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

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

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

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.

30 02-Methods for Structure Determination and Analysis, Computing and Graphics

The further development of the EDH based approach is to consider the set of conditional histograms calculated from the points in the unit cell satisfying some additional restrictions. Such additional constraints may imply position of a point with respect to the molecular region or the values of some other functions connected with the object under investigation.

PS-02.01.08 ELECTRON DENSITY SQUARING METHOD AND NON-CRYSTALLOGRAPHIC SYMMETRY. By A.F.Mishnev, Latvian Institute of Organic Synthesis, Riga, Latvia

Non-crystallographic symmetry imposes restraints on phases of the structure factors. Linear relationships among structure factors due to identical molecules in different crystallographic environment have been obtained by Main & Rossmann (Acta Cryst., 1966, 21, 67-72). For a structure containing like atoms the electron density squaring method (Sayre, Acta Cryst., 1952, 5, 60-65) may be introduced in the analysis of Main & Rossmann, that results in quadratic equations for the structure factors. In the presence of non-crystallographic symmetry the structure factor of the "squared" crystal takes the form

$$G_p = \sum_{n=1}^{N} \int_U \rho_1^2(x_1) \cdot \exp\{2\pi i([C_n]x_1 + d_n)\} dx_1 . \qquad (1)$$

The "squared" structure factor may be expressed by $G_p = g_p/f_p \cdot F_p$. Let $\rho_2(x)$ be the electron density in a second crystal, which contains the same molecule. Since $\rho_2(x_1) = \rho_1(x_1)$ by definition, one can obtain the equations

$$F_p = \frac{f_p}{g_p V_K V_H} \sum_K \sum_H F_K \cdot F_H \cdot S_{KHP}, \qquad (2)$$

where S_{PKH} are functions of molecular envelope, rotation and translation parameters. When the two crystals are identical equation (2) reduces to the Sayre's equation. Numerical test calculations of equations (2) using simulated crystal data will be presented.

PS-02.01.09 DIRECT PHASING FOR MACROMOLECULES BY ENTROPY MAXIMISATION AND LIKELIHOOD RANKING. By G. Bricogne, Department. of Molecular Biology, Biomedical Centre, Box 590, 751 24 Uppsala, Sweden ; and LURE, Bâtiment 209D, 91405 Orsay, France.

A new multisolution phasing method based on entropy maximisation and likelihood ranking, proposed for the specific purpose of extending probabilistic direct methods to the field of macromolecules [Bricogne (1984). *Acta Cryst.* A**40**, 410-445], has been implemented in two different computer programs [Bricogne & Gilmore (1990). *Acta Cryst.* A**46**, 284-297; Bricogne (1993). *Acta Cryst.* D**49**, 37-60] and applied to a wide variety of problems. The latter comprise the determination of small crystal structures from X-ray diffraction data obtained from single crystals [Gilmore, Bricogne & Bannister (1990). *Acta Cryst.* A**46**, 297-308] or from powders [Bricogne (1991). *Acta Cryst.* A**47**, 803-829; Gilmore, K. Henderson & Bricogne (1991). *Acta Cryst.* A**47**, 830-841; Shankland, Gilmore, Bricogne & Hashizume (1993). *Acta Cryst.* A**49**, in the press], and from electron diffraction data partially phased by image processing of electron micrographs [Dong *et al.* (1992). *Nature, Lond.*, **355**, 605-609] or even unphased [Gilmore, Shankland & Bricogne (1993). Submitted to *Proc. R. Soc. London Ser. A.*]; the *ab initio* generation [Bricogne (1993). *Acta Cryst.* D**49**, 37-60] and ranking [Gilmore, A.N. Henderson & Bricogne (1991). *Acta Cryst.* A**47**, 842-846] of phase sets for small proteins; and the improvement of poor quality phases for a larger protein at medium resolution under constraint of solvent flatness [Xiang, Carter, Bricogne & Gilmore (1993). *Acta Cryst.* D**49**, 193-212]. These applications show that the primary goal of this new method – namely increasing the accuracy and sensitivity of probabilistic phase indications compared with conventional direct methods – has been achieved.

