-resolution'' topology that allows the optimal burial of non-polar groups and satisfaction of secondary structure propensities. The native conformation would then correspond to the one that also optimizes interactions that depend on high resolution atomic detail such as close packing, salt-bridges, turn propensities, etc.

Concluding remarks

In the course of writing this article I have been struck by how much of our current understanding of protein folding was anticipated by researchers many years ago. Of course there has also been enormous progress and, indeed, fairly vague and speculative ideas have been replaced by crisp experimental observations and carefully defined theoretical concepts. An essential characteristic of proteins that has been evident since the work by Pauling (Pauling & Corey, 1951) is that the polypeptide backbone is the single most important determinant of protein conformation. This is because the unique ability of polypeptides to form periodically ordered conformations that are internally hydrogen bonded results in the existence of metastable secondary structure elements, α-helices and β-sheets. These are the essential building blocks of protein conformation and we now know that their existence depends only weakly on sequence. Thus, the notion that sequence determines structure might be more precisely formu-

lated with the concept that sequence chooses between the limited number of secondary structure elements available to the polypeptide backbone and determines how they are ordered with respect to one another in space (Honig & Cohen, 1996).

Another feature that I believe to be characteristic of proteins is that the free energy differences between at least a few alternate conformations of the polypeptide chain is not large. Small energy gaps between at least a subset of alternate conformations may provide a natural basis for rapid evolutionary change.

I believe that the central role of secondary structure formation in protein folding constitutes the key element in the resolution of the Levinthal paradox. Long α-helices solve the Levinthal paradox through a nucleation/propagation mechanism, and it seems clear that proteins solve the conformation search problem based in part on the same mechanism. Of course the problem is more complicated when tertiary structure must be formed, but the fact that one or two hydrophobic contacts are large enough to balance the free energy cost of nucleating an α-helix (about 3-4 kcal/mol) provides a clear indication that the simultaneous formation of secondary and tertiary structure is feasible both in terms of energetics and kinetics.

The increased importance of three-dimensional structural information in molecular and cellular biology provides an enormous impetus to improve prediction methods so as to provide structural models that are accurate enough to have biological impact, i.e. to help us understand function. The explosive growth of sequence and structural information has made it increasingly possible to construct accurate homology models in many cases, including some involving low levels of sequence identity, and this area is likely to progress rapidly as more data become available. There has also been continued and steady progress in our understanding of the physical and chemical principles that underlie protein folding and the time is now be ripe to apply these principles to the structure prediction problem. In parallel, sophisticated analysis of the information now available in sequence and structural databases has the potential to significantly enhance our understanding of folding pathways and protein stability. The full integration of these various aspects of the protein folding/prediction problem offers exciting scientific and intellectual challenges for the coming years that are magnified in importance by their potential impact on many areas of modern biology.

Acknowledgements

This article is dedicated to the memory of Cyrus Levinthal who introduced me to the protein folding problem and whose scientific standards and style have guided me for much of my career (Honig, 1991). Much of the work from my laboratory that is summarized in this article and many of the ideas that are expressed were formulated in close collaboration with Dr An-Suei Yang. His important contributions to my thinking in the general area of protein folding and structure prediction are gratefully acknowledged. I am grateful to Diana Murray, Felix Sheinerman and An-Suei Yang for their thoughtful comments on the manuscript. This work was supported in part by grants from the NIH (GM 30518) and the DOE (96ER62265).

References

Alm, E. & Baker, D. (1999). Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* **9**, 189-196.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science,* **181**, 223-230.

Baldwin, R. L. & Rose, G. D. (1999a). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26-33.

Baldwin, R. L. & Rose, G. D. (1999b). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* **24**, 77-83.

Boczko, E. M. & Brooks, C. L. (1995). First-principles calculation of the folding free energy of a three-helix bundle protein. *Science,* **269**, 393-396.

Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. (1996). Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3-10.

Cordes, M. H. J., Walsh, N. P., McKnight, C. J. & Sauer, R. T. (1999). Evolution of a protein fold *in vitro*. *Science,* **284**, 325-327.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry,* **29**, 7133.

Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10-19.

Dobson, C. M. & Karplus, M. (1999). The Fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.* **9**, 92-101.

