

**Protein Structure Determination Using Protein  
Threading and Sparse NMR Data**

Ying Xu, Dong Xu, Oakley. H. Crawford  
and J. Ralph. Einstein  
Life Sciences Division  
Computational Biosciences Section  
Oak Ridge National Laboratory  
Oak Ridge, TN 37831-6480

"The submitted manuscript has been authorized by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

To be presented and published as full paper *The Fourth Annual International Conference on Computational Molecular Biology*, Tokyo Japan, April 8-11, 2000.

---

\*Research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

# Protein Structure Determination using Protein Threading and Sparse NMR Data

(Extended Abstract)

Ying Xu, Dong Xu, Oakley H. Crawford, J. Ralph Einstein  
Computational Biosciences Section, Life Sciences Division  
Oak Ridge National Laboratory  
Oak Ridge, TN 37830-6480, USA, and

Engin Serpersu  
Department of Biochemistry, University of Tennessee  
Knoxville, TN 37996, USA

**Abstract** It is well known that the NMR method for protein structure determination applies to small proteins and that its effectiveness decreases very rapidly as the molecular weight increases beyond about 30 kD. We have recently developed a method for protein structure determination that can fully utilize partial NMR data as calculation constraints. The core of the method is a threading algorithm that guarantees to find a globally optimal alignment between a query sequence and a template structure, under distance constraints specified by NMR/NOE data. Our preliminary tests have demonstrated that a small number of NMR/NOE distance restraints can significantly improve threading performance in both fold recognition and threading-alignment accuracy, and can possibly extend threading's scope of applicability from structural homologs to structural analogs. An accurate backbone structure generated by NMR-constrained threading can then provide a significant amount of structural information, equivalent to that provided by the NMR method with many NMR/NOE restraints; and hence can greatly reduce the amount of NMR data typically required for accurate structure determination. Our preliminary study suggests that a small number of NMR/NOE restraints may suffice to determine adequately the all-atom structure when those restraints are incorporated in a procedure combining threading, modeling of loops and sidechains, and molecular dynamics simulation. Potentially, this new technique can expand NMR's capability to larger proteins.

**Keywords:** protein structure determination, NMR, protein threading, fold recognition, energy minimization.

**Corresponding author:** Ying Xu (xyn@ornl.gov).

# 1 Introduction

The NMR method for protein structure determination is based on (i) a network of distance restraints between spatially close hydrogen atoms, derived from nuclear Overhauser effects (NOEs), and (ii) restraints, calculated from scalar coupling constants, on the dihedral angles defined by the positions of atoms separated by three covalent bonds. The NOE restraints are essential to determine the secondary and tertiary structure of a protein, as they relate hydrogen atoms separated by less than about 5Å in amino acid residues that may not be adjacent in the protein sequence. An NMR structure is typically determined through molecular dynamics simulation/energy minimization under the constraints<sup>a</sup> specified by NMR restraints (Braun and Gö, 1985, Levy *et al.*, 1989, Brünger, 1992, Karimi-Nejad *et al.*, 1998). It typically requires about 15-25 NOE restraints per residue to obtain an accurate (mean) structure (equivalent to a 2-3Å Xray structure).

One problem with the NMR method is that it applies only to “small” proteins. Of the 1558 NMR structures in PDB (release of May 1999) (Bernstein *et al.*, 1977), only 25 are larger than 200 amino acids and the largest one has 269 residues (about 30 kD). This limitation is mainly caused by spectral data crowding and line broadening for larger proteins, which result in reduction in the fraction of spectral peaks that can be identified and assigned.

The goals of our current research are (i) to develop computational methods to fully utilize partial NMR data for protein structure calculation, and (ii) to expand the scope of the NMR method to larger proteins through the application of structural information obtained by a threading method. Our approach consists of two main steps: (a) *NMR-constrained* threading, and (b) *threading-supported* NMR method. In step (a), we construct a backbone structure of a query protein by using a threading method constrained by the NMR data. In step (b), we build a full-atom (or all heavy-atom) model of the query protein using molecular dynamics/energy minimization under the constraints of NMR data and the backbone structure obtained in step (a).

