The Phenix refinement framework

Afonine[#], P.V., Grosse-Kunstleve, R.W. & Adams, P.D.

Lawrence Berkeley National Laboratory, One Cyclotron Road, BLDG 64R0121, Berkeley, CA 94720 USA

[#]e-mail: PAfonine@lbl.gov

1: Introduction

Many questions of biological significance require highly accurate knowledge of structural parameters such as atomic positions, atomic displacement parameters (ADP, also known as "B-factors") and occupancies. The refinement of these structural parameters is therefore an essential step of macromolecular structure determination. As part of the Phenix collaboration (Adams et al., 2004) we have developed new refinement tools to increase the automation of refinement.

Macromolecular structure refinement combines a large number of very diverse steps. The current implementation of the Phenix refinement protocol is shown in Figure 1. Making use of modern software development technology, each of the major building blocks is implemented as a reusable set of modules. Most of the modules are available through the open-source cctbx libraries (Grosse-Kunstleve *et al.*, 2002; http://cctbx.sourceforge.net/) which will be included in future CCP4 releases. Some of the cctbx modules make use of CCP4 developments: the Monomer library (Vagin & Murshudov, 2004; Vagin *et al.*, 2004) and the CMTZ library.

The following sections are a brief description of the practical implementation of the Phenix refinement framework, with pointers to open-source modules that are available to the developer community. An overview of the open source libraries can be found in the series of recent IUCr Computing Commission Newsletter articles, issues 1-5 (http://www.iucr.org/iucr-top/comm/ccom/newsletters/). The pointers are given as the names of Python modules, e.g. `iotbx.pdb`.

2: Refinement framework

2.1: Input processing

To initiate refinement, four major sources of information have to be processed:

- Structural model: coordinates, displacement parameters, occupancies;
- Reflection data: pre-processed observed intensities and optionally experimental phases;
- Parameters determining the refinement protocol;
- Empirical geometry restraints (sometimes referred to a "force field"): bond lengths, bond angles, dihedral angles, chiralities, planarities (Vagin & Murshudov, 2004; Vagin *et al.*, 2004; Grosse-Kunstleve *et al.*, 2004).

The structural model and the reflection data are provided by the user. Default parameters and a library of empirical geometry restraints are provided by the refinement framework but can be customized by the user.

The PDB format (Bernstein *et al.*, 1977; Berman *et al*., 2000) is the most commonly used format for exchanging macromolecular model data and is therefore available as the input format for refinement in Phenix. The iotbx.pdb library module performs the first stage of the PDB interpretation. It is designed to construct a five-deep structural hierarchy of **models** (PDB MODEL keyword), **conformers** (PDB altLoc identifier), **chains**, **residues** and **atoms** in the most robust way. Common simple formatting problems are corrected on the fly.

The second stage of the PDB interpretation is to match the structural data against the CCP4 Monomer library in order to derive geometry restraints, scattering types and nonbonded energy types. This function is performed by the `mmtbx.monomer_library.pdb_interpretation` module. Many common simple formatting and naming problems are considered in this interpretation. The PDB interpretation has been tested with all files found in the PDB database (http://www.pdb.org/). The vast majority of files can be processed without any user intervention. Carefully designed diagnostic messages help the user to quickly identify problems that cannot be automatically corrected.

The experimental data can be given in many commonly used formats, including the MTZ format. Multiple input files can be given simultaneously, e.g. a SCALEPACK file with observed intensities, a CNS (Brünger *et al.*, 1998) file with R-free flags, and a MTZ file with phase information. A complex procedure aims to extract the data most suitable for refinement without user intervention. The underlying core functionality is implemented in the `iotbx.reflection_file_server` module.

The large set of refinement parameters is presented to the user in a novel hierarchical organization specifically designed to be extremely user friendly (Grosse-Kunstleve *et al.*, 2005). This is achieved via a very simple syntax, the option to easily override selected parameters from the command line, and automatic adjustments based on the inputs. This parameter handling framework is completely general and can be reused for other purposes unrelated to refinement.

