PS03.04.08 A NEW APPROACH TO MACROMOLECULAR REFINEMENT VIA BAYES' LAW AND BOLTZMANN'S DISTRIBUTION. Rob Grothe, Washington University, Dept. of Electrical Engineering, St. Louis MO, USA

The refinement of a macromolecular structure from crystal diffraction data can be formulated as follows: find the *most likely* mini-ensemble of structures given molecular energetics and observed data. The (posterior) conditional probability to be maximized can be expressed via Bayes' law as the *multiplicative product* of two distributions, prior and data likelihood. The prior assigns probability to each mini-ensemble, viewed as a single state, via Boltzmann's distribution of states for a canonical ensemble at ambient temperature; the mini-ensemble energy is the mean of the energy values computed for individual structures under an energetic model. For a given mini-ensemble, a virtual asymmetric unit is constructed by averaging the structures and a virtual crystal by symmetric replication. The data likelihood is the probability that the measurement of x-rays (modeled by a random process) diffracted by this crystal results in the observed data.

The Boltzmann relationship converts energy into probability, the common currency through which two disparate information sources, molecular energetics and diffraction data, can be unified via Baye's law. Viewed in reverse, the relationship yields the molecular dynamics interpretation: the most likely ensemble *minimizes* the model energy function derived from the model probability. As energy depends upon the log of probability, the posterior energy is the sum of two terms corresponding to the factors in the posterior probability. *X-PLOR*, a widely used refinement package, minimizes the sum of a model molecular energy and a term penalizing disagreement between structure and data. The user chooses the form of the term along with weighting factors. In this new approach, a model for diffraction data is chosen, and the data-dependent term is derived from it.

A refinement algorithm, using jump-diffusion random sampling, has been implemented on a 16k-processor parallel machine. Preliminary results have been obtained for refining BPTI, using diffraction data from the Protein Data Bank and the published structure as the initial state.

PS03.04.09 HYDROPHILICITY OF CAVITIES IN PROTEINS. Jan Hermans, Li Zhang, Christopher VanDeusen, and Xinfu Xia, Department of Biochemistry and Biophysics University of North Carolina Chapel Hill, NC 27599-7260

Water molecules inside cavities in proteins constitute integral parts of the structure. We have sought a quantitative measure of the hydrophilicity of the cavities by calculating energies and free energies of introducing a water molecule into these cavities. The computations required to survey the atomic coordinates of a protein molecule in terms of low-energy water positions are rapid. A proper assessment of hydration should be based on free energy, not energy; however, much lengthier dynamics simulations are required to obtain free energies of transfer of water molecules into interior sites. These methods are most direct when applied to cavities able to hold a single water molecule. A simple consistent picture of the energetics of isolated buried water molecules has emerged from this study. A threshold value of the water-protein interaction energy at -12 kcal/mol was found to be able to distinguish hydrated from empty cavities. This is nearly the same value as the energy of ice, and, since the threshold must correspond to a free energy of zero, it follows that buried waters in proteins have entropy comparable to that of ice. The results of this study have enabled us to address the reliability of buried waters assigned in experiments.

We have extended this work to two instances of cavities large enough to contain several water molecules. In one case (uteroglobin; 1UTG), the computed energetics support the presence of 8 water molecules, where the x-ray structure reports 12 sites, some of them rather weak. In the other case, interleukin-1beta, the computed energies and free energies of transferring one or two water molecules into the cavity are insufficiently low, and this suggests that the cavity is not hydrated, as reported in crystallographic studies, and at odds with a report based on nmr experiments that the cavity is hydrated.

The program and instructions for rapidly locating possible water interior water positions and discriminating between these on the basis of the energy of transfer are available from the authors (request DOWSER program from xia@femto.med.unc.edu).

PS03.04.10 MULTICOPY MODELING OF THE SOLVENT DISTRIBUTION IN MACROMOLECULAR CRYSTALS. Eduardo I. Howard and J. Raul Grigera, Instituto de Fisica de Liquidos y Sistemas Biologicos (IFLYSIB), Universidad de La Plata, c.c. 565, 1900, La Plata, Argentina; Alberto D. Podjarny and Alexander Urzhumtsev, IGBMC, UPR de Biologie Structurale, B.P. 163, 67404, Illkirch, Cedex, France