In the current study, we are focusing on the NOE restraints, and have formulated the NMR-constrained threading problem as to find the globally optimal threading under the residue-residue distance constraints specified by NOE data. This constrained threading problem is rigorously solved by a generalized version of our previously-published threading algorithm (Xu *et al.*, 1998). By applying this algorithm, we have demonstrated that a small number of NOE restraints can improve threading performance significantly in both fold-recognition and threading-alignment accuracy. In our preliminary tests, we were able to obtain backbone structures with an rmsd of 3-5Å in most cases by using 1-2 NOE restraints per residue.

While a 3-5Å backbone structure may be accurate enough for some functional inferences, it also provides valuable constraints for the full-atom NMR structure determination, and helps to reduce the number of NOE’s typically required for a NMR structure. Also, it helps to avoid entrapment in local minima in the NMR energy minimization procedure (as may occur when starting from a random backbone structure), and hence improve both calculation accuracy and efficiency.

We are currently exploring various NMR techniques to obtain as many NOEs as possible for large proteins, and investigating the potential of this new technique in helping to expand the scope of the NMR method to (significantly) larger proteins.

---

<sup>a</sup>In this paper, a *constraint* is used to refer to optimization algorithms and a *restraint* to NMR data.

## 2 NMR-Constrained Threading

### 2.1 An algorithm for NMR-constrained threading

#### Problem formulation

An *NMR-constrained threading problem* can be defined as to find the optimal<sup>b</sup> alignment between a query sequence and a template structure, which does not violate any distance constraints specified by NOEs. Or more specifically, it is to find an alignment  $((\bar{s}_1, \bar{t}_1), \dots, (\bar{s}_k, \bar{t}_k))$  between a query sequence  $s = s_1 \dots s_n$  and a template structure  $t = t_1 \dots t_m$  to minimize the following function<sup>c</sup>:

$$\sum_{1 \leq i \leq k} E_s(\bar{s}_i, \bar{t}_i) + \sum_{(\bar{t}_i, \bar{t}_j) \in \text{PAIRS}(t)} E_p(\bar{s}_i, \bar{t}_i, \bar{s}_j, \bar{t}_j) \quad (1)$$

**subject to:** if an NOE exists between  $\bar{s}_i$  and  $\bar{s}_j$  then distance  $(\bar{t}_i, \bar{t}_j) \leq D$ .

where  $\bar{s}_i$  (similarly  $\bar{t}_i$ ) is either an element of  $s$  (or  $t$ ) or an alignment gap;  $\max\{n, m\} \leq k \leq n + m$ ;  $E_s(x, y)$  is the *singleton fitness* term, measuring how well residue  $x$  fits the environment of structural position  $y$  if neither  $x$  nor  $y$  is a gap; otherwise it is a gap penalty;  $E_p(x_1, y_1, x_2, y_2)$  is the *pair contact* term, measuring how preferable to have residues  $x_1$  and  $x_2$  in nearby structural positions  $y_1$  and  $y_2$ ;  $\text{PAIRS}(t)$  denotes all the pairs of template positions between which contact potentials may need to be considered (in our current implementation, we use a distance cutoff,  $8\text{\AA}$ , between the  $C_\beta$  atoms to define this);  $D$  is a cutoff distance, and its default value in our program is  $8\text{\AA}$ .

In our current implementation, we assume that alignment gaps appear only in loop regions, and consider pair contacts only between core<sup>d</sup> residues. Also we consider only NOEs between residues that are aligned to core elements of the template.

