Protein Model Determination from Crystallographic Data

K. Edgecombe, A. Ableson, K. Baxter, A. Chiverton, J. Glasgow and S. Fortier
Department of Computing and Information Science
and Department of Chemistry
Queen's University, Kingston, Ontario
Canada K7L 3N6

Abstract

Crystallographic studies play a major role in current efforts towards protein structure determination. However, despite recent advances in computational tools for molecular modeling and graphics, the task of constructing a model of the tertiary structure of a protein from experimental data remains complex and time-consuming, requiring extensive expert intervention. This paper describes an approach to protein model determination that incorporates crystallographic data, along with sequence data. A model is represented as an annotated graph that traces the backbone and side chains for a protein.

The proposed approach incorporates numerical techniques that are applied to construct and analyze an electron density map for a unit cell of a crystal. The purpose of this work is to advance the ability to discern meaningful features of protein structure through the use of topological analysis of the relative density. Experimental results, which demonstrate the viability of the approach, are reported.

1 Introduction

A fundamental goal of research in molecular biology is to understand protein structure. Although, in theory, protein sequence information can be inferred from the ever-growing volume of DNA sequence data [1], the *protein folding problem* — that of predicting the three-dimensional structure of a protein from its sequence — remains an open and important problem [2,3,4]. Currently, there are three main approaches utilized for the prediction of tertiary structure [5]: use of sequence homology; prediction of secondary units followed by assembly of these units; and use of empirical energy functions.

Although recent attempts (such as those of Srinivasan and Rose [6] and Dill *et al.* [7]) show promise in addressing the computational complexity of the protein folding problem[2,3,4], it is doubtful that a solution to this problem will be found in the immediate future. This is due, in part, to the fact that many of the existing techniques rely heavily on our knowledge of previously determined structures, knowledge that is limited by the relatively small (with respect to sequence data) number of known structures.

Crystallography plays a major role in current efforts towards protein structure determination. Building a protein model from crystallographic data, however, is a complex and time-consuming process, which is somewhat assisted by the use of computer graphics for tracing the polypeptide chains, and for viewing and improving the resulting model[8]. Errors in the initial and subsequent models may be corrected using a refinement process, which involves modifying the 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 more sophisticated computational tools[9]. A goal of our research is to improve and accelerate the process of structure determination through the design and development of such automated tools for the purpose of protein model determination. This will result in an expanded repository of available tertiary structures and, in turn, will impact on the practice of structure prediction from sequence alone.

This paper reports on an approach to protein model construction that can be incorporated in a fully automated system for structure determination from crystallographic data. The proposed approach improves upon the previously implemented topological analysis performed in the ORCRIT program[10]. The primary advantage of our new approach is that it uses characteristics of the experimental data to find a path through the tertiary structure of the protein, thus introducing no bias into the data. In particular, it incorporates a *spline interpolation algorithm* to generate a density function for the protein, an *eigenvector following method* to derive critical points corresponding to amino acid residues and side chains, and a *gradient path following method* to connect critical points and trace the backbone for the protein.

The paper is organized as follows. Section 2 sets the context for our model building technique by presenting a general framework for automated protein structure determination from crystallographic data. In Sections 3 and 4 we describe our novel approach to model construction and report on experimental results. The paper concludes with a discussion of related and future research.

2 Molecular Scene Analysis

Our general approach to protein structure determination has been influenced by research in machine vision. The primary task of an image understanding system in artificial intelligence is to derive an underlying scene model from given image data. The research described in this paper is being incorporated into a computational approach to *molecular scene analysis*. A key process for molecular scene analysis is the automated derivation of potential scene models for a protein structure. The input for this process is an electron density map and a primary sequence of amino acid residues. An electron density map, which can be considered as the "protein image", is represented as a three-dimensional array of real values denoting electron density in the unit cell for the crystal. This map of the repeating unit of the crystal is calculated using amplitudes measured in the diffraction experiment and phases estimated from experimental and/or mathematical techniques. The interpretation of the resulting protein image involves finding the polypeptide chain, associating it with the given amino acid sequence and configuring the respective side-chains. This interpretation process is complicated, partly due to errors in the map resulting from noisy data and from the lack of accurate phase information.[a] The quality of the protein image 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 possibly identify regions of secondary structure. At medium resolution (~ 3Å), it is generally possible to trace the backbone of the protein and derive topological properties of the residues along the backbone. Only at high resolution (~ 1 Å) are the individual atoms in the protein observable. For the purpose of this paper we are primarily concerned with protein images at medium resolution from which we can model the polypeptide chain of amino acids.

