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[1], Thomas R. Schneider[1], Zbigniew Dauter[1,2] Keith S. Wilson[1,2], [1] EMBLHamburg Outstation, c/o DESY, Notkestraße 85, 22603 Hamburg, Germany; [2] 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 parent 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 Rfree falls by almost as much as the R factor. Having established the protocol it is unnecessary to assess anisotropy with Rfree for each subsequent refinement. The last cycles must include all data, even those previously omitted for the Rfree.

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 stratergies. 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_2? 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).

* If only the CA coordinates of a model are deposited, nobody will ever be able to validate the model. Actually, in some cases this is possible nowadays.
* One does not need to use non-crystallographic symmetry restraints. The examples to the contrary may make some want to re-do their most recent refinement (7).
* Ramachandran plots are stiflingly boring. On the contrary: they are extremely useful for model validation. We will show some highly entertaining examples from real-life models.

Considering the controversial nature of some aspects of this presentation, the audience is invited to disagree vehemently.

References:
(1) Branden, C.I., & Jones, T.A. (1990). Nature 343, 687-689.
(2) Kleywegt, G.J., & Jones, T.A. (1990). Structure 3, 535-540.
(3) Brunger, A.T. (1992). Nature 355, 472-475.
(4) Jones, T.A., Zou, J.Y., Cowan, S.W., & Kjeldgaard, M. (1991). Acta Cryst. A47, 110-119.
(5) Zou, J.Y., & Mowbray, S.L. (1994). Acta Cryst. D50, 237-249.
(6) Kleywegt, G.J., & Jones, T.A. (1996). Meth. Enzymol., in press.
(7) Kleywegt, G.J. (1996). Acta Cryst. D52, in press.

MS03.04.05 IMPROVED STRUCTURE REFINEMENT THROUGH MAXIMUM LIKELIHOOD.

Randy J. Read and Navraj S. Pannu, Departments of Medical Microbiology & Immunology, and Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.

The least-squares target is not theoretically justified for crystal structure refinement, so it is preferable to use a maximum likelihood target instead. With a maximum likelihood treatment, the need for *ad hoc* weighting schemes and resolution cutoffs is eliminated, observational errors are used appropriately and, above all, the refinement is more successful.

When crystal structures of proteins or small molecules are used to address questions of scientific relevance, the accuracy and precision of the atomic coordinates are crucial. Accordingly, the atomic model is generally improved by refining it to improve agreement with the observed diffraction data. The use of least-squares methods would only be justified (by the principle of maximum likelihood) if the probability distribution relating the observed and calculated diffraction measurements were Gaussian. As the relationship is not Gaussian, the least-squares target is inappropriate.

We have implemented two maximum likelihood targets in the program XPLOR: 1) an amplitude-based Gaussian approximation assuming Gaussian errors in the observed amplitudes; and 2) an intensity-based likelihood function assuming Gaussian errors in the observed amplitudes squared. The amplitude-based target can be implemented easily in any least-squares refinement program, while the intensity-based target has a number of advantages including the ability to use negative observed intensities.

Preliminary tests with protein structures give dramatic results. Compared to least-squares refinement, maximum likelihood refinement can achieve more than twice the improvement in average phase error. The resulting electron density maps are correspondingly clearer and suffer less from model bias.

The authors were supported in these investigations by AHFMR, MRC (Canada), NSERC (Canada) and an International Research Scholars award to RJR from HHMI.

MS03.04.06 DESCRIPTION OF PROGRAM USING MAXIMUM LIKELIHOOD RESIDUAL FOR MACROMOLECULAR REFINEMENT, ILLUSTRATED BY SEVERAL EXAMPLES.

Eleanor J. Dodson and Garib N. Murshudov, Chemistry Department, University of York, Heslington, York, U.K., and Alexei A. Vagin, UCMB-ULB, Free University of Brussels, Avenue Paul Heger cp160/16 - P2 1050 Brussels, Belgium

We illustrate the advantages of the maximum likelihood refinement method over least-squares for macromolecules. Maximum likelihood refinement has been implemented in the program REFMAC.

At each cycle the program performs two steps. First it estimates the overall parameters of likelihood. This is most successful when the parameters are deduced from the FreeR set of reflections. Secondly it uses these parameters to build the likelihood function and refine the atomic parameters.

At the end of a cycle REFMAC also writes weighted map coefficients to give less biased maps for rebuilding, taking care to restore missing data. Absent reflections cause unpredictable noise in map calculations which may lead to errors in interpretation.

Several examples are described. In each case the refinement was carried to convergence from an existing model. Results were compared to maps and phases generated from the final coordinates.

Different parts of structure may be assigned different expected errors and methods for doing this have been explored and implemented. Two important applications for this are being analysed. In the first case the structure contains several U atoms as well as protein atoms. In the second part of the structure has been interpreted from a poor MIR map but the other part is being modelled from the uninterpretable electron density. There is also an option to include available phase information , for example from MIR or MAD calculations.

PS03.04.07 PROTEIN PRECISION RE-EXAMINED: LUZZATI PLOTS DO NOT ESTIMATE FINAL ERRORS.

D W J Cruickshank, Chemistry Department, UMIST, Manchester, M60 1QD, UK

The misuse of Luzzati plots of the residual R versus $\sin\theta/\lambda$ to estimate final coordinate errors has stimulated a re-examination of protein precision. Luzzati (1952, Acta Cryst.) gave a theory for **uncompleted** refinements which estimated the r.m.s. shifts still needed to reach R = 0. His theory assumed no errors in F_{obs} and that **the F_{calc} model was perfect apart from coordinate errors.** The Gaussian error distribution **was the same for all atoms.** These assumptions are invalid for proteins. Quite apart from the dependence on atomic number, it is well established that errors depend very strongly on atomic B values. Nor do Luzzati plots provide an upper limit for $<\Delta r>$.

Restrained refinement will be examined theoretically. As applied to the simplest protein model of 2 like atoms in one dimension, restrained refinement determines a length which is the weighted mean of the diffraction-only length and the geometric-dictionary length.

By extending the order-of-magnitude error formula for small molecules given by Cruickshank (1960, Acta Cryst.), the e.s.d. for protein atom i with $B = B_i$ is, very roughly,

$$\sigma(x_i) = k(N_i/p)^{1/2}\,[g(B_i)/g(B_W)]\,C^{-1/3}\,d_{min}\,R,$$

where k is about 1.0, $N_i = \sum Z_j^2/Z_i^2$,$p = N_{obs} - N_{params}$, [provisionally] $g(B) \approx (1 + 0.04B + 0.003B^2)$, B_w is the Wilson B for the structure, and C is the fractional completeness of the data to d_{min}. For example if $N_i = 1000$, $p = 15000 - 4000$, $B_i = B_w$, $C = 0.9$, $d_{min} = 1.4$Å, and $R = 0.15$, then $\sigma(x_i) = 0.07$Å. This approach reveals the basic statistical flaws in the use of Luzzati plots.

Some authors have been able to invert the full LS matrix, and so obtain proper estimates of e.s.d.'s. Even when this is not possible, determined efforts should be made to use the information in a partial LS matrix.