28 02-Methods for Structure Determination and Analysis, Computing and Graphics

MS-02.01.03 PROGRESS TOWARDS APPLICATION OF THE MINIMAL FUNCTION TO MACROMOLECULES.

Charles M. Weeks*[1], Russ Miller[2], G. David Smith[1], Martha M. Teeter[3] & Herbert A. Hauptman[1]. [1]Medical Foundation of Buffalo, 73 High St., Buffalo, NY 14203, USA; [2]Dept. of Computer Science, State University of New York at Buffalo, Buffalo, NY 14260, USA; [3]Dept. of Chemistry, Boston College, Chestnut Hill, MA 02167, USA.

The *Shake-and-Bake* method of structure determination consists of a phase refinement procedure based on Hauptman's minimal function, R(φ), alternated with Fourier filtering [Weeks, C.M. *et. al.* (1993). *Acta Cryst.* **D49**, 179-181; Miller, R. *et. al.* (1993). *Science*, in press.]. This method provides a powerful and convenient formulation of direct methods, having been used to solve several known and unknown peptide structures containing approximately 100 non-hydrogen atoms in the asymmetric unit. It has also been applied successfully to atomic resolution data for two previously known small proteins, gramicidin A [Langs, D.A. (1988). *Science*, **241**, 188-191.] and crambin [Hendrickson, W.A. & Teeter, M.M. (1981). *Nature*, **290**, 107-113.]. These structures contain approximately 300 and 400 atoms, respectively, after taking disorder and partial occupancy into account.

In the case of gramicidin A, a 0.12Å grid placed in the asymmetric portion of the unit cell was used to obtain initial coordinates for 240,396 1-atom trial structures. The 346 trials having an initial mean phase error ≤ 80° were subjected to the *Shake-and-Bake* procedure, and three solutions were obtained following 450 cycles of refinement and filtering. Thus, in the worst case scenario, there is one solution per 80,000 trials for gramicidin A. In the case of crambin, initial phases were obtained by performing structure factor calculations based on randomly positioned 2-atom trial structures. The success rate was 1.3% following 200 *Shake-and-Bake* cycles. Both the gramicidin A and crambin maps can be easily interpreted either by examination of interpeak distances and angles or by graphical electron density fitting using FRODO. The best maps for manual examination are obtained by terminating the procedure with one cycle of Fourier refinement using all statistically reliable measured data.

	Gramicidin A	Crambin
Resolution	0.86Å	0.83Å
Temperature	120°K	130°K
Space Group	$P2_12_12_1$	$P2_1$
Atoms (approximate)	300	400
Phases	2000	4000
Triplets	20,000	40,000
Negative Quartets	0	0
Cycles	450	200
Trials Generated	240,396	1216
Trials Processed	346	1216
Solutions	3	16

At present, these experiments leave several questions unanswered. For example: What is the random start success rate for gramicidin A? How important was the presence of the six sulphurs in the crambin application? How long will the procedure, which presently relies on peak picking at the Fourier stage, be applicable as the resolution of the data is decreased? These problems are presently under investigation.

This research was supported by NIH grant GM-46733 and by NSF IRI-9108288. The authors also thank Dr. R. H. Blessing for help with data reduction and Dr. R. Jones of the Thinking Machines Corporation for computer time.

MS-02.01.04 MAXIMUM ENTROPY, LIKELIHOOD AND THE CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES

By C.J.Gilmore*, G.Proctor, and J.R.Fryer *Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland,* G.Bricogne, *LURE, Bâtiment 209D, Orsay 91405, France*, S.Xiang and C,W.Carter Jnr. *Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina, USA,* and A.Brisson, G. Moser and M. Schmutz, *Inistut de Chimie Biologique, Strasbourg, France.*

The maximum entropy-likelihood method as formulated by Bricogne (Bricogne, *Acta Cryst.* (1984) **A40,** 410-445) and implemented by Bricogne and Gilmore (*Acta Cryst.* (1990) **A46,** 284-297), and Xiang, Carter, Bricogne and Gilmore (*Acta Cryst.* (1990) **D47,** 193-212), provides a powerful phasing tool for use in macromolecular crystallography including electron crystallography. We report and summarize here the following:

(1) The use of likelihood to select correct phase sets generated by the SAYTAN program for the protein avian pancreatic polypeptide, App. (Gilmore, C.J., Henderson, A.N. & Bricogne, G. *Acta Cryst.* (1991) **A47,** 842-846),.