An approach to molecular scene analysis has previously been proposed [11,12,13] where a scene model is generated using a topological analysis of the protein image data. The research described in this paper extends and improves upon this previous analysis in a number of ways:

- A cubic spline interpolation is performed to approximate the underlying electron density function from the crystallographic data;
- An eigenvector following algorithm is applied to detect the location of individual amino acid residues and side chains for the protein; and
- A Cash-Karp-Runge-Kutta gradient path tracing algorithm is incorporated to connect residues and construct a graph from which potential models of the backbone of the protein can be derived.

Details concerning these techniques and their advantages will be presented in the following sections.

[a] This is referred to as the classic *phase problem* of crystallography.

3 Protein model construction

At medium resolution, we define a protein model as a trace (or subpath) through a graph consisting of critical point nodes (corresponding to amino acid residues), and edges (corresponding to potential polypeptide bonds). A model may also contain branches in the trace, related to observable side chains for the residues. From the crystallographic data we can derive additional environment information for the individual residues (e.g., size, density, distance from solutions, etc.). This information can be used to "thread" a model, i.e., associate individual nodes on the graph with amino acid residues in the given sequence [14]. Thus, a model corresponds to an annotated trace of a protein backbone (or portion of the backbone) along with attribute information (from analysis of the critical point graph and the electron density map) for the residues along the backbone.

The input image data for a molecular scene analysis is represented as an *electron density map.* This uninterpreted, three-dimensional array of real values can be compared to Marr's primal sketch representation for machine vision [15]. As illustrated in Figure 1, our topological analysis of a protein structure consists of four stages. The first stage involves the generation of the electron density image from the experimental data. Once we have approximated the protein's electron density function, we then derive critical points (peaks and passes) in the topological map of the protein. In the next stage we construct a graph that connects the critical points. The final stage involves deriving models of the structure by tracing paths through the critical point graph. In the remainder of this section we describe each of these stages in more detail.

Image generation

Constructing the electron density image for a protein from X-ray diffraction data is an important step in the analysis of crystallographic data. Using standard crystallographic software, density values are provided in a discrete, three-dimensional array for the unit cell. There exists important information (density values off the grid, derivative information, etc.) that is not directly available from these image arrays. By applying an interpolation method with appropriate characteristics, we can approximate the underlying electron density function and regain some of this lost information. The principal characteristics we were concerned with in the development of our interpolation method were robustness and smoothness. The need for robustness is clear; the smoothness of the interpolant is necessary since an accurate topological analysis requires that the first and second derivatives of the density function be continuous.

In our approach to image generation, the crystallographic relative electron density grid is modeled using tri-cubic splines. This provides a smooth

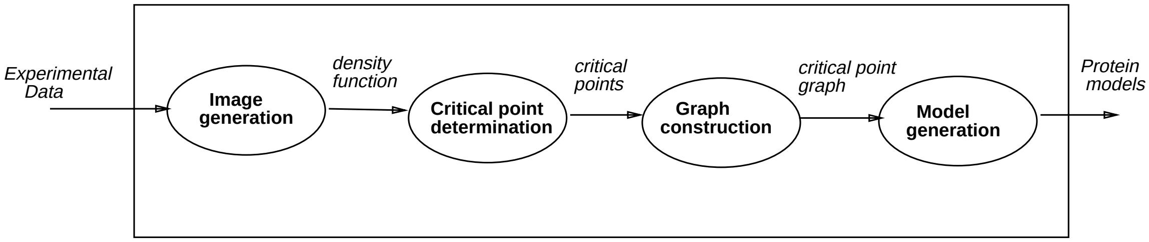

Figure 1: Architecture for model construction module.

function with continuous second derivatives over which values of the relative density, gradient and Laplacians can be calculated. The parameters defining the interpolating splines are determined algebraically from the image array.