#### Our threading algorithm

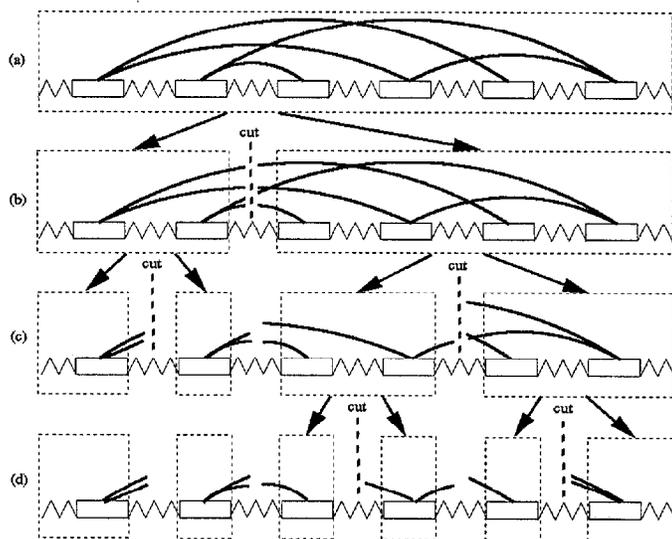
We have previously developed a threading algorithm which guarantees to find an optimal threading as defined in (1) (without the constraints) (Xu *et al.*, 1998), and have implemented the algorithm as a computer program, called PROSPECT (Xu and Xu, 1999). We now give a brief review on how PROSPECT deals with pair contacts, which provides the basic algorithmic framework for handling NOE restraints.

The threading algorithm employs a *divide-and-conquer* strategy for solving the optimal threading problem. It first pre-processes the template structure by repeatedly dividing (bi-partitioning) it into sub-structures until each sub-substructure contains one core secondary structure (see Figure 1). The basic operation of the algorithm is *to calculate the optimal threading score between a sub-structure with links (to the rest of the structure) and a sub-sequence, under the condition that the core secondary structure at the external end of each link is already aligned*. This operation is implemented (recursively) by finding the optimal threading score between even smaller sub-structures and sub-sequences, and combining them optimally. We use Figure 2 (a) to illustrate how to calculate the optimal threading score between  $s[k1, k2]$  and  $t[i, j]$  with links  $\{o1, o2, o3, o4\}$ . We have proved (Xu *et al.*, 1998) that the optimal alignment between  $s[k1, k2]$  and  $t[i, j]$  can be

<sup>b</sup>Throughout this paper, the optimal threading means the globally optimal threading.

<sup>c</sup>Our actual alignment function uses an affine function to penalize alignment gaps. This simplified version is used here to simplify the description of the algorithm.

<sup>d</sup>A core secondary structure is an  $\alpha$ -helix or a  $\beta$ -strand.



**Figure 1:** (a) A template with six core secondary structures and connecting loops. A link between two cores represents that there exists at least one pair contact between the two cores. (b) A cut dividing the template into three parts: two sub-templates inside the left and right dotted boxes, respectively, and the connecting loop. (c) and (d) Further partition of the template.

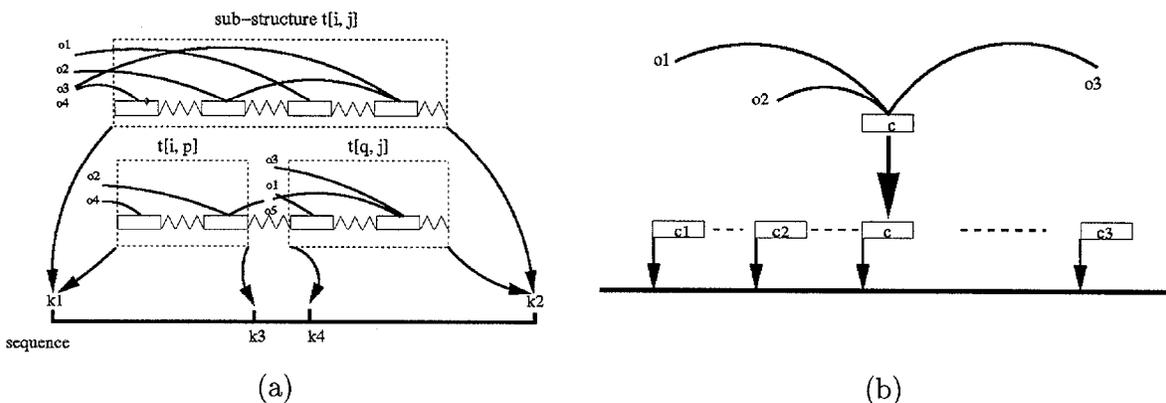
constructed by appending the optimal alignments between (i)  $t[i, p]$  and  $s[k1, k3]$ , (ii)  $t[q, j]$  and  $s[k4, k2]$ , and (iii)  $t[p + 1, q - 1]$  and  $s[k3 + 1, k4 - 1]$ , for some  $k3, k4 \in [k1, k2]$  and some alignment of the core linked by  $o5$ . The optimal alignments (i) and (ii) can be calculated recursively using the same operation on their sub-structures and sub-sequences; and the optimal alignment (iii) can be calculated using a Smith-Waterman type sequence alignment program (note that no pair contacts are considered between loop elements). To determine which of these optimal alignments give the optimal alignment between  $t[i, j]$  and  $s[k1, k2]$ , we need to go through all possible values of  $k3, k4 \in [k1, k2]$  and all possible alignments of the core linked by  $o5$ , and choose the (combined) optimal one. For more details of the algorithm, we refer the reader to (Xu *et al.*, 1998).

