

The Threading Approach to The Inverse Protein Folding Problem

Temple F. Smith*[†] Loredana Lo Conte* & Jadwiga Bienkowska* Bob Rogers*
Chrysanthe Gaitatzes* Richard Lathrop[‡]

*BioMolecular Engineering Research Center, College of Engineering,
Boston University, 36 Cummington Street, Boston, MA 02215.

[†]Recomb97 presenting author. & Department of Computer Science, College of Arts and Sciences,
E-mail: tsmith@darwin.bu.edu. Boston University, Boston, MA 02215.

[‡]Department of Information and Computer Science,
University of California, Irvine, CA 92717.

1 Introduction

The logic behind the threading approach to the prediction of an amino acid sequence's expected three-dimensional fold is almost seductively simple, if not obvious. Given the extreme difficulty of any direct, *de novo*, quantum level approach and our aesthetic sense of science as expressed in Ockam's principle of parsimony, this seduction is understandable. In practice, however, the threading approach has not yet lived up to our expectations (Lemer et al. 1995). Before asking why, let's review what is generally meant by this approach to protein fold prediction.

Threading rests on two basic ideas: first, that there is a limited and rather small number of basic "protein domain" core folds or architectures that need to be modeled; and, second, that a sum over a sequence's amino acid structural environment preferences is a sufficient indicator for the recognition of native-like folds. The first of these ideas arises from our understanding of polymer chemistry, which suggests that there is only a limited number of ways to fold a repetitive polymer of two basic unit types (hydrophilic and hydrophobic), and from the observation that the vast majority of determined structures fall into only a few core architectures. The latter was perhaps first clearly stated by Jane Richardson in her "taxonomy of proteins" (Richardson 1981). The second basic threading idea is also rooted in observation as well as theory. First, protein structures and even their functions are known to be very robust to amino acid substitutions, an obvious requirement of any successfully evolving system. As has been noted many times, the many hundreds of distinct globin amino acid sequences fold to nearly identical structures, thus supporting the idea that some average over a sequence's structural environmental preferences is as important as the atomic level details. Based on these two combined ideas, it has seemed reasonable, given a library of all expected basic fold architectures and an understanding of which average properties of amino acids determine their structural environment preferences,

that we should be able to recognize which basic fold is preferred by which sequence. We should also recognize that this simply involves identifying the fold that places the amino acids in their overall most favored structural environments. (See Figure 1 for a schematic representation of the threading problem.)

The complexity of finding the optimal sequence to structure threading is a function of whether arbitrary length gaps are allowed in the alignment of the sequence to the model fold, and whether or not the score function includes pairwise or higher order amino acid structural environments (Lathrop 1994; Lathrop and Smith 1996). In the general case where both arbitrary alignment gaps and pairwise interactions are allowed the threading problem has been shown to be NP-complete (Lathrop 1994). Our current knowledge of protein structure and evolution supports the need to allow alignment gaps, since single residues or entire independent domains have been observed to have been inserted into the surface loops of many proteins. Second, inclusion of at least pairwise interactions in the definition of amino acid structural environmental preferences is based on the assumption that, at the very least, the compact nature of proteins requires significant contact interactions between spatial neighboring amino acids. In addition, it has long been assumed that the fundamental atomic level pairwise interactions between amino acids help determine the final fold, and that these can be reasonably approximated at the side chain-side chain level.

2 Current Methodologies

There are four components to any practical application of the threading approach to the prediction of the three dimensional fold for an amino acid sequence:

- 1) Construction of a complete as possible library of modeled folds.
- 2) A scoring schema to evaluate any particular threading or alignment of a sequence into such a modeled fold.
- 3) Some means of searching over the vast space of possible alignments between each sequence and each fold for the one with optimal score. (See Figure 1.)
- 4) A means of choosing the best model for a sequence given the optimal scoring alignments of that sequence to all modeled folds.

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.

