Modeling Protein Folding Pathways

Christopher Bystroff, Yu Shao

Dept of Biology
Rensselaer Polytechnic Institute, Troy, NY.
e-mail:{bystrc, shaoy}@rpi.edu

Summary

Proteins fold through a series of intermediate states called a pathway. Protein folding pathways have been modeled using either simulations or a heirarchy of statistical models. Here we present a series of related statistical models that attempt to predict early, middle and late intermediates along the folding pathway. I-sites motifs are discrete models for folding initiation sites. HMMSTR is a model for local structure patterns composed of I-sites motifs. HMMSTR-CM is an approach toward assembling motifs and groups of motifs in a contact map representation, using heuristic rules to predict contact maps either with or without the use of templates. We also discuss the I-sites/ROSETTA server, which is a folding simulation algorithm that uses a fragment library as input. The results of blind structure prediction experiments are discussed. Pathway-based predictions sometimes lead to an unambiguous prediction of the fold topology, even without using templates.

1. Introduction: Darwin versus Boltzmann

All computational models that predict something have certain underlying assumptions that constitute the physical basis for the model. In protein structure prediction, there are two physical/biological processes that can be modeled: the process of evolution, or the process of folding. We may give these two paradigms names, Darwin and Boltzmann, after the scientists who defined the fundamental principles of evolutionary biology and statistical thermodynamics, respectively.

Most of the work in protein structure prediction is Darwin-based, using the well-known premise that sequences that have a common ancestor have similar folds, and they strive to extrapolate this principle to increasingly distant sequence relationships. Methods that use multiple sequence alignment, structural alignment, or "threading potentials" are implicitly searching for a common ancestor. Despite the oft-used “energy-like” scoring functions, these methods do not address the physical process of folding. Evolution happens on the time scale of millions of years, folding on the time scale of fractions of a second.

Protein structure prediction of the Boltzmann kind is perceived to be a very difficult problem. Many have tried their hand at it over the last thirty years, and an equal number have failed to improve upon Darwin-based methods. The problem of predicting folding pathways may be perceived to be even harder, since it *should* depend on first solving the protein folding problem. But this is not true, as we shall see. Prediction of the protein folding pathway may be evaluated by looking at the success in predicting sub-segments or substructures of proteins. If the computational model has the right underlying assumptions about what comes first in the pathway, and what comes next, and so on, then blind predictions, such as those done as part of CASP, the Critical Assessment of Protein Structure Prediction bi-annual worldwide experiment (Moult et al. 2001), may validate that model. And the pathway model that eventually arises from this process will tell us more than just final answer.

In this chapter we present a series of bioinformatics and simulation experiments related to predicting protein structure by modeling the folding pathway. We will conclude that *ab initio* predictions can be done either by simulations or by a rule-based fragment assembly method, and that it is possible to find folds that are not present in the database of structures. We will discuss issues of accuracy and resolution and present some possible directions for the future.

1.1 Protein Folding Pathway 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. 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 somewhat (Baldwin 1995, Colon and Roder 1996, Oliveberg et al. 1998, Pande et al. 1998), in part due to a better understanding of the structure of the so-called “unfolded state" (Dyson and Wright 1996, Gillespie and Shortle 1997, Mok et al. 1999) and to a more detailed description of kinetic and equilibrium folding intermediates (Eaton et al. 1996, Gulotta et al. 2001, Houry et al. 1996). An image of the transition state of folding can now be mapped out by point mutations, or "phi-value analysis" (Fersht et al. 1992, Grantcharova et al. 2000, Heidary and Jennings 2002, Mateu et al. 1999, Nolting et al. 1997). The "folding funnel" model (Chan et al. 1995, Onuchic et al. 1997) has reconciled hydrophobic collapse with the alternative nucleation-condensation model (Nolting and Andert 2000) by envisioning a distorted, funicular energy landscape (Laurents and Baldwin 1998) and a "minimally frustrated" pathway (Nymeyer et al. 2000, Shoemaker and Wolynes 1999) through this landscape. The view remains of a channeled, counter-entropic search for the hole in the funnel as the predominant barrier to folding (Zwanzig 1997).

Simulations using various simplified representations of the protein chain, including lattice models, have clarified the basic nature of folding pathways (Kolinski and Skolnick 1997, Mirny and Shakhnovich 2001, Shakhnovich 1998, Thirumalai and Klimov 1998). The topology of the fold plays a dominant role in defining the critical positions that effect the folding rate (Ortiz and Skolnick 2000, Shea and Brooks 2001). Models that represent the chain in atomistic detail show that minimally frustrated, low-energy pathways may involve the propagation of structure along the chain like a zipper (Alm and Baker 1999, Munoz et al. 1998). All-atom, explicit solvent molecular dynamics simulations have reproduced the experimentally determined conformations for short peptides (Cavalli et al. 2002, Duan and Kollman 1998, Garcia and Sanbonmatsu 2001, Krueger and Kollman 2001, Shao and Bystroff 2003). This large body of work is still inconclusive, but clearly folding is best represented by an ensemble rather than a single pathway.

2. Knowledge-based Models for Folding Pathways

The approach that began with I-sites is an attempt to build a hierarchical series of models mirroring the hierarchy of folding events, from initiation to nucleation to propagation and condensation. The hierarchy can be roughly described as “lo-

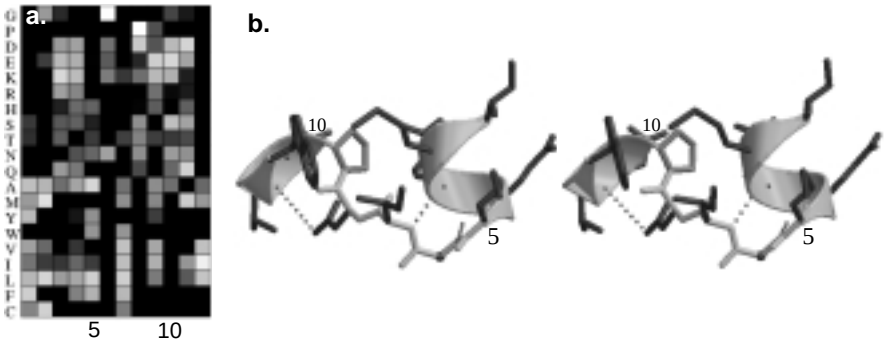

Fig. 1. **a**. I-sites profile for alpha-alpha corner motif. Boxes are shaded lighter in proportion to the log-likelihood ratio of each amino acid at each position relative to the start of the motif. **b**. Stereo image of the alpha-alpha corner motif.

cal to global." Each model builds on the model before it. At each point the results are an ensemble of conformational states.

"Local structure" is a generic term for the conformations of short pieces of the protein chain, usually 3-20 residue pieces. Local structure motifs include the two common forms (alpha helix and beta strand) along with a few dozen turns, half-turns, caps, bulges and coils. The role of local structure motifs with regard to the initiation of folding has been discussed by Baldwin, Rooman and others (Baldwin and Rose 1999, Efimov 1993, Rooman et al. 1990).