2.2: Core refinement tools

The core refinement procedure involves four major objects: the experimental data, the model (atomic model, ordered solvent model, bulk solvent model, coordinate error model, completeness of the atomic model, scale factors), parameterization of prior knowledge (e.g. geometry restraints), and a target function combining all model parameters. Refinement is the process of optimizing the model parameters in order to obtain a model that is most consistent with the experimental data and the prior information. The measure of consistency is the value of the target function. It is designed to decrease as the model parameters improve. For a number of reasons the optimization of the target function cannot be performed in a single step. The most important problems are:

- The target function has many local minima. Therefore sophisticated search algorithms like simulated annealing may need to be applied (Brünger *et al.*, 1987; Adams *et al.*, 1997; Brünger & Adams, 2002).
- Some groups of model parameters are highly correlated, e.g. isotropic displacement parameters and the exponential component of the overall scale factor correction, or displacement parameters and occupancies.
- Different model parameters such as coordinates and ADPs have different behavior (Agarwal, 1978).

Therefore it is common practice to perform refinement iteratively, and to split each iteration into several stages. The Phenix refinement protocol includes the following stages:

Bulk-solvent correction, scaling and error model estimation

Bulk solvent correction and scaling are among the most crucial steps in macromolecular structure refinement (Jiang & Brünger, 1994; Kostrewa, 1997; Badger, 1997; Urzhumtsev, 2000). Experience shows that best results are obtained with the Flat Bulk Solvent model (Phillips, 1980) and anisotropic scaling (Sheriff & Hendrickson, 1987; Murshudov, 1998).

Maximum likelihood target calculations require estimates of model errors and completeness, which in turn depend on the current atomic parameters and bulk solvent model (Lunin & Skovoroda, 1995; Afonine *et al.*, 2005). During refinement the atomic parameters and the bulk solvent model are continuously updated. Therefore it is necessary to also update the maximum likelihood error model. This requires special care since the error model is highly correlated with the bulk solvent model and the anisotropic scaling parameters. Recently we described a robust bulk solvent correction and anisotropic scaling procedure that combines a grid search and LBFGS minimization (Liu & Nocedal, 1989) using either Least-Squares or Maximum-Likelihood scale target functions (Afonine *et al.*, 2005). These algorithms are implemented in the `mmtbx.f_model` library module (Grosse-Kunstleve *et al.*, 2005).

Ordered solvent (water) modeling

We have implemented a completely automated protocol for updating the ordered solvent model during the refinement process (mmtbx.solvent.ordered_solvent module). If requested by the user, waters are updated (added and removed; Badger, 1997; Sheldrick & Schneider, 1997; Lamzin, V.S. & Wilson, K.S., 1997) in each macro cycle as indicated in Figure 1. In the same macro cycle, the complete structure including the waters is subject to coordinate and ADP refinement. Updating the ordered solvent model involves the following steps:

1) Elimination of waters present in the initial model based on user-defined cutoff criteria on ADP, occupancy and inter-atomic distances (water-water, macromolecule-water).
2) Location of peaks in a mFobs-αFcalc maximum likelihood difference map (equivalent to cross-validated σ_A-weighted map; Read, 1986; Urzhumtsev *et al.*, 1996).
3) Confirmation of peaks found in the previous step using a $2m$Fobs-αFcalc difference map.
4) Elimination of peaks in regions occupied by the macromolecule. The bulk-solvent mask is reused for this purpose.

5) Elimination peaks too close to each other (the default cutoff distance is 2.0 Å; the strongest peak is retained).
6) Elimination of peaks too close to macromolecular atoms (the default cutoff distance is 1.8 Å).
7) Elimination of peaks too far away from macromolecular atoms (the default cutoff distance is 6.0 Å).
8) Elimination of peaks based on the evaluation of tabulated empirical distance distributions derived from the analysis of high-resolution models in the PDB (Fig. 2). Distance distributions between water oxygen and macromolecular C, N and O atoms are tabulated. Only peaks with a good fit to at least one distance distribution are retained.

