# Five hierarchical levels of sequence-structure correlation in proteins

**Christopher Bystroff, Yu Shao and Xin Yuan**

Biology Department
Rensselaer Polytechnic Institute
Troy, 12180
518-276-3185., 518-276-2162.


{bystrc,shaoy,yuanx2}@rpi.edu

**Keywords**
protein folding, hidden Markov models, motifs, pathways, contact maps, review

**Abstract**
A review of recent work toward modeling the protein folding pathway using a bioinformatics approach is presented. Statistical models have been developed for sequence-structure correlations in proteins at five levels of structural complexity: (1) short motifs, (2) extended motifs, (3) non-local pairs of motifs, (4) three dimensional arrangements of multiple motifs, and (5) global structural homology. Here we review statistical models, including sequence profiles, hidden Markov models and interaction potentials, for the first four levels of structural detail. The I-sites Library (folding Initiation sites) models local structure motifs. HMMSTR (Hidden Markov Model for STRucture) is a hidden Markov model for extended motifs. HMMSTR-CM (Contact Maps) is a model for pairwise interactions between motifs. And SCALI-HMM (HMMs for Structural Core Alignments) is a set of hidden Markov models for spatial arrangements of motifs. Global sequence models have been extensively reviewed elsewhere and are not discussed here. The parallels between the statistical models and the theoretical models for folding pathways are discussed.

Access to the data used and algorithms presented in this paper are available at http://www.bioinfo.rpi.edu/~bystrc/ or by request to bystrc@rpi.edu. HMMSTR predictions may be obtained from this web site: **http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.html**

## Introduction

Proteins fold through a hierarchical accumulation of order, from short-ranged to long-ranged -- local to global. The folding pathway is somehow encoded in protein sequences. Recurrent patterns in the database of known proteins tell the story of the folding pathway and of its evolutionary history, but these are two entirely different stories. Evolution takes place on the time scale of millions of years, while protein folding happens in milliseconds. Most methods for predicting protein structure implicitly model

the evolutionary process. For example, a typical algorithm would compare two sequences and calculate a score based on the number and type of sequence differences. But we are interested in the folding process, not the evolutionary process. How can we find sequence-structure correlations in proteins so that they tell us about the folding process and not the evolutionary process?

Through a series of statistical models, we have constructed a picture of the protein folding pathway that agrees with theoretical models. Each statistical model measures the degree of a specific type of sequence-structure correlation in the database of known structures. From the perspective of statistical thermodynamics, a database correlation implies that the observed sequence pattern and structural conformation are associated energetically. A strong correlation represents a strong energetic interaction. Sequence structure correlations may therefore represent energetically stable states, such as intermediates along the protein folding pathway.

A folding pathway may be viewed as starting from folding initiation sites, which are local pieces of the chain that have a strong preference to fold into a certain structure. The chain then collapses locally, around the initiation sites – a process called "propagation." Pairs of collapsed structures on the chain may then "condense," or join together, subject to energetic considerations. Finally, in the last stages of the folding pathway, topological constraints predominate in dictating the packing of the preformed units along the chain (Riddle, Grantcharova et al. 1999; Plaxco, Simons et al. 2000), since at this point in folding not all pieces of the chain can reach each other.

This is a working model for the folding pathway. Based on this model, Table 1 lists five hierarchical levels of sequence-structure correlations that should exist in all globular proteins. Each level depends on the one above it.

*Table 1. Five stages of the folding pathway*

| Folding stage | Type of structure | Model(s) |
|---|---|---|
| (1) initiation | local | I-sites[1] |
| (2) propagation | extended local | HMMSTR[2] |
| (3) condensation | non-local pairwise | HMMSTR-CM[3] |
| (4) packing | non-local multibody | SCALI-HMM[4] |
| (5) final | global | Pfam, etc. |

Global sequence-structure correlations, as detected by global sequence alignments or profile hidden Markov models, such as Pfam (Sonnhammer, Eddy et al. 1998), SUPERFAMILY (Gough and Chothia 2002), and SAM (Karplus, Barrett et al. 1998), tell only the story of the last step in folding. To find the parts of the chain that define the earlier steps in folding, we have developed the models listed in the third column. A library of motifs called I-sites (Bystroff and Baker 1997; Bystroff and Baker 1998) consists of sequence-structure motifs that occur frequently in the database and which are thought to be initiation sites for protein folding. The I-sites library was extended and generalized by HMMSTR (Bystroff, Thorsson et al. 2000; Bystroff and Shao 2002), a hidden Markov model of local sequence-structure correlation, to model the prpagation of protein folding along the chain. HMMSTR-CM(Shao and

---

[1] I-sites: initiation sites, a library of short sequence-structure motifs.
[2] HMMSTR: a hidden Markov model for local protein sequence-structure correlations.
[3] HMMSTR-CM: a contact map prediction method using HMMSTR.
[4] SCALI-HMM: a hidden Markov model for protein structure core alignment.

Bystroff 2003), a contact map prediction method using HMMSTR, goes one step further to predict non-local inter-residue contacts. Finally, SCALI (Yuan and Bystroff 2004), a hidden Markov model for protein structure core alignment, models protein folding at the multibody-level. In this paper, we consider how these models work together to build a picture of the folding process.

**Protein Folding Pathways, a Brief History**

The early work of Levinthal and Anfinsen established that a protein chain folds spontaneously and reproducibly to a unique three-dimensional structure when placed in aqueous solution. Levinthal proved beyond the shadow of a doubt that the folding process cannot occur by random diffusion(Levinthal 1969). Anfinsen proposed that proteins must form intermediate structures in a time-ordered sequence of events, or "pathway" (Anfinsen and Scheraga 1975). The nature of the pathways, specifically whether they are restricted to partially native states or whether they might include non-specific interactions, such as an early collapse driven by the hydrophobic effect, was left unanswered.

