SLAC–PUB–9662
March 13, 2003

Protein Structure: Alignment using Mean Field Techniques and Measurement of Isolated Individual Molecules [1]

Richard Blankenbecler [2]
Theory Group

Stanford Linear Accelerator Center
P.O. Box 4349, Stanford
CA 94309, USA

Abstract:

Techniques originally developed in High Energy Physics have been applied to selected problems in genetics with promising results.

First, this talk will briefly review the importance of protein structure from a physics point of view. Then Mean Field Techniques used in detector track fitting algorithms will be applied to the comparison of protein structures. The practical importance of such comparisons will be discussed.

Second, the possibility of measuring the charge structure of "single" isolated molecules using the proposed SLAC Free Electron Laser will be outlined. This involves the development of an algorithm that determines the orientation of each of the many targeted identical molecules, constructs the 3-D transform from the many 2-D patterns, and finally performs an inverse fourier transform when only the magnitude of the transform is known, since the phase is not measurable.

Talk presented at the
Seventh Workshop on Nonperturbative QFT
Villefranche-sur-Mer, France, January 6-10, 2003.
Proceedings to be published by World Scientific.

[1] Work supported by the Department of Energy, contract DE–AC03–76SF00515.
[2] rzbth@slac.stanford.edu

1 Introduction and Motivation

Before starting on the physics and mathematical treatment of the aspects of protein structure that are of interest here, I first must give some motivation as to why it is important. Since the human genome has been completely mapped, are there any more major unknowns? I understand that at a recent proteomics meeting there was a sign that read something like "Human genome mapped – now the real work begins!".

What I want to bring to your notice is that the structure of a protein is essential to its function and understanding proteins involves real work. The amino acid sequence of the protein is determined by the appropriate genetic sequence and this sequence determines how the protein folds. However it is the final shape or structure of the protein that determines whether it operates properly. Consider the following example. There is a gene regulatory protein called $\alpha 2$ which performs the same task in the cells of both yeast and the insect drosophila, for example. These two are separated by over a *billion* years of evolution. The $\alpha 2$ protein contains 58 amino acids in yeast and 55 in the fly. A direct comparison of the two amino acid sequences shows that only 17 of the 58 amino acids match. If the sequences are so different, how can they have the same function? When the structure of these were determined, they were found to be essentially identical, see Figure 1. The 3 extra residues were isolated in a loop that connects two alpha helices. This is an example of the

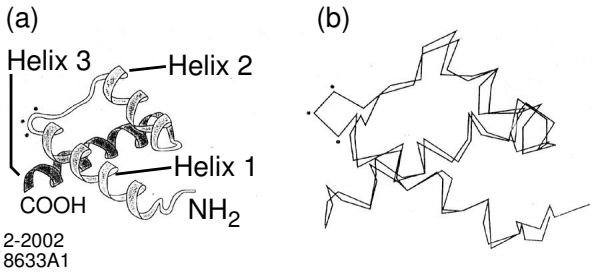

Figure 1: **(a)** The 3-D structures of the $\alpha 2$ protein from yeast and a drosophila. **(b)** A 2-D projection of the superposed proteins.

fact that the structure (shape) of a protein plays an crucial role in defining its function and in its ability to function.

I will assume that you are familiar with the general operation in a living cell of DNA/mRNA/tRMA, etc. Recall that the mRNA emerges from the cell nucleus, finds a ribosome to read its code and a protein chain is produced.

Rates: The ribosome reads the code and adds the required amino acid residue (only 20 different varieties are used in biological systems) in roughly 200 milliseconds. Thus a (short) protein with 300 residues requires ~ 1 minute to be produced. A long (human) muscle protein with 10,000 residues requires ~ 30 minutes!

Folding: The folding of a protein is not well understood and is one of the great unsolved problems. Note however that the problem is very complicated. Since the final shape of the protein is important, nature has provided *chaperons* to assist the proteins in folding properly in the crowded thermal environment of the cell.

