Crystallographic Threading

A. Ableson[1] and J.I. Glasgow[2]
[1]Department of Mathematics
[2]Department of Computing and Information Science
Queen's University, Kingston,
Canada, K7L 3N6
{ableson,janice}@cs.queensu.ca

Abstract

Crystallographic studies play a major role in current efforts towards protein structure determination. Despite recent advances in computational tools for molecular modeling and graphics, the construction of a three-dimensional protein backbone model from crystallographic data remains complex and time-consuming. This paper describes a unique contribution to an automated approach to protein model construction and evaluation, where a model is represented as an annotated trace (or partial trace) of a structure. Candidate models are derived through a topological analysis of the electron density map of a protein. Using sequence alignment techniques, we determine an optimal threading of the known sequence onto the candidate protein structure models. In this threading, connected nodes on the model are associated with adjacent amino acids in the sequence and a fitness score is assigned based on features extracted from the electron density map for the protein. Experimental results demonstrate that crystallographic threading provides an effective means for evaluating the "goodness" of experimentally derived protein models.

Introduction

A fundamental goal of research in molecular biology is to understand the tertiary structure of protein molecules. Protein crystallography is currently at the forefront of methods for determining the three-dimensional conformation of a protein, yet it remains labor intensive in the construction, evaluation and refinement of candidate models for the structure. A protein model represents a hypothesis about the backbone structure of a protein; a good model is one which makes sense, in terms of our knowledge of the chemistry, biology and physics of the molecule, and is consistent with the experimental data and the known protein sequence. Building a protein model is currently a trial-and-error process, in which a crystallographer is assisted by the use of computer graphics tools that trace polypeptide chains and model side chains, and allow them to view and improve the resulting model (Jones *et al.* 1991). Errors in the initial and subsequent models may be corrected with a refinement process that modifies a model to minimize the difference between the experimentally observed data and the data calculated using a hypothetical crystal containing the model. It has been proposed that the process of protein model building could be improved through the development of computational tools (Branden & Jones 1990).

This paper describes a novel computational approach, called *crystallographic threading*, that associates a primary sequence with candidate structure models of the protein backbone. The models are derived through a topological analysis of an electron density map and are represented as three-dimensional traces through a graph consisting of nodes, corresponding to amino acid fragments, and edges, corresponding to polypeptide and side-chain bonds. The threading algorithm assigns a likelihood or score to a given sequence/model alignment, allowing us to determine whether a model is in fact a "good" candidate structure for the protein. Such models can be used in an refinement process to recover important phase information.

Previous and current research on threading has focused on the problem of protein structure prediction and generally rely on the assumption that there is a a one-to-one mapping between the sequence and structure elements. We describe a gapped approach to threading that incorporates techniques from sequence alignment to determine the optimal association between an experimentally derived protein model and the amino acid sequence it is purported to represent.

Model Derivation

The primary task of an image analysis system in artificial intelligence is to derive an underlying scene model from given image data. The research described in this paper contributes to a computational approach to *molecular scene analysis* (Fortier *et al.* 1993; Glasgow, Fortier, & Allen 1993), which refers to the processes associated with the reconstruction, classification and understanding of complex molecular structures. A key process of a molecular scene analysis is the automated determination of potential scene models for a protein structure. The input for this process is an electron density map (the protein image) and a primary sequence of amino acid residues (the image components). The interpretation of the electron density map involves finding the three-dimensional polypeptide chain, associating it with the given amino acid sequence and configuring the respective side-chains. These processes are complicated by noise-based errors and by the lack of accurate phase information.[1] The quality of the electron density map also depends on the resolution of the diffraction data, which is influenced by how well-ordered the crystal is. From the analysis of a map at low resolution (> 5 Å), one can retrieve general shape information and possibly identify regions of secondary structure. At medium resolution ($\sim$ 3Å), it is generally possible to trace the backbone of the protein and derive properties of individual residues along the backbone. Only at high resolution (≤ 1 Å) are the individual atoms in the protein observable. Experiments described in this paper incorporated electron density maps at medium resolution, where components of the protein backbone are potentially identifiable.