2.1. I-sites: A Library of Folding Initiation Site Motifs

I-sites is a library of 262 sequence patterns that map to local structures. A sequence pattern is expressed as a position-specific scoring matrix (PSSM). Recurrent sequence patterns had been previously used for prediction of structural motifs, including the Schellman motif (Schellman 1980), the hydrophobic staple (Munoz et al. 1995), and various types of coiled coil (Woolfson and Alber 1995). Recurrent sequence patterns of various lengths were found by exhaustively clustering short segments of sequence profiles for proteins in a non-redundant database of known structures (Bystroff et al. 1996, Han and Baker 1996, 1995, Han et al. 1997). Bystroff and Baker mapped recurrent sequence patterns to their predominant structural motifs and used reinforcement learning to optimize the sequence-structure correlation (Bystroff and Baker 1998). The resulting I-sites Library (Fig. 1) has been used in various prediction experiments (Bystroff and Baker 1997, Bystroff and Shao 2002) and has inspired numerous experimental studies since its

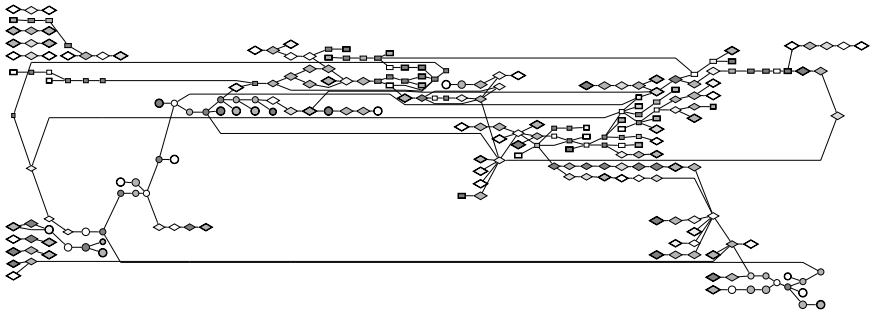

Fig. 2. HMMSTR represented as a directed graph. The symbol shape represents the secondary structure type; circles: helix; rectangles: beta sheet; diamonds: other motifs. Shading represents the amino acid preference; dark grey: non-polar; grey: polar; light-grey: proline; lightest grey: glycine; white: no preference. Only high-probability transitions are shown.

publication (Jacchieri 2000, Mendes et al. 2002, Northey et al. 2002b, Skolnick and Kolinski 2002, Steward and Thornton 2002). I-sites motifs have been linked to local structure stability in both NMR studies (Blanco et al. 1994, Munoz et al. 1995, Viguera and Serrano 1995, Yi et al. 1998) and molecular dynamics simulations (Bystroff and Garde 2003, Gnanakaran and Garcia 2002, Krueger and Kollman 2001). Mutations in high-confidence I-sites motif regions are found to have dramatic effects on folding (Mok et al. 2001, Northey et al. 2002a). About one-third of all residues in all proteins are found in high-confidence (>70%) I-sites motif regions and these sites are predicted to be conformationally stable and early-folding.

2.2. HMMSTR: A Hidden Markov Model for Grammatical Structure

The I-sites library was condensed to a single, non-linear hidden Markov model (HMM), called HMMSTR ("hamster"). This model, trained on a large database of protein structures and multiple sequence alignments, removes the fragment length dependence of I-sites motif predictions, models the adjacencies of motifs in proteins, and puts all of the motifs on the same probability scale. Unlike profile HMMs (Eddy 1996, Gough and Chothia 2002, Karplus et al. 1998), HMMSTR has a highly branched and cyclic connectivity, containing for example a 7-residue cycle of helix states representing the amphipathic helix heptad repeat motif. By modeling the adjacencies of motifs, HMMSTR is a model for the ways that local structure can be arranged along the sequence, similar to the ways that words can

be arranged in a sentence. This is, in a simple way, a model for the grammatical structure of protein sequences, from words to phrases.

The result of a HMMSTR prediction is like that of any HMM, an ensemble of Markov state strings. Each string of states, one state for each position in the sequence, represents a probable arrangement of mutually-compatible local structure motifs. A single prediction may be obtained from the ensemble by either selecting the most probable state string, or better, by a voting procedure over the whole ensemble (Bystroff et al. 2000). HMMSTR improved the overall accuracy in local structure prediction over the I-sites method from 43% to 60% for 8-residue fragments with RMSD < 1.4Å (Bystroff et al. 2000). HMMSTR has been used for local and secondary structure prediction (Bystroff et al. 2000, Rost 2001), inter-residue contact prediction (Zaki et al. 2000), and as the source of a fragment library for Rosetta simulations (Bystroff and Shao 2002). Previous HMMs have modeled proteins globally, not as fragments (Eddy 1996, Gough and Chothia 2002, Karplus et al. 1998).

3. ROSETTA: Folding Simulations Using a Fragment Library

The ROSETTA folding simulation algorithm uses Monte Carlo Fragment Insertion (MCFI) to predict the 3D structures of small proteins or protein fragments without the use of structural templates (Bonneau and Baker 2001, Bonneau et al. 2001, Simons et al. 1999a, Simons et al. 1997, Simons et al. 1999b). MCFI is a mostly downhill search in a knowledge-based energy landscape. Each MCFI move consists of replacing the backbone angles of segments of the chain with fragments in a library. ROSETTA has been successful in prediction experiments (CASP (Moult et al. 2001)) either using fragments from the database, from HMMSTR, or from the I-sites motif library.

In the version of ROSETTA that runs as a public server (www.bioinfo.rpi.edu), the fragment library is derived from I-sites fragment predictions, and the highest confidence I-sites were restrained to their predicted backbone angles to increase efficiency. Fragment insertion was allowed in the restrained regions, but moves were constrained to deviate by more then 60° from the I-sites prediction. Also, long sequences were simulated as overlapping short fragments of approximately 50 residues each, again for efficiency. The resulting predictions are spliced together at the end, using a genetic algorithm in conjunction with the ROSETTA knowledge-based energy function. Detailed descriptions of

each of the algorithms have been previously published (Bystroff and Shao 2002, Simons et al. 1997, Simons et al. 1999b).

3.1 Results of Fully Automated I-SITES/ROSETTA Simulations

3.1.1 Summary

A web server was used to predict 31 protein structures in the CASP4 experiment (2000) and 44 in the CASP5 experiment (2002). The successes and failures of the server may be summarized in a few broad statements. The statistics and conclusions presented here refer to *bona fide* blind predictions sent automatically to the CAFASP site as part of their "Fully-Automated" satellite experiment (Fischer et al. 2001). A more detailed analysis of this and other methods can be obtained from the associated publications (Bystroff and Shao 2002, Shao and Bystroff 2003).

Over the 75 targets, 64% of the residues were found in "topologically correct" large fragments, defined as fragments of 30 residues or more with RMSD < 6Å. At 6A RMSD, the correct overall chain trace has been reproduced, but not the finer details of structure. Occasionally beta strand may be out of order in a sheet, and strands may be substituted for helices.

A smaller percentage of all 30-residue fragments, 44%, were predicted with a 5Å RMSD. At 5Å precision, secondary structure is occasionally mispredicted, loop structures may be wrong in detail, and axial rotations of secondary structure units are possible. However, much or most of the non-local packing interactions are faithfully though roughly reproduced at this level of accuracy, and strand mispairing is not observed.

In practice, the details of the local structure are often correctly predicted when a fragment was globally correct, but the RMSD measure is insensitive to this. Therefore, another measure is used to evaluate the local accuracy of the predictions. The maximum deviation in backbone angles (*mda*) over a window of 8 residues is usually ~180° or small, and serves as a strictly local measure of correctness. 8-residue peptides that have *mda* < 90° and obey all of the stereochemical constraints of a polypeptide, have an RMSD of 1.4Å at most (Bystroff and Baker 1998). Unfortunately, when *mda* is plotted alongside RMSD, it is immediately obvious that the good local structure predictions do not always coincide with the good, large fragment predictions.

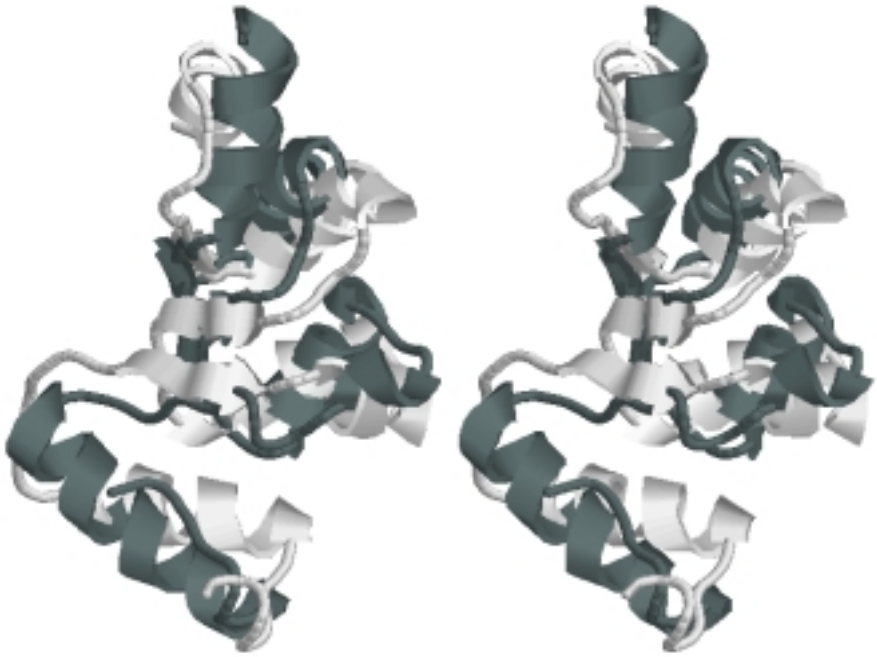

Fig. 3. ROSETTA-predicted (dark grey) and true (light grey) structure of tryptophan synthase alpha subunit from *P. furiosus*, (PDB code 1GEQ) residues 57-153.