Misfolding: The misfolding of a protein can lead to serious consequences. Some misfoldings are innocuous. The cell may simply be starved since the protein cannot perform its proper function. Other misfoldings are toxic. For example, Alzheimer's disease, Cystic fibrosis, Huntington's disease, Parkinson's disease, certain heart diseases and mad cow disease are only a few of the tragic maladies caused by misfolded proteins.

Since protein structure is evidently very important to its function, let us turn to the problem of comparing the shape of two different proteins. The measure of similarity is not obvious, since the proteins may contain a different number of amino acids, and no two molecular structures are likely to be geometrically identical. Finding similarly shaped, and perhaps similar functioning, proteins

are important to the discovery of new drugs.

2 Comparing Proteins

The work [1] in this section was carried out in collaboration with Mattias Ohlsson, Carsten Peterson, and Markus Ringnér of the Complex Systems Division, Department of Theoretical Physics, Lund University, Lund, Sweden.

As has been argued above, it is important to be able to perform detailed protein structure alignment. Previous work in this area can be found in [2] and [3], where other references can be found. Structure alignment enables the study of functional relationships between proteins and is very important for homology and threading methods in structure prediction. Furthermore, grouping protein structures into fold families and subsequent tree reconstruction may shed light on ancestry and evolutionary issues. Nature reuses successful shapes for new purposes.

Structure alignment amounts to matching two 3D structures such that potential common substructures, e.g. α-helices, have priority. The latter is accomplished by allowing for gaps in either of the chains. The key ingredients of our approach are: an error function formulation of the problem simultaneously in terms of binary (Potts) assignment variables and real-valued atomic coordinates, and a minimization of the error function by an iterative method. Each iteration contains two steps: a fuzzy weight matrix is minimized with respect to the assignment variables, and an exact rotation and translation of coordinates weighted with the corresponding assignment variables. The problem is to find the minimum of the energy, or cost function,

$$E_{chain} = \sum_{i=1}^{M} \sum_{j=1}^{N} W_{i,j}\ d_{i,j} \quad \text{with} \quad d_{i,j} = (\mathbf{a} + \mathcal{R}\mathbf{x}_i - \mathbf{y}_j)^2 \ . \tag{1}$$

The elements of the fuzzy matching matrix W are defined such that $0 < W_{i,j} < 1$ is the probability

that atom i in the first chain should be matched to atom j in the second. We anticipate that $W_{i,j} \sim 1$ if the pair (i,j) can be made spatially close by a choice of the displacement $\mathbf{a}$ and the rotation $\mathcal{R}$ and $W_{i,j} \sim 0$ otherwise.

We will briefly rephrase the Needleman-Wunsch algorithm[2] in an implementation that will allow for a straightforward introduction of our approach. Let $\mathbf{X} = (X_1 X_2 \ldots X_M)$ and $\mathbf{Y} = (Y_1 Y_2 \ldots Y_N)$ denote the two chains containing M and N residues in a *dot-matrix* representation, respectively. Every possible alignment of the two chains (not including permutations of atoms in a chain) can be represented as a directed path on the $(M+1)\times(N+1)$ alignment matrix (Figure 2a). Each dot