An approach to molecular scene analysis has previously been applied to generate scene models through a topological analysis of the electron density map (Fortier *et al.* 1993; Glasgow, Fortier, & Allen 1993). Such an analysis consists of two stages. The first stage involves determining a set of *critical points* (points where the gradient of the electron density function is zero). Of particular interest to this paper are the *peak* critical points, which correspond to local maxima of electron density, and *pass* critical points, which correspond to saddle points in the map. Experimental results at medium resolution (Leherte *et al.* 1994) suggest that peaks are useful in identifying the location of amino acids and side chains in the unit cell for the protein crystal; passes generally correspond to peptide or side-chain bonds. The second stage of the topological analysis constructs a graph that links the peak critical points. This work can be related to BONES (Jones *et al.* 1991), a graphical method which has been developed and applied to the interpretation of medium- to high-resolution protein maps. This method incorporates a thinning algorithm and postprocessing analysis for electron density maps. An important difference is that BONES does not segment a graph into individual parts that can be related to amino acid residues, and thus does not lend itself to crystallographic threading.

A *protein model* is constructed as a linear trace of the critical point graph corresponding to a possible backbone (or connected portion of the backbone) structure for the protein. This graph, in general, is an imperfect representation of the protein. Experimental results show that some critical points result from noise in the data, series termination or side-chain density. In addition, large residues may be represented as two or three critical points in the graph (Leherte *et al.* 1997). As discussed in the following section, these errors increase the complexity of threading the known sequence onto candidate models for the protein.

The electron density map and the critical point graph contain other valuable sources of information useful in associating peaks in a trace with individual amino acids. We characterize the properties of a critical point in a trace in terms of a *critical point environment.* Attributes that have been considered for an environment include: maximum peak height, distance to solvent, participation in secondary structure, size of associated side chain, volume, and mass. Thus, a protein model is defined to be a linear trace of the critical point graph where individual nodes (critical points) in the model are annotated with feature characteristics defined by their critical point environment.

Protein Threading

Protein threading is generally associated with the problem of *inverse protein folding*, which asks the question: given a protein structure, what sequences would adopt this structure? Inverse folding research typically involves two main steps performed in a biochemical laboratory or in a computer simulation. First, an ensemble of amino acid sequences is generated from a template representing certain structural or statistical patterns. Second, each of the generated sequences is tested for its propensity to assume a given three-dimensional conformation. In this threading procedure, a protein structure is taken as a starting point, and the sequence is arranged in three-dimensional space by aligning it to the structure template or model. In a *gapped alignment* approach to threading (Bryant & Lawrence 1993; Lathrop & Smith 1996; White, Muchnik, & Smith 1994), loops are removed, resulting in a *core model* of relatively inflexible secondary structure segments. Structural environments, connectivity, and spatial adjacency are all recorded on the core

[1] This is referred to as the classic *phase problem* (Karle 1986) of crystallography.

model. Threading then seeks to align the amino acids in the sequence with the amino acid locations in the model, while maintaining legal loop sizes and not breaking the chain or violating excluded atomic volumes. An *objective function* evaluates each possible sequence/structure alignment in terms of the extent to which the amino acids from the sequence are located in preferred environments. Thus, it can be considered as a score that reflects the "goodness of fit" of a particular sequence in a particular alignment to a particular structure. The main computational work of a threading procedure is therefore the constrained search for a sequence/structure alignment that optimizes the objective function.

We define *crystallographic threading* as the process of aligning amino acid sequences with structure models derived from the analysis of an electron density map of a protein. Models that have a strong sequence/model alignment are serious candidates for being part of the correct protein structure. A goal of our research is to demonstrate that the local characteristics of peaks in the critical point graph, i.e., the critical point environments, contain information that allow us to define an effective objective function for threading. More specifically, the local properties of a peak, resulting from the types and locations of atoms contributing to the peak's electron density, should give an indication of what atoms are near the peak and thus what amino acids are most likely to be associated with the peak.