RECOMB 97, Santa Fe New Mexico USA
Copyright 1997 ACM 0-89791-882-8/97/01 ..\$3.50

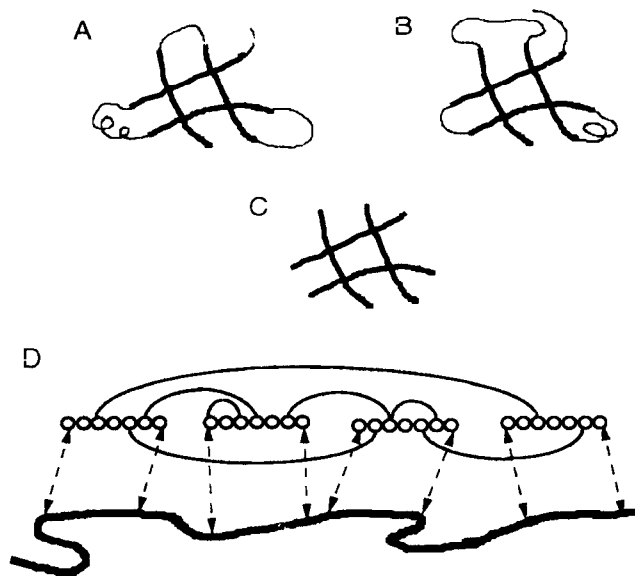


Figure 1: An illustration of the gapped protein threading methodology (Bryant and Lawrence, 1993, Greer, 1990, Jones et al., 1992) used in this work. (A and B) Conceptual drawing of two structurally similar proteins and a common core of four secondary structure segments (dark lines, I-L). To form the structural models used here, side-chains are replaced by a methyl group and loops are removed. (C) Abstract structural model showing spatial adjacencies (interactions). Small circles represent amino acid positions (core elements), and thin lines connect neighbors in the folded core. The structural environments and spatially neighboring positions will be recorded for later use by the score function. (D) One possible threading with a novel sequence. A sequence is threaded through the model by placing successive sequence amino acid residues into adjacent core elements, indexes the sequence residue placed into the first element of segment X. Sequence regions between core segments become connecting turns or loops. (Lathrop and Smith, 1996)

The first three of these four components have proposed solutions. There are a number proposed for the first and second (Greer 1990, Jones et al. 1992, Bryant and Lawrence 1993, White et al. 1994, Stultz et al. 1996) and the third has recently received a practical algorithmic solution (Lathrop & Smith 1996). However, while the amino acid structural environment scoring schema has the most proposed solutions, they are the least satisfying solutions. Finally, a statistically rigorous solution to the fourth is still problematic.

The branch-and-bound search algorithm described by Lathrop & Smith (1996), which solves the third subproblem above, accepts as input the sequence, a near arbitrary scoring schema, and compatible fold models. The global optimal score and the entire pseudo-energy or score function landscape are completely determined by that input, although unknown in general. The task of the algorithm is to identify the global score and the globally optimal alignment or threading that instantiate it. The branch-and-bound algorithm described is unusual in that, in any given case, it either finds the mathematically exact answer or it exhausts time or space resources. Importantly it never returns an approximate or inexact result. Different landscapes affect the time required by such an algorithm (and hence, whether or not it converges within a specific time limit), but they cannot change the fact that the algorithm always finds the mathematically exact global score or fails to converge. It is easy to see that the landscape topology is a sensitive function of the scoring schema. Note that unless there is considerable differentiation between the different amino acids' preferences for different structural environments (including preferences for neighboring amino acids), the landscape will be relatively flat and a branch-and-bound search would be expected to converge only slowly.

The general form of various scoring schemata proposed thus far consists of sets of pseudo energies associated with the placement of each amino acid type into each defined structural environment. These energies are generally derived from Gibbs' relationship between energy and probability by estimating from known structures the probability of observing each amino acid type in each modeled structural environment. The structural environments proposed have included the degree of solvent exposure, the type of secondary structure, and various measures of "distance" between neighbors. The simplest way of including the latter neighborliness has been to count the frequency at which the different types of amino acids are observed as neighbors in a set of determined native structures and then convert them to pseudo Gibbsian energies or log likelihood ratios. As discussed below, we believe that these simple definitions of neighbor and/or pairwise environment is a likely source of some of the limitations of the current threading approaches.

A set of fold models and a scoring schema capable of recognizing the correct structure should be unbiased toward the particulars of the native sequences. This is obviously important if one is testing scoring schema by the recognition of native sequences by their native-derived modeled structures. The simplest side chain independent definition of neighbors depends only on the physical distance between the alpha or beta carbon pairs. Even this simple definition is not obviously related to the physical forces expected to directly affect protein folding. Note, two amino acids on opposite sides of a beta sheet might satisfy the distance conditions, but fail to make any physical contact. More to the point is the fact that two close amino acids, even on the same side of a beta sheet or alpha helix, may make little or no physical energetic contact due to the positioning of their side chain

rotamers. Thus any simple spatial definition of neighborliness used in calculating pseudo energies directly from frequencies will be a mixture of non interacting or "chance" neighbors and truly interacting neighbors.