Our threading algorithm applies this basic operation, starting with the whole query sequence and whole template structure, and continuing until each sub-structure contains one core secondary structure. Pair contact potentials are calculated when links are considered. Since no alignment gaps are allowed within a core, the contact potential between any pair of residues aligned to nearby structural positions can be calculated based on the starting alignment positions of the corresponding cores.

### Dealing with NOE restraints

Based on the above discussion, we can rigorously prove that if no core alignment with links violates any NOE restraints then the optimal alignment between the whole template and the whole query sequence will not violate any NOE restraints. Our NMR-constrained threading algorithm uses the algorithmic framework outlined above with one addition that checks for any NOE violations when doing a core alignment.

We use Figure 2(b) to explain how to conduct such a check when doing a core alignment. Let  $c$  be a core with links  $\{o1, o2, o3\}$ , and these links connect to cores  $c1, c2$ , and  $c3$ , respectively.



**Figure 2:** (a) A schematic of the basic operation of the divide-and-conquer algorithm. (b) Core alignments.

For each possible alignment of these cores  $\{c1, c2, c, c3\}$ , we check if any NOEs are violated. That is, we check whether, for any NOE that relates two residues aligned to some of these cores, the corresponding distance is more than  $D$  (see the objective function (1)). If a violation is found, this particular alignment will not be further considered when building larger alignments in our divide-and-conquer algorithm.

One way to implement this is to go through the list of all NOEs each time we examine a new arrangement of core alignments. But this simple strategy may not be computationally feasible when there are hundreds of NOEs or more. The running time of our threading algorithm is essentially determined by the number of alignments we have to consider. To examine hundreds to thousands of NOEs for each such alignment may increase the running time of the algorithm by that many times. Our solution to this problem is to examine only the relevant NOEs when examining a particular arrangement of core alignments.

While examining NOEs may increase the computational time, the NOEs can also help to reduce the search space size and hence the computational time, as they will rule out alignments violating NOEs without explicitly examining them. The total effect on computational time of using the NOEs is a rather complicated issue, and will not be addressed further in this abstract. We now present an algorithm for finding all relevant NOEs, given a core and its links.

We can formulate this problem as follows. Let  $\mathcal{N} = \{(l_1, r_1), \dots, (l_p, r_p)\}$  be the list of NOE restraints with  $l_i$  and  $r_i$  being the left and right positions (in the query sequence) of the  $i^{th}$  NOE. Let  $\mathcal{A} = \{(c_l^1, c_r^1), \dots, (c_l^q, c_r^q)\}$  be a list of left and right aligned positions of cores<sup>e</sup>  $c^1, \dots, c^q$ . The problem is to find all  $(l_i, r_i)$  such that  $c_l^j \leq l_i \leq c_r^j$  and  $c_l^k \leq r_i \leq c_r^k$ , for some  $j, k \in [1, q]$ . To do this fast, we pre-process the NOE restraint list  $\mathcal{N}$  to facilitate fast searches. We use a *segment tree* data structure (Preparata and Shamos, 1985) to achieve this.

A segment tree is defined on a list of consecutive integers  $[1, p]$  (from 1 to  $p$ ). A segment tree consists of  $p \log(p)$  *standard intervals* of  $[1, p]$ , e.g., an interval could be  $[5, p - 3]$ . A key property of these specially chosen intervals is that any arbitrary interval  $[a, b] \subseteq [1, p]$  can be represented as the union of at most  $\log(p)$  non-overlapping standard intervals.