3.1.2 Topologically correct large fragment predictions are found

Figure 3 shows a 97-residue fragment prediction with 5.9Å RMSD. At this level of precision, the residues found in the core are correct and their 3D arrangement is roughly correct. In fragments that contained helices, the N and C capping residues were usually but not always correctly located, and the direction of the chain coming off of the helix was generally correct. The orientation of parallel sheets to helices was reproduced to within about 60°, and the axial orientation of the helices with respect to strands was almost always correct, even though rolling the helix would not greatly effect the RMSD value.

Some characteristics of even the "correct" fragment predictions suggested ways in which the algorithm could be improved. The most obvious of these is the distortion of alpha helices. True native helices retain very straight helix axes despite

variability in the backbone angles. Helices in the predictions, however, were often distorted, sometimes bending the axis by 90° over its length. A combination of factors produces these errors. ROSETTA has no energy penalty for helix distortion, while it gives a large energetic bonus for packing hydrophobic residues in the core and for maintaining a low radius of gyration. Bent helices are found to replace helix kinks and alpha-alpha corners. Adding a penalty for helix distortion might fix this problem.

Topological correctness is a weak criterion for usefulness, since it means that only the handedness of the chain reversals and most of the secondary structure are right. However, these fragmentary predictions may narrow the search space for a structural analog or remote homolog, and may therefore be useful in combination with other methods. The I-sites Server correctly identified the overall anti-parallel β topology of one of the CASP5 targets, F-actin capping protein (PDB code 1IZN), a new fold at the time.

3.1.3. Good local structure correlates weakly with good tertiary structure

If the ROSETTA simulations followed a "local structure first" pathway, then we would expect to see good super-secondary structure predictions coinciding with good local structure predictions. However, this is not always the case. Frequently, the topologically correct large fragments have the wrong local structure. This is true despite the fact that at least 90% of the target sequences are covered by at least one fragment with the correct local structure in the fragment library.

Three-state secondary structure (SS) predictions were made using a version of HMMSTR that was trained on a large dataset of proteins of known structure with SS states assigned using DSSP (Kabsch and Sander 1983). The accuracy of these predictions over the 31 targets was 73.3%, only slightly lower than the state of the art in SS prediction (Jones 1998). SS predictions based on tertiary structure (TS) predictions from ROSETTA had the potential of benefiting from the added TS information, however this did not improve the prediction accuracy.

Using SS assignments derived from the TS predictions using DSSP or STRIDE (Frishman and Argos 1995), the prediction accuracy was low (50-60% Q3) because these programs depend on precise positioning of the hydrogen-bonding residues in assigning the strand state (E). Instead, the SS predictions were derived from the fragments in the fragment library, using SS assignments from their native proteins. Using this method, the overall Q3 score improved to 72.4%, but this is still no better than the SS predictions that use sequence alone without running a simulation.

If the simulation were reproducing the folding process, one might expect that the correctly-predicted tertiary interactions would add information to the secondary structure prediction. One explanation for the lack of improvement in secondary structure, despite some success in tertiary packing, is that topologically correct tertiary structures are possible even when the wrong local structure is used to build it.

3.1.4. Average contact order is too low.

Relative contact order (Plaxco et al. 1998) is calculated from the coordinates as follows:

$$CO = \frac{1}{L \bullet N} \sum^{N} \Delta S_{ij}, \qquad (1)$$

where ΔS_{ij} is the sequence separation $|i\text{-}j| \geq 5$, for residues, *ij*, that are in contact (Cα-Cα distance < 8Å). N is the number of contacts, and L is the length of the sequence. The overall average *CO* in the targets was 0.252, while the *CO* for the 32 predictions was 0.119. The lower *CO* is mostly the result of an increased number of beta hairpins. Contacts that are local, such as those in beta hairpins, are easier to find in a conformational search, and thus may represent kinetic intermediates, trapped at the end of the simulation. Kinetic trapping may be exacerbated by the more computationally efficient server protocol. A possible solution is to do more replicates and rely on cluster analysis to identify the global energy minimum. Practical limitations currently stand in the way of implementing this.

Alternatively, the predominance of beta hairpins may reflect an error in the energy function with regard to the backbone angles. Positive ϕ angles, favored only in glycine residues and usually required for turns, are found in the same proportion in the targets (8%) and in the predictions (7%), but in the targets, 44% of these turn residues are glycines, while in the prediction only 16% are glycines. This suggests that a larger energetic penalty for positive ϕ angles in non-glycine residues might correct the overabundance of hairpin turns.

3.1.5 How could Automated ROSETTA be improved?

Our results suggest that a combination of improvements in efficiency may increase the potential of the ROSETTA algorithm as a high-throughput engine for tertiary structure prediction at the 30-100 residues length scale. We suggest that a combination of structure comparison metrics be used for the evaluation of correct-

ness; a low RMSD in the context of low backbone angle deviations is shown to identify predictions that were "correct for the right reasons".

Secondary structure assignments were not improved by the use of tertiary structure predictions, partly because it was possible to obtain a globally correct tertiary structure prediction by inserting fragments of the wrong local structure.

An overall low contact order was observed in the predictions relative to the true structures. This is at least partly due to the absence of an energetic penalty for unfavorable backbone torsion angles. These may also represent kinetically trapped intermediate structures from a simulation that was too short.

4. HMMSTR-CM: Folding Pathways Using Contact Maps

HMMSTR-CM is a pathway-based method for predicting protein structure using contact maps. Contact maps are square symmetrical Boolean matrices that represent protein tertiary structures in a two-dimensional format. The 2D format has simplified the process of developing a rule-based algorithm for folding pathways. Contact maps may be projected into three-dimensions using existing methods (Aszodi et al. 1997, Brunger et al. 1986, Crippen 1988, Vendruscolo et al. 1997).

Two-dimensional flat images are more readily discernable to the eye and more memorable than complex, rotating three-dimensional images. With only a little training, a student can learn to quickly distinguish a contact map for an α/β barrel from a 3-layer α/β fold, different topologies which are very similar in their secondary structures. Similarities between distant homologs or analogs of α/β and all β folds can be seen easily in contact maps, even when the 3D structures superimpose poorly. It makes sense that if our eyes can recognize protein folds from 2D patterns, we should be able to program a computer to do so, and thereby create a new tool for learning the rules of folding.

Previous contact map prediction methods have used neural nets (Fariselli and Casadio 1999, Pollastri and Baldi 2002), correlated mutations (Olmea and Valencia 1997, Ortiz et al. 1998, Singer et al. 2002), and association rules (Hu et al. 2002, Zaki et al. 2000). Neural net based predictions had an average accuracy of about 21% overall (Fariselli et al. 2001), while higher accuracies were reported for local contacts (Pollastri and Baldi 2002), but the accuracy is lower for all-α proteins.