The main components of the method as implemented in the computer program BUSTER [Bricogne (1993). *Acta Cryst.* D**49**, 37-60] are (1) a tree-directed search through a space of trial phase sets; (2) the saddlepoint method for calculating joint probabilities of structure factors, using entropy maximisation; (3) likelihood-based scores to rank trial phase sets and prune the search tree; (4) a new method for optimising the choice of reflexions so as to maximise the sensitivity of the likelihood to their phases; (5) efficient schemes, based on error-correcting codes, for sampling trial phase sets; (6) a statistical analysis of the scores for automatically selecting reliable phase indications by multidimensional Fourier techniques coupled with tests of statistical significance. This program has been successfully tested on two small structures and has been applied to data from two small proteins.

The mathematical techniques now available in BUSTER bring closer a number of major enhancements of standard macromolecular phasing methods proposed earlier [Bricogne (1988). *Acta Cryst.* A**44**, 517-545] as an extension of the initial theory. In the molecular replacement method, for instance, the detection and placement of a known fragment described in a reference position and orientation by a density ρ^M with transform F^M can be accomplished by calculating the log-likelihood gain :

$$\mathrm{LLG}(\mathbf{R}, \mathbf{t}) = \log \frac{\mathcal{P}\left(\left|F_{\mathbf{h}}\right| = \left|F_{\mathbf{h}}\right|^{\mathrm{obs}} \text{ for all } \mathbf{h} \,\middle|\, (\mathcal{H}_1[\mathbf{R}, \mathbf{t}]) \right)}{\mathcal{P}\left(\left|F_{\mathbf{h}}\right| = \left|F_{\mathbf{h}}\right|^{\mathrm{obs}} \text{ for all } \mathbf{h} \,\middle|\, (\mathcal{H}_0) \right)}$$

where $(\mathcal{H}_0)$ denotes the null hypothesis that all atoms are uniformly distributed in the asymmetric unit while $(\mathcal{H}_1[\mathbf{R}, \mathbf{t}])$ denotes the alternative hypothesis that the known fragment is placed in the asymmetric unit with orientation $\mathbf{R}$ at position $\mathbf{t}$, and the rest of the atoms are distributed at random. A drastic simplification of LLG yields a sum of (1) a Patterson correlation (PC) - based rotation function in which a sum of point-group symmetry-related copies of the self-Patterson of the rotated fragment is correlated with the origin-removed self-Patterson of the whole structure; and (2) a PC-based translation function, expressed as a Fourier series with argument $\mathbf{t}$ itself. This function is already an improvement on the PC functions used in XPLOR [Brünger (1990). *Acta Cryst.* A**46**, 46-57], yet it is in general a poor approximation to LLG. It will be shown how the systematic use of LLG and of its relations to Bayesian statistical methods yields a new procedure for the detection and accurate placement of a known molecular fragment and of its recycling into the phasing process which overcomes every single limitation of the current methodology [Bricogne (1993). In *The Molecular Replacement Method*, edited. by W. Wolf, E.J. Dodson & S. Gover. Warrington: SERC Daresbury Laboratory, in the press].

PS-02.01.10 FOURIER-TRANSFORM-BASED METHODS FOR PHASE EXTENSION AND REFINEMENT AND PERHAPS THE SOLUTION OF MACROMOLECULES. By L Refaat, C Tate and M M Woolfson*, Department of Physics, University of York, UK.

The Sayre equation is known to be effective for phase extension and refinement, either alone (Sayre, D, 1972, Acta Cryst A28, 210-212) or in conjunction with other constraints, as in the SQUASH procedure (Main, P, 1990, Acta Cryst A46, 372-377).

A method is described by which the coefficients of a set of linear equations are derived, solely from FFT operations, leading to phase