Although a single environment is not sufficient to uniquely determine what amino acid is associated with a given peak, we can formulate a probabilistic mapping from a peak to a subset of possible amino acids. Once we have even a short sequence of such mappings we greatly restrict the number of possible alignments between the structure and the sequence. Models that do not correspond to a correct subtrace of the protein backbone should be unalignable, i.e., the fitness function should return a low value for the alignment of the model with all possible subsequences of the primary structure. The probabilities for the mapping are estimated using a conditional classifier applied to a database containing peaks, their environments and their previously identified amino acid class. Given these derived probabilities, it is then straightforward to express crystallographic threading as a *local sequence alignment* exercise using a standard pair-wise scoring function.

Figure 1 illustrates a simple example of a conditional classification of peaks in a model, as well as the most likely assignment of the model to a given sequence. The alignment of the amino acid classes BCA with the three peaks returns the highest probability for any subsequence in the given sequence (the probability of BAA is higher but is not a valid subsequence).

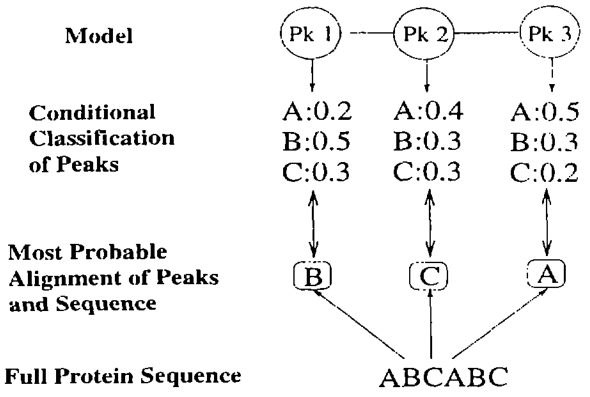

Figure 1: Example of an optimal model/sequence alignment.

As stated earlier, a goal of our research to demonstrate that crystallographic threading can be treated using classic sequence alignment techniques. Most local alignments are scored based on a log-odds score function comparing the alignment probabilities in optimal alignments to a random alignment model(Altschul 1991):

$$M_{i,j} = -\frac{1}{\lambda} log\left(\frac{q_{ij}}{p_i p_j}\right). \qquad (1)$$

We can model the probability densities of peak characteristics to build a function that plays exactly the same role as the probability q_{ij} in the alignment score function for residues. The probabilities p_i and p_j represent the natural probabilities of occurrence of a given peak and amino acid, respectively. The probability of a given amino acid is determined by a frequentist approach; the probability of a given peak is approximated from the distribution of all peak characteristics. The net result is that we are able to construct a probabilistic scoring function for the alignment of a given amino acid to a given peak in the model. We then determine how well a model matches a given subsequence of amino acids by applying a local alignment algorithm. An interesting difference between our approach and standard alignment algorithms is that we do not require a substitution matrix setting scores for the mapping between amino acids; instead we use a more general probability model.

Due to errors in the critical point graph, resulting from noisy and incomplete data, the alignment of peaks to amino acids may not be one-to-one. Experimental results indicate that a single peak may correspond to two amino acids (missing peak), or that two peaks may relate to a single element of the sequence (extra peak) (Leherte *et al.* 1997). Figure 2 illustrates some possible alignments based on no gaps, gaps in the primary sequence (one peak/two amino acids) or gaps in the model (two peaks/one amino acid). Fortunately, allowing gaps in the sequence or model requires a simple modification of the alignment dynamic program if there is a fixed penalty for a gap (Waterman 1995).

Another difficulty in crystallographic threading is that critical point environment information along the backbone may not be sufficient for characterizing individual peaks in the model. Since the backbone atoms are the most homogeneous groupings of atoms in a protein, their peak characteristics are also homogeneous and thus different amino acid peaks may appear with similar environments in the model. Although this is a problem with individual amino acid/peak alignments, our studies show that as the model grows the number of possible sequence/model alignments diminishes. This issue is also addressed by considering the critical points (and their environments) that may correspond to side chains for the model. Since side chains may differ significantly in their chemical constituents, their environments are generally more discerning that those of the backbone critical points. Each peak in a model generally has at most one or two neighboring peaks that could correspond to a side chain, thus considering side-chain environments does not add greatly to the complexity of our algorithm.