Over the years, the theoretical models for folding have converged (Baldwin 1995; Colon and Roder 1996; Oliveberg, Tan et al. 1998; Pande, Grosberg et al. 1998) due to a better understanding of the structure of the "unfolded state" (Dyson and Wright 1996; Gillespie and Shortle 1997; Mok, Kay et al. 1999) and to a more detailed description of kinetic and equilibrium folding intermediates (Eaton, Thompson et al. 1996; Houry, Rothwarf et al. 1996; Gulotta, Gilmanshin et al. 2001). An image of the transition state of folding can now be mapped out by point mutations, or "phi-value analysis" (Fersht, Matouschek et al. 1992; Nolting, Golbik et al. 1997; Mateu, Sanchez Del Pino et al. 1999; Grantcharova, Riddle et al. 2000; Gromiha and Selvaraj 2002; Heidary and Jennings 2002; Garbuzynskiy, Finkelstein et al. 2004). The "folding funnel" model (Chan, Bromberg et al. 1995; Onuchic, Luthey-Schulten et al. 1997) has been reconciled with the "hydrophobic collapse" and "nucleation-condensation" models (Nolting and Andert 2000) by envisioning a distorted, funicular energy landscape (Laurents and Baldwin 1998) and a "minimally frustrated" pathway (Shoemaker and Wolynes 1999; Nymeyer, Socci et al. 2000), where the rate limiting step is a counter-entropic search for the hole in the funnel ((Zwanzig 1997). As such, the important role of the protein topology, especially as measured by the contact order, in determining the rate of folding is understood(Plaxco, Simons et al. 1998; Miller, Fischer et al. 2002; Ivankov, Garbuzynskiy et al. 2003). Our set of statistical models is consistent with current thinking about folding pathways.

**Systems and Methods**

Knowledge-based methods for protein structure prediction assume that the frequency of an observed property in the database is a measure of its free energy, provided the database has been properly corrected for redundancy and over-counting. For example, the knowledge-based free energy of a contact between a glycine and an alanine is found by counting the frequency of finding those two amino acids in contact over a database of known protein structures. In the four models presented here, we have derived Bayesian conditional probabilities for local sequence motifs (I-sites ((Bystroff and Baker 1998)), for sequential strings of multiple motifs (HMMSTR (Bystroff, Thorsson et al. 2000)), for non-local pairwise contacts between motifs (HMMSTR-CM (Shao and Bystroff 2003)), and for non-sequential three-dimensional packing arrangements of multiple motifs (SCALI (Yuan and Bystroff 2004)). The hierarchy of models can be roughly described as "local to global", mirroring the nucleation/condensation
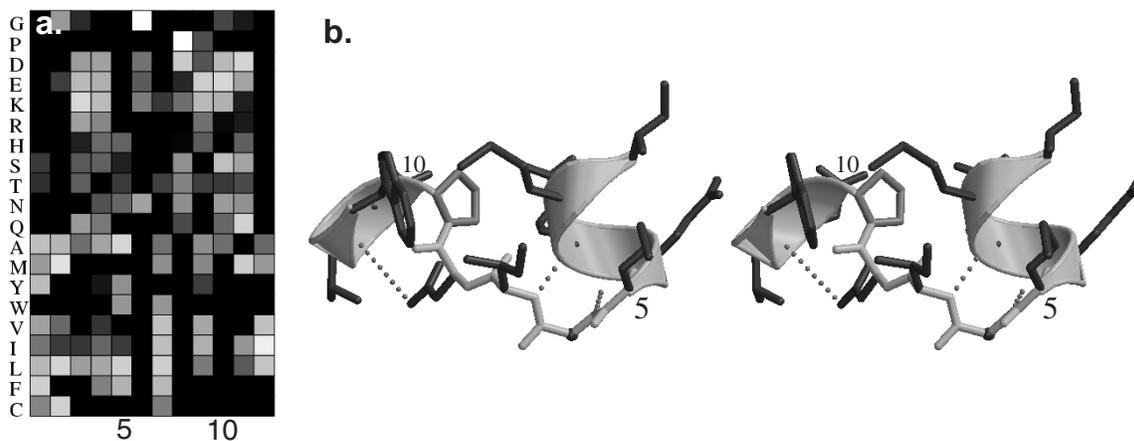
**Fig. 1.** **a**. I-sites profile for alpha-alpha corner motif. Boxes are shaded lighter in proportion to the log-likelihood ratio of each amino (Y axis) acid at each position (X axis) relative to the start of the motif. **b**. Stereo image of the alpha-alpha corner motif showing conserved H-bonds and sidechains interactions.

mechanism for protein folding. For each model the prediction results are an ensemble of conformational states. "Local " in this context means nearby along the chain.

Each of these models has been described elsewhere. Here we will review each model and discuss how they are related, and how the structure of the models may reflect the structure of the physical process that they are intended to represent.

**The I-sites Library of Folding Initiation Site Motifs**

The role of local structure motifs with regard to the initiation of folding has been discussed by Baldwin, Rooman and others (Rooman, Rodriguez et al. 1990; Efimov 1993; Baldwin and Rose 1999). Recurrent local structure motifs might exist because they fold into a specific structure independent of their context, and since the structure is small and local, the folding is fast. I-sites is a library of 262 local sequence/structure motifs. A motif is expressed as a position-specific scoring matrix, a structural model, and a "confidence curve" which maps the sequence score to a probability or "confidence." Recurrent sequence patterns of various lengths were found by first exhaustively clustering short segments of sequence families in a non-redundant database of known structures (Han and Baker 1995; Bystroff, Simons et al. 1996; Han and Baker 1996; Han, Bystroff et al. 1997), then optimizing the sequence structure correlation using reinforcement learning (Bystroff and Baker 1998). I-sites motifs have been used in blind prediction experiments (Bystroff and Baker 1997; Bystroff and Shao 2002) and have inspired several experimental studies (Jacchieri 2000; Mendes, Guerois et al. 2002; Northey, Maxwell et al. 2002; Skolnick and Kolinski 2002; Steward and Thornton 2002).

The confidence of an I-sites prediction is defined as the probability of the prediction being correct, given the sequence score. A score-to-confidence mapping was found by making predictions on a large test set of proteins that were not used to build the model. Only about one-third of all residues in all proteins are found in high-confidence (>70%) I-sites motif regions. But nearly all residues in all proteins (98%) belong to one or more of the I-sites motif structures, although many are difficult to predict.