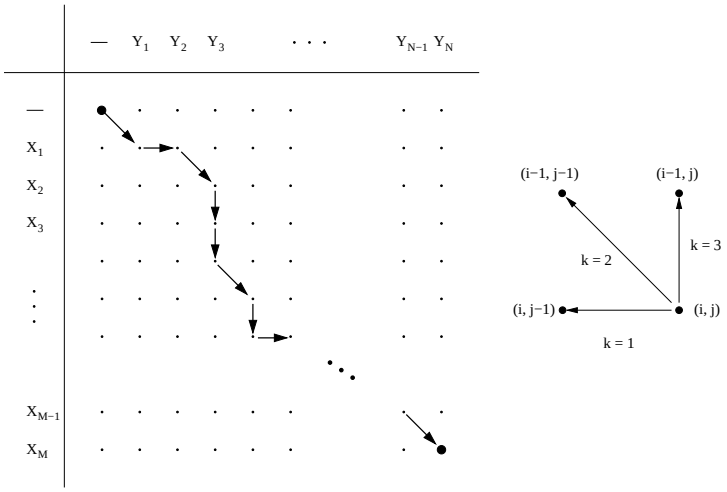

Figure 2: **(a)** The alignment matrix for an alignment between the two chains $\mathbf{X} = (X_1 X_2 ... X_M)$ and $\mathbf{Y} = (Y_1 Y_2 .... Y_N)$. **(b)** Unit vectors connecting to the three possible predecessors to a dot (i,j).

(i,j) has, excluding obvious boundary restrictions, three possible predecessors along the alignment path, which are denoted by $k = 1, 2, 3$ (see Fig. 2b)[10]. As a mnemonic, identify the step directions as k=1=horizontal, k=2=diagonal, and k=3=vertical.

The *alignment cost* $\mathcal{D}_{i,j}$ for the optimal alignment of sub-chains $(X_1 X_2 ... X_i)$ and $(Y_1 Y_2 ... Y_j)$ is

given by

$$\mathcal{D}_{i,j} = \min_k \{\widetilde{\mathcal{D}}_{i,j;k}\} \;=\; \sum_k s_{i,j;\,k}\, \widetilde{D}_{i,j;\,k} \;, \tag{2}$$

where $\widetilde{\mathcal{D}}_{i,j;k}$ is the alignment cost if the alignment path is forced to pass through the preceding node given by k and

$$s_{i,j;\,k} = \begin{cases} 1 & \text{if } \widetilde{D}_{i,j;\,k} = \min_{k'}\{\widetilde{D}_{i,j;\,k'}\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We next introduce fuzzy alignment paths that will finally lead to a fuzzy matching matrix. We replace the binary variable $s_{i,j;\,k}$ by the continuous variable $v_{i,j;\,k}$, with the property that $\sum_k v_{i,j;\,k} = 1$. This allows for the interpretation that $v_{i,j;\,k}$ is the probability that an optimal alignment path that passes through (i,j) also passes through the preceding node specified by k. The replacement

$$s_{i,j;\,k} \to v_{i,j;\,k} = \frac{e^{\widetilde{D}_{i,j;\,k}/T}}{\sum_{k'} e^{\widetilde{D}_{i,j;\,k'}/T}} \;. \tag{4}$$

can be viewed as a soft implementation of the 'min' function in Eq. (2) where the parameter $T > 0$ controls the fuzziness.

We restrict ourselves to position dependent linear gap penalties of the following type,

$$\lambda_a^{(n)} + (l-1) l_{ext} \;, \tag{5}$$

where $\lambda_a^{(n)}$ is the penalty for opening a gap in chain n at position a, l_{ext} is the extension penalty and l is the gap length. The problem can now be expressed in terms of $v_{i,j;\,k}$, and the $\widetilde{\mathcal{D}}_{i,j;k}$ can be calculated using the following recursive relation,

$$\begin{aligned} \widetilde{D}_{i,j;\,k=1} &= D_{i,j-1} + \lambda_{j-1}^{(2)}(1 - v_{i,j-1;1}) + l_{ext} v_{i,j-1;1} \;, \\ \widetilde{D}_{i,j;\,k=2} &= D_{i-1,j-1} + d_{i,j} \;, \\ \widetilde{D}_{i,j;\,k=3} &= D_{i-1,j} + \lambda_{i-1}^{(1)}(1 - v_{i-1,j;3}) + l_{ext} v_{i-1,j;3} \;. \end{aligned} \tag{6}$$

