Published in final edited form as:

Comput Chem Eng. 2009 December ; 33(12): 2055–2062. doi:10.1016/j.compchemeng.2009.06.006.

Optimization techniques in molecular structure and function elucidation

Nikolaos V. Sahinidis*

Department of Chemical Engineering Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

This paper discusses recent optimization approaches to the protein side-chain prediction problem, protein structural alignment, and molecular structure determination from X-ray diffraction measurements. The machinery employed to solve these problems has included algorithms from linear programming, dynamic programming, combinatorial optimization, and mixed-integer nonlinear programming. Many of these problems are purely continuous in nature. Yet, to this date, they have been approached mostly via combinatorial optimization algorithms that are applied to discrete approximations. The main purpose of the paper is to offer an introduction and motivate further systems approaches to these problems.

Keywords

Bioinformatics; Molecular structure and function; Optimization

1 Introduction

Structural biology has witnessed an unprecedented influx of mathematical and computational techniques over the past two decades. These techniques have transformed research in a field that has become increasingly data-driven. As part of this revolution, optimization techniques are now used routinely to solve problems in structure elucidation and the study of structure-function relationships. For instance, the list of The ten most wanted problems in protein bioinformatics (Tramontano, 2005) includes:

1. protein sequence alignment,
2. predicting protein features from the sequence,
3. function prediction,
4. protein structure prediction,
5. membrane protein structure prediction,
6. functional site identification,
7. protein-protein interaction,

© 2009 Elsevier Ltd. All rights reserved.

*Tel.: +1 412 268 3338; Fax: +1 412 268 7139 sahinidis@cmu.edu (Nikolaos V. Sahinidis)..

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

8. protein-small molecule interaction,
9. protein design, and
10. protein engineering.

Most problems on this list require the use of optimization algorithms to solve a variety of predictive models. Research on the first problem has led to the development of dynamic programming algorithms that are now central in the BLAST code, a software that has received over 46,000 citations over the past two decades (Altschul et al., 1990; Altschul and Schaffer, 1997). In general, optimization problems and algorithms have become indispensable problem-solving tools in modern biology.

The current paper presents optimization approaches to three problems of structure and function elucidation. All three problems are continuous in nature, in the sense that the unknowns are, in general, continuous quantities. All three problems have been approached via combinatorial optimization techniques, following suitable approximations, in order to arrive at fast algorithms and cope with the underlying complexity. We begin with the protein side-chain prediction problem, a simplified version of the protein folding problem, that plays a major role in structure prediction and protein design. Then, we address the problem of protein structural comparison, which can be regarded as a natural extension of sequence comparison and is currently intensely pursued by the structural biology community. Finally, the third problem involves the analysis of X-ray diffraction measurements for the purpose of determining crystal structures.

The three problems addressed here comprise but a small fraction of structural informatics problems addressed by optimization techniques. However, they are highly representative of, and central in, a wealth of important problems in structural bioinformatics and chemical informatics.

The current paper is an extended version of one that appeared in the Proceedings of FOCAPO 2008 (Sahinidis, 2008). Section 2 presents a combined enumeration-domain-reduction technique for the protein side-chain conformation problem. Section 3 is devoted to a combinatorial approach for the structural alignment of proteins. Section 4 describes discrete and nonlinear optimization approaches for the problem of structure determination from single-crystal X-ray diffraction data. Finally, conclusions are drawn in Section 5.

2 Protein side-chain conformation problem

2.1 Problem statement and motivation

The protein-side chain conformation problem (SCP) is that of predicting the 3D structure of the side chains of a protein, assuming that the 3D structure of the protein backbone is known. The problem is a simpler version of the protein folding problem, which calls for predicting the entire folded structure of a protein. In addition to serving as a subproblem in protein structure prediction, SCP is also central in the context of designing new proteins. In particular, when one or a few amino acid residues of a given protein are substituted, one would expect small structural changes to the protein backbone and would seek to determine the side-chain angles using the known backbone structure as a starting point in the analysis. The latter problem is relevant to drug design and, as such, of particular importance to the biotechnology and pharmaceutical industries (Dahiyat and Mayo, 1997).

2.2 Rotamer approximation

The amino acid residues of the protein side chain interact energetically with the protein backbone as well as with other side-chain amino acid residues. It is commonly assumed that proteins fold to a structure that corresponds to a global minimum of the total system energy

due to all of these interactions. Then, SCP calls for finding the backbone dihedral angles that minimize the total energy of these interactions. Whereas the dihedral angles can take any continuous value, it is commonly assumed that the search problem can be restricted to values from a discrete set. In particular, a set of statistically significant conformations, known as *rotamers*, is commonly used to identify the folded structure (Ponder and Richards, 1987; Dunbrack and Karplus, 1993). The energetically minimal protein conformation may, in reality, involve dihedral angles that do not equal any of the rotamers. However, the rotamer approximation simplifies the search problem considerably, as it reduces SCP to a combinatorial optimization problem that can be solved with combinatorial optimization tools in a variety of ways. From now on, we will assume that SCP involves n amino acid residues and an average of R rotamers per residue. Therefore, the problem can be represented on a graph with nR nodes, where each node corresponds to a specific residue-rotamer combination. Edges will be present in the graph if and only if two specific residue-rotamers interact energetically. The problem is illustrated in Figure 1. In this figure, we use circles, ellipses, and edges to represent rotamers, residues, and energetic interactions between rotamers, respectively. Exactly one rotamer must be selected for each residue. Once rotamers are selected, all energetic interactions are determined (red portion of the figure). Solving SCP requires finding a minimum edge-weighted clique in this graph.