Critical point determination

The first step in the analysis of an electron density image of a protein involves finding the *critical points* (points in the electron density map where the gradients vanish). At such points, local maxima, minima and saddle points are defined by computing first and second derivatives. The first derivatives of the density function characterize the zero-crossings, and the second derivatives are used to identify the type of critical points. In particular, we are interested in *peak* (local maxima) and *pass* (saddle point) critical points. Experiments at medium resolution have suggested that peaks and passes above a particular density cutoff generally correspond to the location of amino acid residues and peptide bonds respectively[13].

In order to calculate the critical points, a grid of initial starting points is chosen and a search for peaks is initiated using the *eigenvector following method* implemented by Popelier[16]. Next, a check for additional peaks is performed by initiating the search algorithm beginning mid-way between previously derived peaks. This is followed by a search for passes in the region between peaks.

The possible complexity of the topology in unknown densities makes it difficult to assess where a good starting point for other methods, such as the Newton-Raphson, would be. However, utilizing the eigenvector following method and a good grid for search initialization, one can have a degree of confidence that the features of interest, in this case peaks and passes, can be located.

Graph construction

The next step in our model generation process is to connect peak critical points in order to trace the backbone of the protein structure. This is achieved by applying the Cash-Karp-Runge-Kutta gradient path tracing algorithm[17] to connect a pass with its associated peaks (representing residues and side chains). The resulting trace determines the lines of interaction (lines of maximum density) between two peaks. The methodology is implemented in a program based on Poplier's MORPHY algorithm[18]. Unlike MORPHY, our version of the program utilizes cubic splines, rather than a Gaussian basis set and a quantum mechanical wave function. As well, MORPHY requires sets of nuclear coordinates which are assumed to be peaks; our program utilizes a thorough search of the image space to locate peaks.

Following the peak-pass-peak gradient path, at some points the trace may branch. We do know that there are disulphide bridges and we also know that these bridges have peaks that are of the highest density, perhaps only matched or exceeded by those of other sulphur peaks or heteroatoms peaks. Figure 2 illustrates a disulphide bridge that was formed between two continuous chains in the protein. The disulphide bridges are identified in our analysis because of the role they play in determining the protein structure. These bridges serve as anchors in our analysis with their characteristic binding of distinct segments of the trace.

Model generation

The result of the graph construction program is a connected graph that may contain cycles. The final stage of our analysis involves finding traces through this graph that correspond to potential models for the protein structure. The original implementation of ORCRIT was used to construct a single model derived by constructing the minimal spanning tree for the graph. Rather than considering a single model, at this stage we will use the critical point graph to predict multiple models - one for each possible trace of a backbone through the critical point graph.

4 Experimental results

The protein bovine pancreatic phospholipase A2 (henceforth referred to as BP2) was chosen to test our techniques since its structure has been resolved and used in previous studies. It contains 123 residues and its crystalline form