The optimal alignment cost at node (i, j) is

$$D_{i,j} = \sum_k v_{i,j;\ k}\, \widetilde{D}_{i,j;\ k}\ . \tag{7}$$

From the probabilities $v_{i,j;\ k}$ it is straightforward to calculate $P_{i,j}$. The probability $P_{i,j}$ that node (i, j) is part of the optimal path can be calculated with a recursion relation. With the obvious initial value $P_{MN} = 1$ one finds

$$\begin{aligned} P_{i,j} = \ &= \ v_{i,j+1;1} P_{i,j-1} \\ &+ \ v_{i+1,j+1;2} P_{i-1,j-1} \\ &+ \ v_{i+1,j;3} P_{i+1,j}\ . \end{aligned} \tag{8}$$

By construction this leads to the necessary condition $P_{1,1} = 1$. Finally, the fuzzy matching matrix can now be calculated as

$$W_{i,j} = P_{i,j} v_{i,j;\ 2}\ . \tag{9}$$

We have tested and compared our alignments of protein pairs with results from other common procedures on a wide variety of protein families with good results. The advantages of the method include a probabilistic interpretation of the match, a local reliability index for the match of each residue pair, fast convergence, and the ability to add additional user constraints of a general form.

3 Measuring Individual Molecules

The problem of reconstructing the charge distribution of an object from the measured magnitude of its scattering fourier transform in the Fraunhofer diffraction regime is discussed here. This, in turn, involves reconstructing the unknown phases based on general properties of the image such as positivity and finite extent[4]. The terms object, image, and charge distribution will be used interchangeably in this note.

The determination of structure from a set of patterns measured from a 3-D object at known orientations has been well discussed and has been treated by a number of authors. Papers on this topic can be found in references [5] to [9], where earlier citations are given.

The possibility of determining the structure of a single molecule by measurements at many different orientations has been proposed by Miao, Hodgson and Sayre[10] (for background see also [11] to [14]). In these papers, the phase iteration method was developed and applied to solve the unknown phase problem. In these papers the oversampling ratio σ was discussed as was the uniqueness of the solution. These important works sparked my personal interest in this subject[15].

The 3-D diffraction pattern from a coherent x-ray beam is given in the Fraunhofer scattering region as

$$F[\mathbf{k}] \ = M[\mathbf{k}] \ \exp(-i\phi[\mathbf{k}]) = \sum_{\mathbf{r}} \ \exp[-i\mathbf{k}\cdot\mathbf{r}\,] \ v[\mathbf{r}\,] \ , \tag{10}$$

where $M[\mathbf{k}]$ is the magnitude of the pattern, $\phi[\mathbf{k}]$ its phase, and $v[\mathbf{r}\,]$ is the charge distribution, which will also be termed the image or the object. If both $M[\mathbf{k}]$ and $\phi[\mathbf{k}]$ were measurable, then $v[\mathbf{r}\,]$ can be computed directly from the inverse transform. However the full 3-D transform cannot be directly measured. It must be constructed from a series of 2-D patterns measured from the target at many different (and in the present case, unknown) orientations. A final problem is that the phase of the 2-D patterns cannot be measured and are unknown. They must be inferred from the data and known physical properties of the image.

Since all the phases are unknown in fourier space, additional information must be added in order to define the problem. The fact that the object is of finite extent and is zero outside some chosen volume V_o which is contained in the full volume V ($V > V_o$) provides the additional constraints. Thus we have a mixed problem with some of the information given in coordinate space, and the remainder given in fourier space, in particular the magnitudes of the 2-D projections.

Rotation Determination The relative orientation angles between the sources that give rise to 2-D patterns can be determined for a sufficient number of patterns. The fourier pattern from a source rotated[16] by R is

$$\begin{aligned} F_R[k_\perp] &= \sum_{\mathbf{r}} \exp[-ik_\perp \cdot \mathbf{r}\,]\; v[R^T r] \\ &= \sum_{\mathbf{r}} \exp[-ik_\perp \cdot R \cdot \mathbf{r}\,]\; v[\mathbf{r}\,] \; . \end{aligned} \tag{11}$$