(2) The use of entropy maximisation combined with solvent flattening applied to cytidine deaminase.(Xiang, S., Carter, C.W., Bricogne, G. and Gilmore C.J.(*Acta Cryst.* (1990) **D47,** 193-212).

(3) Two structures of biological macromolecules studied using electron diffraction, and phased image data from high resolution electron microscopy:

(a) Purple membrane (*Halobium Halobactorium*) data (Baldwin, J.M., Henderson,R., Beckman, E., & Zemlin, F. *J. Mol. Biol.* (1988) **202,** 585-591). This was as a test of the method, and has produced some controversial results concerning resolution enhancement.

(b) Cholera toxin. Here we are phasing data to 4Å from 56 unique phased reflections at 8.8Å resolution using the ME method incorporating the application of five-fold non-crystallographic symmetry, and solvent flattening.

In both (a) and (b) a low resolution basis set of phased reflections had been derived from the Fourier transform of optical image data suitably averaged, and used to phase the high resolution diffraction data *via* a process of entropy maximisation and likelihood evaluation coupled with the building of phasing trees.

The maximum entropy method is ideal in these circumstances because:

(1) It will work with projection data.

(2) It is stable regardless of data resolution.

(3) It can utilize non-crystallographic symmetry, and solvent flattening in a wholly natural and relatively simple way.

(4) it uses non-uniform atomic distributions which are constantly updated.

(5) Likelihood can be used to determine an effective unit cell contents that reflects the data resolution.

MS-02.01.05 ELECTRON-DENSITY HISTOGRAMS AND THE PHASE PROBLEM.

By V Yu Lunin, Institute of Mathematical Problems of Biology, Pushchino, Moscow Region, Russia.

The spectra of frequencies (histograms) of different values in Fourier syntheses provide the most adequate representation of information on 'what values may be found in a good Fourier synthesis and how frequent they are'. These Electron-Density Histograms (EDH)

02-Methods for Structure Determination and Analysis, Computing and Graphics 29

depend on synthesis resolution and are sensitive to errors in structure factors modules and phases. Data-bank based methods have been developed which allow one to predict the true EDH for a protein under investigation before the phase problem has been solved. Such histograms may be used as additional constraints when solving the phase problem.

Three ways of using EDH as an additional restriction are suggested. The first one consist in direct minimisation of the discrepancy between the true histogram and one calculated from current phases values. Any additional requirements formulated as a minimal principle may be incorporated in this process. The second way is iterative electron density modification restoring the true histogram alternating with replacement of structure factors modules in the modified synthesis by experimental ones. One more way to exploit the information contained in a histogram is to use the similarity of the true and a calculated histogram as an additional criterion for the check of generated variants in Monte-Carlo based approaches.

There are many density-modification methods which imply some form of histogram matching and that such methods are reviewed and related to the EDH methods. Parallels with other classical approaches are also found.

MS-02.01.06 INCORPORATION OF DIRECT METHODS WITH ALL THE PROTEIN-CRYSTALLOGRAPHY PHASING TOOLS AVAILABLE. A NEW PROBABILISTIC EXPRESSION FOR TRIPLETS AND QUARTETS. By Christos Kyriakidis[*], René Peschar and Henk Schenk. Laboratory for Crystallography, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.

Hitherto, the use of direct methods for solving structures from single-crystal data seems to have been limited to small structures. The reason for this is clear: the joint probability distribution (j.p.d.) of three structure factors depends in first approximation on $N^{-1/2}$ so the j.p.d. gets increasingly flattened if N becomes large. On the other hand, large structures such as proteins have been solved with use of SIRNAS and/or SAS. This raises the question why direct methods fails while other techniques succeed. An efficient way to improve the applicability of direct methods is to reduce the number of variables (N) involved. In the case of isomorphous data, as present in techniques such as SIRNAS, SIRAS, SAS and 2DW, this reduction can be achieved in a very simple way. It has been shown recently that the concept of isomorphous structure factors can be useful for estimating the doublet and triplet phase-sums present amongst them (Kyriakidis, Peschar and Schenk, *Acta Cryst.*, 1993, **A49**, March issue). From the tests it appeared that for too low diffraction ratios, *i.e.* almost perfectly isomorphous structures, no useful estimates could be obtained, even for small structures. Analyses showed that in these cases the reliability indicators were no longer properly defined. If the differences between isomorphous structure factors become too small, the normal mathematical procedure more or less fails. It seems that the very small quantities cannot be expressed in terms of the usual variables. This suggested that a different type of random variable should be defined: the single difference of isomorphous structure factors, $F_\nu{}^d$ which is the difference between two isomorphous structure factors $F_\nu{}^\ell$ and $F_\nu{}^m$. The subscript ν refers to a particular reflection and the superscripts ℓ, m and d denote dependence on the isomorphous data sets ℓ, m and both ℓ and m respectively. We have