Hydration models of biomacromolecular crystals, as obtained by crystallographic diffraction studies, usually position water molecules on precisely defined sites. Other experimental results, such as NMR, indicate that a large part of the water content has high mobility and is delocalized. The objective of this work is to find a hydration model that describes these mobile water molecules, while keeping the agreement with the observed diffraction amplitudes. A multicopy water model is proposed to describe the mobility. A set of water molecules, positioned by conventional methods, is used to generate several non-interacting copies. The system is set to an initial temperature (typically 400° K) through assigning different initial velocities to each water molecule of the different copies. Then the system is very slowly cooled (5° steps) until it reaches the desired temperature (300° K). This dynamics simulation, implemented in XPLOR, includes the usual modeling forces and an X-ray term. The free R-factor is used to monitor the validity of the process. The method was applied to X-ray diffraction data from BPTI and RNA crystals. The results show that while some water molecules are highly localized (the different copies remain clustered in specific hydration sites), the rest are more widely distributed, sometimes forming water channels. The shape of the multicopy distribution agrees precisely with the Fo-Fc difference maps; in the BPTI case, simulations and difference maps using neutron data were used to cross-check the results. The obtained models agree with the crystallographic data and are more compatible with other experimental observations than the ones with single fixed sites.

This work is supported by the CNRS through the UPR 9004, by the CONICET through the IFLYSIB and by the EU through the collaborative project CII-CT 93 0014 (DG 12 HSMU); JRG is an invited fellow financed by the MENESRIP (France).

PS03.04.11 A TEST OF MAXIMUM-LIKELIHOOD REFINEMENT OF MACROMOLECULAR STRUCTURES WITH BUSTER & TNT. John Irwin and Gérard Bricogne, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

The Bayesian viewpoint [1,2,3,4] has long suggested that structure refinement should be carried out by maximising the log-likelihood gain (LLG) rather than by minimising the conventional least-squares residual, as only the maximum-likelihood (ML) method can take into account the uncertainty of the phases associated to model incompleteness and model imperfection by suitably downweighting the corresponding amplitude constraints. It was predicted [3] that ML refinement would allow the refinement of an incomplete model by using the structure factor statistics of randomly distributed scatterers to represent the effects of the missing atoms, in such a way that the latter would not be wiped out; and that the final would then provide indications about the location of these missing atoms.

These predictions have now been confirmed by actual tests carried out by combined use of BUSTER [4] and TNT [5] on an incomplete (60% of molecule) and imperfect (1Å rms positional error) model. The maximum-likelihood result is more accurate than that from least-squares, and the final LLG gradient map is much more informative than the usual difference map, thus greatly increasing

the chances of success in "bootstrapping" from an unpromising molecular replacement starting point to a complete structure.

We will also discuss the two main concerns at the moment in the fields of structure refinement and validation where Bayesian methods have much to offer, namely (1) getting better reliability indicators for the final results of structure refinement, and (2) ensuring that these indicators are effectively optimised during refinement.

References

[1] G. Bricogne (1984). *Acta Cryst.* A40, 410-445.
[2] G. Bricogne (1988). *Acta Cryst.* A44, 517-545.
[3] G. Bricogne (1992). In *The Molecular Replacement Method*, edited by W. Wolf, E.J. Dodson and S. Gover, pp. 62-75. Warrington: Daresbury Laboratory.
[4] G. Bricogne, *Acta Cryst.* D49, 37-60 (1993).
[5] D.E. Tronrud (1987). L.F. Ten Eyck, and B.W. Matthews, *Acta Cryst.* A43, 489-501.

PS03.04.12 MACROMOLECULAR ANISOTROPIC TEMPERATURE-FACTOR REFINEMENT WITH STRICT CONSTRAINTS. Jennifer A. Kelly and Todd O. Yeates. The Department of Chemistry and Biochemistry and The Molecular Biology Institute, University of California at Los Angeles, USA

Macromolecular models typically have high crystallographic residuals; the discrepancy between observed and calculated structure factor amplitudes is usually from 15% to 25%. A significant component of the residual is due to inadequate representations of atomic motion in protein models. Data collected from most protein crystals are insufficient to refine individual atomic anisotropic temperature-factors. Anisotropic b-factor refinement introduces six times the number of parameters as isotropic b-factor refinement, and most often results in over-fitting macromolecular models to the data. In order to reliably model protein motion as anisotropic, strict constraints are required. Our method introduces a new form for constraints on temperature-factors described by Fourier series, and uses an FFT-based algorithm for refining the relevant values. Anisotropic temperature-factors are defined by a position-dependent B matrix whose elements are constrained to vary smoothly over the unit cell in a crystal. Each of the six matrix elements is represented by a Fourier series with few terms. Individual anisotropic temperature-factors are determined by refining the coefficients of the six Fourier series which comprise the B matrix. Gradients are computed using FFT's by a modification of the method of Agarwal*. Since each Fourier series has only a few terms, the number of refinement parameters is kept low and over-fitting is avoided.