The recent structure determination of a number of beta propeller proteins has provided an excellent test for the comparison of conserved physical contacting neighbors versus "chance" neighbors. In particular, the recent determination of the G-protein's beta subunit structure (Wall et al. 1995, Lambright et al. 1996) has proved very informative. This protein is a member of the so-called WD-repeat family of functionally diverse proteins (Neer et al. 1996). Here there are seven sequence variable repeats correlating one to one with the seven small beta sheet structural "blades". All propeller blades in this very symmetric structure are nearly identical at the peptide backbone level. By comparing equivalent amino acid position contacts in each of the seven blades two facts are seen: first, less than half of the neighboring position pairs make energetic contacts; and, second, only about half of those contacts are made in the majority of the seven blades (see Figure 2). The latter is particularly interesting since the amino acid types involved in the apparently conserved contacts are often not the same. They may be the key structural determinants since many of the equivalent contacts are found to exist as interacting amino acids in totally unrelated non WD-repeat beta propeller proteins. If these interpretations are correct, there are important implications for any threading approach to fold prediction.

3 New directions

Our recent experiences and those of others (Lemer et al. 1995; Lathrop & Smith 1996) on current threading model environment and scoring function definitions, in combination with some successful sequence pattern alphabets (Adams et al. 1996; Henikoff & Henikoff 1993; Smith & Smith 1990), suggested the need for much better modeled structure characterizations. Secondary structure, the degree of solvent exposure, and the various types of pairwise "neighborliness" proposed, clearly reflect important structural characteristics expected to be common between distant homologs. However, either they are not very robust over the observed range of homologue variation, or we have not been able yet to exploit them fully in the design of sequence-to-structure alignment scoring schemes. We have therefore begun to investigate a new set of structural environment descriptors based on two native protein properties. The first is the ideal tetrahedral geometry of the beta carbon reflected in the various libraries of side chain gamma carbon rotamers. The second is the close packing of internal side chains. The latter results in few internal empty spaces or internal water molecules in most proteins. This leads to two contradictory ideas: one, many neighboring contacts are purely incidental and are not fold discriminating; two, all neighbors significantly contribute to the compatibility of any amino acid to its local environment. Nearly all past threading work has assumed the second idea.

A number of environmental variables can be defined relative to this beta carbon tetrahedral geometry. One needs only to place one's self at the beta carbon, and ask what can be "seen" through each of the tetrahedral windows. These windows are formed by each face of a tetrahedron constructed by taking the normal to the three ideal beta-gamma carbon vectors (the bottom face which is normal to the alpha-beta vector is not considered). These can be