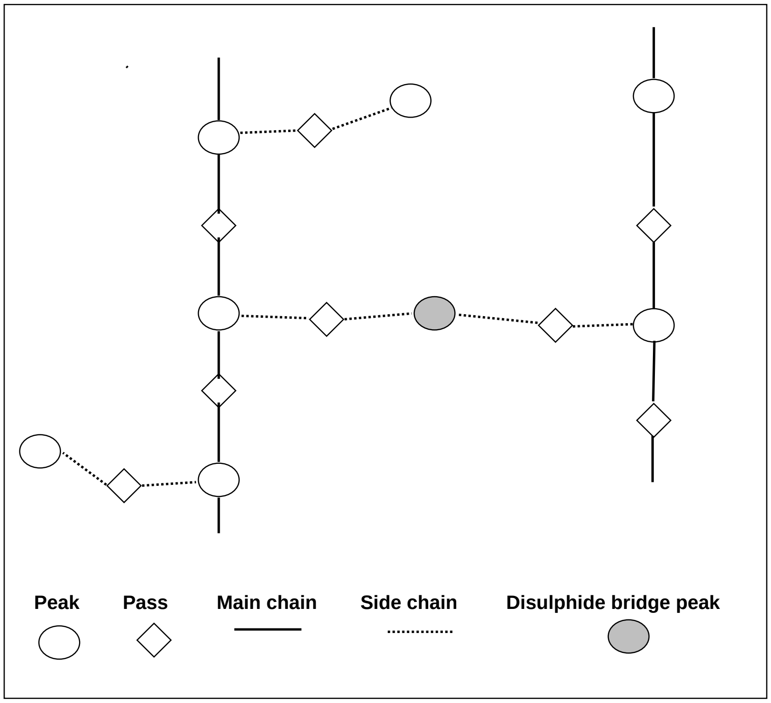

Figure 2: Illustration of disulphide bridge between main chains.

is a member of the $P2_12_12_1$ space group. Our analysis was performed on a portion of the electron density map that contains the connected protein molecule. In order to discern the effects of topological features outside the area being considered, the analysis was extended by 5 Å outside each side in this region.

The results of our experiment were evaluated in three stages:

1. *Proximity of peaks to backbone atoms and backbone connectivity:*
   The location of the peaks with respect to the protein backbone atoms was considered. A peak was considered as a correct assignment if it had a relative density greater than 0.8 and was located within 2 Å of a backbone atom. Of the 123 residues, only 8 did not have an associated peak. Next, the results of the gradient path tracing were considered. In total there were 3 missing edges (bonds). One is due to an intervening peak that was further than 2 Å from the backbone, another is due to the insertion of

a side chain peak in the chain resulting from the distance criterion, and a third is due possibly to oscillation during the path tracing procedure. Thus, there is only one break in a continuous chain from start to end.

Eight residues do not have a peak within 2Å of a backbone atom at this resolution. This is not a concern as in those areas we have a continuous backbone trace. That is, the peaks forming our topological backbone are connected through the gradient path tracing in these areas. Thus, the peaks corresponding to these "missing" residues have been absorbed into other peaks. Note that in some cases a residue may have two peaks associated with its backbone atoms, usually towards each end.

2. *Correctness of assignment of peaks to residues:*
   The position of the peaks with a relative density value greater than 0.8 were compared with the positions of all the non-hydrogen atoms in the protein. Peaks within 2 Å of such an atom were assigned to the respective residue. Only one residue, number 32 (GLY), was not assigned a peak. As well, a peak was assigned to the Ca^{2+} ion associated with the protein in the asymmetric unit. Of note is the high number of side chains represented by peaks at this resolution (3 Å).

3. *Correctness of connectivity in trace of backbone:*
   Using the peak identified as residue 1 (ALA), the best trace is created by incorporating the highest density passes and discounting side chains where the trace stops. Points at which there appear to be a fork are explored. From previous work in this area[13], we are aware that the highest peaks usually represent disulphide bridges and heteroatoms, such as Ca^{2+}, at this resolution. We follow the peak-pass-peak path until the trace leads to the break point. Picking up the trail and continuing on after tracing through the entire chain, we find seven disulphide bonds and also two locations where relatively large passes (greater than 0.6 relative density) bridge sidechains from one portion of the chain to another. The size of the peaks, however, allow us to discount them as disulphide bridges. In addition, we find a large peak that acts as a bridge with 3 large passes as well as one that is close to the cutoff. This peak is identified as the Ca^{2+} ion.

The derived critical point graph for BP2 is illustrated in Figure 3. The trace (in grey) has been superimposed on the protein backbone (in black) to illustrate the correlation between the critical point graph and the known structure. Also note the presence of disulphide bridges that connect portions of the main chain.