2.3 Literature review

SCP is NP-hard to solve (Pierce and Winfree, 2002) and even to approximate within a constant times the number of edges of the underlying graph (Chazelle et al., 2004). As a result, two types of algorithms have been developed for SCP: exact optimization algorithms (Desmet et al., 1992; Lasters and Desmet, 1993; Goldstein, 1994; Gordon and Mayo, 1998; Pierce et al., 2000; Looger and Hellinga, 2001; Gordon et al., 2003; Xie and Sahinidis, 2006) and heuristics (Lee and Levitt, 1991; Hellinga and Richards, 1994; Dahiyat and Mayo, 1994).

Exact optimization algorithms yield conformations that are provably global energy minima. Because of the NP-hardness of the problem, any global optimization approach for SCP may, in the worst case scenario, run in exponential computing time. Heuristics are faster but do not come with any theoretical guarantee on the accuracy of their predicted conformation. It has been reported that the quality of heuristic solutions for SCP in practice degrades with increasing problem size (Voigt et al., 2000).

The first exact algorithm to be proposed for SCP is dead-end-elimination (DEE) (Desmet et al., 1992), which is still widely used, by itself (Looger and Hellinga, 2001; Gordon et al., 2003) or within more complex algorithms (Gordon and Mayo, 1999; Xie and Sahinidis, 2006). DEE is a domain filtering technique that prunes parts of the search space via domination arguments. In particular, certain rotamers or combinations of rotamers are ruled out by proving the existence of better alternatives. DEE is simple and very efficient at the same time, even though its efficiency deteriorates for more complex energetic interactions.

A mixed-integer linear programming (MILP) approach is also possible for SCP. In particular, consider the residue-rotamer graph of Figure 1 and define binary variables associated with selecting a rotamer from the nodes corresponding to each residue. The energetic interactions with the backbone can be represented via linear terms involving these binary variables. Binary products of these variables are needed for modeling energetic interactions between amino acid residues as follows:

$$\begin{aligned}
\min \quad & \sum_{(i,r): r \in \mathcal{R}_i} c_{ir} y_{ir} + \sum_{(i,r,j,s): i < j, j \in \mathcal{N}(i), r \in \mathcal{R}_i, s \in \mathcal{R}_j} p_{irjs} y_{ir} y_{js} \\
\text{s.t.} \quad & \sum_{r \in \mathcal{R}_i} y_{ir} = 1, \quad i \in N \\
& y_{ir} \in \{0, 1\}, \quad i \in N, r \in \mathcal{R}_i.
\end{aligned} \tag{1}$$

Here, $\mathcal{R}_i$ denotes the set of rotamers for residue i , N denotes the set of side-chain residues, and $\mathcal{N}(i)$ denotes the set of residues that interact with residue i . The binary variable y_{ir} takes the value of 1 if and only if $r \in \mathcal{R}_i$ is selected as the rotamer for residue i . Constraint (1) requires each residue to assume exactly one rotamer position. In the objective function, c_{ir} denotes the energetic interaction between rotamer r of residue i and the backbone, while p_{irjs} represents the energy of interaction between rotamer r of residue i and rotamer s of residue j .

Using a standard linearization of these quadratic terms results in a standard MILP model, for which a solution can be obtained using a standard MILP solver (Eriksson et al., 2001; Althaus et al., 2002; Kingsford et al., 2005):

$$\begin{aligned}
\min \quad & \sum_{(i,r): r \in \mathcal{R}_i} c_{ir} y_{ir} + \sum_{(i,r,j,s): i < j, j \in \mathcal{N}(i), r \in \mathcal{R}_i, s \in \mathcal{R}_j} p_{irjs} x_{irjs} \\
\text{s.t.} \quad & \sum_{r \in \mathcal{R}_i} y_{ir} = 1, \quad i \in N \\
& \sum_{r \in \mathcal{R}_i} x_{irjs} = y_{js}, \quad i \in N, j \in \mathcal{N}(i), s \in \mathcal{R}_j \\
& x_{irjs} \in \{0, 1\}, \quad i \in N, r \in \mathcal{R}_i, j, s \in \mathcal{R}_j \\
& y_{ir} \in \{0, 1\}, \quad i \in N, r \in \mathcal{R}_i.
\end{aligned} \tag{2}$$

Here, $x_{irjs} = 1$ if and only if $r \in \mathcal{R}_i$ is selected as the rotamer for residue i and $s \in \mathcal{R}_j$ is selected as the rotamer for residue j . Constraint (2) ensures that $x_{irjs} = y_{ir} y_{js}$.

Another exact approach to SCP uses graph-theoretic properties to decompose the underlining residue graph and eliminate inferior combinations, including residues with a single rotamer or a single neighbor (Canutescu et al., 2003) and residues with two neighbors (Xu, 2005).