Methodology and Results

The hypothesis of our research is that crystallographic threading can be applied to determine whether a given model is a likely candidate structure for the protein. To test our hypothesis we implemented a common technique for local sequence alignment, where the objective function (score) for an alignment is calculated as the sum of scores for matching each pair of aligned residues (Waterman 1995). An optimal alignment is found by searching over the space of all possible alignments, allowing for single gaps in both the sequence and the model. This section describes scoring and fitness functions for crystallographic threading and reports on the results of applying an implementation of these to the threading of protein sequences onto computer generated models.

Scoring Function

Our fitness function incorporates a log-odds scoring function Q (Altschul 1991). This standard scoring formula normally involves comparing two similar objects, amino acid or nucleic acid types. For the problem of crystallographic threading, however, function Q determines a score for associating a peak critical point i with an amino acid class C:

$$Q(i, C) = log\frac{P(\overline{x_i}, C)}{P(\overline{x_i})P(C)} \;=\; log\frac{P(\overline{x_i}|C)}{P(\overline{x_i})},$$

where $\overline{x_i}$ is a vector that denotes the critical point environment characteristics for peak i. $P(\overline{x_i}, C)$ is the probability that an amino acid class C occurs with characteristics $\overline{x_i}$. Thus, the function application $Q(i, C)$ computes the log of the relative probability that the characteristics $\overline{x_i}$ for peak i are a result of peak i being a member of class C.

Three peak characteristics are used to define critical point environments, $\overline{x_i}$, in the scoring function: the peak density, the difference between the peak density and its highest pass density, and the *log* of the volume associated with the peak. These attributes were chosen based on their ease of computation and on their ability to discern among amino acids (Sunderji 1996). To model the probability distribution of the peak environments, we need a continuous probability density model. We chose to model the distribution of peak characteristics using kernel functions (Silverman 1986) because of their general robustness. The probability functions were built based on a database of known peaks and their environments. A kernel smoothing parameter was determined separately for each class by minimizing a least-squares error approximation. The peaks in the critical point graph were broken into 41 possible classes based on the amino acids they represented: 20 ***backbone classes*** (B_j denotes the backbone class for residue j), 20 ***side-chain*** classes (S_j denotes the side-chain class for residue j), and a ***non-protein*** (denoted NP) class.

We improved our scoring function by considering side-chain critical points in the critical point graph. Figure 3 presents a simple critical point graph and a model derived as a linear trace (through the sequence of peaks 1, 3, 5, 8) of the graph. Since there are no potential side-chain peaks connected to peak 1, the alignment score for peak 1 and an amino acid j is calculated as before as: $Q(1, B_j)$. There are two side-chain peaks (peak 2 and peak 4) that could be associated with the backbone peak 3. Under a maximal support philosophy, the score for the alignment of peak 3 with amino acid j is calculated as:

$$MAX\left[Q(3, B_j), Q(2, S_j), Q(4, S_j)\right].$$

Informally, this states that peak 3 should be a given a high alignment score if can be associated with amino acid j as a backbone peak, or if either peak 2 or 4 can be associated with j as a side-chain peak. In general, we define the score function for a residue j and a peak i in the model as:

$$score(i, j) = MAX\left[Q(i, B_j), Q(k, S_j)\right]$$

where k ranges over all possible side-chain peaks connected to peak i.[2]

[2] A peak k is considered a potential side-chain peak for backbone peak i in the model if and only if there is an edge connecting i and k in the critical point graph, but not in the model. The dotted lines in Figure 3 illustrate the possible connections to side-chain peaks.