<sup>e</sup> $q$  is typically a small number ( $\leq 6$ ) as a core is typically “in contact” with (at most) 4 - 5 other cores.

First we sort all NOE restraints in the increasing order of their left positions, i.e.,  $\{l_i\}$ 's, and represent this sorted list as a segment tree data structure. For each standard interval, we sort its NOEs by their right positions, i.e.,  $\{r_i\}$ . We use the following procedure to find all relevant NOEs for a given arrangement,  $\mathcal{A}$ , of core alignments.

---

**Procedure SEARCH\_NOE** ( $\mathcal{N}, \mathcal{A}$ )

1. **for**  $1 \leq a \leq q$  **do**
  2.     find left and right boundaries,  $l$  and  $r$ , of  $\mathcal{N}$ 's sublist whose left positions are within  $[c_l^a, c_r^a]$  of  $\mathcal{A}$ ;
  3.     retrieve the  $v$  standard intervals of  $[l, r]$ ;
  4.     **for** each of the  $v$  standard intervals,  $\mathcal{J}$  **do**
  5.         find left and right boundaries of  $\mathcal{J}$ 's sublist whose right positions are within  $[c_l^b, c_r^b]$ , for some  $b \in [a, q]$
  6.         "output" all NOEs within the boundary.
- 

Note that the algorithm spends  $O(1)$  time on each relevant NOE, and in addition it spends  $O(\log^2(p))$  time to search the segment tree and  $O(q)$  time to go through all the involved cores. So the total time spent is  $O(q + e + \log^2(p))$ , compared to  $O(q + p)$  time using the straightforward method, where  $e$  denotes the number of relevant NOEs for a given arrangement of core alignments.

## 2.2 A preliminary study on NMR-constrained threading

NMR-constrained threading uses an additional scoring term  $E_{\text{NMR}}$  to reward the use of NOEs and to penalize deviations from the NOE-specified distance within the cutoff distance  $D$ . We use two types of distance restraints: (a) NOE restraints between backbone hydrogens, and (b) estimated  $C_\beta$  distance restraints based on NOEs involving sidechains. An estimated  $C_\beta$  distance restraint is used only when it is at most  $7\text{\AA}$ .  $E_{\text{NMR}}(x, y)$  will be applied<sup>*f*</sup> only when an NOE's two ends are aligned to structural positions  $x$  and  $y$  and satisfy the  $C_\beta$  distance cutoff.

$$E_{\text{NMR}}(x, y) = \begin{cases} -150, & \text{if distance between } x \text{ and } y \text{'s backbone hydrogens} \\ & \leq 3.7\text{\AA}, \text{ and an NOE aligned to } x, y \text{'s hydrogens;} \\ -300, & \text{if } C_\beta\text{-distance, } D_{cb}(x, y), \text{ between } x \text{ and } y \leq 7\text{\AA}; \\ 300 \times (D_{cb}(x, y) - 7)^2, & \text{if } 7\text{\AA} < D_{cb}(x, y) \leq 8\text{\AA}. \end{cases} \quad (2)$$

Using this function, we have conducted preliminary analyses on (i) how NOEs affect the threading-alignment accuracy; and (ii) how NOEs affect the accuracy of fold recognition by our NMR-constrained threading algorithm. For this study, we selected from the FSSP database (Holm and Sander, 1996) 17 sequences as the query sequences and 667 proteins as the template structures. The query sequences and templates are selected randomly within the following constraints: (1) the query sequences have NOE data in the PDB database; (2) the query sequences should evenly represent three classes of proteins: (2a) all- $\alpha$ , (2b) all- $\beta$ , and (2c)  $\alpha$  and  $\beta$  mixed; and (3) each query sequence has a native-like structure in the template set.