2.4 The residue-rotamer reduction (R3) algorithm

Since many residues are too distant to have any energetic interactions, the residue graph introduced above can be expected to be sparse. For an SCP with $n = 2000$ backbone amino acid residues with an average of $R = 8$ rotamers per residue, a typical scenario for the protein folding problem, the underlying residue-rotamer graph would involve 16,000 nodes. Assuming a density of 1 to 10% for this graph, one would end up with anywhere from 640,000 to 6,400,000 arcs in the graph. Consequently, a standard optimization approach would involve an MILP with hundreds of thousands to millions of binary variables, and would be difficult to apply. The residue-rotamer-reduction (R3) algorithm (Xie and Sahinidis, 2006) circumvents this difficulty by integrating residue and rotamer reduction in a domain filtering approach that carefully limits its memory requirements. By eliminating dominated residues and rotamers, this approach simplifies the underlying residue graph, thus yielding more opportunities for further reduction based on optimality arguments that involve one or a few pairs of residues or rotamers. When no such further reduction is possible, the algorithm relies on a partial enumeration step that combines two residues, again yielding more opportunities for reduction.

The main components of this algorithm are three nested algorithmic loops executed repeatedly in the following order:

1. *Residue reduction.* Residue reduction eliminates residues from the residue graph by creating an equivalent but simpler problem. Four types of residue reduction techniques are incorporated:
 - Isolated residues, *i.e.*, residues with no incident arcs, are simply eliminated from the problem.
 - Residues with a single rotamer are eliminated after adding the corresponding residue-rotamer costs to the objective function.
 - Residues with a single neighboring residue can be eliminated after a suitable modification of the energy terms and neighbor lists involved (Canutescu et al., 2003).
 - Residues with only two residue neighbors can be eliminated after a suitable modification of the energy terms and neighbor lists involved (Xu, 2005).
2. *Rotamer reduction.* Rotamer reduction is done via the following two DEE arguments:
 - A rotamer can be eliminated if it can be shown that its best-case energetic interactions are worse than another rotamer's worst-case energetic interactions (Goldstein, 1994).
 - Two rotamers cannot coexist in the global minimum energy conformation if their coexistence would lead to worse energetic interactions than another pair of rotamers (Pierce et al., 2000).
3. *Residue unification.* Any two residues may be combined into a single residue by creating a new rotamer for each pair of rotamers of the original two residues and summing up their corresponding energy contributions.

The process of unification is illustrated in Figure 2. Unification of two residues simplifies the topology of the residue-rotamer graph and may create opportunities for additional residue and rotamer reduction. However, immediately after unification the memory requirements of the algorithm are typically increased due to an increase in the size of the underlying residue-rotamer graph. For this reason, unification is kept in the inner-most loop of the algorithm. The other two steps, residue reduction and rotamer reduction, are executed repeatedly until neither type of reduction is possible. The overall algorithm is illustrated in Figure 3. Residue reduction comes first in this setting because it is simpler than rotamer reduction, simplifies rotamer reduction operations considerably, and often solves SCP entirely on its own. Even though residue reduction is executed first, the algorithm returns to this step if rotamer reduction simplifies the residue-rotamer graph. When no further reduction of any type is possible, the algorithm unifies the residue pair with the last number of resulting rotamer pairs so as to minimize the size of the resultant graph.

The implementation of the R3 algorithm relies on the backbone-dependent library (Dunbrack and Karplus, 1993) to generate rotamers for a specific SCP. Energy terms are generated in a manner similar to (Canutescu et al., 2003). The BALL library (Kohlbacher and Lenhof, 2000) is used to read PDB files, generate rotamers, compute energy terms, and write out predicted structures into PDB files.

Computational results with R3 have been reported for problems with up to 1000 residues and 10000 rotamers (Xie and Sahinidis, 2006). For these problems, R3 determined the globally minimal energetic conformation in a few seconds on a standard computer workstation. Its running time was up to two orders of magnitude less than that of the competing approaches by MILP and the popular SCP software SCWRL 3.0, with comparable prediction accuracy. An

on-line server implementing the algorithm can be accessed through <http://archimedes.cheme.cmu.edu>.

3 Protein structural alignment

3.1 Importance and approaches

Sequence alignment methods (Needleman and Wunsch, 1970; Smith and Waterman, 1981) have found wide applications in protein similarity detection, database search, and function prediction. *Structural* alignment (Bourne and Shindyalov, 2003) has been emerging recently as a technique that provides even more meaningful information for comparing proteins than sequence-based approaches. In fact, structural alignment is now used to assess various sequence alignment methods (Vogt et al., 1995; Orengo et al., 1997).

Qualitatively, the structural alignment problem requires assigning amino acid residues of a given protein to amino acid residues of another given protein, in a way that highlights similarities between the two proteins. Several structural alignment algorithms, together with measures of similarity have been proposed (Shindyalov and Bourne, 1998; Holms and Sander, 1993; Zemla, 2000; Orengo et al., 1997; Gerstein and Levitt, 1996). Two major modeling paradigms have been developed in this context. If the atomic coordinates of the two proteins are known, one can fix one of the two proteins in space and rotate and translate the second protein until a correspondence between amino acid residues is identified that optimizes the chosen measure of similarity. The second major approach to this problem requires that each protein be described by a contact map, *i.e.*, a graph in which nodes correspond to amino acid residues and arcs are present if and only if the corresponding nodes are “within contact,” *i.e.*, if their distance is below a certain threshold. In this, so-called contact map overlap (CMO) approach, one begins with the contact maps-graphs of two given proteins and would like to assign nodes of one graph to nodes of the second graph in a way that maximizes the number of common contacts of the aligned residues (Godzik et al., 1992; Godzik and Skolnick, 1994).