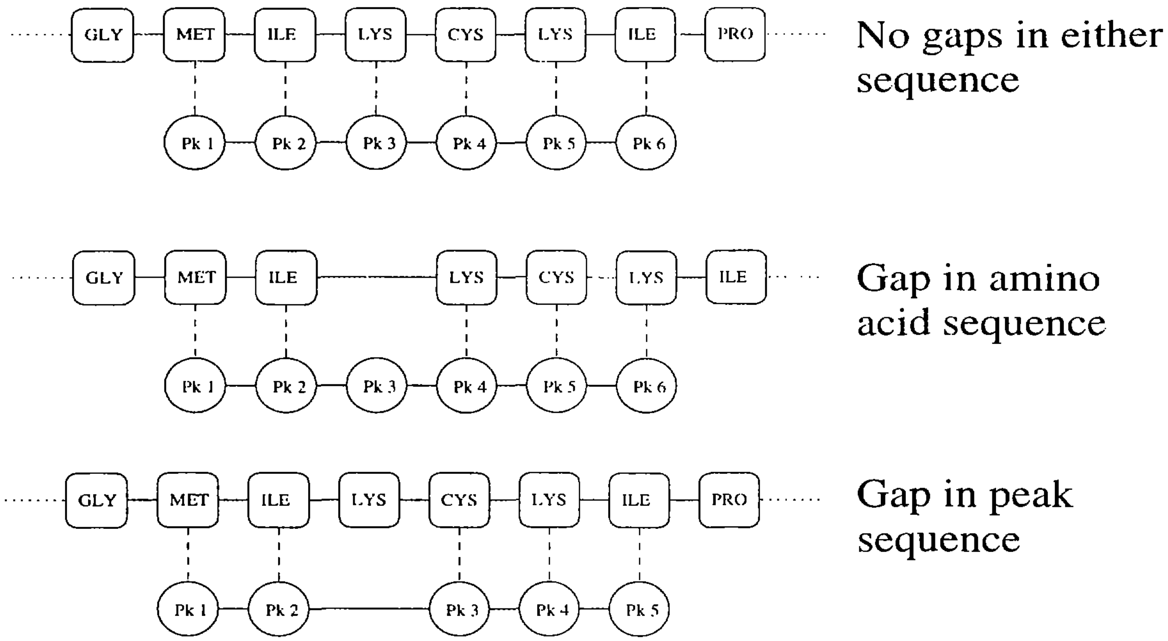

Figure 2: A variety of sequence-to-peak threadings/alignments

To determine the scoring for an alignment between a model and the protein sequence we use the scoring function described above along with a standard local alignment dynamic program (Waterman 1995). To allow for irregularities in the mapping of peaks to the sequence, we permit single insertions/deletions in the dynamic program and penalize those with a small negative score. Thus, we define an *objective function*, F, for an alignment of a model m and a sequence s, as the sum of the pairwise peak/amino acid alignments:

$$F(m,s) = \sum_{i} score(i,j) - \sum_{\#gaps} gap\_penalty,$$

such that a peak i is aligned with amino acid j for all i in model m not skipped with gaps. Given a primary structure P, we say that a subsequence s_i of P is an *optimal alignment* for a model m if:

$$F(s_i, m) \geq F(s_j, m)$$

for all subsequences s_j of P.

Results

An experiment was designed and carried out to test:

- whether the optimal threading of a correct model is the correct threading;
- whether the optimal threading score for correct models is higher than the optimal threading for incorrect (or partially correct) models; and
- whether the accuracy of crystallographic threading is greater for longer models.

70 non-homologous proteins (50 as a training set and 20 as a test set) were used in the experiment. Ideal critical point graphs for all proteins were generated from data deposited in the Protein Data Bank (Abola *et al.* 1997). Using Bayes' rule for conditional probability and the given training set, probability density functions for the scoring function were generated. The sequence of peaks that make up the true backbone (i.e., the correct model) was determined for each critical point graph. Given this backbone structure, 5,000 correct models were randomly generated at varying lengths. For comparison, 5,000 partially correct models were generated using a heuristic based on peak connectivity.

The first step in our evaluation of crystallographic threading was to determine whether correct models threaded better than incorrect models. Figure 4 presents the relationship between the the optimal scores of correct and partially correct models as the length of the model varies between 5 and 50 critical points. As anticipated, the scores of the correct models improved more rapidly with model length than those of partially correct models. When the models have reached a length of 30-35 peaks, there is almost a total separation between the scores of correct models and partially correct models. These results indicate that for long models, threading scores are capable of separating correct from incorrect models. Even at shorter lengths, the scores of correct alignments are generally higher than those of incorrect models, showing that the quality of short, automatically generated models