$$F_\nu{}^d \equiv F_\nu{}^\ell - F_\nu{}^m = \sum_{j=1}^{N} f_{j\nu}^\ell \exp(2\pi i \mathrm{H}_\nu \mathbf{r}_j) - \sum_{j=1}^{N} f_{j\nu}{}^m \exp(2\pi i \mathrm{H}_\nu \mathbf{r}_j)$$

$$= \sum_{j=1}^{n} (f_{j\nu}^\ell - f_{j\nu}{}^m) \exp(2\pi i \mathrm{H}_\nu \mathbf{r}_j) = |F_\nu{}^d| \exp(i\phi_\nu{}^d) \qquad (1)$$

Expression (1) shows that $F_\nu{}^d$ depends only on the number of atoms (n) for which the atomic scattering factors differ in $F_\nu{}^\ell$ and $F_\nu{}^m$ while, in contrast, $F_\nu{}^\ell$ and $F_\nu{}^m$ depend on all N variables. Both the magnitude $|F_\nu{}^d|$ and the phase $\phi_\nu{}^d$ of $F_\nu{}^d$ are functions of the magnitudes and phases of $F_\nu{}^\ell$ and $F_\nu{}^m$.
Based on the use of the $F_\nu{}^d$ as random variables an efficient procedure will be presented for the derivation of j.p.d.s of isomorphous data sets. It will be shown that the usual probabilistic techniques, applied to these random variables, finally results in the j.p.d. of three, four and seven single differences of isomorphous structure factors comprising three doublets, eight triplets and sixteen quartet phase sums. A major advantage of the new technique is that the inherent correlation between the isomorphous data sets is removed if a mathematical procedure is set up for the small difference itself. An important goal of the present contribution is the derivation of a new expression for estimating the triplet and quartet phase sums present among isomorphous data. It will be shown that the new procedure, supplemented by optimal doublet phase-sum estimates that use difference Patterson information (Kyriakidis, Peschar & Schenk, *Acta Cryst.*, 1993, **A49**, March issue) leads to far better results than obtainable by other j.p.d.-based expressions (Hauptman, *Acta Cryst.*, 1982, **A38**, 289-294; 632-641; Giacovazzo, *Acta Cryst.*, 1983, **A39**, 585-592; Giacovazzo, Cascarano & Zheng, *Acta Cryst.*, 1988, **A44**, 45-51; Fortier & Nigam, *Acta Cryst.*, 1989, **A45**, 247-254; Peschar & Schenk, *Acta Cryst.*, 1991, **A47**, 428-440) in particular if the diffraction ratio is small. E.g. for the protein Cytochrome *c* from *Paracoccus Denitrificans* (Timcovich & Dickerson, *J. Biol. Chem.*, 1976, **251**, 4033-4046) the error reduction for the triplets and quartets is more than 50% compared to previous techniques. This reduction leads to a phase error small enough for direct methods applications without the knowledge of the heavy atom substructure (Kyriakidis, Peschar & Schenk, *Acta Cryst.*, 1993, **A49**, May issue).
It should be noted that the above procedure may be used also for complicated small structures when the traditional direct methods are not sufficient to give any solution. Applications for small and protein structures will be communicated.

PS-02.01.07 MOLECULAR SCENE ANALYSIS: THE INTEGRATION OF DIRECT METHODS AND ARTIFICIAL INTELLIGENCE STRATEGIES FOR SOLVING PROTEIN STRUCTURES. By S. Fortier[*&+] and J. Glasgow[+], Depts. of Chemistry[&] and Computing and Information Science[+], Queen's University, Kingston, Canada K7L 3N6 and F.H. Allen, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1E2, England.

A progress report on the development of a knowledge-based system for protein crystal structure determination will be presented. The approach integrates direct methods and artificial intelligence strategies to rephrase the structure determination process as an exercise in *scene analysis*. A general joint probability distribution framework, which allows the incorporation of isomorphous replacement, anomalous scattering and *a priori* structural information, forms the basis of the direct methods strategies. The accumulated knowledge on crystal and molecular structures is exploited through the use of artificial intelligence strategies, which include techniques of knowledge representation, search and machine learning. Progress on the construction of a protein knowledge base, the implementation of routines for the automated interpretation of protein electron density maps and the development of conceptual clustering techniques for application to crystallographic data will be reported.