Define two patterns labelled by a and b

$$\begin{aligned} F_a[k_\perp] &= \sum_{\mathbf{r}} \exp[-ik_\perp \cdot R_a \cdot \mathbf{r}\,]\; v[\mathbf{r}\,] \\ F_b[k_\perp] &= \sum_{\mathbf{r}} \exp[-ik_\perp \cdot R_b \cdot \mathbf{r}\,]\; v[\mathbf{r}\,] \; . \end{aligned} \tag{12}$$

Search for lines in the $k_x - k_y$ plane along which the patterns are equal. Define a match line in each pattern with a tilt angle A where $c = \cos A$ and $s = \sin A$. Then $k_\perp = k\, t_\perp$, with $t_x = c$, $t_y = s$ and $t_z = 0$ with $(-K < k < K)$. Assume that the following relation holds

$$F_a(k\, c_a, k\, s_a) = F_b(k\, c_b, k\, s_b) \; , \tag{13}$$

where $c_a = \cos a$, $s_a = \sin a$, $c_b = \cos b$, and $s_b = \sin b$. Note that along these lines, both the magnitude and the phase of the patterns are separately equal. In the real data, only the magnitude of the patterns will be available.

The equality of the a and b transform patterns along this line then implies that the phases agree identically for all $\mathbf{r}$ and k. That is,

$$t^a_\perp \cdot R_a = t^b_\perp \cdot R_b \tag{14}$$

$$t^a_\perp = t^b_\perp \cdot (R_b\; R^T_a) \;=\; t^b_\perp \;\cdot\; R_{ba} \; . \tag{15}$$

Explicitly this is of the form

$$c_a = c_b R(x,x) + s_b R(y,x) \qquad s_a = \; c_b R(x,y) + s_b R(y,y) \tag{16}$$

$$\text{and} \quad 0 = c_b R(x,z) + s_b R(y,z) \; . \tag{17}$$

In terms of the Euler angles, $R(x,z) = \sin\psi \sin\theta$ and $R(y,z) = \cos\psi \sin\theta$. The last condition requires $\sin(\psi + b) = 0$, while the first two conditions become

$$c_a = \cos\phi \ \cos(\psi + b) \tag{18}$$

$$s_a = \sin\phi \ \cos(\psi + b) \ . \tag{19}$$

Therefore there are two discrete solutions

$$\text{soln 1} \ : \qquad \psi = \ -b \ , \qquad \phi = a \tag{20}$$

$$\text{soln 2} \ : \qquad \psi = \ \pi - b, \quad \phi = \pi + a \ . \tag{21}$$

The second solution is the Necker reversal of the object and corresponds to $\theta \to 2\pi - \theta$. Solution 1 will be chosen as the canonical one, that is $R_{ba} = R(-b, \theta, a)$. The Euler angle θ is not determined by these conditions. The matching of two patterns along a line cannot determine the angle θ between the planes. Theta measures the angle of intersection and must be determined by comparing more than two patterns. I shall not go into detail here but given the 2-D match line angles from 3 nondegenerate patterns (the three R matrices must be closed), all of the θ's can be expressed analytically. Therefore the rotation matrices between all pairs of patterns can be fully determined just from the 2-D patterns themselves.

Phase Determination The data is given by the series of patterns $1 \le n \le N$

$$F_n[k_\perp] = M_n[k_\perp] \ \exp[-i\phi_n[k_\perp]] = \sum_{\mathbf{r}} \ \exp[-ik_\perp \cdot R_n \cdot \mathbf{r}\,] \ v[\mathbf{r}\,] \ , \tag{22}$$

where the phases $\phi_n[k_\perp]$ are *not* measured.

The Problem: From the measured magnitude $M_n[k_\perp]$ and the fact that the image has compact support, find the set of phases $\phi_n[k_\perp]$ that yield a positive semidefinite real image $v[\mathbf{r}]$ and determine that image.