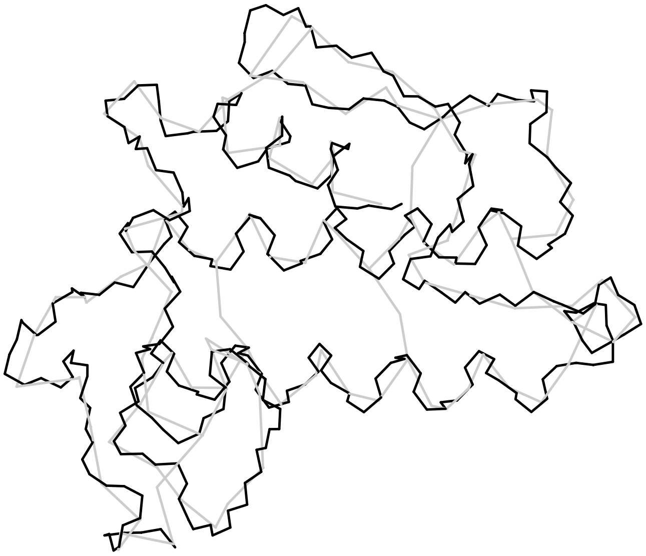

Figure 3: Critical point graph for protein BP2.

5 Discussion

The original ORCRIT program utilizes a relative density cutoff to prune the list of peaks and passes that would be candidates for the protein chain. Then a distance criterion is applied to cut down on the number of candidates to be used to generate the connectivity of the peaks and passes. So, for example, using a density cutoff of 0.8 and 0.6, respectively, for peaks and passes and a distance cutoff of 4 Å would result in 4824 edges and an average number of associations per vertex of 5.05. Increasing the distance cutoff to 4.5 Å raises the number of edges to 7249 and the average number of associations to 7.6 (associations/vertex). The weighting function is then applied which gives rise to a minimal spanning tree. Thus, should a side chain peak satisfy a distance criteria, it may be included in the spanning tree despite the lack of interactions of an appropriate nature. The application of the eigenvector following method and subsequent path tracing from the passes to the peaks eliminates any uncertainty as to the assignment of peaks and passes. Side chain peaks will almost always have only one strong interaction, that to the backbone peak representing the residue along the chain. A disulphide bridge peak will, however, display two strong connections. In this manner the decision making formerly handled by the weighting function is reduced tremendously. The eigenvector following method reduces the possible number of edges to 1828 and the average number of possible associations per vertex to 1.9. Thus, the complexity of finding the correct model of the protein structure is drastically reduced.

Hydrogen bonding and van der Waals interactions can also be characterized through analysis of the density and the use of the gradient path tracing method. These interactions are not as strong as normal 'bonding' interactions but play a vital role in the understanding of protein structure. The strength of these interactions is reflected in the magnitude of the relative density observed at the pass on the interaction line.

An evaluation of the derived models is the next step in our research in molecular scene analysis. Assessing the quality of a protein model involves determining whether or not it makes sense (in terms of our knowledge of the chemistry, biology and physics of the molecule) and whether it is consistent with the primary sequence information. Research in model evaluation for experimentally derived structures has previously focussed on verifying that the final protein model is correct [19].[b] Recently, Kleywegt and Jones [20] have proposed some quality control criteria for the assessment of intermediate protein models. The tools they suggest assume a single model, which is evaluated to determine what parts of the structure need to be revised or rebuilt.

We propose to extend previous work in protein model evaluation in a number of ways. First, we are considering a comprehensive set of general criteria that will be considered in the evaluation process. Although implementations exist for some of these criteria, many of them are still applied through expert visual inspection of a graphics display for a model. In our approach, individual criteria are implemented as expert agents, whose potentially conflicting advice is combined using a learning and problem solving architecture that was developed for the purpose of integrating multiple sources of expertise [21]. Our research is novel in the sense that it provides for the prediction and comparative assessment of multiple, computationally derived models, and can be used to guide a heuristic search towards a fully-interpreted protein structure.

The methodology for protein model construction described in this paper is aimed at assisting structure determination at low to medium resolution,

[b] Even with techniques for evaluating the final protein model, incorrect models have been published and entered into the protein database.

enabling faster interpretation of crystallographic data than present methods allow. This, in turn, allows the crystallographer to improve the phase estimates and consequently the protein image. Note that as the resolution of the experimental data is increased from low to medium to high, individual peaks in the electron density map evolve into multiple peaks representing backbone and sidechain portions of each residue and finally into peaks representing the atoms themselves. The point at which these events occur is dependent on such factors as temperature, quality of the crystal and the quality of data derived from the diffraction experiment.