Table 1 summarizes how the number of NOEs affects the performance of our algorithm. In this study, NOEs are selected randomly (and uniformly) from the corresponding PDB files. The

---

<sup>*f*</sup>  $E_{\text{NMR}}(x, y)$  may be used twice if both their hydrogen distance and  $C_\beta$  distance satisfy the conditions of (2).

alignment accuracy is the highest accuracy over ten runs, and the fold recognition accuracy is based on a single run<sup>9</sup>. When using 2 NOEs per residue in this test, the threading-alignment accuracy (the number of residues aligned within a 4-residue shift from the correct positions versus the total number of alignable residues) improved from 70% to 92.7%.

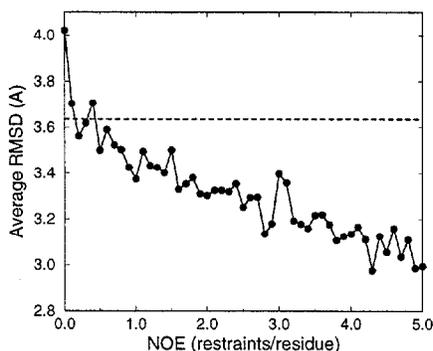
**Table 1. Threading accuracy versus number of NOE restraints.**

class	query	nres (a.a.)	temp	iden (%)	rmsd (Å)	RMSD (Å)/rank vs. NOE/a.a.						
						0	0.5	1.0	1.5	2.0	2.5	3.0
$\alpha$	1bbn	133	1cnt1	8	2.6	<b>16.2/29</b>	6.6/-	6.6/1	6.6/-	5.3/1	5.3/-	<b>5.3/1</b>
	1itl	130	3inkC	12	2.2	<b>13.1/7</b>	4.3/-	4.2/1	4.2/-	3.6/3	3.6/-	<b>3.6/3</b>
	1ner	74	1lmb3	14	2.4	<b>3.8/57</b>	3.8/-	3.8/57	3.8/-	3.8/49	3.8/-	<b>3.8/39</b>
	1il6	166	1bgc	15	2.1	3.0/1	3.0/-	3.0/1	3.0/-	3.0/1	3.0/-	3.0/1
	1ocd	104	1cyj	28	2.0	3.4/1	3.4/-	3.4/1	3.4/-	3.4/1	3.4/-	3.4/1
$\beta$	1maj	113	1agdA	7	3.0	<b>8.8/37</b>	8.0/-	6.9/1	6.9/-	6.6/1	6.6/-	<b>6.6/1</b>
	1nct	98	1bec	12	2.5	<b>14.7/1</b>	13.8/-	5.4/1	5.4/-	5.4/1	5.4/-	<b>5.4/1</b>
	1bla	155	1hce	14	2.1	<b>7.7/1</b>	4.0/-	4.0/1	4.0/-	4.0/1	4.0/-	<b>4.0/1</b>
	1vhp	117	1cd8	15	2.9	<b>4.5/1</b>	4.4/-	4.4/1	4.4/-	4.4/1	4.4/-	<b>4.4/1</b>
	1ghj	79	1iyu	24	1.8	2.0/1	2.0/-	2.0/1	2.0/-	2.0/1	2.0/-	2.0/1
	1a7i	60	1qli	33	2.5	<b>2.8/1</b>	2.8/-	2.8/1	2.6/-	2.6/1	2.6/-	<b>2.6/1</b>
$\alpha/\beta$	1afi	72	2acy	7	2.8	<b>6.8/376</b>	5.5/-	5.6/26	5.5/-	3.8/3	3.8/-	<b>3.8/2</b>
	3trx	105	1a8y	12	2.7	<b>4.1/1</b>	3.1/-	3.1/1	3.1/-	3.2/1	3.0/-	<b>3.0/1</b>
	1ikm	69	1dokB	13	2.1	<b>2.1/2</b>	2.1/-	2.1/2	2.1/-	2.1/1	2.1/-	2.1/1
	3phy	125	1bv6	15	2.3	<b>12.2/41</b>	8.8/-	3.6/31	3.6/-	3.6/6	3.6/-	<b>3.6/4</b>
	1crp	166	1byuB	24	1.7	2.6/1	2.6/-	2.6/1	2.6/-	2.6/1	2.6/-	2.6/1
	1fht	116	2ula	34	2.1	<b>4.5/1</b>	4.5/-	4.5/1	4.5/-	4.5/1	4.5/-	<b>2.1/1</b>