Hamiltonian Formulation In order to simplify the equations, the following definitions are introduced, recall Eq[22], as sums inside V_o

$$C_n[k_\perp] = \sum_{\mathbf{r}} v[\mathbf{r}\,]\cos[k_\perp \cdot R_n \cdot \mathbf{r}\,] \tag{23}$$

$$S_n[k_\perp] = \sum_{\mathbf{r}} v[\mathbf{r}\,]\sin[k_\perp \cdot R_n \cdot \mathbf{r}\,]\ , \tag{24}$$

and the measured magnitude of the pattern that is to be fit by an appropriate choice of $v[\mathbf{r}\,]$ is given by

$$M_n[k_\perp]^2 = C_n[k_\perp]^2 + S_n[k_\perp]^2\ .$$

An energy functional that at its minimum determines $v[\mathbf{r}\,]$ is

$$H(v[\mathbf{r}\,]) = \frac{1}{2}\sum_n \sum_{k_\perp} \{\sqrt{C_n[k_\perp]^2 + S_n[k_\perp]^2} - M_n[k_\perp]\}^2 - \sum_{\mathbf{r}} \lambda(\mathbf{r}\,)v[\mathbf{r}\,]\ .$$

Recall that the constraints on an allowable image are that

$$v[\mathbf{r}\,] = 0 \qquad \text{for } \mathbf{r} > V_o \qquad \text{and} \qquad v[\mathbf{r}\,] \geq 0 \qquad \text{for } \mathbf{r} < V_o\ , \tag{25}$$

where V_o is a region that is large enough to definitely contain the object. Thus the integral over $\mathbf{r}$ in Eq[23], Eq[24] and the Hamiltonian are *only* over the interior of V_o. The positivity constraint has been enforced by adding the term

$$\delta H = -\sum_{\mathbf{r}} \lambda(\mathbf{r}\,)v[\mathbf{r}\,]\ , \tag{26}$$

where $\lambda(\mathbf{r}\,)$ is a positive semidefinite inequality multiplier[17]. Consider the variation of the energy with respect to the image value at the point $\mathbf{r}$

$$\frac{\delta H}{\delta v[\mathbf{r}\,]} = \sum_n \sum_{k_\perp} j_n(k_\perp, \mathbf{r}\,) - \lambda(\mathbf{r}\,)\ , \tag{27}$$

where the auxiliary quantity j_n has been introduced as

$$j_n(k_\perp, \mathbf{r}\,) = \frac{\sqrt{C_n[k_\perp]^2 + S_n[k_\perp]^2} - M_n[k_\perp]}{\sqrt{C_n[k_\perp]^2 + S_n[k_\perp]^2}} \times \{C_n[k_\perp]\cos[k_\perp \cdot R_n \cdot \mathbf{r}\,] + S_n[k_\perp]\sin[k_\perp \cdot R_n \cdot \mathbf{r}\,]\}\ . \tag{28}$$

If $C_n[k_\perp] = S_n[k_\perp] = 0$ at an isolated value of $k_\perp$, then a direct evaluation leads to $j_n(k_\perp, \mathbf{r}) = -M_n[k_\perp]$. It is convenient to normalize the overall image by replacing $v[\mathbf{r}] \to wv[\mathbf{r}]$ and finding the minimum of H with respect to w.

Minimization Assume that there is a tentative (guess) image $v_0[\mathbf{r}]$. Improvements to this image are computed by using steepest descent iteration. If the derivative of the Hamiltonian is negative (positive) then one increases (decreases) the image at the point $\mathbf{r}$. If the tentative image is negative then it is always increased to zero by the appropriate choice of the positive definite parameter $\lambda(\mathbf{r})$.

Schema: Guess an image $v_0[\mathbf{r}]$ that satisfies the requisite conditions. Initialize the restraint parameter $\lambda(\mathbf{r})$ to zero and choose an iteration step parameter $\eta > 0$. Compute the 2-D quantities $C_n[k_\perp]$ and $S_n[k_\perp]$ for $(1 \le n \le N)$ using $v_0[\mathbf{r}]$ and the known R_n.