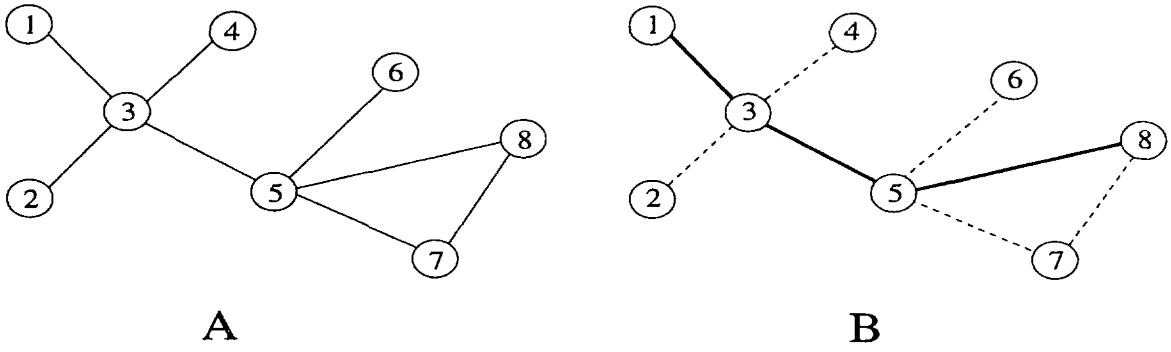

Figure 3: A) a two-dimensional projection of a three-dimensional critical point graph, with peak numbers indicated. B) a model (indicated by the dark lines) traced through the graph, with potential side-chain peaks connected to the model indicated by the dashed lines.

can be roughly evaluated and compared using crystallographic threading.

Our experiment also measured whether the optimal threading score for correct models corresponds to the alignment with the correct subsequence. Figure 5 shows the percentage of correct models for which the optimal threading corresponded to the true structure. Again, we see an increase in accuracy as the model length increases. At length 50 the accuracy is approximately 80%. We might expect this value to be even higher since the number of possible amino acid subsequences of length 50 in a given protein is quite small and the chance that two subsequences matching a single model appears correspondingly tiny. On observing the pairwise scores that lead to these incorrect threadings, most of the large contributions to the overall score were made by side-chain peaks. This is perhaps an indication that the approach of treating any peak neighbouring the model as a side-chain peak might be too liberal. A possible improvement to our scoring function might be to weigh side-chain scores less than those for backbone peaks.

Discussion

Current approaches to protein model construction and evaluation rely heavily on an expert's ability to build and judge the quality of a potential structure. This process is greatly aided by sophisticated graphics programs, which allow the user to display, rotate and compare potential models, and by software designed to assist in the computation of individual criteria (e.g., R factors) (Jones *et al.* 1991). Our research is unique in that it takes as input a computer generated model and determines how well the structure can be associated with a subsequence of the primary sequence of amino acids. This represents a fundamental step towards a fully automated approach to protein structure determination from experimental data. Crystallographic threading results can be applied at intermediate steps of protein structure determination to determine the "best candidate models", which can then be used to recover phase information and subsequently refine the electron density map. In the final stages of interpretation, crystallographic threading can be applied to determine a full mapping of the sequence onto the structure as an initial step towards determining an atomic-level model for the protein.

The reported experimental results demonstrate the effectiveness of a gapped sequence alignment approach to crystallographic threading. Gapped alignment threading algorithms have previously been proposed and developed for the inverse folding problem (Lathrop *et al.* 1998). Such algorithms are able to find the global minimum three-dimensional threading between a protein sequence and a core motif. A fundamental difference between these classic approaches to threading for structure prediction and our work is that we are not using known structures, but rather models derived from experimental data. Although previous threading algorithms have considered gaps in alignment, these are generally large gaps corresponding to non-core regions of the known structure, rather than small gaps resulting from experimental error.