“Query” and “temp” represent the PDB codes of the query and template proteins, respectively. “Nres” is the number of alignable residues between the query and template. “Iden” denotes the sequence identity between the query and template sequences. “Rmsd” is the C $\alpha$ -RMSD between the structurally equivalent residues of the query and template structures. “RMSD/rank vs. NOE” are the C $\alpha$ -RMSD between the experimental structure and the predicted structure (alignable portions), and the rank of the correct template structure among 667 templates. 0, 0.5, 1.0, ... represent the averaged number of NOEs used per residue. “-” indicates that no fold recognition test is conducted.

We have also used simulated data to test our program on a number of large proteins to show that it is applicable to large proteins. We now outline one such test. The query protein is 1klna (an X-ray structure of DNA polymerase I with 605 residues), and the template is 1t7pa (T7 DNA polymerase with 698 residues). Without using any NOE restraint, PROSPECT aligned 145 residues correctly (within 4-residue shift to the correct positions) among the 405 structurally alignable ones. We then generated all the hydrogen atoms based on the coordinates of the heavy atoms using X-plor (Brünger, 1992), and constructed distance restraints for all hydrogen pairs within 4.0 Å. Then we randomly and uniformly selected 1.0 distance restraint per residue. Using these simulated restraints, our constrained threading program aligned 219 residues correctly.

<sup>9</sup>We only performed a single run for this abstract due to the tight submission deadline. More runs will be performed before the publication of this work. The small sampling size (single run) partially explains the performance fluctuation in our fold recognition test.



**Figure 3:** The average RMSD of all heavy atoms between the model and the X-ray structure (2igd) over ten runs versus the number of NOE restraints per residue used. Each solid dot represents the structure accuracy for the corresponding number of NOEs used.

### 3 Threading-Supported NMR Method

We have conducted a preliminary study on how an approximate backbone structure predicted by NMR-constrained threading can help reduce the number of NOEs required for accurate structure determination. We now outline our study result on the third IGG-binding domain of protein G (with 61 amino acids).

This protein has both an NMR structure (2igh) and a high-resolution (1.1Å) X-ray crystallographic structure (2igd). The RMSD of all heavy atoms between 2igh and 2igd is 3.6Å. 2ptl is the template structure. The 2ptl and 2igh sequences have 17% identity; and the  $C_{\alpha}$ -RMSD is 4.0Å between their aligned portions. Our threading program finds the alignment between the two correctly. We then applied MODELLER (Sali and Blundell, 1993) to generate an all-atom structure of 2igh based on the structure of 2ptl and NOE restraints. The NOEs used in this test are selected randomly and uniformly from the whole set of NOEs. Ten runs are performed and the averaged structure accuracy is plotted in Figure 3. The structure is generally becoming more accurate as the number of NOEs increases. The small fluctuations in the averaged RMSD are presumably due to the small sampling size (ten) for averaging.

The above examples show that NMR-constrained threading, followed by modeling of side chains and possibly loops, gives an approximate representation of the true target structure. We are investigating a method for improving the accuracy of the determined structure by minimizing an objective function consisting of a linear combination of the conformational energy (the CHARMM22 potential (MacKerrell, Jr., et al., 1998)) and a pseudo-energy of violation of NMR restraints. Currently we implement this with simulated annealing by molecular dynamics, using the CNS program (Brünger *et al.*, 1998). In this method, numerous calculations are performed, taking the approximate structure determined above as the starting point, but with different random atomic velocities. We have found that almost every single such calculation performed in our preliminary investigation results in a reduced rms difference between the model and the true structure.

## 4 Discussion

To fully take advantage of the capability of this new technique, we have considered how to obtain as many NOEs as possible using various NMR techniques. Our ultimate goal is to expand the scope of the NMR method to significantly larger proteins. Currently we are working on a 46 kD enzyme PGK (yeast phosphoglycerate kinase) for its structure determination. We use PGK as an example to briefly explain how we will extract NOE data from this large protein.