In conclusion, numerical techniques have been implemented and tested for the derivation of protein models from protein image data. These represent a significant advancement in our ability to automatically analyze molecular scenes. In particular, they allow us to construct a critical point graph from which we can derive and evaluate potential tertiary structures for the protein. These techniques can be used to accelerate the process of structure determination, resulting in the expansion of the the repository of known protein structures.

Acknowledgements

The authors would like to thank Evan Steeg and Ed Willis for their contributions to the research described in this paper. Funding for the research was provided by the Natural Science and Engineering Research Council of Canada.

1. C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
2. G.N. Reeke, Jr. Protein folding: computational approaches to an exponential-time problem. *A. Rev. Comput. Sci.*, 3:59–84, 1988.
3. R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is NP-hard problem: proof and implications. *Bull. Math. Biol.*, 55:1183–1198, 1993.
4. A.S. Fraenkel. Complexity of protein folding. *Bull. Math. Biol.*, 55:1199–1210, 1993.
5. J.W. Ponder and F.M. Richards. Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193:775–791, 1987.
6. R. Srinivasan and G.D. Rose. Linus: a hierarchical procedure to predict the fold of a protein. *Proteins*, 22:81–99, 1995.
7. K.A. Dill, K.M. Fiebig, and H.S. Chan. Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci.*, 90:1942–1946, 1993.

8. T.A. Jones, J.Y. Zou, S.W. Cowan, and M. Kjeldgaard. Improved methods for building protein models in electron-density maps and the location of errors in those models. *Acta Crystallographica*, A47:110–119, 1991.
9. C.I. Branden and T.A. Jones. Between objectivity and subjectivity. *Nature*, 343:687–689, February 1990.
10. C.K. Johnson. ORCRIT. The Oak Ridge critical point network program. Technical report, Chemistry Division, Oak Ridge National Laboratory, USA, 1977.
11. et al. Fortier S. Molecular scene analysis: The integration of direct methods and artificial intelligence strategies for solving protein crystal structures. *Acta Crystallographica*, D49:168–178, 1993.
12. J.I. Glasgow, S. Fortier, and F.H. Allen. Molecular scene analysis: crystal structure determination through imagery. In L. Hunter, editor, *Artificial Intelligence and Molecular Biology*, pages 433–458. AAAI Press, Menlo Park, California, 1993.
13. L. Leherte, S. Fortier, J. Glasgow, and F.H. Allen. Molecular scene analysis: A topological approach to the automated interpretation of protein electron density maps. *Acta Crystallographica D*, D50:155–166, 1994.
14. K. Baxter, E. Steeg, R. Lathrop, J. Glasgow, and S. Fortier. 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*, pages 25–33. AAAI/MIT Press, Menlo Park, California, 1996.
15. D. Marr and H.K. Nishihara. Representation and recognition of the spatial organisation of three-dimensional shapes. *Proceedings of the Royal Society of London*, B200:269–294, 1978.
16. P.L.A. Popelier. A robust algorithm to locate all types of critical points in the charge density and its Laplacian. *Chem. Phys. Lett.*, 228:160–164, 1994.
17. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes, 2nd edition.* Cambridge Press, Cambridge, 1992.
18. P.L.A. Popelier. MORPHY, a program for an automated atoms in molecules analysis. *Comput. Phys. Comm.*, 93:212, 1996.
19. R. Luthy, J.U. Bowie, and D. Eisenber. Assessment of protein models with three-dimensional profiles. *Nature*, 356:83–85, March 1992.
20. G.J. Kleywegt and T. A. Jones. Good model-building and refinement practice. In R.M. Sweet and C.W. Carter Jr., editors, *Methods in Enzymology.* to appear.
21. S. Epstein. For the right reasons: The FORR architecture for learning in a skill domain. *Cognitive Science*, 18(3):479–511, 1994.