Prediction of protein structure by emphasizing local side-chain/backbone in ensembles of turn fragments

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JANUARY 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10541 · Source: PubMed

2 AUTHORS, INCLUDING:

Qiaojun Fang
Fred Hutchinson Cancer Research Center
8 PUBLICATIONS **131** CITATIONS

PROTEINS: Structure, Function, and Genetics 53:486–490 (2003)

Prediction of Protein Structure by Emphasizing Local Side-Chain/Backbone Interactions in Ensembles of Turn Fragments

Qiaojun Fang and David Shortle*
Department of Biological Chemistry, The Johns Hopkins University School of Medicine Baltimore, Maryland

ABSTRACT **The prediction strategy used in the CASP5 experiment was premised on the assumption that local side-chain/backbone interactions are the principal determinants of protein structure at low resolution. Our implementation of this assumption made extensive use of a scoring function based on the propensities of the 20 amino acids for 137 different sub-regions of the Ramachandran plot, allowing estimation of the quality of fit between a sequence segment and a known conformation. New folds were predicted in three steps: prediction of secondary structure, threading to isolate fragments of protein structures corresponding to one turn plus flanking helices/strands, and recombination of overlapping fragments. The most important step in this fragment ensemble approach, the isolation of turn fragments, employed 2 to 6 sequence homologues when available, with clustering of the best scoring fragments to recover the most common turn arrangement. Recombinants formed between 3 to 8 turn fragments, with cross-overs confined to helix/strand segments, were selected for compactness plus low energy as estimated by empirical amino acid pair potentials, and the most common overall topology identified by visual inspection. Because significant amounts of steric overlap were permitted during the recombination step, the final model was manually adjusted to reduce overlap and to enhance protein-like structural features. Even though only one or two models were submitted per target, for several targets the correct chain topology was predicted for fragment lengths up to 100 amino acids. Proteins 2003;53:486–490.** © 2003 Wiley-Liss, Inc.

Key words: structure prediction; propensities; Boltzmann hypothesis; Ramachandran map; threading; turns; secondary structure; fragments

In the standard approach to protein structure prediction, unique three-dimensional conformations are generated that position the sequence of amino acids in space, and the energy of each conformation is estimated by applying a scoring function. These two steps, conformational sampling and scoring, are repeated a large number of times, with the conformation having the best score after N trials declared the best prediction. Usually the conformations sampled involve all of the residues of the protein, or of one domain for multi-domain proteins, because the energy is known to be minimized over all chemical interactions involving the protein chain. And usually the most significant term of the scoring function assesses the energetics of long range contacts between side-chains, especially those involving hydrophobic residues.

As documented in the previous CASP experiments, this standard approach has enjoyed very limited success in the prediction of new folds until recently. With the extensive use of large amounts of bioinformatics-type information in both secondary structure prediction and in fragment selection, real progress in structure prediction has been made .[1] The most common explanation for how sequence information from known homologues can produce such major improvements in prediction methods is improved sampling: by restricting the fragment search to chain segments that have similar sequences and similar amino acid substitution profiles, more time is spent in promising regions of conformation space.

Recent experimental studies from our laboratory call into question the conventional wisdom that long range contacts play a dominant role in establishing the global fold, or topology, of globular proteins. Application of a new NMR structural parameter, the residual dipolar coupling, to denatured staphylococcal nuclease has demonstrated that, in 8 M urea, a "native-like topology" still persists in the denatured state.[2] Even after replacing 10 large hydrophobic residues with similarly-shaped polar side chains plus addition of 8 M urea, obvious features of this "native-like topology" are still present.[3] While at least 4 other proteins show similar evidence of persistent structure under strongly denaturing conditions, it remains to be determined if they also display the same overall topologies of their native states. The unavoidable conclusion appears to be that local side-chain/backbone interactions play a dominant, perhaps the dominant role in specifying the low resolution structure of proteins.