We will first conduct multidimensional NMR experiments using uniformly  $^{13}\text{C}/^{15}\text{N}$ -labeled enzyme. We expect to make assignments for a limited number of residues. Distance restraints will be derived from  $^{15}\text{N}$  and  $^{13}\text{C}$  edited  $^1\text{H}$ - $^1\text{H}$  NOE experiments. Particular attention will be directed to observe NOEs involving protons of the protein backbone. We believe that this should yield a limited number of distance restraints. If the uniformly labeled PGK fails to provide sufficient number of distance restraints for computation, we will then use selectively isotope-labeled PGK to obtain additional distance restraints by isotope filtered NOE experiments. We have developed procedures for specific labeling of this enzyme at designated residues such as histidines or tyrosines (Pappu and Sempersu, 1994). In addition, we plan to supplement the distance restraints with longer distances determined by paramagnetic probe-T1 method (Mildvan and Gupta, 1978). Earlier, we were able to determine distances up to 14Å using paramagnetic CrATP in selectively isotope labeled PGK (Pappu and Sempersu, 1994).

Our preliminary study has strongly suggested that a small number of NOEs can help extend the scope of threading to structural analogs. As in the cases of 1afi-2acy, 3phy-1bv6, and 1bla-1hce (all are analogous pairs with low sequence identities; see Table 1), our program was able to achieve high performance on both fold recognition and threading alignment.

To summarize, we have demonstrated that (1) a small number of NOEs can significantly improve the threading performance, and (2) the use of threading can greatly reduce the requirement of NOEs for an accurate NMR structure determination. We expect this approach to be extremely useful in cases where experimental procedures can provide only incomplete NMR data. It should also be useful even when structure determination by NMR methods is feasible, by allowing substantial reductions in the number of labeling experiments and the NMR-data collection time – i.e., by achieving equivalent results more rapidly.

## Acknowledgements

We thank Dr. Ed Uberbacher for his helpful discussion related to this work. This research was sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation. Engin Sempersu's work was supported by Petroleum Research Fund (PRF# 32874-AC4).

## References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: a computer

- based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
- Braun, W. and Gö, N. (1985). Calculation of protein conformations by proton-proton distance constraints: a new efficient algorithm. *J. Mol. Biol.* **186**, 611 – 626.
- Brünger, A. T. (1992). *X-PLOR, Version 3.1, A System for X-ray Crystallography and NMR*. The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven.
- Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L., Simonson, T., and Warren, G. L. (1998). Crystallography and NMR system (CNS): a new software system for macromolecular structure determination. *Acta Cryst.* **D54**, 905–921.
- Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–602.
- Karimi-Nejad, Y., Warren, G. L., Schipper, D., Brünger, A. T., and Boelens, R. (1998). NMR structure calculation methods for large proteins - application of torsion angle dynamics and distance geometry/simulated annealing to the 269-residue protein serine protease PB92. *Mol. Phys.* **95**, 1099–1112.
- Levy, R. M., Bassolino, D. A., Kitechen, D. B., and Pardi, A. (1989). Solution structures of proteins from NMR data and modeling: alternative folds for neutrophil peptide 5. *Biochemistry*, **28**, 9361 – 9372.
- MacKerrell, Jr., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem.* **B102**, 3586–3616.
- Mildvan, A. S. and Gupta, R. K. (1978). Nuclear relaxation measurements of the geometry of enzyme-bound substrates and analogs. *Methods in Enzymology*, **46G**, 322–359.
- Pappu, K. M. and Serpersu, E. H. (1994). Proton NMR studies of a large protein. pH, substrate titrations, and NOESY experiments with perdeuterated yeast phosphoglycerate kinase containing  $^1\text{H}$  histidine residues. *J. Magn. Reson. (Series B)*, **105**, 157–166.
- Preparata, F. P. and Shamos, M. I. (1985). *Computational Geometry: An Introduction*. New York: Springer-Verlag.
- Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
- Xu, Y. and Xu, D. (1999). Protein threading using PROSPECT: Design and evaluation. Submitted.
- Xu, Y., Xu, D., and Uberbacher, E. C. (1998). An efficient computational method for globally optimal threading. *J. Comp Biol.* **5** (3), 597–614.