An approach to threading experimentally derived models was previously described in (Baxter *et al.* 1996). In this work, a gapped threading algorithm (Lathrop & Smith 1996) was customized to the problem of threading a hypothesized structural model where the scoring function relates to information available from the interpretation of an electron density map. Similar to our work, this research attempted to associate peaks in a critical point graph with amino acids in a sequence. However, it only considered alignment over the entire sequence and full structure, and assumed that the relationship between backbone and side-chain criti-

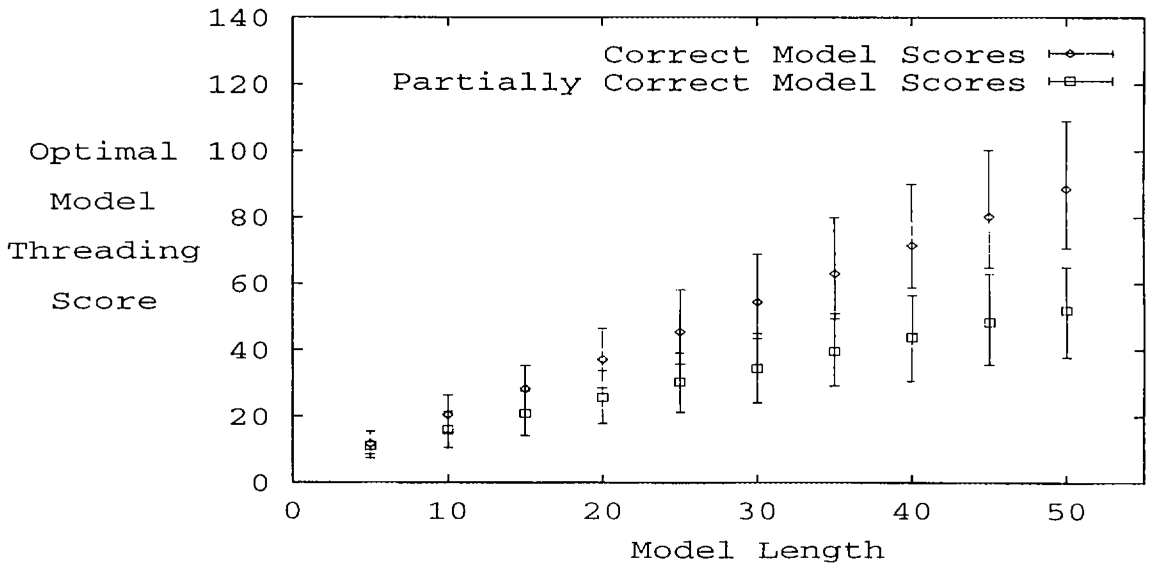

Figure 4: Optimal threading scores for correct and partially correct models. The error bars cover 90% of the scores for each class of models.

cal points was known.[3] As well, peaks were characterized by a single attribute (density) and the given fitness function did not allow for error-induced gaps in the sequence or the model. The implications of these differences is that our current approach provides for a more accurate, flexible and robust system for threading and evaluating protein structure models.

Future research in crystallographic threading includes an investigation of *thread synchronization*. Using techniques from constraint satisfaction, thread synchronization involves searching through the set of possible sequence/model alignments to determine a maximally consistent threading for the entire protein structure.

Acknowledgements. Financial support for the research reported in this paper was provided by the Natural Science and Engineering Research Council of Canada, the Communications and Information Technology Ontario Center of Excellence, the Institute for Robotics and Intelligent Systems and the PENCE Network Center of Excellence.

References

Abola, E. E.; Sussman, J. L.; Prilusky, J.; and Manning, N. O. 1997. Protein data bank archives of three-dimensional macromolecular structures. In *Methods in Enzymology*. Academic Press. 556–571.

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* 291:555–65.

Baxter, K.; Steeg, E.; Lathrop, R.; Glasgow, J.; and Fortier, S. 1996. From electron density and sequence to structure: Integrating protein image analysis and threading for structure determination. In *Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology*, 25–33. AAAI/MIT Press, Menlo Park, California.

Branden, C., and Jones, T. 1990. Between objectivity and subjectivity. *Nature* 343:687–689.

Bryant, S., and Lawrence, C. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16:92–112.

Fortier, S.; Castleden, I.; Glasgow, J.; Conklin, D.; Walmsley, C.; Leherte, L.; and Allen, F. 1993. Molecular scene analysis: The integration of direct methods and artificial intelligence strategies for solving protein crystal structures. *Acta Crystallographica* D49:168–178.