Additional support for the energetic importance of side-chain/backbone interactions comes from a quantitative analysis of the propensities of the 20 amino acids for different sub-divisions of the Ramachandran map.[4] If the

*Correspondence to: David Shortle, Department of Biological Chemistry, The Johns Hopkins University School of Medicine, 725 N. Wolfe St., Baltimore, MD 21205-2185. E-mail: dshortl1@jhmi.edu

Received 12 February 2003; Accepted 11 April 2003

propensity of amino acid type x for a set of phi/psi values y is calculated as:

$$p(x,y) = \frac{number\ of\ aa(x)\ in\ region(y)/number\ of\ aa(x)}{number\ of\ 20\ aa\ in\ region\ (y)/number\ of\ 20\ aa'} \quad (1)$$

the resulting value can be used in scoring functions that are surprisingly successful at identifying the wild-type conformation of sequence fragments with length of 20-40 amino acids when assayed against more than 300,000 alternative conformations by threading.

From data such as these, we have adopted the premise that local side-chain/backbone interactions are the major determinant of protein structure. From this assumption it follows that the global structure of proteins, at low resolution, should be predictable by recombining short fragments with highly favorable local interactions to form larger chain segments. This is the strategy we explored in CASP5. Fragments were selected from a very large set of protein structures, based on functions that scored local energetics, and then recombined with little attention to long range interactions. Only in large recombinant molecules was the energy of long range side-chain/side-chain contacts used in the selection of structures. While the success achieved in CASP5 is modest, evidence cited below suggests that this approach can be significantly improved by moving beyond sampling by threading to *de novo* conformational sampling.

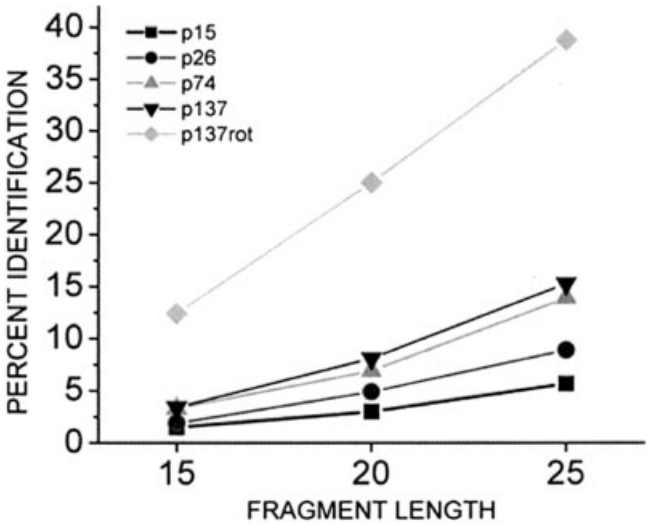

Fig. 1. Graph of the fraction of sequence fragments for which the correct native conformation could be identified on the basis of backbone dihedral angle propensity scores. 1100 arbitrary protein sequence fragments of a given length (x-axis) were threaded through a set of 1052 high resolutions protein structures, which included the protein from which the sequence fragment was taken (described in more detail in Shortle, 2002). The best scoring structure out of ~250,000 alternative conformations was determined and the fraction of the time this proved to be the correct "wild-type" structure is plotted on the y-axis. The scoring functions are pN, where N is the number of subdivisions of the Ramachandran plot (Shortle, 2002) and p137rot, where each of the 137 phi/psi bins has been further subdivided for the 3 rotamers.

METHODOLOGY/RESULTS

Scoring Functions

Three scoring functions were used in the prediction of secondary structure and the selection of turn-containing fragments for recombination. The most important of these, p137, is an extension of the series of phi/psi propensities described previously.[4] When the number of Ramachandran map sub-divisions is increased from 15, to 26, 74, and 137, a simple sum of the negative logarithm of the propensity becomes an increasingly accurate scoring function. As shown in Figure 1, the correct conformation for fragments of protein sequence (15 to 25 amino acids in length) can be identified at a fairly high frequency from among more than 250,000 alternative conformations by threading a set of protein structures that include the source of the sequence fragment.

If each of the 137 phi/psi bins is further subdivided into 3 bins for the trans, gauche+, and gauche- rotamers and the propensities for each phi/psi/chi1 bin calculated, a very large improvement in fragment identification results (Fig. 1). However, this p137rot scoring function was only used in CASP5 for the prediction of secondary structure. For fragment selection, the number of turn fragments identified by threading with favorable p137rot scores was almost always very small or zero. It appears the number of conformations sampled by threading is often too small to find even one conformation that has all three dihedral angles optimized for the target sequence, once the turn segment is longer than 4 to 8 residues.