I-sites motifs include alpha helix, helix caps, beta strands, beta hairpins and other loop structures. **Fig. 1** shows one of the I-sites motifs, the alpha-alpha corner, and the sequence pattern that predicts it.
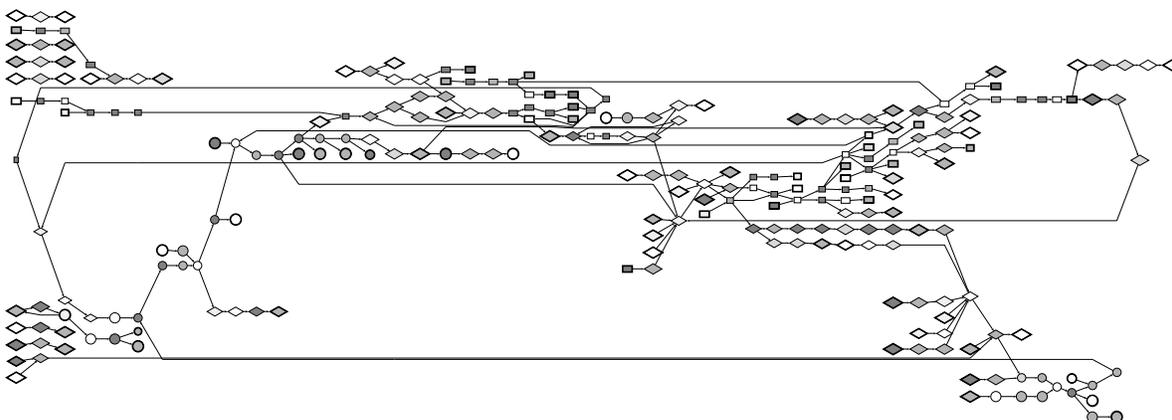
**Fig. 2.** HMMSTR model "R" represented as a directed graph. The symbol shape represents the secondary structure type; circles: helix; rectangles: beta sheet; diamonds: other motifs. Shading represents the amino acid preference; dark grey: non-polar; grey: polar; light-grey: proline; lightest grey: glycine; white: no preference. Only high-probability transitions are shown.

Peptide sequences that match the I-sites motifs have been shown to be structurally stable in isolation in both NMR studies (Blanco, Rivas et al. 1994; Munoz, Blanco et al. 1995; Viguera and Serrano 1995; Yi, Bystroff et al. 1998) and in molecular dynamics simulations (Krueger and Kollman 2001; Gnanakaran and Garcia 2002; Bystroff and Garde 2003). Mutations in high-confidence I-sites motif regions are found to have dramatic effects on folding (Mok, Elisseeva et al. 2001; Northey, Maxwell et al. 2002). These experiments are consistent with I-sites being early folding intermediates, or initiation sites of folding.

**HMMSTR: A model for propagation**

Folding initiation sites are marginally stable, but they are stable enough to provide a starting place for the propagation of structure up and down the chain. The sequence dependence of these extensions can be found by labeling all of the I-sites motifs in the database and drawing connections between them wherever they occur adjacent to each other in the sequence. I-sites motifs that are adjacent to each other "extend" each other. For example, the amphipathic alpha helix motif extends the alpha-alpha corner motif (Fig. 1).

All adjacencies of I-sites motifs in known structures were found, counted, and the motif-motif transition probabilities were condensed into a single, non-linear hidden Markov model (HMM) called HMMSTR ("hamster") (Bystroff, Thorsson et al. 2000). The sequence preferences and the inter-motif transitions were trained on a non-redundant database of protein structures. HMMSTR models the ways that local structure can be arranged along the sequence, modeling the way an initiation site motif is likely to affect the conformation of residues that are adjacent to it in the sequence. **Fig. 2** shows the highly branched and cyclic state connectivity. Note the region containing a cycle of helix states, representing the well-known heptad repeat motif of the amphipathic alpha-helix structure.

Each state in HMMSTR represents the structure (as backbone angles) of one residue. An unbroken string of states represents a local structure motif. A branching chain of states represents two or more alternative adjacent motif structures. The result of a HMMSTR prediction is a matrix of Markov state probabilities. We may use the model to sample from this distribution, or we can choose a single structure prediction for each position by a voting procedure as described previously (Bystroff, Thorsson

et al. 2000). HMMSTR predictions average 60% accuracy in predicting 8-residue fragments with RMSD < 1.4Å.  HMMSTR has been used for local and secondary structure prediction (Bystroff, Thorsson et al. 2000; Rost 2001), inter-residue contact prediction (Zaki, Shan et al. 2000; Shao and Bystroff 2003), and as the source of a fragment library for Rosetta (Bystroff and Shao 2002) folding simulations.



(a)

(b)

(c)

(d)
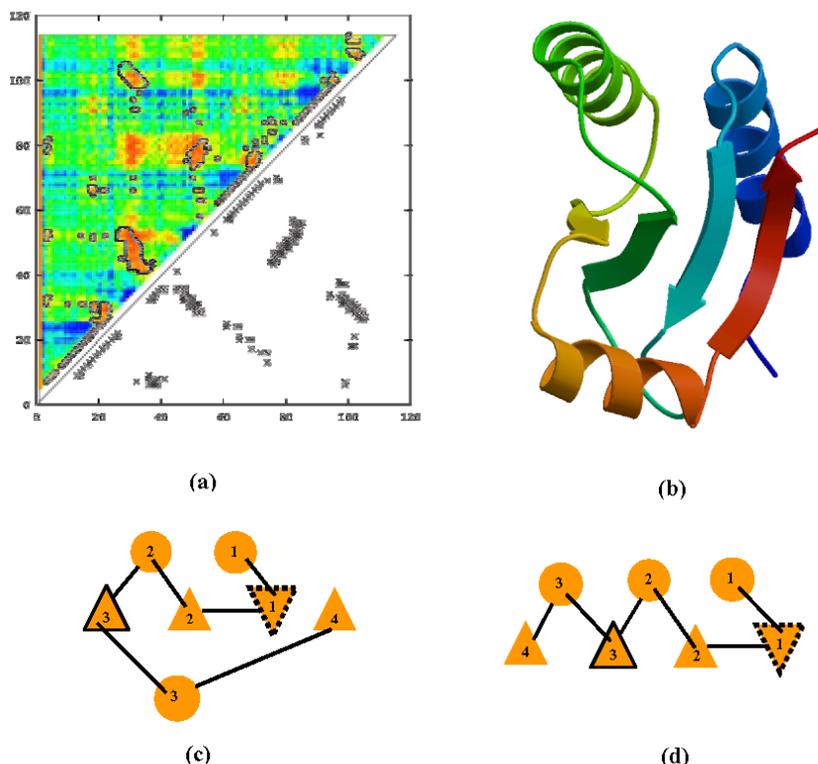
**Fig. 3.  HMMSTR-CM prediction of a CASP5 target.**