3.2 Contact map overlap maximization

Given the contact graphs of two proteins, CMO maximization calls for identifying two maximal common subgraphs. This problem is NP-hard to solve (Goldman et al., 1999). It is even APX-hard (Goldman et al., 1999), thus refuting the possibility for the existence of polynomial-time approximation algorithms with arbitrarily good constant approximation ratios. Solution approaches for CMO maximization include heuristics (Krasnogor et al., 2005), approximation algorithms with constant approximation ratios for certain special cases of the problem (Goldman et al., 1999), a parameterized polynomial-time approximation scheme (Xu et al., 2006), and several exact optimization algorithms. All exact algorithms for this problem are variants of branch-and-bound. The integer programming approach of (Carr et al., 2000; Lancia et al., 2001) is based branch-and-cut. More recent exact algorithms have relied on Lagrangian relaxation (Caprara and Lancia, 2002; Caprara et al., 2004), reformulation of CMO as a maximum clique problem (Strickland et al., 2005), and dynamic-programming bounds (Xie and Sahinidis, 2007). The branch-and-reduce algorithm of (Xie and Sahinidis, 2007) relies on dynamic programming to compute lower and upper bounds for CMO and represents the current state-of-the-art for the problem. This algorithm will be detailed below after the presentation of an illustrative example.

3.3 Illustrative example

Figure 4 presents the symmetric units of two proteins (2NPV and 2OKX) whose structures are to be compared. The contact graph of 2NPV has 47 residues and 178 edges, while the contact graph of 2OKX has 82 residues and 218 edges. The contact graphs and alignment of the two

structures are provided in Figure 5. Amino acid residues correspond to nodes of the two contact graphs, while dashed lines show the final alignment between residues of the two proteins. The edges of the two graphs are colored in yellow to indicate common overlap and purple indicating no common contacts. As seen, two parts of the first protein are near perfectly aligned with two parts of the second protein. A total of 121 edges are part of the maximal alignment.

3.4 A branch-and-reduce approach to CMO (Xie and Sahinidis, 2007)

The algorithm seeks a maximum-size common subgraph of the contact graphs of two given proteins in a way that preserves the order of residues in the original sequences. This latter requirement reduces the search space and eliminates certain biologically meaningless alignments, although it may certainly eliminate some meaningful alignments, too. The problem calls for determining a correspondence between the node sets of the two contact graphs. The algorithm employs the principles of branch-and-bound to ensure an exact solution for CMO. At the root node of the branch-and-bound tree, we allow all possible residue pairs in the final solution. Lower and upper bounds on the maximal number of overlaps are then computed and used to eliminate inferior parts of the search space. There are two crucial parts of this algorithm: reduction and bounding. Reduction involves the use of dominance-based arguments to eliminate certain parts of the search tree from further consideration. Once certain residue pairs have been matched, potential new matches can be eliminated after calculating the worst-case change to the objective that will result from one potential match, and comparing it to the best-case change that will result from another match. Several involved reduction steps are possible and the most involved ones invoke dynamic programming. The computation of an upper bound for CMO can also utilize dynamic programming to identify a longest path of an acyclic directed graph determined by the as yet unmatched residue pairs. The reduction criteria eliminate certain unmatched residue pairs, simplify the contact graphs, accelerate computation on each node, strengthen lower and upper bounds, and expedite convergence of the algorithm. Whenever branching decisions must be made, a potential match of one residue from one protein to a residue of the second protein is selected in a way that aims to induce a balanced search tree or lead to maximal objective function improvement. Once the branching pair is chosen, the assignment is enforced in one child node and disallowed in the second child node. The detailed algorithmic description of all these steps can be found in Xie and Sahinidis (2007) and an on-line server implementing the algorithm can be accessed at <http://archimedes.cheme.cmu.edu>.

Computational experiments on four benchmark test sets with the above algorithm suggest that this algorithm represents the current state-of-the-art in the CMO literature, in the sense that it can solve more, in some cases many more, problems than competing algorithms can solve within a given amount of allowable CPU time. Yet, a comparison between all pairs of proteins currently in the Protein Data Bank is still not feasible for any exact CMO algorithm in a reasonable computing time. Nonetheless, when the algorithm described above is interrupted before it closes the gap between the lower and upper bounds, it still provides a lower bound that constitutes a feasible alignment. The gap between this alignment and the maximum number of all possible overlaps for the given pair of proteins can be used as a measure of similarity of the two proteins. It turns out that, even when the above-described algorithm does not terminate in a reasonable computing time, the similarity measure it thus provides leads to very meaningful classification of protein domains for a set of 40 large protein domains described in Lancia et al. (2001). This suggests that the contact map overlap model is a biologically meaningful one.