The table of distance distributions used in the last step is located in `mmtbx.max_lik` module.

Determination of target weights

As mentioned before, crystallographic refinement is the process of model improvement through the optimization of a target function. Depending on the input parameters, the target function in Phenix is defined as $T_{xyz} = w_{x\_chem} * E_{xray} + E_{chem}$ for coordinate refinement, or $T_{adp} = w_{x\_adp} * E_{xray} + E_{adp}$ for ADP refinement. E_{chem} is a sum over six types of empirical geometry restraints as described by (Grosse-Kunstleve *et al.*, 2004). The weights w_{x_chem} or w_{x_adp} are introduced to balance the contributions from the experimental observations (E_{xray}) and the empirical a priori information (E_{chem} or E_{adp}). The automatic weight estimation procedure is implemented as described in (Brünger *et al.*, 1989; Adams *et al.*, 1997) and used by default since experience shows that it is very robust. However in a few cases it was found to produce poor results. For such cases, the more time-consuming automatic weight optimization procedure as described by Brünger (1992) is also available. The underlying core algorithms for the weight determination are implemented in the `mmtbx.dynamics.cartesian_dynamics` module.

Simulated Annealing refinement

Simulated annealing is a powerful tool for escaping local minima in crystallographic refinement (Brünger *et al.*, 1987; Adams *et al.*, 1997; Brünger & Adams, 2002). Depending on the model and data quality, simulated annealing can be performed during Phenix refinement. This is supported by the `mmtbx.dynamics.simulated_annealing` module.

Coordinate refinement

Coordinate refinement is performed by LBFGS minimization of the target T_{xyz} w.r.t. atomic coordinates, while keeping all other parameters fixed. T_{xyz} can be the Least-Squares target (LS, as defined in Afonine *et al.*, 2005), the amplitude-based Maximum-Likelihood target (ML, as defined in Afonine *et al.*, 2005) or the Phased Maximum-Likelihood target (MLHL, Pannu *et al.*, 1998). Some other target functions are available for research purposes, for example the quadratic approximation of ML (LS*, Lunin & Urzhumtsev, 1999) or LS with different types of weighting and scaling schemes. The underlying core algorithms can be found in the `cctbx.xray.target_functors` module.

ADP refinement

In the refinement of Atomic Displacement Parameters (ADP) the target T_{adp} is minimized w.r.t. isotropic *ADP*s while all other model parameters are fixed. E_{adp} is defined as:

$$E_{adp} = \sum_{i=1}^{N_{atoms}} \left[\sum_{j=1}^{M_{atoms}} \frac{1}{r_{ij}^{k}} \frac{(B_i - B_j)^2}{B_i + B_j} \right]$$

Here N_{atoms} is the total number of atoms in the model, the inner sum is extended over all M_{atoms} in the sphere of radius R around atom i, r_{ij} is a distance between two atoms i and j, B_i and B_j are the corresponding isotropic ADPs and k is user-defined constant. By default, R and k are fixed at 5.0Å and 1.0, respectively, but they can also be refined. This "3 in 1" target function makes use of the following ideas:

- A bond is almost rigid, therefore the ADPs of bonded atoms are similar (Hirshfeld, 1976);
- ADPs of spatially close (non-bonded) atoms are similar (Schneider, 1996);
- The bond rigidity, and therefore the difference between the ADPs of bonded atoms, is related to the absolute values of the ADPs. Atoms with higher ADPs can have larger differences (Ian Tickle, CCP4 Bulletin Board, letter from March 14, 2003).

3: Conclusion

The Phenix refinement framework is a rapidly growing set of modular, reusable refinement tools, designed for future development of ever more integrated, highly automated structure determination methods. To enable collaboration among all developers, the core libraries are made available to the community as open source.

4: Acknowledgments

We gratefully acknowledge the financial support of NIH/NIGMS through grants 5P01GM063210, 5P50GM062412 and 1R01GM071939. Our work was supported in part by the US department of Energy under Contracts No. DE-AC03-76SF00098 and DE-AC02-05CH11231.