Our earlier work (Zaki et al. 2000) led us to believe that two important factors were missing in contact map predictions. First, typical predicted contact maps were structurally ambiguous or physically impossible, representing either multiple

Table 1. Physicality and Propagation Rules
1. **Maximum neighbor rule**: One residue can have at the most 12 contacts.
2. **Maximum mutual contact rule**: If residue i and j are in contact, there are at the most 6 residues in contact with both i and j.
3. **Beta pairing rule**: A beta strand can be in contact with at the most 2 other beta strands.
4. **Beta sheet rule**: any two pairing strands are either parallel or antiparallel.
5. **Helix mutual contact rule**: No residue can be in contact at the same time with the residues on the opposite sides of a helix.
6. **Helix rule**: Only the contacts between residues i and i+4 is allowed in a helix.
7. **Beta rule**: No contacts (|j-i|>3) are allowed within any strand
8. **Right-hand crossover rule**: Crossovers between parallel strands of the same sheet (paired or not) are right-handed. especially if the crossover contains a helix.
9. **Helix crowding rule**: If a helix can go to either side of a sheet, it picks the side with fewer crossovers.
10. **Strand burial rule**: If a strand can pair with either of two other strands, it chooses the one that is more non-polar.
11. **Propagation rule**: A contact cannot be assigned between i and j if there are more than 8 residues in the intervening sequence that have no assigned contacts.

or zero possible folds when projected into three dimensions. Second, the order of appearance of contacts (i.e. the pathway) was not considered, even though much is known about the general character of folding pathways (Baldwin 1995, Fersht 1995, Galzitskaya et al. 2001, Nolting and Andert 2000). In the new approach we tried to enforce “physicality” and protein-like characteristics by using protein templates and simple rules. The rules consist of common sense facts for the packing of secondary structures (Table 1). Rules for the order of appearance were derived from the general assumptions of a nucleation/propagation pathway (Nolting and Andert 2000).

4.1 A knowledge-based potential for motif-motif interactions

The first step in predicting a contact map is to assign an energy to each potential contact. The energy in this case is the database-derived likelihood of contact between any two local structure motifs. This implies that local structure forms first, then these sub-structures condense to form larger units, subject to a free energy of interaction, similar to a binding energy. But like its predecessors I-sites and HMMSTR, HMMSTR-CM is a Bayesian ensemble approach; each residue is represented as a probability distribution of motifs, rather than as a single motif.

Thus, each contact potential models a pair of flickering local structures, interacting in proportion to their structural content.

The energetic interaction potential of two motifs is modeled as the statistical interaction potential between two corresponding Markov states of the HMMSTR model. Knowledge-based Markov state "pair potentials" were summed from the CATH database of protein domains. Each domain was first preprocessed into Markov state probability distributions using the Forward/Backward algorithm (Rabiner 1989) to get the position-dependent Markov state probability distribution γ (Eq. 2).

$$\gamma(i,q) = P(q \mid i) \tag{2}$$

The pairwise contact potential between any two HMMSTR states *p* and *q* (*G(p,q,s)*) was calculated as the log of the mutual probability of these two states in contacting residues (Cα-Cα distance < 8Å), for proteins in the PDBselect database (Hobohm and Sander 1994) (Eq. 3).

$$G(p,q,s) = -\log \frac{\sum_{PDBSelect} \sum_{i \ni D_{i,i+s} < 8\text{Å}} \gamma(i,p) * \gamma(i+s,q)}{\sum_{PDBSelect} \sum_{i} \gamma(i,p) * \gamma(i+s,q)} \tag{3}$$

The sensitivity of discriminating contacts from non-contacts improved greatly by calculating *G* as a function of the sequence separation *s*=|*j*-*i*| (4 ≤ s ≤ 20. For sequence separations greater than 20, s=20 was used.) The total number of potential functions *G* was 1037153, one for every pair of 247 Markov states in HMMSTR and every separation distance from 4 to 20. *G* may be viewed as the knowledge-based energy of contacts between local structure motifs.

The target contact potential map *E* (Eq. 4) is the matrix of contact potentials between every two residues in the target sequence. The contact potential between residues *i* and *j* (*E(i,j)*) in the target was calculated as the probability-weighted sum of the pairwise potential functions *G*.

$$E(i,j) = \sum_{p} \sum_{q} \gamma(i,p) * \gamma(j,q) * G(p,q,s), \tag{4}$$

where *s* = |*i*-*j*|. In general, the contact potential map readily identifies possible contacts between β strands, and also finds super-secondary structure motifs such as the right-handed parallel $\beta\alpha\beta$ motif and the $\alpha\alpha$-corner.

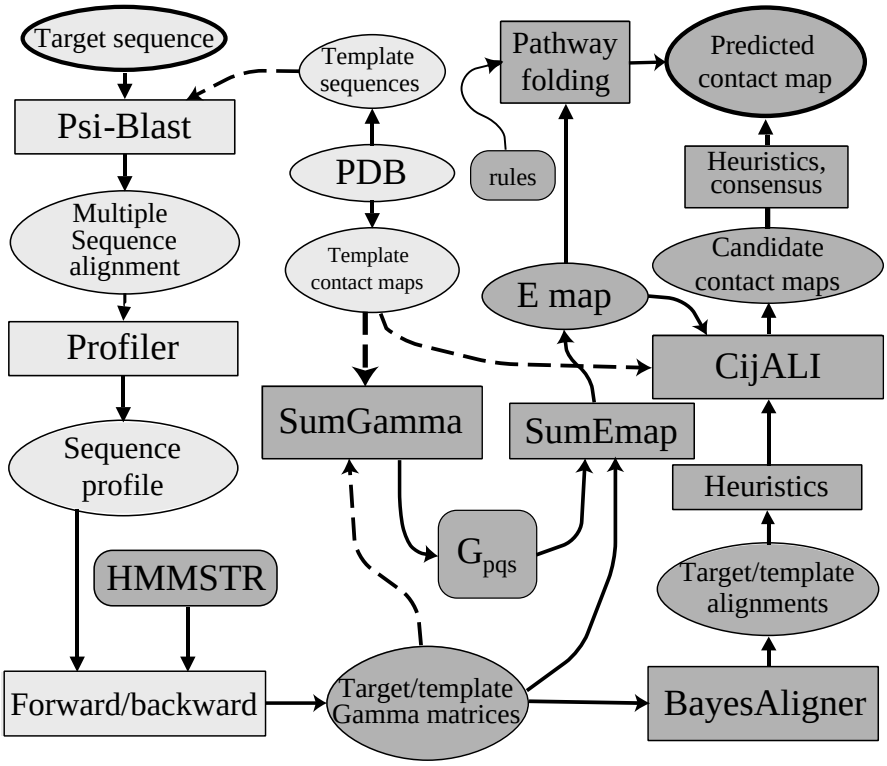

Fig. 4. Flowchart for HMMST-CM contact map prediction. Rectangles represent algorithms, ovals are data, and rounded rectangles are models. Dashed lines apply to training set data (templates) and solid lines apply to both templates and targets. Light gray items are describe in referenced material Dark gray items are described in this text as follows: HMMSTR, section 2.2; Gamma matrices, Eq. 2; SumGamma, Gpqs, Eq. 3; SumEmap, E map, Eq. 4; Rules,, Pathway folding, section 4.4, Table 1; BayesAligner, Target/template alignments, section 4.2, Eq. 5, Fig. 5a; Heuristics, Eq. 6; CijALI, section 4.2, Eq. 7; Heuristics, consensus, section 4.3, Fig. 6.

4.2. Fold recognition using contact potential maps

The flowchart in Fig. 4 summarizes the steps in a contact map prediction using HMMSTR-CM. Target sequences were aligned to database sequences using PSI-BLAST (Altschul et al. 1997). The resulting multiple sequence alignment was converted to an amino acid probability distribution or sequence profile, as described previously (Bystroff and Baker 1998). The target sequence profile and

1239 template profiles from the PDBselect database (Hobohm and Sander 1994) were converted to HMMSTR γ-matrices (Eq. 2), and γ^{target} was aligned against each $\gamma^{template}$ using Bayesian adaptive alignment (Zhu et al. 1998). The alignment matrix in this case was the sum over all joint probabilities of Markov states (Eq. 5). The alignments were evaluated using contact potential maps to choose the best template.

$$A_{ij} = \sum_q \gamma_{iq}^{target} \gamma_{jq}^{template} \tag{5}$$

Candidate target contact maps were generated for each alignment, and each was evaluated by the contact free energy (CFE), as described below, and other measures. The BayesAligner produced a single score and any number of alignments. Templates with low alignment scores were rejected. Otherwise, 100 alignments were selected at random for further evaluation.