4 Structure determination from single-crystal X-ray diffraction measurements

Since the mid nineteen hundreds, analysis of X-ray diffraction data of crystals has been used extensively for the determination of molecular structure and properties. While diffraction is employed almost on a routine basis worldwide, it is often a major challenge to identify the three-dimensional structure that best fits the diffraction data. A key obstacle, in particular, is

the identification of the phases of the diffracted rays from measurements of intensities alone. This problem is known as the *phase problem* in crystallography and its solution represents a key obstacle towards structure and function elucidation.

Methods developed for the phase problem in X-ray crystallography include the tangent formula (Karle and Hauptman, 1956), maximum entropy (Bricogne, 1984), the minimal principle (Debaerdemaeker and Woolfson, 1983), the minimum charge (Elser, 1999), and variants of the above. These methods make use of a merit function to score potential structures based on how well they match the experimental data. The complexity of the resulting phase estimation problem is significant because of the existence of multiple local optima in the underlying optimization formulations. In the sequel, we present the phase problem and outline an optimization approach to this problem.

4.1 The phase problem

Direct methods in crystallography rely on the assumption that the electron density is everywhere nonnegative and electrons are strongly concentrated around the centers of the atoms. The electron density, $\rho(\mathbf{r})$, at any point $\mathbf{r}$ in space can be expressed in terms of X-ray diffraction quantities as:

$$\rho(\mathbf{r}) = \sum_{m=1}^M F_m \exp(-2\pi i \mathbf{h}_m \cdot \mathbf{r}) \quad (3)$$

where F_m and $\mathbf{h}_m$ are the structure factor and reciprocal-lattice vector of the m th reflection. The normalized structure factor, E_m , which is directly obtainable from F_m , can be expressed in terms of its corresponding magnitude, $|E_m|$, and phase, ϕ_m , as:

$$E_m = |E_m| \exp(i\phi_m) \quad m=1, \dots, M \quad (4)$$

From the X-ray experiment, we directly obtain only the reciprocal-lattice vectors and structure-factor magnitudes. If the phases were also known, then (3) and (4) could be used to determine atomic positions by identifying electron density maxima. However, the phases are not measurable experimentally. This gives rise to the phase problem, the objective of which is to deduce the values of the phases from structure-factor amplitudes. The normalized structure factor under the assumption of a unit cell consisting of n discrete, nonvibrating point atoms is:

$$E_m = \frac{1}{\sigma_2^{1/2}} \sum_{j=1}^n Z_j \exp(2\pi i \mathbf{h}_m \cdot \mathbf{r}_j) \quad (5)$$

where $\sigma_2 = \sum_{j=1}^n Z_j^2$, Z_j is the atomic number of atom j , and $\mathbf{r}_j = (x_j, y_j, z_j)$ is the position vector of atom j .

Equations (4) and (5) show that there is a direct relation between atomic positions and phases. Therefore, since the position vectors depend on the selection of the origin, the phases also depend on the selection of the origin. Hauptman and Karle (1953) have shown that certain linear combinations of the phases, the *structure invariants*, are independent of the choice of origin. The most important of these invariants are the triplets:

$$\omega_t = \phi_{m_t} + \phi_{m'_t} + \phi_{m''_t} \quad t=1, \dots, T$$

where the indices m_t , m'_t , and m''_t correspond to reciprocal-lattice vectors $\mathbf{h}_{m_t}$, $\mathbf{h}_{m'_t}$, and $\mathbf{h}_{m''_t}$, respectively, such that $\mathbf{h}_{m_t} + \mathbf{h}_{m'_t} + \mathbf{h}_{m''_t} = 0$ for all T triplet invariants. Under the assumption that all atom position vectors are random variables, uniformly and independently distributed in the unit cell, the conditional probability distribution of the triplet ω_t is given by (Cochran, 1955):

$$P(\omega_t | |E_{m_t}|, |E_{m'_t}|, |E_{m''_t}|) = \frac{1}{2\pi I_0(A_t)} \exp(A_t \cos \omega_t) \quad t=1, \dots, T$$

where, I_0 is the zeroth order modified Bessel function and, when all n atoms are identical, $A_t = 2n^{-1/2} |E_{m_t}| |E_{m'_t}| |E_{m''_t}|$. From the conditional probability distribution, it readily follows that the expected value of ω_t is zero. Thus, an estimate of ω_t is:

$$\omega_t = \phi_{m_t} + \phi_{m'_t} + \phi_{m''_t} \approx 0 \quad t=1, \dots, T$$

and is valid only for large values of A_t . However, as n increases, A_t decreases and the estimate $\omega_t \approx 0$ is not accurate. This limitation has motivated Debaerdemaeker and Woolfson (1983) to suggest a least squares minimal principle involving the cosine of the invariants instead of the invariants themselves. For non-centrosymmetric structures, the conditional expected value of this cosine is (Germain et al., 1970):

$$\langle \cos \omega_t \rangle_t = \frac{I_1(A_t)}{I_0(A_t)} > 0 \quad t=1, \dots, T \quad (6)$$

where I_1/I_0 is the ratio of the modified Bessel functions of first and zeroth order. When the structure possesses a center of symmetry, the expectation of the cosine is (Woolfson, 1954):

$$\langle \cos \omega_t \rangle_t = \tanh(A_t/2) > 0 \quad t=1, \dots, T \quad (7)$$

4.2 Minimal principle

The minimal principle approach estimates the phases of the diffracted rays by solving a least squares optimization problem that requires the triplet invariants to be as close as possible to the theoretical prediction in (6) or (7). The optimization problem with respect to the triplet invariants and phases can be cast as follows (Debaerdemaeker and Woolfson, 1983; Hauptman, 1988; DeTitta et al., 1991, 1994).