The structure of hypothetical protein HI0073 from *H. influenzae* was successfully predicted using the HMMSTR-CM method. HI0073 (PDB code 1JOG, true structure shown in (b)) has 116 residues arranged in a three-layer all-parallel $\alpha/\beta$ sandwich. The contact potential map **(a)** shows that most of the true contacts were assigned favorable (red) contact potentials. However, there are also favorable regions that are non-contacts. After selecting a nucleation site, $\beta_2\alpha_2\beta_3$, contacts were assigned or erased in a 4 step pathway, as follows:

(1) Parallel $\beta$ contacts were assigned for $\beta_2$ to $\beta_3$.

(2) Anti-parallel $\beta$ contacts were assigned for $\beta_1$ and $\beta_2$. All other $\beta$ contacts to $\beta_2$ were erased (Rule 3).

(3) There were two choices for a right-handed crossover from $\beta_3$ to $\beta_4$, as shown in **(c)** and **(d)**. Since $\beta_1$ was more hydrophobic than $\beta_3$, we paired $\beta_1$ and $\beta_4$. All other $\beta$ contacts to $\beta_1$ were erased, and contacts between $\alpha_2$ and $\alpha_3$ were erased (Rules 8, 10).

(4) $\alpha_1$ was placed on the opposite side of the sheet from $\alpha_3$, since $\alpha_3$ extends across the sheet. Contacts were assigned between $\alpha_1$ and $\alpha_2$ and erased between $\alpha_1$ and $\alpha_3$ (Rule 9).

**(c)** The completed TOPS diagram and contact map (outlines) match the true structure. The contact map prediction has 42% contact coverage and 29% accuracy, or if we count near misses ($\pm 1$ residue), then the coverage is 75% and the accuracy is 57%. (d) The wrong choice at step (3) would give this structure.

**HMMSTR-CM: Pairwise condensation of motif structures**

HMMSTR-CM is a model for pairwise interactions between local structure motifs. Pairwise interactions are represented as a probabilistic contact map. Contact maps are square symmetrical Boolean matrices that represent pairs of residue positions that close in space (within 8Å). A contact map may be projected into three-dimensions if it obeys certain mathematical constraints (Brunger, Clore et al. 1986; Crippen 1988; Aszodi, Munro et al. 1997; Vendruscolo, Kussell et al. 1997; Wintjens, Wodak et al. 1998; Miyazawa and Jernigan 2003; Selvaraj and Gromiha 2003). Previous contact map prediction methods have used neural nets (Fariselli and Casadio 1999; Pollastri and Baldi 2002), correlated mutations (Olmea and Valencia 1997; Ortiz, Kolinski et al. 1998; Singer, Vriend et al. 2002), and association rules (Zaki, Shan et al. 2000; Hu, Shen et al. 2002). Neural net based predictions had an average accuracy of about 21% overall (Fariselli, Olmea et al. 2001), while higher accuracies were reported for local contacts (Pollastri and Baldi 2002). Contact potential has been used for many protein structure methods(Jones, McGuffin et al. 1999; Skolnick, Zhang et al. 2003).

The first step in predicting a contact map is to assign a probability to each potential contact. The probability in this case is the database-derived likelihood of contact between any two local structure motifs. This implies that the local structure motif forms first, then these sub-structures condense to form larger units. The database statistics give us a free energy of interaction, similar to a binding energy. But each residue is represented not as a single motif but as a probability distribution of motifs. We may envision pairs of flickering local structures, interacting in proportion to their structural content.

The interaction potential between any two motifs is modeled as the statistical interaction potential between two corresponding Markov states. Knowledge-based Markov state "pair potentials" (actually log-likelihood ratios) were summed from the CATH database of protein domain structures(Orengo, Michie et al. 1997). Each domain was first preprocessed into HMMSTR Markov state probability distributions using the Forward/Backward algorithm (Rabiner 1989) to get the position-dependent Markov state probability distribution of states $\gamma$ (Eq. 1).

$$\gamma(i,q) = P(q \mid i) \hspace{4cm} \text{Eq. 1}$$

The pairwise contact potential between any two HMMSTR states $p$ and $q$ ($G(p,q,s)$) was calculated as the negative log of the mutual probability of these two states in contacting residues (defined as having C$\alpha$ -C$\alpha$ distance < 8Å) (Eq. 2).

$$G(p,q,s) = -\log \frac{\displaystyle\sum_{CATH} \sum_{i \ni D_{i,i+s} <8\text{Å}} \gamma(i,p)\gamma(i+s,q)}{\displaystyle\sum_{CATH} \sum_{i} \gamma(i,p)\gamma(i+s,q)} \hspace{1cm}, \hspace{2cm} \text{Eq. 2}$$