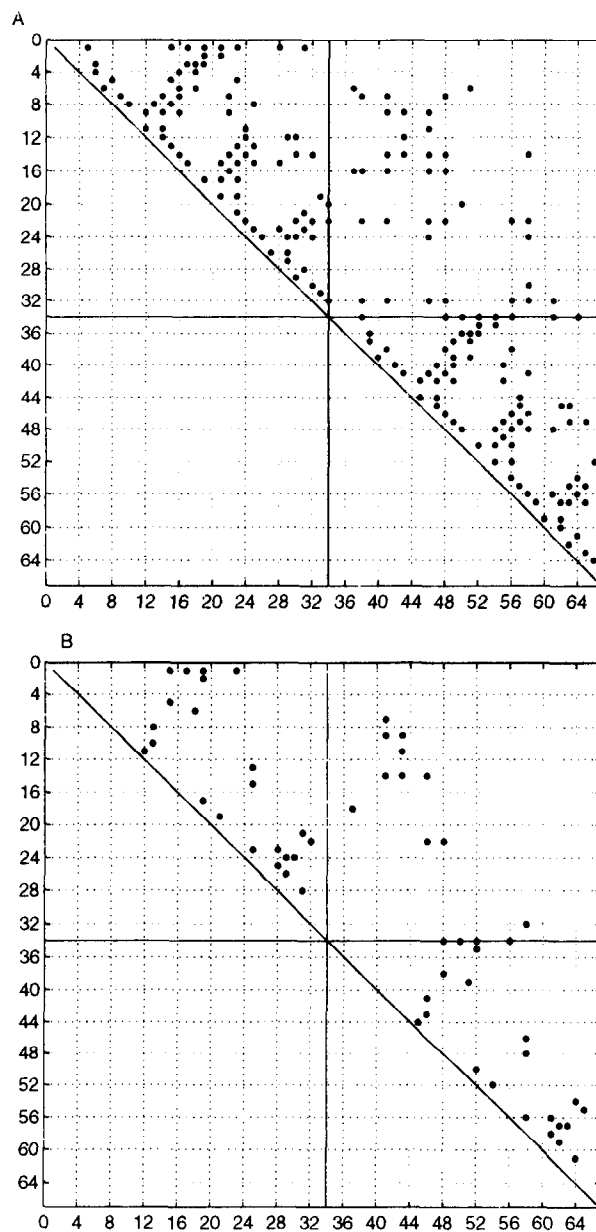


Figure 2: Plot of inter and intra G-protein beta subunit contacts. (A) Shows line of sight beta carbon pairs common to five out of seven blades. (B) Shows energetic amino acid side chain contacts common to five out of seven blades.

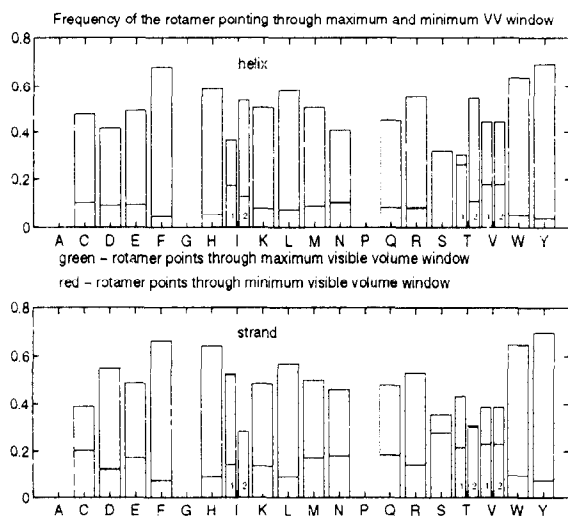


Figure 3: The observed frequencies of the rotamer state (tetrahedral window through which the side chain points outward) being the same as the maximum available visible volume window (green bars) and minimum available visible volume window (red bars). For isoleucine, threonine and valine, which have two distinct gamma branched atoms, we have indicated two rotamer states. For isoleucine first rotamer state corresponds to the direction of the vector between atoms CB-CG1 (which indicates the direction of the long arm of isoleucine). For threonine first rotamer corresponds to the direction of the vector CB-OG1. For valine the maximal or minimal visible volume windows for both CG1 and CG2 are taken into account and the frequency is calculated as the average value.

simply defined by the geometry of the backbone's atom coordinates of a model or real structure, independent of any particular set of observed side chains (including Glycine and Proline). Clearly in a surface position one will be able to see the solvent out of at least one of these windows. Depending on the local as well as the neighboring secondary structure, one will see different beta carbon and/or backbone atoms at different distances through each of the tetrahedral windows. Among the various possible "line-of-sight" tetrahedral characterization of any modeled position is a vector of "visible volume". This is defined as the volume visible about each beta carbon that is not occluded by any other beta carbon or backbone atom (Lo Conte and Smith 1997). This three components vector (the lower window is totally occluded by the backbone) is expected to correlate with solvent exposure, large side chain rotamer preferences and various pairwise packing preferences. The latter two have the potential to help distinguish between significant and incidental pairwise neighbors. See for example, in Figure 3 the strong correlations observed in nearly two hundred non homologous proteins, between the maximal visible volume component and the window containing the observed gamma carbon rotamer.