A second scoring function used in our CASP5 predictions is the propensities of the 20 amino acids to have their "stretched" CB atom (CA-CB bond length of 2.6 angstroms, which approximates the average side chain centroid) surrounded by N neighboring stretched CB atoms within 10 Angstroms. These burial propensities, p(aa, N) were calculated, using equation 1, and incorporated into a separate scoring function by taking the sum of the negative logarithm of the propensity for each residue in the sequence to find itself within that side chain environment. In addition, long range side-chain/side-chain contacts were evaluated using the empirical pair potential function of Bryant and Lawrence,[5] which also makes use of stretched CB atoms. In almost all cases, the burial propensity and B-L energy were used conservatively, by rejecting conformations with scores above a specified cutoff value. Consequently, the lowest value of the summed p137 score was the principal selection criteria in predicting secondary structure and in selecting turn-containing fragments for recombination.

Conformational Sampling

Using the scoring functions above, fragments were selected by threading large sets of proteins of known structure, taken from the culled pdb lists of Wang and Dunbrack[6] (http: //www.fccc.edu/research/labs/dunbrack/pisces/). For prediction of secondary structure, the set consisted of approximately 1500 proteins with less than 20

percent sequence identity and x-ray structures of resolution greater than 2.2 angstroms. For selection of turn-containing fragments, approximately 5000 proteins were used that had less than 60 percent identity and a structure determined by either x-ray crystallography or by NMR. Each protein structure was reduced to a linear template for threading, consisting of a list of phi/psi and chi1 bins, the secondary structure (turn, helix, or strand), the number of neighboring CB atoms within 10 angstroms, plus a matrix of CB-CB distances.

When more than one homologue was used for threading, the residue positions in the homologue were renumbered to correspond to those of the target protein. Threading of each homologue was then conducted, permitting insertions and deletions of up to three residues; fragments that contained larger insertions/deletions were not used. Importantly, the sequence of the template was not used as a parameter in threading. This constitutes a fundamental difference between our "quasi physico-chemical" approach and that of other new fold methods described in this volume, which employ a variety of strategies for recognizing patterns that correlate sequence and structure.

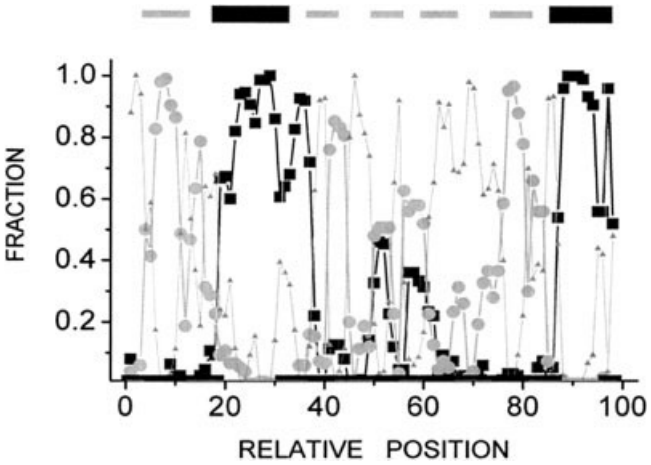

Fig. 2. Profile of the predicted secondary structure for T0149_2, with position 1 corresponding to residue 221. The correct secondary structure of T0149_2 is shown across the top: helix = wide black bar; strand = thin grey bar. The y-axis is the fraction of approximately 100 overlapping fragments with the best p137 scores with secondary structure of turn (triangles), helix (squares), or strand (circles) at that position in the sequence.

Prediction of Secondary Structure

Although a number tactics for predicting secondary structure were evaluated, the most accurate consisted of threading sequence segments from 6 to 12 residues in length and selecting for the 25 fragments with best p137 or p137rot scores, with rejection of fragments with positive burial propensity scores or B-L energies greater than +0.4. The target sequence was divided into sets of constant length segments shifted by 2 or 3 residues, and the secondary structure at each residue position for the best 25 fragments was recorded. To this running tally was added the results from the next set of partially overlapping fragments. Consequently, approximately 75-150 fragments, with a range of end-points contributed to the secondary structure profile at each position in the sequence. One example of this type of turn/helix/strand profile is shown in Figure 2 for the second domain of target T0149.