BayesAligner produces a probability distribution over all possible alignments with no more than *k* gaps (*k* depends on the sequence lengths). The quality of the alignment distribution (see Fig. 5a) was a strong indicator of the quality of the template. Templates and/or alignments were removed from this set if they were highly fragmented. This was assessed using a "*compactness score*" which is simply the length of the longest contiguously aligned region, ignoring small gaps (≤ 3 residues). The template distance at the ends of the aligned blocks was enforced to be physically possible values (Eq. 6) by trimming the aligned blocks if necessary.

$$D_{i'j'} \le 3.8\text{Å} \times |i - j| \tag{6}$$

Candidate contact maps (*C*) were generated using the alignments and the contact maps of each of the templates that had the top 10 compactness scores was scored using the "contact free energy" (*CFE*, Eq. 7). CFE was calculated by summing the relative contact potential *E* over all contacts, *C*. Contacts with sequence separations |*j*-*i*| less than 4 were ignored.

$$CFE = \sum_{i,j \ni C_{ij}=1 \cap (j>(i+3))} E(i,j) - \langle E \rangle, \tag{7}$$

where <*E*> is the mean contact potential for the target. For each template, we calculated the *CFE* for all contact map candidates and chose the one with the best energy as the best alignment to that template.

After we carried out the above procedure for every template in our dataset, we usually accumulated several hundred target contact map predictions. How to evaluate them and choose one as the final prediction became a problem itself. The

decision was made by considering four parameters: CFE, the BayesAligner score, the *compactness score* and the similarity between sequence lengths of the target and the template. The primary parameter was the CFE since it represented the free energy of the sequence when folded to the template structure. But we observed that better alignments and similar lengths improved the perceived prediction quality.

The automated selection of templates was sometimes overridden by our *ab initio* analysis, described below. If the propagation rules favored one topology over another and a template of the favored topology was present in our list of top scorers, we would select that template over a higher scoring one.

4.3. Consensus and composite contact map predictions

Often several of the top-scoring templates contained the same fold or substructure. Consensus was considered a strong indicator, especially if the fold was uncommon. Multiple candidates were sometimes used to construct a single composite map. In practice, consensus similarity between many structures is difficult to see in a 3D multiple superposition, but is easy to see in superimposed contact maps.

This prediction can be done in different ways when the top scoring templates share a similar fold. When they disagree on some contacts, the consensus contacts (not necessarily those from the best scoring template) are used; when some templates aligned well in one region and other templates aligned well in another region, the predictions from these templates were spliced to maximize the coverage. For some recurrent contact patterns, e.g. the parallel βαβ motif, the parallel β contacts or the helix contacts were sometimes incomplete because of misalignment of the template. By combining the top scoring predictions, we could "grow" the incomplete pattern into a complete one.

Simply combining the contact maps introduces "noise" – contacts that make the prediction non-physical. (A "non-physical" contact map cannot be projected into 3-dimensions.) Manual post-processing, including pathway-based editing (discussed next) was needed to enforce the physicality of the final contact map.

4.4. *Ab initio* rule-based pathway predictions

The fold-recognition methods described above have their roots in evolution, but contact maps as a representation of protein structures were chosen not with the

intention of building a Darwin-based prediction strategy, but with the intention of modeling the folding pathway. Contact maps simplify the conformational search. However, as we have pointed out, not all contact maps represent physically-possible three-dimensional objects. Therefore, external information about proteins must be included. A set of aligned templates is one source of external information. Here we present a set of fundamental rules (Table 1) and energies (Eq. 4) that serve the same purpose – to restrict the conformational search to contact maps that are physically possible and protein-like.

A rule-based structure propagation model was used either in conjunction with templates or *ab initio* (without templates). In CASP5, *ab initio* predictions were sometimes done on targets found later to be remote homologs by CASP5 assessors, but because our alignment method was not always able to recognize remote homology, we treated them as potential new folds. The procedure is as follows.

Starting from a contact potential map, *E*, we kept the contacts that were better than a cutoff value. The cutoff value was chosen such that blocks of contacts were found between most secondary structural units, especially between β strands. As a result, the initial contact map was often characterized by dense blocks of contacts between β strands and sparse contacts to helices and between helices.

If we kept all of these contacts, clearly the map would be physically impossible. For example, a β strand element cannot be paired with more than two other β strands. A set of common-sense rules were compiled to weed out the possible contacts from the impossible or unlikely, and to enforce protein-like characteristics, such as right-handed crossovers and exposed reverse turns (Table 1). These rules were enforced as the prediction was propagated.

The folding pathway consisted of "assigning" or "erasing" contacts. Contacts were assigned if the energy E(i,j) passed a threshold and the corresponding contact $C_{ij} = 1$ did not violate any of the rules, otherwise they were erased. Blocks of potential contacts were considered together, and the order in which blocks were considered depended on their proximity to previously assigned blocks of contacts (Table 1, Rule 11), following the principles of the nucleation/condensation folding mechanism.

To start the folding pathway, we selected one or more local regions with high confidence contacts as the "nucleation site(s)". We then propagated the prediction in both directions by assigning or erasing blocks of contacts around and between the nucleation site(s), subject to our set of rules. TOPS diagrams (Sternberg and Thornton 1976) were drawn for the growing structure as a visual aid, since some rules applied to the topology. The pathway, and the prediction, was complete

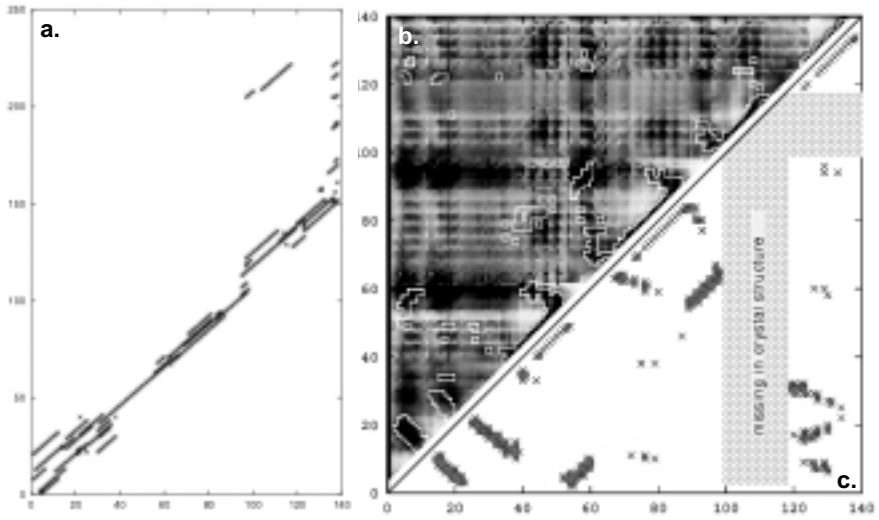

Fig. 5. a. BayesAligner summary of most probable alignments between YqgF (X-axis) and 1HJR (Y-axis). **b.** Contact potential map for YqgF; darker is lower energy *E(i,j)*. Predicted contacts are outlined in white. **c.** Contact map from crystal structure of YqgF, hypothetical protein from E. coli.

when all of the remaining contacts were rejected. The method is best described using examples, as in the next section.

4.5. Selected Results of HMMSTR-CM Blind Structure Predictions

HMMSTR-CM was used to predict contact maps as part of the CASP5 experiment. Targets in the FR (fold recognition) and NF (new fold) categories were predicted using the three methods described above: threading, consensus and *ab initio*, collectively called HMMSTR-CM. In all these three methods, the overall accuracy of the contact map prediction depends on the accuracy of the secondary structure prediction, which was done using HMMSTR.

4.5.1. A prediction using templates and a pathway

YqgF, a hypothetical protein from E. coli, was successfully predicted using the template-based approach in conjunction with a pathway prediction. All visible secondary structure units are correctly predicted (note that the 17 residues from 102 to 118 are missing in the crystal structure), and all of the true contacts have better-than-average *E(i,j)* score. After aligning the contact potential matrix, *E*, to

each of the 1258 templates, a consensus contact map was plotted using the top-scoring six templates. This map was used to construct a folding pathway. Nucleating the pathway at $\beta_4\alpha_2\beta_5$ and propagating produced a TOPS diagram that agreed with only one of the templates, 1HJR, and this template was therefore chosen to construct the consensus contact map. 1HJR had the third highest CFE score. In the prediction based on 1HJR, the N-terminal 3-strand β meander is slightly under-predicted, and a contact between helices 1 and 2 is slightly over-predicted. Nonetheless, the topology is correct throughout (Fig. 5b). The two higher-scoring templates that were not chosen had very different, and incorrect, topologies.