*Ramesh C. Agarwal, *Acta Cryst.* (1978). A34, 791-809

PS03.04.13 PREDICTING AND ANALYZING DETERMINANTS OF WATER-MEDIATED LIGAND RECOGNITION. Leslie A. Kuhn[1], Michael L. Raymer[1,2], William F. Punch[2], Paul C. Sanschagrin[1], and Erik D. Goodman[3], Depts. of [1]Biochemistry, [2]Computer Science, and [3]Case Center for Computer-Aided Engineering and Manufacturing, Michigan State University, East Lansing, MI 48824, USA

Protein recognition of ligands, from nucleic acids to small molecule inhibitors, is usually mediated by bound water molecules bridging the protein-ligand interface. These water molecules influence both the shape and chemistry of interaction. A barrier to appropriately incorporating active-site bound water to improve molecular simulations and ligand design has been the absence of a method for determining which water sites are likely to be conserved upon ligand binding. Our *Consolv* technique, using a hybrid k-nearest-neighbor classifier/genetic algorithm, predicts which water molecules will mediate ligand binding by examining the structural and chemical environment of each water molecule in the free protein structure, without knowledge of the ligand. After training on 13 non-homologous proteins, *Consolv* correctly predicted conservation or displacement of 74.6% of the active-site water molecules in 7 new proteins. Moreover, water sites mispredicted to be conserved typically were displaced by a polar (oxygen or nitrogen) atom from the ligand. Overall accuracy for predicting conserved water or polar ligand atom binding was 89.6%. The ability to predict water-mediated interactions from the free protein structure implies that the majority of conserved active-site water binding is independent of the ligand, and that the protein micro-environment of each water molecule is the dominant influence. We are now using genetic algorithms with linear weighting and genetic programs allowing non-linear scaling to evaluate the relative importance of several features of the site - temperature factor, hydrogen-bonding capacity, local atomic packing, and atomic hydrophilicity - for determining conserved water binding.

PS03.04.14 AN INTENSITY-BASED LIKELIHOOD FUNCTION FOR STRUCTURE REFINEMENT. Navraj S. Pannu and Randy J. Read, Department of Mathematical Sciences and Medical Microbiology & Immunology, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

In order to improve the quality of a model, structural refinement attempts to optimize the agreement between the observed and the calculated diffraction measurements. The process of structural refinement is commonly based on a least-squares analysis. Since the probability distribution of the observed structure factor amplitude given the calculated structure factor amplitude is not a Gaussian centered at k|Fc|, where k is a scale factor, least-squares is not theoretically justified. Accordingly, least-squares refinement does not make optimal use of the discrepancies between observed and calculated diffraction measurements. Therefore, in order to improve the process of structural refinement, a more general maximum likelihood analysis is considered.

A maximum likelihood target function has been derived and implemented in XPLOR. This function takes into account errors both in the model and in the measurements. Furthermore, this function is intensity-based allowing the use of negative intensities derived from the values observed in the diffraction experiment.

Preliminary tests show that the intensity-based likelihood function can achieve more than twice the improvement in average phase error compared to a conventional least-squares refinement. As a result, the electron density maps are clearer and suffer less from model bias.

Research supported by AHFMR, MRC, NSERC and HHMI.

PS03.04.15 CRYSTAL STRUCTURE AT 1.0Å RESOLUTION OF HUMAN p56*lck* SH2 DOMAIN IN COMPLEX WITH A SHORT PHOSPHOTYROSYL PEPTIDE. L. Tong, T. C. Warren, J. King, R. Betageri, J. Rose and S. Jakes, Boehringer Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road, P. O. Box 368, Ridgefield, CT 06877, U.S.A

SH2 domains are modules of about 100 amino acid residues and bind to phosphotyrosine-containing motifs in a sequence-specific manner. They play important roles in intracellular signal transduction and represent potential targets for pharmacological intervention. The protein tyrosine kinase p56*lck* is a member of the *src* family and is involved in T-cell activation. The crystal structure of its SH2 domain with an 11-residue high-affinity peptide [1] showed that the phosphotyrosine (pY) and the Ile residue at the pY+3 position are recognized by the SH2 domain. We have determined the crystal structure of this SH2 domain in complex with the short phosphotyrosyl peptide Ac-pTyr-Glu-Glu-Ile (pYEEI peptide) at 1.0Å resolution [2]. The structural analysis at atomic resolution reveals that residue Arg 134 (αA2), which interacts with the phosphotyrosine side chain, is present in two conformations in the complex. This structure will be compared with those of other complexes.

[1] Eck, M.J., Shoelson, S.E., Harrison, S.C. (1993). *Nature*, **362**, 87.
[2] Tong, L., Warren, T.C., King, J., Betageri, R., Rose, J., Jakes, S. (1996). *J. Mol. Biol.* **256**, 601.