where $s=|j-i|$ is sequence separation. There is one $G$ value for each pair of HMMSTR's 282 states and each sequence separation, from $s=4$ to 20, a total of 1037153 potential functions. (For s > 20, we used $G$ for s=20) Using $G$ and a target sequence, we may sum the contact potential map $E(i,j)$, which is a matrix of contact potentials between every residue pair $ij$ in a target sequence. **Fig. 3** shows a contact potential map, $E$, for a protein that was one of the targets in the CASP5 prediction experiment. In this map we see patterns for super-secondary structure motifs and possible $\beta$ strand pairings.
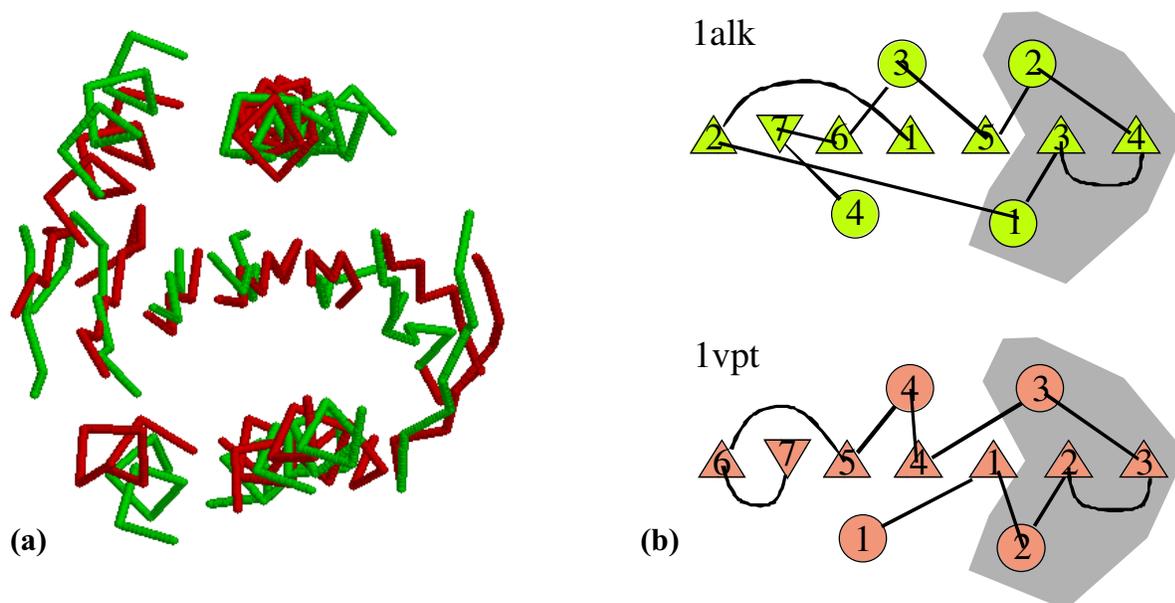
**Fig. 4.** Non-sequential alignment of conserved core packing arrangement. Two proteins of very different overall topology, Alkaline phosphatase (1alk, green) and Vp39 from vaccinia virus (1vpt, red) share eleven superimposable secondary structure elements despite having no sequence similarity and different topologies. **(a)** Superimposed secondary structure elements. **(b)** TOPS diagrams showing topological connections. Topology is conserved only in the shaded region.

Nearly all of the true pairings are given a high score (i.e. low energy) by HMMSTR-CM, but too many high scores are given. Spatial constraints prevent many of the predicted contacts from happening. For example, at most two beta strands may pair with any one beta strand. Other rules enforce the physically possible density of contacts and mutual contacts, and the triangle inequality. Common sense rules such as these were used to extract a self-consistent set of contacts from the high scoring ones. Simple rules were sufficient to extract the correct set of contacts for some but not all of the CASP5 targets(Shao and Bystroff 2003). However, these rules only approximate the complex topological constraints on the protein chain. A multibody model is needed to capture this higher level of organization.

**SCALI: Multibody packing arrangements of local structure motifs**

HMMs have been used to predict protein structure at both the local sequence level (Bystroff, Thorsson et al. 2000) and the global level (Karplus, Barrett et al. 1998; Sonnhammer, Eddy et al. 1998; Gough and Chothia 2002). But there are interesting recurrent features in proteins that are neither global nor local, specifically, three-dimensional packing arrangements in the hydrophobic core regions. Common secondary structure types have characteristic ways of arranging themselves in globular proteins (Murzin and Finkelstein 1988; Efimov 1994; Murzin, Lesk et al. 1994; Ruczinski, Kooperberg et al. 2002). If we take away the connections between the secondary structure units, we find that virtually all ways of packing secondary structures have already been seen in the protein database. When proteins are discovered to have a "new fold", it is often found to be a permuted version of an "old" fold. If so, then a method for finding non-sequential (i.e. permuted) alignments would be useful for structure prediction.

Recently, we described SCALI, a new algorithm for aligning structures without sequential constraints. SCALI was compared to several structure-based alignment programs, including CE, DALI, and

8

KENOBI, but none of these programs were able to find conserved core packing arrangements consistently when the component parts were not arranged sequentially along the chain. For example, SCALI was able to align the structures 1ALK and 1VPT and found that the positions of its secondary structure elements were superimposible with a RMSD of 4.3Å (**Fig. 4**) even though they are arranged very differently along the chain.

The topological constraints on folding, given the pairwise energies of interaction, may be modeled by finding all of the common ways that local structure units are arranged in space. That is, we identify the commonly recurring core packing arrangements (i.e. superimposable sets of secondary structure elements). By looking at each example protein that contains a common packing arrangement, we may trace the sequential order of the secondary structure elements by drawing connections wherever they are adjacent in the sequence. The result is a hidden Markov model. In this model a state has a specific location in space, relative to all other states. A pathway through this HMM is a structure prediction.

Recurrent core packing arrangements were found by clustering SCALI non-sequential structure-based alignments. Using a simple greedy algorithm, we found regions of proteins that occurred in multiple SCALI alignments, each protein having a different connectivity. For example, we did an all-against-all pairwise structure comparison for the 61 representative structures of the 3-layer $\alpha/\beta/\alpha$ class, a total of 1830 alignments. 56 out of 61 structures were clustered into four subclasses using a simple greedy algorithm. The structures that clustered together conserved the same core. **Fig. 5** shows two of the resulting HMMs.