1. Evaluate the scale factor w and renormalize $v[\mathbf{r}]$, $C_n[k_\perp]$ and $S_n[k_\perp]$.

2. Choose a lattice point $\mathbf{r}_i = (i_x, i_y, i_z)$.

3. Compute $H'(\mathbf{r}_i)$, the derivative of H w.r.t. $v_0[\mathbf{r}]$, using Eq[27].

4. If $H'(\mathbf{r}_i)$ is not 0, tentatively replace

$$v[\mathbf{r}_i] \to v[\mathbf{r}_i] + \delta v[\mathbf{r}_i] \quad \text{where} \quad \delta v[\mathbf{r}_i] = -\eta H'(\mathbf{r}_i) \; . \tag{29}$$

   If the new $v_0[\mathbf{r}] \ge 0$ with $\lambda(\mathbf{r}_i) = 0$, set $\lambda(\mathbf{r}_i) = 0$.

   If not, choose $\lambda(\mathbf{r}_i)$ so that the new $v[\mathbf{r}_i] = 0$.

5. If $v[\mathbf{r}_i]$ has changed at $\mathbf{r}_i$, update $C_n[k_\perp]$ and $S_n[k_\perp]$ at all $k_\perp$ via

$$C_n[k_\perp] \;\; \to \;\; C_n[k_\perp] + \delta v[\mathbf{r}_i] \cos[k_\perp \cdot R_n \cdot \mathbf{r}_i] \tag{30}$$

$$S_n[k_\perp] \;\; \to \;\; S_n[k_\perp] + \delta v[\mathbf{r}_i] \sin[k_\perp \cdot R_n \cdot \mathbf{r}_i] \; . \tag{31}$$

6. Return to Step 2 and repeat for a new lattice point.

7. After all lattice points have been examined, return to step 1.

Alternative orderings of these iteration steps can be used. For example, the order of step 5 and 6 can be reversed. For stability, the trial change in the image, $\delta v[\mathbf{r}_i]$, should be bounded.

Once the minimum of H has been found, the image is given by $v[\mathbf{r}]$, while the phase of the transform is given in terms of $C_n[k_\perp]$ and $S_n[k_\perp]$.

Orientation Check: The scheme described here requires that the orientation of each pattern be known. In order to check these angles, which are subject to errors arising from pattern intensity and noise, one may add to the above steps an examination of the error in each individual pattern. If this quantity is large for a particular pattern n, after the iteration has had a chance to partially converge, one can test if it is due to an poor determination of the pattern orientation by redetermining the euler angles in $R_n(\psi, \theta, \phi)$ to minimize this pattern's error (as measured from the pattern from the tentative image at that iteration stage), and then resume the full iteration scheme.

Results: The above formulation was implemented in a program written in C++ and a few small test cases were run. A typical case used a full lattice of size $N = 15$, i.e., $V = 3375$, and an image size of $N_o = 9$, i.e., $V_o = 729$ with $\sigma = 4.6$. Only a few shapes were examined and no noise was added to the pattern. These objects were sufficiently small and simple that there was no indication of locking into a local minima.

The input data utilized ~ 10 patterns in $k_\perp$ calculated from the input object density distribution at known orientations. The program ran at a rate of ~ 10 iteration per minute. The initial image was assumed to be either a constant or random. The constant initial start converged faster than

the random start and the results below are for this case. After 15 iterations, the fractional RMS error in the density was typically

$$\sqrt{< (v_{exact} - v)^2 >}/ < v > \sim 0.03 \text{ to } 0.10 \ , \tag{32}$$

where $< \ >$ indicates an average over the volume V_o. For 45 iterations, the error ratio dropped by a factor of 2.

