

## Computing IV

### Macromolecular Refinement/Water/High Resolution Structures

**MS03.04.01 OVERVIEW OF REFINEMENT AND LEAST-SQUARES METHODS.** Dale E. Tronrud, Dept. of Chemistry, University of Oregon, Oregon, USA

The process of refinement is a large problem in function minimization. To reduce the amount of computation the methods chosen to minimize the function incorporate a number of assumptions. When these assumptions break down special procedures must be used.

The methods of minimization used in macromolecular refinement span the range from Simulated Annealing to Full-Matrix Least-Squares. The properties of Simulated Annealing, it being a stochastic method, are difficult to characterize and will be only touched upon. The other methods commonly used are classified as gradient descent and include Steepest Descent, Conjugate Gradient, and Preconditioned Conjugate Gradient (also known as Conjugate Direction). The Full-Matrix method can only be applied to small proteins whose crystals diffract to high resolution because of the huge amount of computer resources it requires.

Each of the gradient descent procedures are derived by making specific assumptions about the nature of the function being minimized. Because these assumptions usually are not valid for the crystallographic residual the methods will fail unless special precautions are taken by the crystallographer.

Many of these precautionary procedures are commonly known, such as rigid body refinement, but an understanding of the details of the methods themselves allows one to know when and what procedure to apply.

This talk will describe the various minimization methods used today and their relationships to one another. The assumptions and resulting limitations of each method will be discussed along with, where they exist, suggestions for diagnostics which should be monitored. Where there are no diagnostics for certain limitations, procedures will be given which must be applied blindly to prevent the refinement from "hanging up".

**MS03.04.02 REFINEMENT OF PROTEINS AT ATOMIC RESOLUTION.** Victor S. Lamzin<sup>1</sup>, Thomas R. Schneider<sup>1</sup>, Zbigniew Dauter<sup>1,2</sup> Keith S. Wilson<sup>1,2</sup>, <sup>1</sup> EMBL Hamburg Outstation, c/o DESY, Notkestraße 85, 22603 Hamburg, Germany; <sup>2</sup> Department of Chemistry, University of York, Heslington, York YO1 5DD, UK

For small molecules X-ray data can be recorded to atomic resolution and positions of ordered atoms identified with an error of about 0.002 Å. Particular problems for proteins involve their bigger size and disorder. Lack of data causes difficulties at all stages of structure analysis. Advances in recent years, area detectors, synchrotron sources and cryogenic freezing, allow recording of atomic resolution data for at least a subset of protein crystals. Currently data, extending to at least 1.2 Å, have been collected by visitors and in-house for about 40 proteins at EMBL Hamburg alone. No longer are these only small tightly packed systems such as rubredoxin: the list includes alcohol dehydrogenase with 80 kDa in the asymmetric unit.

There has generally been a model available giving an initial R factor about 30 %. The model is refined with stereochemical restraints and isotropic temperature factors. Subjective inspection and building of water structure becomes increasingly time consuming as more potential sites emerge. Semi-objective criteria for water selection on the basis of distance and electron density have been adopted. Introduction of hydrogen atoms riding on their par-

ent atoms reduces the R factor by about 1 %. The isotropic models typically have R factors of 14 to 18 %. Anisotropic atomic thermal parameters are then refined leading to final values of R factors of 8 to 12 %. A final cycle of block-matrix minimisation provides a reliable estimate of coordinate error from inversion of the normal matrix.

At atomic resolution anisotropic refinement of thermal motion is clearly valid. The improvement in the maps allows easier identification of solvent and disordered residues. The main chain atoms in the ordered parts have a coordinate error of about 0.03 Å, the average for the whole structure is 0.05 Å. On introduction of anisotropy R<sub>free</sub> falls by almost as much as the R factor. Having established the protocol it is unnecessary to assess anisotropy with R<sub>free</sub> for each subsequent refinement. The last cycles must include all data, even those previously omitted for the R<sub>free</sub>.

The number of atomic resolution protein structures will increase and within the next years provide a phenomenal data base for detailed analysis. Preliminary comparison of protein stereochemistry has already showed significant deviations from parameters derived from small molecules.

**MS03.04.03 CRYSTALLOGRAPHIC STRUCTURE REFINEMENT USING MOLECULAR DYNAMICS CONSTRAINED TO TORSION ANGLES.** Luke M. Rice and Axel T. Brünger, Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, 06520, USA

A reduced variable conformational sampling strategy based on molecular dynamics constrained to torsion angles has been implemented and applied to crystallographic refinement (L. M. Rice and A. T. Brünger, *PROTEINS* 19:277-290, 1994). This formulation reduces the number of adjustable parameters by approximately tenfold, and allows for significantly higher simulation temperatures by eliminating high frequency bond and angle vibrations. Refinement protocols using torsion angle dynamics with constant temperature searching typically have a greater radius of convergence compared to conventional refinement strategies. Applications to refinements of very poor initial models and to refinements at low resolution will be discussed. Preliminary results suggest that this reduced variable method will allow refinement at lower resolution than is currently possible with existing approaches.

**MS03.04.04 Au OR FeS<sub>2</sub>? VALIDATION OF PROTEIN MODELS AND REFINEMENT PROTOCOLS.** Gerard J. Kleywegt & T. Alwyn Jones, Department of Molecular Biology, Biomedical Centre, Uppsala University, Box 590, S-751 24 Uppsala, Sweden

At low resolution, it is often non-trivial to produce a model which is an accurate representation of the protein one has collected data on. Even data with a nominal resolution of ~2-2.5 Å is no guarantee for a good model, or even a correctly traced one (1,2).

Methods that may help in preventing serious errors and over-fitting of the data while the refinement is in progress will be discussed. These include: the use of the free R-factor to monitor the refinement and to optimise the refinement protocol (3), the use of databases during model rebuilding (4,5), and the use of "quality control" as an integral part of the refinement process (6). In addition, a number of caveats with respect to the use of the free R-factor will be discussed.

Subsequently, a number of popular myths and wide-spread misconceptions with respect to the validation of final models will be debunked. These include:

\* A low R-factor and small r.m.s. deviations from ideal geometry prove that a model is correct. In fact, these are necessary, but hopelessly insufficient conditions (2).