For each SCALI-HMM, certain topological connections are observed and others are not, probably reflecting the physical constraints on secondary structure packing. Certain recurrent substructures are found in these models, for example the right-handed parallel $\beta\alpha\beta$ motif, and the shaded region in Fig. 4, a helix-hairpin-helix motif. SCALI-HMMs have been built for several recurrent classes of proteins including the "up-down bundle" alpha proteins, beta "sandwich" proteins, beta "jelly-roll" and 3-layer $\alpha\beta\alpha$ proteins. Refinement and analysis of these conserved core packing arrangements is ongoing.
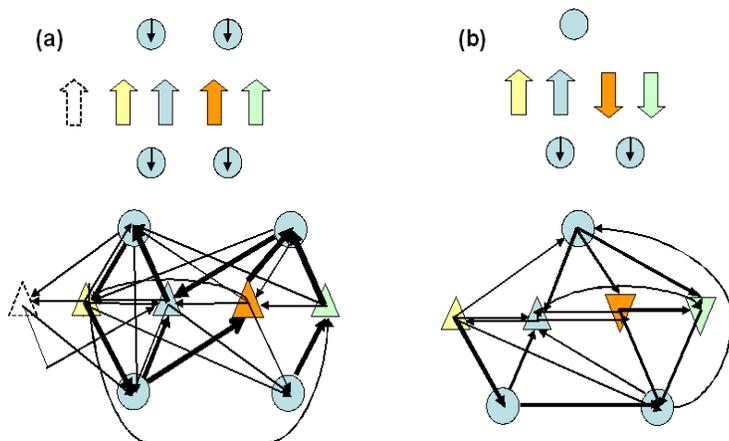


**Fig. 5.** Diagrammatic hidden Markov models for the two sub-classes of the 3-layer $\alpha\beta\alpha$ class of proteins based on SCALI alignments In each subclass, the upper panel shows the topology diagram without connectivities for that core structure. (strands are arrows pointing up or down, helices are circles. Circles with arrows are helices pointing down) The lower panel is the hidden Markov model for that core, drawn as a TOPS diagram. Thicker lines indicate more frequent connections. (a) Largest sub-class 37 proteins (b) Next largest, 9 proteins.

9

## Relevance of Approaches and Results

Using database statistics in a heirarchical way, we have shown that specific recurrent themes exist at different levels of structural generality. It is reasonable to assume that an underlying physical model exists for these statistical observations. A physical model for folding has been proposed that explains the speed with which proteins fold (Nolting and Andert 2000), despite the enormous theoretical size of the conformational space to be searched. They fold fast because there is no combinatorial explosion in the number of conformational states, because at each stage of the folding pathway only a small number of options are available. The situation is similar when considering the four statistical models together.

At the stage of initiation, there are only about 40 choices. This is approximately the number of different sequence-structure patterns in I-sites, once overlap is accounted for (Bystroff and Baker 1998; Bystroff, Thorsson et al. 2000).

At the second stage, propagation, there are even fewer choices. HMMSTR (Bystroff, Thorsson et al. 2000), the model that represents the ways that structure can propagate up and down the chain, is a sparsely branching HMM, reflecting the paucity of choices of different ways of adding structure onto structure, locally.

At the third stage of folding, condensation, the number of possibilities is not the square of the number of local motifs, because some pairs of motifs cannot physically fit together due to differences in their shape and in their surfaces. HMMSTR-CM models this by assigning a near-zero probability for many of the potential pairwise interactions (Shao and Bystroff 2003).

And at the last stage of folding, the number of ways that pairwise interactions can be combined is much less than the number of secondary structure elements "choose two", because the interaction of any two elements restricts the possibilities for the third and so on, in a process of elimination. The number of possible tracings through the HMMs produced by SCALI(Yuan and Bystroff 2004) is fewer than expected because only "self-avoiding" paths are physically possible. Each secondary structure element has an assigned location in space, and therefore that element cannot occur twice along the chain.

The absence of a combinatorial explosion in the physical folding model may explain the absence of a sparse data problem in our hierarchy of statistical models. Although the database is a fixed size, we still see recurrence as we model increasing larger pieces of protein chain. Even at the global level, there are recurring themes in the database (Orengo, Jones et al. 1994; Russell, Sasieni et al. 1998; Zhang and DeLisi 1998; Gough and Chothia 2002). For example, the 8-stranded alpha-beta barrel ("TIM barrel") seems to have independently evolved many times, as has the 7-helix topology called the "globin fold." In many cases, there is no support for these analogous proteins having a common ancestor. Rather, they are likely to have arisen independently, by "convergent evolution." If so, if certain topologies have been sampled many times in evolutionary history, then perhaps the total number of ways that a protein chain can fold is not so large as we have previously supposed.

## Acknowledgements

# References

Anfinsen, C. B. and H. A. Scheraga (1975). "Experimental and theoretical aspects of protein folding." <u>Adv Protein Chem</u> 29: 205-300.

Aszodi, A., R. E. Munro, et al. (1997). "Distance geometry based comparative modelling." <u>Fold Des</u> 2(3): S3-6.

Baldwin, R. L. (1995). "The nature of protein folding pathways: the classical versus the view." <u>J Biomol NMR</u> 5(2): 103-9.

Baldwin, R. L. and G. D. Rose (1999). "Is protein folding hierarchic? I. Local structure and peptide folding." <u>Trends Biochem Sci</u> 24(1): 26-33.

Blanco, F. J., G. Rivas, et al. (1994). "A short linear peptide that folds into a native stable beta-hairpin in aqueous solution." <u>Nat Struct Biol</u> 1(9): 584-90.

Brunger, A. T., G. M. Clore, et al. (1986). "Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin." <u>Proc Natl Acad Sci U S A</u> 83(11): 3801-5.

Bystroff, C. and D. Baker (1997). "Blind predictions of local protein structure in CASP2 targets using the I-sites library." <u>Proteins</u> Suppl 1: 167-71.

Bystroff, C. and D. Baker (1998). "Prediction of local structure in proteins using a library of sequence-structure motifs." <u>J Mol Biol</u> 281(3): 565-77.

Bystroff, C. and S. Garde (2003). "Helix propensities of short peptides: Molecular dynamics versus bioinformatics." <u>Proteins</u> 50(4): 552-62.

Bystroff, C. and Y. Shao (2002). "Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA." <u>Bioinformatics</u> 18 Suppl 1: S54-61.