One of the advantages of this approach is that the fit to the data does not require any interpolation. This is performed implicitly, see Eq[23] and Eq[24]. One related disadvantage is that the formulation does not admit the simple use of fast fourier transforms to spend up the algorithm. However it should be noted that a full 3-D iteration of the phase iteration method requires the transforms of N^3 lattice points in addition to the interpolation stage. The Hamiltonian method here requires the transform of $M \times N^2$ lattice points. Normally, one expects that M is considerably smaller than N. A test of the method on realistic data and a comparison with other methods is planned.

ACKNOWLEDGEMENTS

I wish to thank Dr. Jianwei Miao, who introduced me to this subject, and Professor Richard Haymaker for helpful discussions.

References

[1] For an earlier version of the method see M. Ohlsson, C. Peterson Markus Ringner and R. Blankenbecler, "A Novel Approach to Structure Alignment", LU TP 00-07, SLAC-PUB-8429, April (2000).

[2] S. B. Needleman and C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *J. Mol. Biol.* **48**, 443-453, 1971.

[3] M. Gerstein and M. Levitt, "Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the SCOP Classification of Proteins", *Prot. Sci.* **7**, 445–456, 1998.

[4] Fienup, J. R., "Phase Retrieval Algorithms: A Comparison", Applied Optics **21**, 2758-69,1982.

[5] A. Krug and R. A. Crowther, "Three-dimensional Image Reconstruction from the Viewpoint of Information Theory", *Nature* **238**, 435-440,1972.

[6] S. D. Fuller, S. J. Bitcher, R. H. Cheng and T. S. Baker, "Three-dimensional Reconstruction of Icosahedral Particles - the Uncommon Line", *J. Struc. Biol.* **116**, 48-55,1996.

[7] R. A. Crowther, D. J. DeRosier and A. Krug, "The reconstruction of a three-dimensional structure from projections and its application to electron microscopy", *Phil. Trans. Roy. Soc. Lond.* **317**, 319-340,1990.

[8] R. A. Crowther, "Procedure for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs", *Phil. Trans. Roy. Soc. Lond.* **261**, 221-230,1971.

[9] M. Lindahl, "Strul - a method for 3D alignment of single particle projections based on common line correlation in Fourier space", *Ultramicroscopy* **87**, 165-176,2001.

[10] J. Miao, K.O. Hodgson, and D. Sayre, "An Approach to 3-D Structures of Biomolecules by using Single Molecule Diffraction Images", Proc.Nat.Acad.Sci. **98**,6641,2001.

[11] J. Miao, T. Ishikawa, Erik H. Anderson and K. O. Hodgson. "Phase retreival of diffraction patterns from non-crystalline samples by using the oversampling method", submitted to Phys. Rev. B, (2003).

[12] J. Miao, D. Sayre, and H. N. Chapman. "Phase retreival from the magnitude of the fourier transforms of nonperiodic objects.", *J. Opt. Soc. Am. A***15**, 1662 - 1669 (1998).

[13] J. Miao, T. Ishikawa, B. Johnson, E. H. Anderson, B. Lai, and K.O. Hodgson. "High Resolution 3D X-Ray Diffraction Microscopy", (SLAC, SSRL). Aug 2002, Phys. Rev. Lett. **89**, 088303, 2002.

[14] J. Miao, T. Ohsuna, O. Terasaki, K.O. Hodgson, M.A. O'Keefe. "Atomic Resolution Three-Dimensional electron Diffraction Microscopy", (SLAC, SSRL). Oct 2002, Phys. Rev. Lett. **89**, 155502, 2002.

[15] Richard Blankenbecler, "3D Image Reconstruction - Determination of Pattern Orientation", SLAC-PUB-9645, March (2003), and "3D Image Reconstruction - Hamiltonian Method for Phase Recovery", SLAC-PUB-9646, March (2003). Both submitted to Phys. Rev. B.

[16] George Arfken, "Mathematical Methods for Physicists", Academic Press, New York and London, for properties of the rotation matrices.

[17] See, for example, Martin B. Einhorn and Richard Blankenbecler, "Bounds on Scattering Amplitudes", *Annals of Physics* **67**, 480-517 (1971). Many earlier references on inequality constraints are given here.