As is evident in Figure 2, somewhat more than half of residues appear to be clearly and correctly predicted, while the remainder are either wrong or ambiguous. Addition of 2 to 10 sequence homologues typically did not increase the clarity of the results. Although such turn/helix/strand profiles for individual homologues often supported predictions based on the target sequence, in general composite profiles formed by summing the profiles for a set of homologues invariably contained more noise with less pronounced peaks of helix and strand structure. Consequently, the secondary structure predicted by PSIPRED[7] and other servers was consulted to arbitrate the ambiguous segments. Obviously, this human intervention completely compromises the results obtained by our fragment ensemble approach, so we are unable to provide a quantitative estimate of its success rate. Although the rate is considerably lower than PSIPRED, we anticipate that this approach will become more competitive by CASP6 as local side-chain/side-chain interactions (i, i+3 and i, i+4 for helices; i, i+2 for strands) are included in an expanded scoring function.

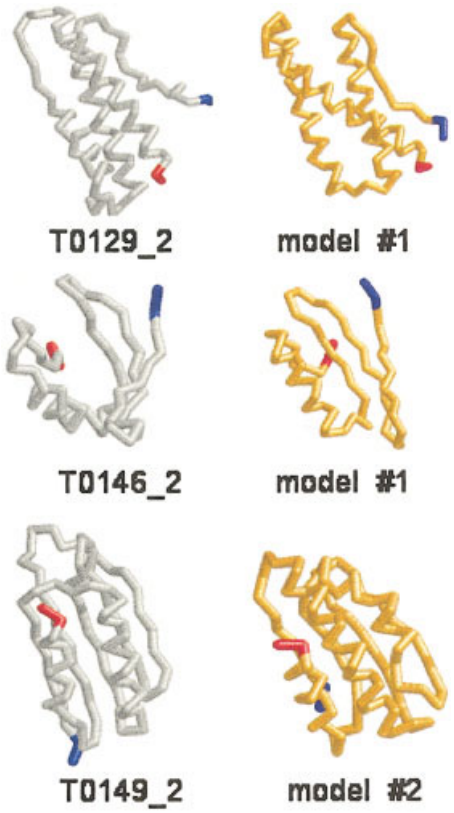

Fig. 3. RasMol backbone representations of the experimental structure (silver) along side the predicted model (gold): T0129_2, residues 77-171, RMSD (CA coordinates) = 8.5; T0146_2, residues 55-113, RMSD = 4.0; T0149_2, residues 223-317, RMSD = 7.7.

Selection of Helix/Strand-Turn-Helix/Strand Fragments

After the secondary structure had been assigned, the sequence of the target protein was divided into turn-containing segments that extend from the beginning of a helix or strand, through a turn, to the end of the next helix or strand. These segments of target sequence were threaded with selection for "non-helix, non-strand" secondary structure for the predicted turn residues, and for the correct secondary structure of the 2 or 3 helical/strand residues nearest the turn region. The secondary structure at the ends of the fragment, which is always either helix or strand, was not selected; instead these residues in the recovered fragments were given canonical helix or strand phi/psi values as necessary. This tactic makes the turn the primary consideration, so that the sample size encountered during threading in not limited by the lengths of the flanking helices/strands.

In most cases, from 3 to 6 sequence homologues were used in addition to the target sequence, choosing those homologues with the fewest and smallest insertions/deletions. Approximately 30 of the best p137 scoring fragments were saved for each homologue (rejecting those with positive burial propensity scores and positive B-L energies), and the saved fragments were compared by pair-wise RMSD of CA-CA distances; those fragments with the nearest average distance to their 10 closest neighbors were kept. Typically this generated a list of 30 to 50 non-redundant fragments. Of these, approximately 25 were saved for each turn-containing segment by selecting those with the lowest burial propensity scores and/or lowest B-L energies. Only coordinates for backbone atoms (N, CA, C, O) and the CB atoms were saved.