Indices

m index used for reflections ($m = 1, \dots, M$)

t index used for triplet invariants ($t = 1, \dots, T$)

Variables

ϕ_m phase of the m th reflection

ω_t triplet invariant defined by $\omega_t = \phi_{m_t} + \phi_{m'_t} + \phi_{m''_t}$, where $\mathbf{h}_{m_t} + \mathbf{h}_{m'_t} + \mathbf{h}_{m''_t} = 0$

Parameters

M number of reflections

n number of atoms in the unit cell

T number of invariants

$|E_m|$ normalized structure-factor amplitude associated with reflection $\mathbf{h}_m$

A_t constant equal to $2n^{-1/2}|E_{m_t'}||E_{m_t''}|$

$\bar{\omega}_t$ conditional expected value of the cosine of the triplet invariant from the right-hand-side of (6) or (7)

Model MP

$$\min_{\omega} f(\omega) = \frac{\sum_{t=1}^T A_t \left(\cos \omega_t - \bar{\omega}_t \right)^2}{\sum_{t=1}^T A_t} \quad (8)$$

$$\begin{aligned} \text{s.t.} \quad & \omega_t = \phi_{m_t} + \phi_{m_t'} + \phi_{m_t''} \quad t=1, \dots, T \\ & \phi_m \in [0, 2\pi] \quad m=1, \dots, M \\ & \omega_t \in [0, 6\pi] \quad t=1, \dots, T \end{aligned} \quad (9)$$

In model MP, $f(\omega)$, the objective function in (8), is minimized subject to satisfying the relationships between phases and triplet invariants (9). This model constitutes the starting point of many popular approaches to the phase problem, including the crystallographic software Shake-and-Bake (SnB) (Miller et al., 1993). The challenges associated with MP are primarily two. First, because of nonconvexities in the objective function, this model possesses multiple local minima. Second, even if a global minimum is found, this does not necessarily solve the phase problem because the model does not enforce atomicity constraints. In other words, solutions of MP do not guarantee that no two atoms will coincide. Nonetheless, recent progress on the phase problem suggests that the modeling of phases with integer variables can facilitate the use of linear and integer programming approaches to resolve the multiple minima difficulty (Vaia and Sahinidis, 2003,2005; Smith et al., 2007; Smith, 2008), while the addition of constraints that enforce atomicity requirements via limiting electron density over a grid provides a complete formulation that leads to correct structures (Smith, 2008). Models and algorithms have been developed for centrosymmetric and non-centrosymmetric structures. For the former case, these developments have already led to software that is distributed to practitioners of crystallography (Xu et al., 2008).

5 Conclusions

The three problems discussed in this paper constitute three of the most important computational problems in the area of molecular structure and function elucidation. The protein side-chain conformation prediction problem is a subproblem of the protein folding problem, as well as an important problem on its own in the context of protein design. The protein structural alignment problem is key in terms of identifying structural and functional similarities of proteins, and can be used for structure comparisons as well as classification. Finally, the phase problem in crystallographic computing arises routinely in the use of diffraction measurements to determine structures of crystals.

All three problems represent approaches to different aspects of the puzzle of protein structure and function elucidation. All three problems are routinely approached via optimization techniques. The techniques outlined above rely on the solution of linear, integer, and nonlinear optimization problems with multiple local optima.

Significant progress has been made over the past decade towards the solution of these problems, and software implementations of highly successful algorithms are already in place. Nonetheless, new approaches are needed in order to solve larger versions of the underlying optimization models and increase the size of tractable structures. The astute reader must have noticed that posing the first two problems required virtually no mathematical notation and very little domain knowledge. This is not to imply that understanding of macromolecular interactions and atomic-level properties cannot be meaningfully employed to solve these problems. However, side-chain conformation prediction and structural alignment of proteins can be posed as graph problems that can be analyzed from the mathematical and optimization points of view, independent of the underlying biological problem. Thus, an entirely data-driven solution approach can be developed for these problems. This is not the case for the problem of using diffraction data to determine crystal structures. In the context of the latter problem, new formulations are still needed to bring diffraction physics under a complete optimization formalism that can be subsequently analyzed for developing suitable algorithms.

Acknowledgments

I am grateful to my former doctoral students A. B. Smith, A. Vaia, and W. Xie for extensive discussions on the topics addressed in this paper. Partial financial support from the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology under NIH award GM072023 is gratefully acknowledged.