4.5.2. A prediction using several templates

Ycdx, another hypothetical protein from *E. coli*, was successfully predicted using multiple templates. The threading approach found 4 templates that had high CFE scores and also shared common structural components. Three of those templates were 8-stranded α/β barrels and the other consisted of two parallel α/β domains. Ycdx turned out to be an αβ barrel with 7 parallel β strands (PDB code 1M65). Templates with good CFE scores existed but none of them predicted all of the first five helices and the parallel β strand contacts correctly. However, by combining the results from the top scoring templates, we made a consensus prediction that was better than any of the contact maps made from the single templates. In particular, we correctly found parallel contacts between the first 6 β strands (Fig. 6).

The sixth helix and the contacts between the sixth and the seventh strands were predicted but misaligned. Our method mispredicted the C-terminus to be a parallel βαβ motif, as in a standard 8-stranded TIM barrel, but the true structure is three short helices connected by loops. Visual inspection of the templates confirmed that they share the same topology, and a consensus fold prediction would have been obvious given this result. But finding structural similarity and combining structures is more easily automated in the 2D contact map format than in 3D coordinate space. Consensus in contact maps provides a way to merge and "grow" the incomplete contact maps of different targets into a more complete contact map.

Ycdx also revealed a weakness of the method. HMMSTR, which is trained to recognize recurrent super-secondary motifs, does not recognize the unusual substructure at the C-terminus of this protein, 3 short helices instead of the usual βαβ motif. The consensus method, as we have defined it, tends to bias the prediction toward the more common folds. In fact, this is a problem with any template-based method.

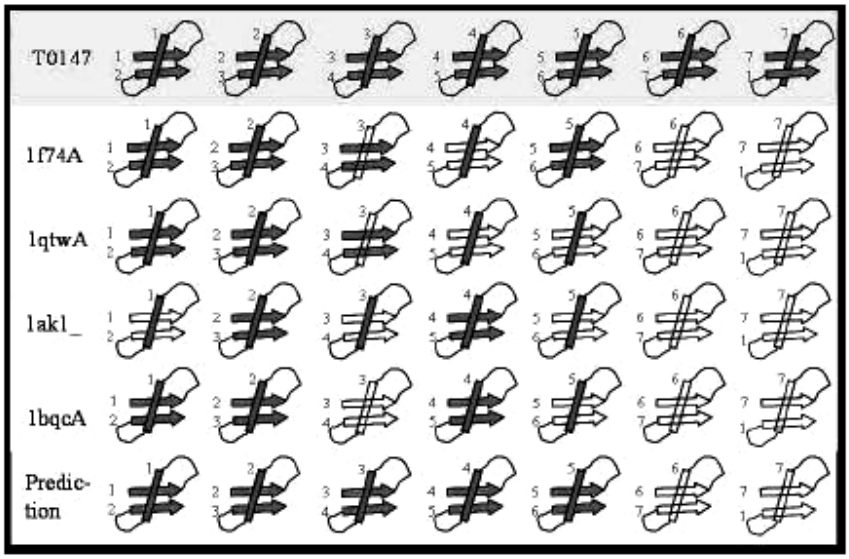

Fig. 6. Summary of strand-strand (arrows) contacts and helix predictions for four templates aligned to Ycdx (T0147). Shaded symbols represent contacts that were correctly predicted using the template specified in the margin. The last line shows contacts that were correctly predicted after combining the four templates and using the consensus set.

4.5.3. Correct prediction using only the folding pathway

Hypothetical protein HI0073 from *H. influenzae* is an example of a successful *ab initio* prediction. It has 116 residues arranged in a three-layer all-parallel α/β sandwich. The contact potential map (Fig. 7a) shows that most of the true contacts are assigned favorable (darker) contact potentials. However, many other favorable regions are also correctly predicted as non-contacts. Depending on the choice of nucleation sites, there was more than one way to derive a physically possible and high scoring topology. In this case, the nucleation site was selected to be $\beta_2\alpha_2\beta_3$. Contacts were assigned or erased in 4 steps, as follows:

(1) Parallel β contacts were assigned between β_2 and β_3.

(2) Anti-parallel β contacts were assigned to β_1 and β_2. All other β contacts to β_2 were erased.

(3) There were two ways to make a right-handed crossover from β_3 to β_4, as shown in figure 3 (c) and (d). Since β_1 was more hydrophobic and β_3 more polar, we paired β_1 and β_4. All other β contacts to β_1 and contacts between α_2 and α_3 were erased.

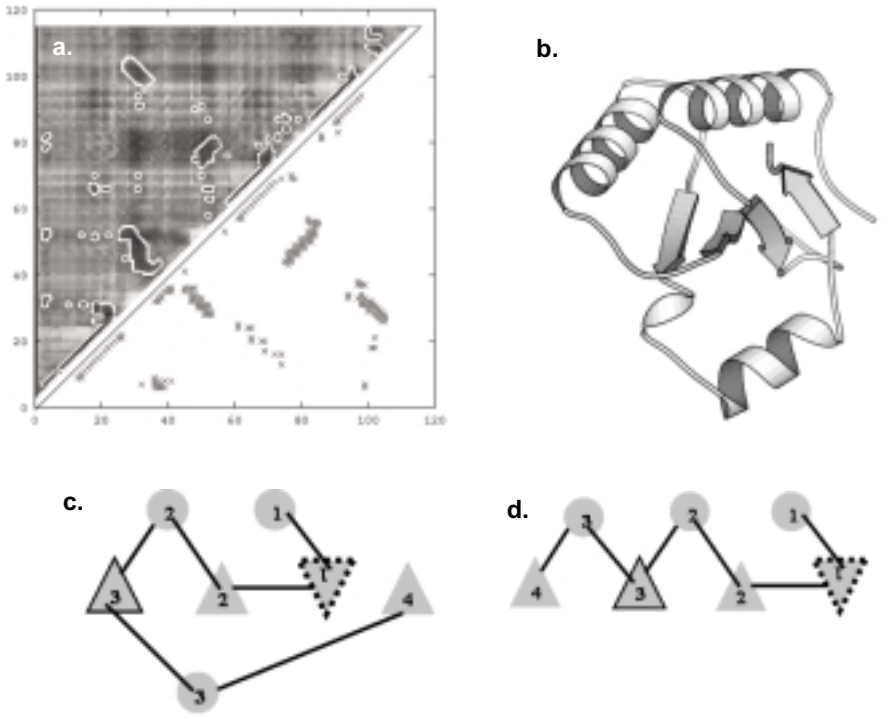

Fig. 7. a. Upper triangle: contact potential map for HI0073 showing predicted contacts as white outlines. Darker means lower energy, *E(i,j)*. Lower triangle: true contacts. **b.** Molscript drawing of the crystal structure of HI0073, a hypothetical protein from *haemophilus influenzae*. **c.** Correct TOPS diagram showing non-polar strand (dashed) buried. **d.** Incorrect TOPS diagram, consistent with all rules except strand burial rule.

(4) α_1 must be on the opposite side of the sheet from α_3, since α_3 extends across the sheet. Therefore, contacts were assigned between α_1 and α_2 and erased between α_1 and α_3.

The completed TOPS diagram and contact map accurately match the true structure (Fig. 7b). The contact map prediction has 42% contact coverage and 29% accuracy. However, accuracy and coverage are not good measures of the quality of a contact map prediction, since near-contacts and gross errors are counted equally. Most of the false positive contacts in the HI0073 prediction are adjacent to true contacts. If we count near misses (±1 residue), then the coverage is 75% and the accuracy is 57%. Note that the long range contacts between the β_1 and β_4

were correctly predicted, which speaks to the power of rule-based methods over raw statistics.

Identification of the folding nucleation site is the critical step in this approach. Once the nucleation site is chosen, the subsequent contact assignments are often unambiguous. After assigning secondary structures and choosing $\beta_2\alpha_2\beta_3$ as the nucleation site, only one folding pathway was possible, and it leads to the correct structure (Fig. 7c). It is interesting to note that this pathway also predicts a possible misfolded state (Fig. 7d). At step (3) in the pathway, a critical decision is made that depends on the sequences of strands 1 and 3. If strand 1 was more polar and strand 3 more hydrophobic, then the alternative structure would be predicted. A simple mutation experiment might tell us whether our model is on the right track.