Current studies of the potential for characterizing the essential conserved features of any family of structures by a set of visible volume vectors have been quite suggestive. For example it has long been recognized that one of the key characteristics of any globular structure is the correlation

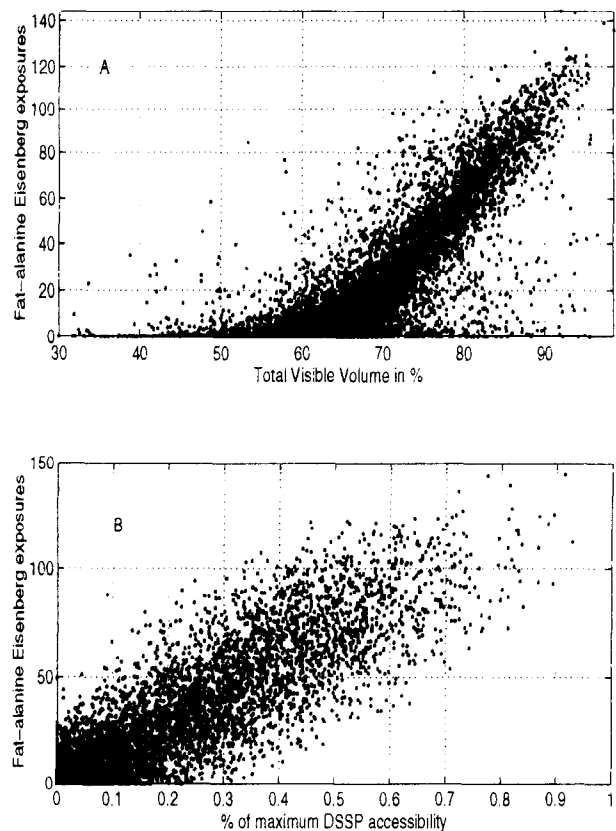


Figure 4: (A) A plot of the exposure about each protein core modeled positions as defined by the total visible volume versus the solvent accessibility of that position when occupied by a beta carbon only. Here the beta carbon or equivalently, Fat Alanine, effective radius has been set to 3.5 Å. (B) A plot of the above "Fat Alanine" versus the percentage of native side chain exposure as calculated by DSSP.

between amino acid hydrophobicity and its fold position exposure. As seen in Figure 4, the total visible volume clearly contains this information.

In fact the visible volume, which contains no direct specific side chain information, appears as sensitive in detecting the exposure/buried periodicity of an alpha helix as the side chain hydrophobicity or side chain exposure (Lo Conte and Smith 1997). In addition, the visible volume vector appears capable of discriminating between extended or helical local backbone environments. This arises directly from the fact that the alpha beta vector has a very different angle relative to the overall helical versus beta sheet surfaces (Lo Conte and Smith 1997). In combination with the number of beta carbons "visible" through each tetrahedral window, this visible volume vector has the potential to encode allowed side chain packings and thus provide a better method of scoring potential pairwise contacts in sequence to structure alignments.

We have been able to identify one of the probable factors limiting threading predictability using the above defined beta carbon tetrahedral window line-of-sight views. We have defined neighboring positions in our threading fold models as any two beta carbons on direct line of sight at less than 8.5 Å. It appears that the majority of these line

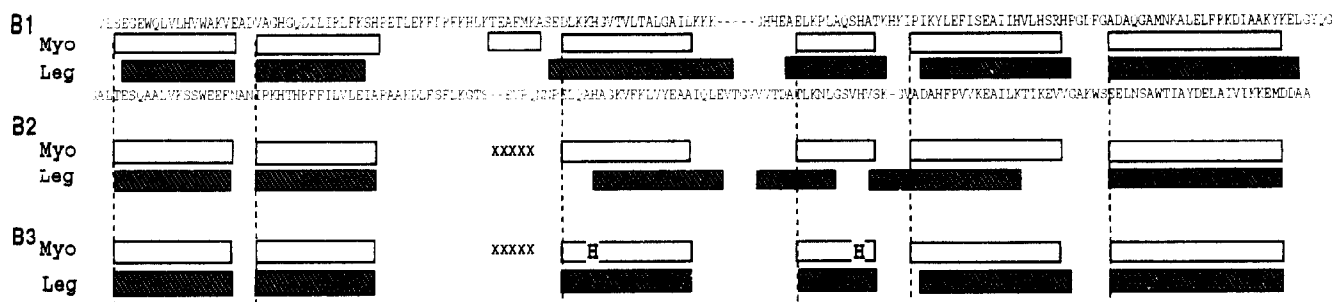


Figure 5: Leg-hemoglobin sequence threaded through the myoglobin structural model. (B1) Reference alignment. (B2) Short "D" helix removed, no active site constraints. (B3) Short "D" helix removed, requiring model elements at positions E7 and F8 to be occupied by histidine, H. from the sequence. (Lathrop and Smith 1996)