References

- Althaus E, Kohlbacher O, Lenhof H, Müller P. A combinatorial approach to predict protein docking with flexible side chains. *Journal of Computational Biology* 2002;9:597–612. [PubMed: 12323095]
- Altschul SF, Gish W, Miller W. Basic local alignment search tool. *Journal of Molecular Biology* 1990;215:403–410. [PubMed: 2231712]
- Altschul SF, Schaffer AA. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 1997;25:3389–3402. [PubMed: 9254694]
- Bourne, PE.; Shindyalov, IN. Protein structure comparison and alignment. In: Bourne, PE.; Weissig, H., editors. *Structure Bioinformatics*. Wiley-Liss; 2003. p. 321-337.
- Bricogne G. Maximum entropy and the foundations of direct methods. *Acta Cryst* 1984;A40:410–445.
- Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Engineering* 2003;12:2001–2014.
- Caprara A, Carr R, Istrail S, Lancia G, Walenz B. 1001 optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap. *Journal of Computational Biology* 2004;11:27–52. [PubMed: 15072687]
- Caprara, A.; Lancia, G. Structural alignment of large-size proteins via Lagrangian relaxation. *Proceedings of the International Conference on Computational Biology (RECOMB)*; 2002. p. 100-108.
- Carr, RD.; Lancia, G.; Istrail, S. Branch-and-cut algorithms for independent set problems: Integrality gap and an application to protein structural alignment. 2000. Tech. rep., Sandia National laboratories, sandia Report SAND2000-2171
- Chazelle B, Kingsford C, Singh M. A semidefinite programming approach to side-chain positioning with new rounding strategies. *INFORMS Journal on Computing* 2004;16:308–392.
- Cochran W. Relations between the phases of structure factors. *Acta Cryst* 1955;8:473–478.
- Dahiyat BI, Mayo SL. Protein design automation. *Protein Sci* 1994;5:895–903. [PubMed: 8732761]
- Dahiyat BI, Mayo SL. De novo protein design: Fully automated sequence selection. *Science* 1997;278:82–87. [PubMed: 9311930]

- Debaerdemaeker T, Woolfson MM. On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement. *Acta Cryst* 1983;A39:193–196.
- Desmet J, De Maeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 1992;346:539–542.
- DeTitta, GT.; Miller, R.; Thuman, P.; Weeks, CM.; Hauptman, H.; Sabin, T.; Pagels, M. Parallel solutions to the phase problem in X-ray crystallography: An update. *IEEE Proceedings of the Sixth Distributed Memory Computing Conference*; 1991. p. 587–594.
- DeTitta GT, Weeks CM, Thuman P, Miller R, Hauptman HA. Structure solution by minimal-function phase refinement and Fourier filtering. I. Theoretical basis. *Acta Cryst* 1994;A50:203–210.
- Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins—Application to side-chain prediction. *Journal of Molecular Biology* 1993;230:543–574. [PubMed: 8464064]
- Elser V. X-ray phase determination by the principle of minimum charge. *Acta Cryst* 1999;A55:489–499.
- Eriksson O, Zhou Y, Elofsson A. Side chain-positioning as an integer programming problem. *Lecture Notes in Computer Science* 2001;2149:128–141.
- Germain G, Main P, Woolfson MM. On the application of phase relationships to complex structures II. Getting a good start. *Acta Cryst* 1970;B26:274–285.
- Gerstein, M.; Levitt, M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. In: States, D.; Agarwal, P.; Gaasterland, T.; Hunter, L.; Smith, R., editors. *Proceedings of international conference on intelligent systems in molecular biology*. AAAI Press; 1996. p. 59–67.
- Godzik A, Skolnick J. Flexible algorithm for direct multiple alignment of protein structures and sequences. *Computer applications in biosciences: CABIOS* 1994;10:587–596.
- Godzik A, Skolnick J, Kolinski A. A topology fingerprint approach to inverse protein folding problem. *Journal of Molecular Biology* 1992;227:227–238. [PubMed: 1522587]
- Goldman, D.; Papadimitriou, C.; Istrail, S. Algorithmic aspects of protein structure similarity. *Proceedings of the 40th annual symposium on foundations of computer science (FOCS)*; IEEE Computer society; 1999. p. 512–522.
- Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J* 1994;66:1335–1340. [PubMed: 8061189]
- Gordon DB, Hom GK, Mayo SL, Pierce NA. Exact rotamer optimization for protein design. *Journal of Computational Chemistry* 2003;23:232–243. [PubMed: 12497602]
- Gordon DB, Mayo SL. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *Journal of Computational Chemistry* 1998;19:1505–1514.
- Gordon DB, Mayo SL. Branch-and-terminate: A combinatorial optimization algorithm for protein design. *Structure* 1999;7:1089–1098. [PubMed: 10508778]
- Hauptman, HA. *Proceedings of the American Crystallographic Association Meeting*; Philadelphia. 1988. Abstract R4
- Hauptman, HA.; Karle, J. *ACA Monograph 3. Solution of the Phase Problem. I. The Centrosymmetric Crystal*. American Crystallographic Association; Michigan: 1953.
- Hellinga HW, Richards RM. Optimal sequence selection in proteins of known structure by simulated evolution. *Proceedings of the National Academy of Sciences USA* 1994;91:5803–5807.
- Holms L, Sander C. Protein-structure comparison by alignment of distance matrices. *Journal of Molecular Biology* 1993;233:123–138. [PubMed: 8377180]
- Karle J, Hauptman H. A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Cryst* 1956;9:635–651.
- Kingsford C, Chazelle B, Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* 2005;21:1028–1039. [PubMed: 15546935]
- Kohlbacher O, Lenhof H. BALL—Rapid software prototyping in computational molecular biology. *Bioinformatics* 2000;16. [PubMed: 10812473]
- Krasnogor, N.; Lancia, G.; Zemla, A.; Hart, W.; Carr, R.; Hirst, J.; Burke, E. A comparison of computational methods for the maximum contact map overlap of protein pairs. working paper. 2005. URL {<http://citeseer.ist.psu.edu/659931.html>}