The choice of the nucleation site in HI0073 was relatively easy. Only one of the three potential $\beta\alpha\beta$ units had a high score. The hairpin between β_1 and β_2 would also be a correct choice, but the selection of $\beta_2\alpha_1\beta_3$ eliminated more of the potential incorrect folding pathways.

4.5.4. False prediction using the folding pathway. What went wrong?

The KaiA N-terminal domain from *S. elongatus* (PDB code 1M2E) is an example where we chose the wrong nucleation site. KaiA is 135 residues long and has five β and five α units. From its contact potential, two possible nucleation sites could be identified, $\beta_2\alpha_2\beta_3$, or $\beta_3\alpha_3\beta_4$. We chose $\beta_2\alpha_2\beta_3$ as the nucleation site instead of the correct, and higher scoring, $\beta_3\alpha_3\beta_4$ unit in order to favor a region of non-local high confidence contacts between β_1 and β_3 and between β_1 and β_4. Our mistake was in assigning non-local contacts before assigning local ones. If we had chosen the correct nucleation site, $\beta_3\alpha_3\beta_4$, there would be an unambiguous choice of the N-terminal $\beta\alpha\beta\alpha\beta$ segment. This sequence of five secondary structures is most commonly found in a three stranded parallel sheet, and since in this case β_2 is polar and β_3 already pairs with another strand, only β_1 could be placed in the middle of the sheet. This would have given the correct 2134 strand order (Fig. 8a), and the helices would have been correctly placed according to our propagation rules (particularly the Right-handed Crossover Rule). Our erroneous choice of the nucleation site leads to the incorrect strand order 2314 (Fig. 8b), instead of 2134. For the record, here is the correct pathway for KaiA using HMMSTR-CM:

(1) Nucleation site at $\beta_3\alpha_3\beta_4$

(2) The N-terminal parallel $\beta\alpha\beta\alpha\beta$ unit must have β_1 in the middle, since β_2 is polar and β_3 cannot be in the middle. To satisfy right-handed crossover rule, α_2 must be on the same side of the sheet as α_3.

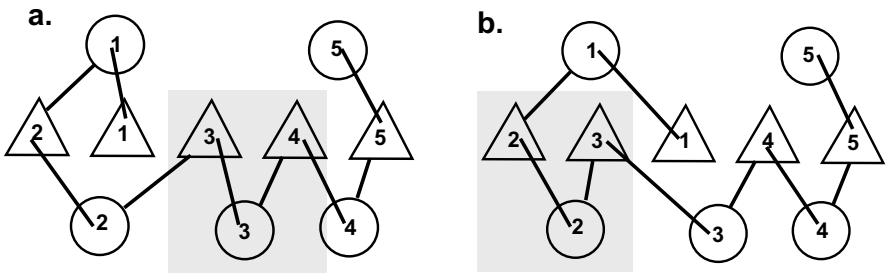

Fig. 8. a. Correct TOPS diagram for KaiA, generated using the pathway described in the text using the shaded βαβ unit as the nucleation site**. b.** Incorrect TOPS diagram, similar to the actual prediction, generated using a similar pathway but starting with the wrong nucleation site (shaded).

(3) β_5 must pair with β_4 since cannot pair with β_2, due to crossovers on both sides of the sheet.

4) α_5 must go on the same side of the sheet as α_1, due to helix crowding on the other side.

For other targets, pathway construction and CFE score alignment methods failed if the secondary structure prediction was inaccurate. In several targets, including HIP1R N-terminal domain from rat, an all-helix protein, secondary structure prediction by HMMSTR significantly under-predicted the helices. The wrong secondary structure pattern led to the wrong assignment of contact potentials, and therefore the wrong assumption of possible topologies. Under-prediction of helices was identified as a problem in HMMSTR.

4.6. Future Directions for HMMTR-CM

By gaining insight about how different parts of the protein pack together, we can improve the accuracy of the *ab initio* method. This will be necessary to make the whole prediction process automatic. The rule-based pathway approach depends on the correct assignment of the fold class of the target (all-α, α/β, α+β or all β (Zhou 1998)), since the rules of propagation depend on choices of the final topology. Generally this assignment is not difficult. So far, it has been applied only to the α/β class, but a different set of rules may be envisioned for the packing of helices and all β proteins.

The difficulty of choosing the correct nucleation site increases with protein size, since there are more to choose from. For larger proteins, more than one cor-

rect choice may be required. One possible approach could be a recursive algorithm to exhaust all the possible topologies by starting with each potential nucleation site, and then evaluate the topologies using the contact potential.

Another improvement might be to attempt to make the contact map prediction more protein like. Our predictions have many false contacts adjacent to true contacts, e.g. a "fat" β-hairpin prediction – even though it is predicted at the right position. Rules to prune this type of false contacts – in other words, to beautify the predicted contact blocks – would increase the accuracy of our prediction. This will require better secondary structure predictions.

5. Conclusions

We have developed methods for calculating an inter-residue contact potential map for a protein sequence, for aligning that map to templates, and for pruning that map using a folding pathway model. Results on CASP5 targets reveal that the folding pathways for some α/β proteins are unambiguous given the correct choice of the folding nucleation site. Pathway predictions improved the selection of a remote homolog for one threading target. Consensus contact maps are more complete than maps from single templates. The contact map representation of protein structure is a useful intermediate-level of detail that facilitates rule-based algorithm development.

References

Alm E & Baker D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc Natl Acad Sci U S A* 96, 11305-10.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.

Anfinsen CB & Scheraga HA. (1975). Experimental and theoretical aspects of protein folding. *Adv Protein Chem* 29, 205-300.

Aszodi A, Munro RE & Taylor WR. (1997). Distance geometry based comparative modelling. *Fold Des* 2, S3-6.

Baldwin RL. (1995). The nature of protein folding pathways: the classical versus the view. *J Biomol NMR* 5, 103-9.

Baldwin RL & Rose GD. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem Sci* 24, 26-33.

Blanco FJ, Rivas G & Serrano L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Biol* 1, 584-90.

Bonneau R & Baker D. (2001). Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* 30, 173-89.

Bonneau R, Strauss CE & Baker D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* 43, 1-11.

Brunger AT, Clore GM, Gronenborn AM & Karplus M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc Natl Acad Sci U S A* 83, 3801-5.

Bystroff C & Baker D. (1997). Blind predictions of local protein structure in CASP2 targets using the I-sites library. *Proteins* Suppl 1, 167-71.

Bystroff C & Baker D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 281, 565-77.

Bystroff C & Garde S. (2003). Helix propensities of short peptides: Molecular dynamics versus bioinformatics. *Proteins* 50, 552-62.

Bystroff C & Shao Y. (2002). Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* 18 Suppl 1, S54-61.

Bystroff C, Simons KT, Han KF & Baker D. (1996). Local sequence-structure correlations in proteins. *Curr Opin Biotechnol* 7, 417-21.

Bystroff C, Thorsson V & Baker D. (2000). HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology* 301, 173-90.

Cavalli A, Ferrara P & Caflisch A. (2002). Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins* 47, 305-14.

Chan HS, Bromberg S & Dill KA. (1995). Models of cooperativity in protein folding. *Philos Trans R Soc Lond B Biol Sci* 348, 61-70.

Colon W & Roder H. (1996). Kinetic intermediates in the formation of the cytochrome c molten globule. *Nat Struct Biol* 3, 1019-25.

Crippen GM, Havel, T.F. (1988). *Distance Geometry and Molecular Conformation*. Chemometrics Series, 15, John Wiley & Sons.

Duan Y & Kollman PA. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution [see comments]. *Science* 282, 740-4.

Dyson HJ & Wright PE. (1996). Insights into protein folding from NMR. *Annu Rev Phys Chem* 47, 369-95.

Eaton WA, Thompson PA, Chan CK, Hage SJ & Hofrichter J. (1996). Fast events in protein folding. *Structure* 4, 1133-9.

Eddy SR. (1996). Hidden Markov models. *Curr Opin Struct Biol* 6, 361-5.

Efimov AV. (1993). Standard structures in proteins. *Prog Biophys Mol Biol* 60, 201-39.

Fariselli P & Casadio R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Eng* 12, 15-21.