Fragment Recombination

Sets of overlapping turn-containing fragments were recombined across every peptide bond along the length of the overlapped helix or strand. Those recombinants displaying steric clashes between atoms with reduced van der Waals radii were rejected. Typically 3 to 8 sets of overlapping fragments were recombined, yielding recombinants of length 40 to 100 amino acids. Final recombinants scoring in the top 50 percentile for B-L energy and the top 50 percentile for radius of gyration were rejected, and the remainder clustered based on RMSD of CA-CA distance.

In most cases a range of sizes of recombinants with different endpoints were generated and inspected visually by both of us. Patterns in turn direction and super-secondary structural features common to the cluster centers of different sets of recombinants were noted in an attempt to infer the preferred directions of individual turns and overall topology of super-secondary structures. (Since only 1 or 2 computer processors were used for fragment recombination, very limited conformational sampling was achieved, putting this approach at a great disadvantage for constructing a single full length recombinant molecule that was free of steric overlap and that displayed all of the turn preferences obvious in the smaller recombinants. Extensive human intervention was used in this and the next step as a rational alternative to greater computer power.) After evaluation of observed patterns of turn directions, a decision was made on the global topology implied by the ensemble of recombinant fragments.

Manual Refinement

The longest recombinant displaying the inferred global topology was identified and refined manually. If necessary, one or more non-overlapping fragments from other parts of the protein were added to the coordinate file and moved into position. Single residues were deleted from the center of turn regions, enabling individual pieces of the model to be translated and rotated to eliminate the most serious steric clashes and to align segments of secondary structure in ways that are "protein-like". All manipulations were carried in the graphics package of HYPERCHEM v5.1, using an in-house program to monitor changes in steric overlap, radius of gyration, and B-L energy as the model was being adjusted.

When the manually reworked model displayed the inferred topology and appeared reasonably "protein-like", the turns containing missing residues were repaired using the lego_loop function of the O software package of Alwyn Jones.[8] In a small number of cases, more than one topology remained as distinct possibilities after viewing recombinants, and so a second (and in one case a third) model was refined in this way. A total of twenty-nine models for 20 different target proteins were submitted. Of the 13 domains assessed in the new fold category, models were submitted for 12 (Fig. 3); the remaining 8 were fold recognition targets that we were unable to recognize with PSI-BLAST.[9]

DISCUSSION/CONCLUSIONS

The results reported here reflect our first implementation of a strategy based predominantly on scoring functions that estimate the energy of side-chain/backbone interactions. While the results were less successful than we had expected, we interpret the limited success of some of our models as an indication that the emphasis on local interactions instead of long range contacts has significant promise, much of which remains to be explored.

During the step of recovering turn-containing fragments for recombination, it became quite obvious that the PDB represents a very limited sample of turn conformations. In many cases in which turn regions included more than 6 to 8 amino acids, no fragment was recovered with a negative p137 score, indicating that the fragments actually used for building some of our models had little chance of being correct. In almost no cases could the much more information-rich p137rot scoring function be used, again for the simple reason that no fragments with scores below zero could be identified. In view of this finding, the one modification of our protocol most likely to lead to future improvements is the development of *de novo* turn generation methods that yield a greater range of phi/psi/chi1 angles, allowing more of the information in p137 and p137rot to be used in specifying details of turn directions.

ACKNOWLEDGMENTS

We thank Michael Ackerman for help with lego_loop and Satoshi Ohnishi for encouragement and helpful discussions.

REFERENCES

1. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;Suppl 5: 119-126.
2. Shortle D, Ackerman MS. Persistence of native-like topology in a denatured protein in 8 M urea. Science 2001;293: 487-489.
3. Ackerman MS, Shortle D. Robustness of the long-range structure in denatured staphylococcal nuclease to changes in sequence. Biochemistry 2002;41: 13791-13797.
4. Shortle D. Composites of local structural propensities: Evidence for local encoding of long range structure. Prot Science 2002;11: 18-26.
5. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16: 92-112.
6. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics, submitted in 2002.
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292: 195-202.
8. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47: 110-119.
9. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-3402.