- Lancia, G.; Carr, R.; Walenz, B.; Istrail, S. 101 optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem. *Proceedings of Annual International Conference on Computational Biology (RECOMB)*; 2001. p. 193-202.
- Lasters I, Desmet J. The fuzzy-end elimination theorem—Correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem. *Protein Engineering* 1993;6:717–722.
- Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* 1991;352:448–451. [PubMed: 1861725]
- Looger LL, Hellenga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structure genomics. *Journal of Molecular Biology* 2001;307:429–445. [PubMed: 11243829]
- Miller R, DeTitta GT, Jones R, Langs DA, Weeks CM, Hauptman H. On the application of the minimal principle to solve unknown structures. *Science* 1993;259:1430–1433. [PubMed: 8451639]
- Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol* 1970;48:443–453. [PubMed: 5420325]
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—A hierarchic classification of protein domain structures. *Structure* 1997;5:1093–1108. [PubMed: 9309224]
- Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: A more powerful criterion for dead-end elimination. *Journal of Computational Chemistry* 2000;21:999–1009.
- Pierce NA, Winfree E. Protein Design is *NP*-hard. *Protein Engineering* 2002;15:779–782.
- Ponder JW, Richards FM. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology* 1987;193:775–792. [PubMed: 2441069]
- Sahinidis, NV. Optimization techniques in molecular structure and function elucidation. In: Ierapetritou, M.; Bassett, M.; Pistikopoulos, EN., editors. *Proceedings of FOCAPO 2008*. CACHE Corporation; Austin, TX: 2008. p. 193-202.
- Shindyalov I, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* 1998;11:739–747. [PubMed: 9796821]
- Smith, AB. Ph.D. thesis. Department of Chemical and Biomolecular Engineering, University of Illinois; Urbana, IL: May. 2008 Optimization Techniques for Phase Retrieval based on Single-crystal X-ray Diffraction Data.
- Smith AB, Xu H, Sahinidis NV. An integer minimal principle and triplet sieve method for phasing centrosymmetric structures. *Acta Crystallographica A* 2007;63:164–17.
- Smith T, Waterman M. Identification of common molecular subsequences. *J. Mol. Biol* 1981;147:195–197. [PubMed: 7265238]
- Strickland DM, Barnes E, Sokol JS. Optimal protein structure alignment using maximum cliques. *Operations Research* 2005;53:389–402.
- Tramontano, A. *The Ten Most Wanted Solutions in Protein Bioinformatics*. Chapman & Hall/CRC; 2005.
- Vaia A, Sahinidis NV. An integer programming approach to the phase problem for centrosymmetric structures. *Acta Crystallographica A* 2003;59:452–458.
- Vaia A, Sahinidis NV. Polynomial-time algorithms for the integer minimal principle for centrosymmetric structures. *Acta Crystallographica A* 2005;61:445–4528.
- Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *Journal of Molecular Biology* 1995;249:816–831. [PubMed: 7602593]
- Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology* 2000;299:789–803. [PubMed: 10835284]
- Woolfson MM. The statistical theory of sign relationships. *Acta Cryst* 1954;7:61–64.
- Xie W, Sahinidis N. A reduction-based exact algorithm for the contact map overlap problem. *Journal of Computational Biology* 2007;14:637–654. [PubMed: 17683265]
- Xie W, Sahinidis NV. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* 2006;22:188–194. [PubMed: 16278239]

- Xu H, Smith AB, Sahinidis NV, Weeks CM. SnB version 2.3: Triplet sieve phasing for centrosymmetric structures. *Journal of Applied Crystallography* 2008;644–646. [PubMed: 18497890]
- Xu J. Rapid protein side-chain packing via tree-decomposition. RE-COMB 2005, Lecture Notes in Computer Science 2005;3500:423–439.
- Xu, J.; Jiao, F.; Berger, B. A parametric algorithm for protein structure alignment. In: Apostolico, A.; Guerra, C.; Istrail, S.; Pevzner, P.; Waterman, M., editors. RECOMB 2006, Lecture Notes in Computer Science. Vol. Vol. 3909. 2006. p. 488-499.
- Zemla A. LGA program: A method for finding 3d similarities in protein structures. *Nucleic Acids Research* 2000;31:3370–3374. [PubMed: 12824330]

Fig. 1.
Residue graph for protein side-chain problem

Fig. 2.
Unification of two residues in SCP

Fig. 3.
Outline of rotamer-residue-reduction algorithm

Fig. 4.
Symmetric units of 2NPV (left) and 2OKX (right)

Fig. 5.
Contact maps and maximal alignment for 2NPV (top) and 2OKX (bottom)