Fariselli P, Olmea O, Valencia A & Casadio R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins* Suppl 5, 157-62.

Fersht AR. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci U S A* 92, 10869-73.

Fersht AR, 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, 771-82.

Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR & Dunbrack RL, Jr. (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* Suppl 5, 171-83.

Frishman D & Argos P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23, 566-79.

Galzitskaya OV, Ivankov DN & Finkelstein AV. (2001). Folding nuclei in proteins. *FEBS Lett* 489, 113-8.

Garcia AE & Sanbonmatsu KY. (2001). Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins* 42, 345-54.

Gillespie JR & Shortle D. (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, 170-84.

Gnanakaran S & Garcia AE. (2002). Folding of a Highly Conserved Diverging Turn Motif from the SH3 Domain. *Biophys J*.

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

Grantcharova VP, Riddle DS & Baker D. (2000). Long-range order in the src SH3 folding transition state. *Proc Natl Acad Sci U S A* 97, 7084-9.

Gulotta M, Gilmanshin R, Buscher TC, Callender RH & Dyer RB. (2001). Core formation in apomyoglobin: probing the upper reaches of the folding energy landscape. *Biochemistry* 40, 5137-43.

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

Han KF & Baker D. (1995). Recurring local sequence motifs in proteins. *J Mol Biol* 251, 176-87.

Han KF, Bystroff C & Baker D. (1997). Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci* 6, 1587-90.

Heidary DK & Jennings PA. (2002). Three topologically equivalent core residues affect the transition state ensemble in a protein folding reaction. *J Mol Biol* 316, 789-98.

Hobohm U & Sander C. (1994). Enlarged representative set of protein structures. *Protein Sci* 3, 522-4.

Houry WA, Rothwarf DM & Scheraga HA. (1996). Circular dichroism evidence for the presence of burst-phase intermediates on the conformational folding pathway of ribonuclease A. *Biochemistry* 35, 10125-33.

Hu J, Shen X, Shao Y, Bystroff C & Zaki MJ. (2002). *BIOKDD 2002, Edmonton, Canada*.

Jacchieri SG. (2000). Mining combinatorial data in protein sequences and structures. *Molecular Diversity* 5, 145-152.

Jones DT. (1998). *Critical Assessment of Protein Structure Prediction 3, Asilomar, CA.*

Kabsch W & Sander C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637.

Karplus K, Barrett C & Hughey R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846-56.

Kolinski A & Skolnick J. (1997). High coordination lattice models of protein structure, dynamics and thermodynamics. *Acta Biochim Pol* 44, 389-422.

Krueger BP & Kollman PA. (2001). Molecular dynamics simulations of a highly charged peptide from an SH3 domain: possible sequence-function relationship. *Proteins* 45, 4-15.

Laurents DV & Baldwin RL. (1998). Protein folding: matching theory and experiment. *Biophys J* 75, 428-34.

Mateu MG, Sanchez Del Pino MM & Fersht AR. (1999). Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. *Nat Struct Biol* 6, 191-8.

Mendes J, Guerois R & Serrano L. (2002). Energy estimation in protein design. *Current Opinion in Structural Biology* 12, 441-446.

Mirny L & Shakhnovich E. (2001). Protein folding theory: from lattice to all-atom models. *Annu Rev Biophys Biomol Struct* 30, 361-96.

Mok YK, Elisseeva EL, Davidson AR & Forman-Kay JD. (2001). Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states. *J Mol Biol* 307, 913-28.

Mok YK, Kay CM, Kay LE & Forman-Kay J. (1999). NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J Mol Biol* 289, 619-38.

Moult J, Fidelis K, Zemla A & Hubbard T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins* Suppl 5, 2-7.

Munoz V, Blanco FJ & Serrano L. (1995). The hydrophobic-staple motif and a role for loop-residues in alpha- helix stability and protein folding. *Nat Struct Biol* 2, 380-5.

Munoz V, Henry ER, Hofrichter J & Eaton WA. (1998). A statistical mechanical model for beta-hairpin kinetics. *Proc Natl Acad Sci U S A* 95, 5872-9.

Nolting B & Andert K. (2000). Mechanism of protein folding. *Proteins* 41, 288-98.

Nolting B, Golbik R, Neira JL, Soler-Gonzalez AS, Schreiber G & Fersht AR. (1997). The folding pathway of a protein at high resolution from micro-seconds to seconds. *Proc Natl Acad Sci U S A* 94, 826-30.

Northey JG, Di Nardo AA & Davidson AR. (2002a). Hydrophobic core packing in the SH3 domain folding transition state. *Nat Struct Biol* 9, 126-30.

Northey JGB, Maxwell KL & Davidson AR. (2002b). Protein folding kinetics beyond the Phi value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. *Journal of Molecular Biology* 320, 389-402.

Nymeyer H, Socci ND & Onuchic JN. (2000). Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. *Proc Natl Acad Sci U S A* 97, 634-9.

Oliveberg M, Tan YJ, Silow M & Fersht AR. (1998). The changing nature of the protein folding transition state: implications for the shape of the free-energy profile for folding. *J Mol Biol* 277, 933-43.

Olmea O & Valencia A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 2, S25-32.

Onuchic JN, Luthey-Schulten Z & Wolynes PG. (1997). Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 48, 545-600.

Ortiz AR, Kolinski A & Skolnick J. (1998). Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol* 277, 419-48.

Ortiz AR & Skolnick J. (2000). Sequence evolution and the mechanism of protein folding. *Biophys J* 79, 1787-99.

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

Plaxco KW, Simons KT & Baker D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277, 985-94.

Pollastri G & Baldi P. (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 LR. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc IEEE* 77, 257-286.

Rooman MJ, Rodriguez J & Wodak SJ. (1990). Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* 213, 327-36.

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

Schellman C. (1980). *Protein folding: The proceedings of the 28th Conference of the German Biochemical Society, University of Regensburg, Sept 10-12, 1979, Regensburg, West Germany*.

Shakhnovich EI. (1998). Folding nucleus: specific or multiple? Insights from lattice models and experiments. *Fold Des* 3, R108-11; discussion R107.

Shao Y & Bystroff C. (2003). Predicting inter-residue contacts using templates and pathways. *Proteins, Structure, Function and Genetics* in press.

Shea JE & Brooks CL, 3rd. (2001). From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu Rev Phys Chem* 52, 499-535.

Shoemaker BA & Wolynes PG. (1999). Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. *J Mol Biol* 287, 657-74.

Simons KT, Bonneau R, Ruczinski I & Baker D. (1999a). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* Suppl 3, 171-6.

Simons KT, Kooperberg C, Huang E & Baker D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268, 209-25.

Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C & Baker D. (1999b). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34, 82-95.

Singer MS, Vriend G & Bywater RP. (2002). Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng* 15, 721-5.

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

Sternberg MJ & Thornton JM. (1976). On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit. *J Mol Biol* 105, 367-82.

Steward RE & Thornton JM. (2002). Prediction of strand pairing in antiparallel and parallel beta- sheets using information theory. *Proteins-Structure Function and Genetics* 48, 178-191.

Thirumalai D & Klimov DK. (1998). Fishing for folding nuclei in lattice models and proteins. *Fold Des* 3, R112-8; discussion R107.

Vendruscolo M, Kussell E & Domany E. (1997). Recovery of protein structure from contact maps. *Fold Des* 2, 295-306.

Viguera AR & Serrano L. (1995). Experimental analysis of the Schellman motif. *J Mol Biol* 251, 150-60.

Woolfson DN & Alber T. (1995). Predicting oligomerization states of coiled coils. *Protein Sci* 4, 1596-607.

Yi Q, Bystroff C, Rajagopal P, Klevit RE & Baker D. (1998). Prediction and structural characterization of an independently folding substructure in the src SH3 domain. *J Mol Biol* 283, 293-300.

Zaki MJ, Shan J & Bystroff C. (2000). *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering, Arlington, VA, USA.*

Zhou GP. (1998). An intriguing controversy over protein structural class prediction. *J Protein Chem* 17, 729-38.

Zhu J, Liu JS & Lawrence CE. (1998). Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 14, 25-39.

Zwanzig R. (1997). Two-state models of protein folding kinetics. *Proc Natl Acad Sci U S A* 94, 148-50.