Bystroff, C., K. T. Simons, et al. (1996). "Local sequence-structure correlations in proteins." <u>Curr Opin Biotechnol</u> 7(4): 417-21.

Bystroff, C., V. Thorsson, et al. (2000). "HMMSTR: A hidden markov model for local sequence-structure correlations in proteins." <u>Journal of Molecular Biology</u> 301(1): 173-90.

Chan, H. S., S. Bromberg, et al. (1995). "Models of cooperativity in protein folding." <u>Philos Trans R Soc Lond B Biol Sci</u> 348(1323): 61-70.

Colon, W. and H. Roder (1996). "Kinetic intermediates in the formation of the cytochrome c molten globule." <u>Nat Struct Biol</u> 3(12): 1019-25.

Crippen, G. M., Havel, T.F. (1988). <u>Distance Geometry and Molecular Conformation</u>, John Wiley & Sons.

Dyson, H. J. and P. E. Wright (1996). "Insights into protein folding from NMR." <u>Annu Rev Phys Chem</u> 47: 369-95.

Eaton, W. A., P. A. Thompson, et al. (1996). "Fast events in protein folding." <u>Structure</u> 4(10): 1133-9.

Efimov, A. V. (1993). "Standard structures in proteins." <u>Prog Biophys Mol Biol</u> 60(3): 201-39.

Efimov, A. V. (1994). "Favoured structural motifs in globular proteins." <u>Structure</u> 2(11): 999-1002.

Fariselli, P. and R. Casadio (1999). "A neural network based predictor of residue contacts in proteins." <u>Protein Eng</u> 12(1): 15-21.

Fariselli, P., O. Olmea, et al. (2001). "Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations." <u>Proteins</u> 45(S5): 157-62.

Fersht, A. R., A. Matouschek, et al. (1992). "The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding." J Mol Biol 224(3): 771-82.

Garbuzynskiy, S. O., A. V. Finkelstein, et al. (2004). "Outlining folding nuclei in globular proteins." J Mol Biol 336(2): 509-25.

Gillespie, J. R. and D. Shortle (1997). "Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures." J Mol Biol 268(1): 170-84.

Gnanakaran, S. and A. E. Garcia (2002). "Folding of a Highly Conserved Diverging Turn Motif from the SH3 Domain." Biophys. J.

Gough, J. and C. Chothia (2002). "SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments." Nucleic Acids Res 30(1): 268-72.

Grantcharova, V. P., D. S. Riddle, et al. (2000). "Long-range order in the src SH3 folding transition state." Proc Natl Acad Sci U S A 97(13): 7084-9.

Gromiha, M. M. and S. Selvaraj (2002). "Important amino acid properties for determining the transition state structures of two-state protein mutants." FEBS Lett 526(1-3): 129-34.

Gulotta, M., R. Gilmanshin, et al. (2001). "Core formation in apomyoglobin: probing the upper reaches of the folding energy landscape." Biochemistry 40(17): 5137-43.

Han, K. F. and D. Baker (1995). "Recurring local sequence motifs in proteins." J Mol Biol 251(1): 176-87.

Han, K. F. and D. Baker (1996). "Global properties of the mapping between local amino acid sequence and local structure in proteins." Proc Natl Acad Sci U S A 93(12): 5814-8.

Han, K. F., C. Bystroff, et al. (1997). "Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns." Protein Sci 6(7): 1587-90.

Heidary, D. K. and P. A. Jennings (2002). "Three topologically equivalent core residues affect the transition state ensemble in a protein folding reaction." J Mol Biol 316(3): 789-98.

Houry, W. A., D. M. Rothwarf, et al. (1996). "Circular dichroism evidence for the presence of burst-phase intermediates on the conformational folding pathway of ribonuclease A." Biochemistry 35(31): 10125-33.

Hu, J., X. Shen, et al. (2002). Mining Protein Contact Maps. BIOKDD 2002, Edmonton, Canada.

Ivankov, D. N., S. O. Garbuzynskiy, et al. (2003). "Contact order revisited: influence of protein size on the folding rate." Protein Sci 12(9): 2057-62.

Jacchieri, S. G. (2000). "Mining combinatorial data in protein sequences and structures." Molecular Diversity 5(3): 145-152.

Jones, D. T., L. J. McGuffin, et al. (1999). "GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences." J Mol Biol 287(4): 797-815.

Karplus, K., C. Barrett, et al. (1998). "Hidden Markov models for detecting remote protein homologies." Bioinformatics 14(10): 846-56.

Krueger, B. P. and P. A. Kollman (2001). "Molecular dynamics simulations of a highly charged peptide from an SH3 domain: possible sequence-function relationship." Proteins 45(1): 4-15.

Laurents, D. V. and R. L. Baldwin (1998). "Protein folding: matching theory and experiment." Biophys J 75(1): 428-34.

Levinthal, C. (1969). How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems, Allerton House, Monticello, Illinois, University of Illinois Press.

**Mateu, M. G., M. M. Sanchez Del Pino, et al. (1999). "Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53." <u>Nat Struct Biol</u> 6(2): 191-8.**

**Mendes, J., R. Guerois, et al. (2002). "Energy estimation in protein design." <u>Current Opinion in Structural Biology</u> 12(4): 441-446.**

**Miller, E. J., K. F. Fischer, et al. (2002). "Experimental evaluation of topological parameters determining protein-folding rates." <u>Proc Natl Acad Sci U S A</u> 99(16): 10359-63.**

**Miyazawa, S. and R. L. Jernigan (2003). "Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space." <u>Proteins</u> 50(1): 35-43.**

**Mok, Y. K., E. L. Elisseeva, et al. (2001). "Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states." <u>J Mol Biol</u> 307(3): 913-28.**

**Mok, Y. K., C. M. Kay, et al. (1999). "NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions." <u>J Mol Biol</u> 289(3): 619-38.**

**Munoz, V., F. J. Blanco, et al. (1995). "The hydrophobic-staple motif and a role for loop-residues in alpha- helix stability and protein folding." <u>Nat Struct Biol</u> 2(5): 380-5.**

**Murzin, A. G. and A. V. Finkelstein (1988). "General architecture of the alpha-helical globule." <u>J Mol Biol</u> 204(3): 749-69.**