Dyson, H. J. & Wright, P. E. (1991). Defining solution conformations of small linear peptides. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 519-538.

Dyson, H. J., Merutka, G., Waltho, J. P., Lerner, R. A. & Wright, P. E. (1992). Folding of peptide fragments comprising the complete sequence of protein models for initiation of protein folding I. Myohemerythrin. *J. Mol. Biol.* **226**, 795-817.

Englander, S. W., Sosnick, T. R., Mayne, L. C., Shtilerman, M., Qi, P. X. & Bai, Y. (1998). Fast and slow folding in Cytochrome c. *Acc. Chem. Res.* **31**, 737-744.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.,* **7**, 3-9.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771-782.

Hendsch, Z. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatics analysis. *Protein Sci.* **3**, 211-226.

Honig, B. (1991). In Memorium: Cyrus Levinthal. *Proteins: Struct. Funct. Genet.* **11**, 239-241.

Honig, B. & Cohen, F. E. (1996). Adding backbone to protein folding: Why proteins are polypeptides. *Fold. Des.* **1**, R17-R20.

Honig, B. & Hubbell, W. L. (1984). Stability of ''salt bridges'' in membrane proteins. *Proc. Natl Acad. Sci. USA,* **81**, 5412-5416.

Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science,* **268**, 1144-1149.

Honig, B. & Yang, A.-S. (1995). The free energy balance in protein folding. *Advan. Protein Chem.* **46**, 27-58.

Honig, B., Ray, A. & Levinthal, C. (1976). Conformational flexibility and protein folding: rigid structural fragments connected by flexible joints in subtilisin BPN. *Proc. Natl Acad. Sci. USA,* **73**, 1974-1978.

Hughson, F. M., Barrick, D. & Baldwin, R. L. (1991). Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry,* **30**, 4143-4148.

Ingwall, R. T., Scheraga, H. A., Lotan, N., Berger, A. & Katchalski, E. (1968). Conformational studies of poly-L-alanine in water. *Biopolymers,* **6**, 331-368.

Jennings, P. A. & Wright, P. E. (1993). Formation of a molten globular intermediate early in the kinetic folding pathway of apomyoglobin. *Science,* **262**, 892-896.

Karplus, M. & Weaver, D. (1976). Protein-folding dynamics. *Nature,* **260**, 404-406.

Katz, L. & Levinthal, C. (1972). Interactive computer graphics and representation of complex biological structures. *Annu. Rev. Biophys. Bioeng.* **1**, 465-504.

Laziridis, T. & Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477-487.

Lazaridis, T., Archontis, G. & Karplus, M. (1995). Enthalpic contribution to protein stability: Insights from atom-based calculations and statistical mechanics. *Advan. Protein Chem.* **47**, 231-306.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44-45.

Levitt, M. & Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269-279.

Lifson, S. & Roig, A. (1961). On the theory of the helix coil transition in polypeptides. *J. Chem. Phys.* **34**, 1963-1974.

Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature,* **339**, 31-36.

Makhatadze, G. I. & Privalov, P. L. (1993). Contribution of hydration to protein folding thermodynamics. I The enthalpy of hydration. *J. Mol. Biol.* **232**, 639-659.

Makhatadze, G. I. & Privalov, P. L. (1995). Energetics of protein structure. *Advan. Protein Chem.* **47**, 307-425.

Marqusee, S., Robbins, V. H. & Baldwin, R. L. (1989). Unusually stable helix formation in short alanine-based peptides. *Proc. Natl Acad. Sci. USA,* **86**, 5286-5290.

Matthews, C. B. (1993). Pathways of protein folding. *Annu. Rev. Biochem.* **62**, 653-683.

McCammon, J. A., Wolynes, P. G. & Karplus, M. (1979). Picosecond dynamics of tyrosine side chain in proteins. *Biochemistry,* **18**, 927-942.

Minor, D. L. & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature,* **380**, 730-734.

Miranker, A. D. & Dobson, C. M. (1996). Collapse and cooperativity in protein folding. *Curr. Opin. Struct. Biol.* **6**, 31-42.

Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. *Proteins: Struct. Funct. Genet.* **20**, 301-311.

Murphy, K. P., Privalov, P. L. & Gill, S. J. (1990). Common features of protein unfolding and dissolution of hydrophobic compounds. *Science,* **247**, 559-561.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281-296.

Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.

Pauling, L. & Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl Acad. Sci. USA,* **37**, 251-256.

Phelps, D. K., Rossky, P. J. & Post, C. B. (1998). Influence of an antiviral compound on the temperature dependence of viral protein flexibility and packing: a molecular dynamics study. *J. Mol. Biol.* **276**, 331-337.

Privalov, P. & Gill, S. J. (1988). Stability of protein structure and hydrophobic interaction. *Advan. Protein Chem.* **39**, 191-234.

Ptitsyn, O. B. (1973). Stage mechanism of the self-organization of protein molecules. *Dokl. Acad. Nauk.* **210**, 1213-1215.

Radford, S. E., Dobson, C. M. & Evans, P. A. (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Science,* **358**, 302-307.

Richards, F. M. & Lim, W. (1993). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423-498.

Rost, B. & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA,* **90**, 7558-7562.

Scheraga, H. A. (1968). Calculations of conformations of polypeptides. *Advan. Phys. Org. Chem.* **6**, 103-184.

Sheinerman, F. B. & Brooks, C. L., III (1998). Molecular picture of folding of a small α/β protein. *Proc. Natl Acad. Sci. USA,* **95**, 1562-1567.

Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry,* **31**, 725-732.

Sippl, M. J. (1995). Knowledge based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229-235.

Tanford, C. (1962). Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **84**, 4240-4247.

Tobias, D. L. & Brooks, C. L. (1991). Thermodynamics and mechanism of α helix inititation in alanine and valine peptides. *Biochemistry,* **30**, 6059-6070.

Vorobjev, Y. N., Almagro, J. C. & Hermans, J. (1998). Discrimination between native and intentionally misfolded conformations of proteins. *Proteins: Struct. Funct. Genet.* **32**, 399-413.

Waldburger, C. D., Schildbach, J. F. & Sauer, R. T. (1995). Are buried salt bridges important for protein stability and conformational specificity?. *Nature Struct. Biol.* **2**, 122-128.

Warwicker, J. & Watson, H. C. (1982). Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J. Mol. Biol.* **157**, 671-679.

Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995). Navigating the folding routes. *Science,* **267**, 1619-1620.

Wright, P. E., Dyson, J. & Lerner, R. A. (1988). Conformation of peptide fragments of proteins in aqueous

solution: implications for initiation of protein folding. *Biochemistry,* **27**, 7167-7175.

Xiao, L. & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.* **289**, 1435-1444.

Yan, Y., Tropsha, A., Hermans, J. & Erickson, B. W. (1993). Free energies for refolding of the common β turn into the inverse-common β turn: simulation of the role of D/L chirality. *Proc. Natl Acad. Sci. USA,* **90**, 7898-7902.

Yang, A.-S. & Honig, B. (1993). On the pH dependence of protein stability. *J. Mol. Biol.* **231**, 459-474.

Yang, A.-S. & Honig, B. (1994). Structural origins of pH and ionic strength effects on protein stability: acid denaturation of sperm whale apomyoglobin. *J. Mol. Biol.* **237**, 602-614.

Yang, A.-S. & Honig, B. (1995a). Free energy determinants of secondary structure formation: I. α-helices. *J. Mol. Biol.* **252**, 351-365.

Yang, A.-S. & Honig, B. (1995b). Free energy determinants of secondary structure formation: II. antiparallel β-sheets. *J. Mol. Biol.* **252**, 366-376.

Yang, A.-S. & Honig, B. (1999). Sequence to structure alignment in comparative modeling using PrISM. *Proteins: Struct. Funct. Genet.* In the press.

Yang, A.-S., Sharp, K. & Honig, B. (1992). Analysis of the heat capacity dependence of protein folding. *J. Mol. Biol.* **227**, 889-900.

Yang, A.-S., Hitz, B. & Honig, B. (1996). Free energy determinants of secondary structure formation: III. β-turns and their roles in protein folding. *J. Mol. Biol.* **259**, 873-882.

Zimm, B. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**, 526-535.

Zwanzig, R. (1995). Simple model of protein folding kinetics. *Proc. Natl Acad. Sci. USA,* **92**, 9801-9804.