Glasgow, J.; Fortier, S.; and Allen, F. 1993. Molecular scene analysis: crystal structure determination through imagery. In Hunter, L., ed., *Ar-*

[3] Such information is not normally available from the analysis of the electron density map.

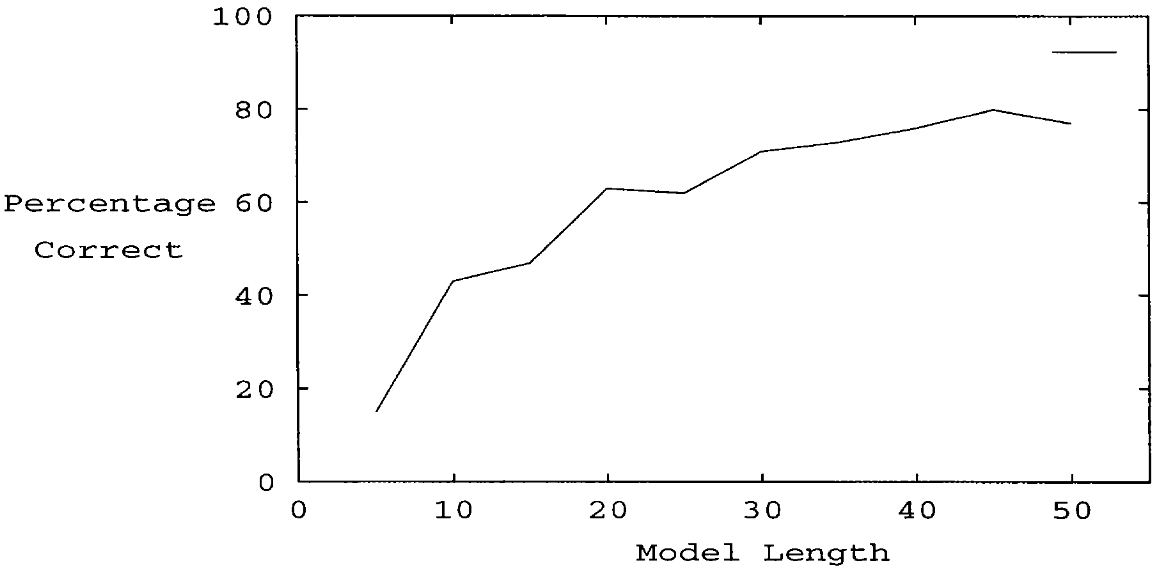

Figure 5: Percentage of correct models which were correctly threaded

tificial Intelligence and Molecular Biology. AAAI Press, Menlo Park, California. 433–458.

Jones, T.; Zou, J.; Cowan, S.; and Kjeldgaard, M. 1991. Improved methods for building protein models in electron-density maps and the location of errors in those models. *Acta Crystallographica* A47:110–119.

Karle, J. 1986. Recovering phase information from intensity data. *Science* 232:837 – 843.

Lathrop, R., and Smith, T. 1996. Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology* 255(4):641–665.

Lathrop, R. H.; Jr., R. R.; Bienkowska, J.; Bryant, B.; Buturovic, L.; Gaitatzes, C.; Nambudripad, R.; White, J.; and Smith, T. 1998. Analysis and algorithms for protein sequence-structure alignment. In Salzber, S.; Searls, D.; and Kasif, S., eds., *Computational Methods in Molecular Biology.* Elsevier Press. section 12.

Leherte, L.; Fortier, S.; Glasgow, J.; and Allen, F. 1994. Molecular scene analysis: A topological approach to the automated interpretation of protein electron density maps. *Acta Crystallographica D* D50:155–166.

Leherte, L.; Glasgow, J.; Baxter, K.; Steeg, E.; and Fortier, S. 1997. Analysis of three-dimensional protein images. *Journal of Artificial Intelligence Research (JAIR)* 125–159.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.

Sunderji, A. 1996. Protein database analysis. Undergraduate thesis, Queen's University, Kingston, Canada.

Waterman, M. S. 1995. *Introduction to Computational Biology.* London: Chapman and Hall.

White, J.; Muchnik, I.; and Smith, T. 1994. Modeling protein cores with Markov random fields. *Math. Biosci.* 124:149–179.