of sight neighbors are equivalent to those that contribute to the pairwise scores in most other current threading approaches (data not shown). Yet, one would assume that for all amino acids with no gamma branched atoms only the neighbors seen through the rotamer occupied window would make major contacts. Two neighboring amino acids having rotamers pointing in opposite direction surely make insignificant contacts. Thus, it is nearly obvious that two amino acids need not be in structurally meaningful contact just because two corresponding fold positions are close neighbors. Most existing pairwise scoring schemata might therefore be expected to result in about two parts noise and one part signal. This is in agreement with our initial study of the fraction of amino acid pairs making significant energetic contact in the G-protein beta subunit. This in turn suggests that if in any particular threading such pairwise score function noise happens to average out, the prediction may be very accurate; if not, it may be totally off the mark.

It would appear that one needs to include rotamer preferences and to associate them when possible to the various tetrahedral line-of-sight properties. Potential neighbors or pairwise contacts contribute to the scoring of a given threading depending on the particular amino acid type aligned to each tetrahedral line-of-sight characterized pair of positions, and this in turn is expected to complicate the current bounds calculations used in the Lathrop and Smith (1996) Branch and Bound optimal threading algorithm. In Figure 1, this means that different subsets of arcs or potential pairwise contacts in the modeled fold contribute to the alternate threading scores. However, the pairwise score may still be only a function of the amino acid pairs aligned to any two potentially interacting positions.

Finally, we would note additional information has already been shown to greatly increase the potential accuracy of a threading alignment (Lathrop and Smith 1996). This is done by restricting all allowed threadings to place a particular type of amino acid within a very restricted part of the modeled fold architecture. Figure 5 shows, for example, the result of requiring the conserved Histidines in the threading of Leg hemoglobin into a myoglobin derived model fold, the alignment of which is rather poor otherwise. This figure shows that: the accuracy of all the modeled positions is greatly improved by the included pattern; and of course, the specificity of such a minimal pattern is increased by the context of the modeled structural fold. In particular, this means a reduction in false positives with little or no loss in true positives.

While the inclusion of a require amino acid type in a given modeled position is very different from including a vec-

tor of visible volume and/or numbers of visible beta carbons about a modeled position, we conclude that with the inclusion of such additional structural information the threading approach may yet live up to our expectations.

Acknowledgments

This work was supported in part by grant #P41 LM05205-13 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the granting agency.

REFERENCES

- Adams, R. Mark, Das, Sudeshna and Smith, Temple F. (1996). Multiple Domain Protein Diagnostic Patterns. *Protein Science* 5, 1240-1249.
- Bryant, S.H. and Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Func. Genet.* 16, 92-112.
- Greer, J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins: Struct. Func. Genet.* 7, 317-333.
- Henikoff, S. and Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* 17, 49-61.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* 358, 86-89.
- Lambright, D.G., Sondek, J., Bohm, A., Skiba, N.P., Hamm, H.E. and Sigler, P.B. (1996). The 2.0 Å crystal structure of a heterotrimeric G protein. *Nature* 379, 311-319.
- Lathrop, Richard H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* 7, 1059-1068.
- Lathrop, Richard H. and Smith, Temple F. (1996). Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* 255, 641-665.
- Lemer, Christian, M.-R., Rooman, Marianne J. and Wodak, Shoshana J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Func. Genet.* 23 (3), 337-355.
- Lesk, A.M and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136, 225-270.
- Lo Conte, Loredana and Smith, Temple F. (1997). Visible Volume: A robust measure for protein structure characterization. Manuscript in preparation.

Neer, Eva J., Schmidt, Carl J., Nambudripad, Raman and Smith, Temple F. (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature* 371, 297-300.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structures. *Advan. Protein Chem.* 34, 157-339.

Smith, Randall F. and Smith, Temple F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, 87, 118-122.

Stultz, C.M., Nambudripad, R., Lathrop, R.H., and White, J.V. (1996). "Predicting Protein Structure with Probabilistic Models." In: *Protein Structural Biology in Bio-Medical Research* (N. Allewell and C. Woodward, editors). JAI Press, Greenwich (in press).

Wall, M.A., Coleman, D.E., Lee, E., Iniguez-Lluhi, J.A., Posner, B.A., Gilman, A.G. and Sprang, S.R. (1995). The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2. *Cell* 83, 1047-1058.

White, James V., Muchnik, Ilya and Smith, Temple F. (1994). Modeling protein cores with Markov random fields. *Math. Biosci.* 124, 149-179.