**Murzin, A. G., A. M. Lesk, et al. (1994). "Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures." <u>J Mol Biol</u> 236(5): 1382-400.**

**Nolting, B. and K. Andert (2000). "Mechanism of protein folding." <u>Proteins</u> 41(3): 288-98.**

**Nolting, B., R. Golbik, et al. (1997). "The folding pathway of a protein at high resolution from microseconds to seconds." <u>Proc Natl Acad Sci U S A</u> 94(3): 826-30.**

**Northey, J. G. B., K. L. Maxwell, et al. (2002). "Protein folding kinetics beyond the Phi value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state." <u>Journal of Molecular Biology</u> 320(2): 389-402.**

**Nymeyer, H., N. D. Socci, et al. (2000). "Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration." <u>Proc Natl Acad Sci U S A</u> 97(2): 634-9.**

**Oliveberg, M., Y. J. Tan, et al. (1998). "The changing nature of the protein folding transition state: implications for the shape of the free-energy profile for folding." <u>J Mol Biol</u> 277(4): 933-43.**

**Olmea, O. and A. Valencia (1997). "Improving contact predictions by the combination of correlated mutations and other sources of sequence information." <u>Fold Des</u> 2(3): S25-32.**

**Onuchic, J. N., Z. Luthey-Schulten, et al. (1997). "Theory of protein folding: the energy landscape perspective." <u>Annu Rev Phys Chem</u> 48: 545-600.**

**Orengo, C. A., D. T. Jones, et al. (1994). "Protein superfamilies and domain superfolds." <u>Nature</u> 372(6507): 631-4.**

**Orengo, C. A., A. D. Michie, et al. (1997). "CATH--a hierarchic classification of protein domain structures." <u>Structure</u> 5(8): 1093-108.**

**Ortiz, A. R., A. Kolinski, et al. (1998). "Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments." <u>J Mol Biol</u> 277(2): 419-48.**

**Pande, V. S., A. Grosberg, et al. (1998). "Pathways for protein folding: is a new view needed?" <u>Curr Opin Struct Biol</u> 8(1): 68-79.**

**Plaxco, K. W., K. T. Simons, et al. (1998). "Contact order, transition state placement and the refolding rates of single domain proteins." <u>J Mol Biol</u> 277(4): 985-94.**

Plaxco, K. W., K. T. Simons, et al. (2000). "Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics." Biochemistry 39(37): 11177-83.

Pollastri, G. and P. Baldi (2002). "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners." Bioinformatics 18 Suppl 1: S62-S70.

Rabiner, L. R. (1989). " A tutorial on Hidden Markov Models and selected applications in speech recognition." Proc IEEE 77(2): 257-286.

Riddle, D. S., V. P. Grantcharova, et al. (1999). "Experiment and theory highlight role of native state topology in SH3 folding." Nat Struct Biol 6(11): 1016-1024.

Rooman, M. J., J. Rodriguez, et al. (1990). "Automatic definition of recurrent local structure motifs in proteins." J Mol Biol 213(2): 327-36.

Rost, B. (2001). "Review: protein secondary structure prediction continues to rise." J Struct Biol 134(2-3): 204-18.

Ruczinski, I., C. Kooperberg, et al. (2002). "Distributions of beta sheets in proteins with application to structure prediction." Proteins 48(1): 85-97.

Russell, R. B., P. D. Sasieni, et al. (1998). "Supersites within superfolds. Binding site similarity in the absence of homology." J Mol Biol 282(4): 903-18.

Selvaraj, S. and M. M. Gromiha (2003). "Role of hydrophobic clusters and long-range contact networks in the folding of (alpha/beta)8 barrel proteins." Biophys J 84(3): 1919-25.

Shao, Y. and C. Bystroff (2003). "Predicting interresidue contacts using templates and pathways." Proteins 53 Suppl 6: 497-502.

Shoemaker, B. A. and P. G. Wolynes (1999). "Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble." J Mol Biol 287(3): 657-74.

Singer, M. S., G. Vriend, et al. (2002). "Prediction of protein residue contacts with a PDB-derived likelihood matrix." Protein Eng 15(9): 721-5.

Skolnick, J. and A. Kolinski (2002). A unified approach to the prediction of protein structure and function. Computational Methods for Protein Folding. 120: 131-192.

Skolnick, J., Y. Zhang, et al. (2003). "TOUCHSTONE: a unified approach to protein structure prediction." Proteins 53 Suppl 6: 469-79.

Sonnhammer, E. L., S. R. Eddy, et al. (1998). "Pfam: multiple sequence alignments and HMM-profiles of protein domains." Nucleic Acids Res 26(1): 320-2.

Steward, R. E. and J. M. Thornton (2002). "Prediction of strand pairing in antiparallel and parallel beta- sheets using information theory." Proteins-Structure Function and Genetics 48(2): 178-191.

Vendruscolo, M., E. Kussell, et al. (1997). "Recovery of protein structure from contact maps." Fold Des 2(5): 295-306.

Viguera, A. R. and L. Serrano (1995). "Experimental analysis of the Schellman motif." J Mol Biol 251(1): 150-60.

Wintjens, R., S. J. Wodak, et al. (1998). "Typical interaction patterns in alphabeta and betaalpha turn motifs." Protein Eng 11(7): 505-22.

Yi, Q., C. Bystroff, et al. (1998). "Prediction and structural characterization of an independently folding substructure in the src SH3 domain." J Mol Biol 283(1): 293-300.

Yuan, X. and C. Bystroff (2004). "Non-sequential Structure-based Alignments Reveal Topology-independent Core Packing Arrangements in Proteins." Proc Nat Acad Sci submitted.

Zaki, M. J., J. Shan, et al. (2000). **Mining residue contacts in proteins using local structure predictions**. Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering, Arlington, VA, USA.

Zhang, C. and C. DeLisi (1998). "Estimating the number of protein folds." <u>J Mol Biol</u> 284(5): 1301-5.

Zwanzig, R. (1997). "Two-state models of protein folding kinetics." <u>Proc Natl Acad Sci U S A</u> 94(1): 148-50.