5: References

- Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl. Acad. Sci*. **94**, 5018-5023.
- Adams P.D., Gopal K., Grosse-Kunstleve R.W., Hung L.-W., Ioerger T.R., McCoy A.J., Moriarty N.W., Pai R.K., Read R.J., Romo T.D., Sacchettini J.C., Sauter N.K., Storoni L.C. and Terwilliger T.C. (2004). *J. Synchrotron Rad.* **11**, 53-55.
- Afonine, P.V., Grosse-Kunstleve, R.W. & Adams, P.D. (2005). *Acta Cryst.* D**61**, 850-855.
- Agarwal, R.C. (1978). *Acta Cryst.* A34, 791-809.
- Badger, J. (1997). *Methods Enzymol*. **277**, 344-352.

- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). *Nucleic Acids Research*, **28**, 235-242.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.
- Brünger, A. T., Kuriyan, J., Karplus, M. (1987). *Science*. **235**, 458- 460.
- Brünger, A.T., Karplus, M. & Petsko, G.A. (1989). *Acta Cryst.* A**45**, 50-61.
- Brünger, A.T. (1992). *Nature (London),* **355**, 472-474.
- Brünger, A.T., Adams, P.D., Clore, G.M., DeLabo, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. & Warren, G.L. (1998). *Acta Cryst.* **D54**, 905-921.
- Brünger, A. T & Adams, P. D. (2002). *Acc. Chem. Res*. **35**, 404-412.
- Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760-763.
- Grosse-Kunstleve, R.W., Sauter, N.K., Moriarty, N.W. & Adams, P.D. (2002). *J. Appl. Cryst.* **35**, 126-136.
- Grosse-Kunstleve, R.W., Afonine, P.V., Adams, P.D. (2004). *Newsletter of the IUCr Commission on Crystallographic Computing*, **4**, 19-36.
- Grosse-Kunstleve, R.W., Afonine, P.V., Sauter, N.K., Adams, P.D. (2005). *Newsletter of the IUCr Commission on Crystallographic Computing*, **5**, 69-91.
- Hirshfeld, F.L. (1976). *Acta Cryst.* A**32**, 239-244.
- Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol*. **243**, 100-115.
- Kostrewa, D. (1997). *CCP4 Newsl*. **34**, 9-22.
- Lamzin, V.S. & Wilson, K.S. (1997). *In Methods in Enzymology*. (Carter, C. & Sweet, B. eds.) **277**, 269-305
- Liu, D.C. & Nocedal, J. (1989). *Mathematical Programming*, **45**, 503-528.
- Lunin, V.Y. & Skovoroda, T.P. (1995). *Acta Cryst*., A**51**, 880-887.
- Lunin, V.Y., Urzhumtsev, A.G. (1999). *CCP4 Newsletter on Protein Crystallography*, **37**, 14-28.
- Murshudov, G.N., Davies, G.J., Isupov, M., Krzywda, S., Dodson, E.J. (1998). *CCP4 Newsletter on Protein Crystallography*, **35**, 37-43.
- Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst*. **D54**, 1285-1294.
- Phillips, S. E. V. (1980). *J. Mol. Biol*. **142**, 531-554.
- Read, R.J. (1986). *Acta Cryst*. A**42**, 140-149.
- Schneider, T. (1996). *Proceedings of the CCP4 Study Weekend*. SERC Daresbury Laboratory, Daresbury, U.K., 133-144.
- Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol*. **277**, 319-343.
- Sheriff, S. & Hendrickson, W. A. (1987). Acta Cryst. A**43**, 118-121.
- Urzhumtsev, A.G. (2000). *CCP4 Newsl*. **38**, 38-49.
- Urzhumtsev, A., Skovoroda, T.P. & Lunin, V.Y. (1996). *J.Appl.Cryst*., **29**, 741-744.
- Vagin, A.A. & Murshudov, G.N. (2004). *Newsletter of the IUCr Commission on Crystallographic Computing*, **4**, 59-72.
- Vagin, A.A., Steiner, R.A., Lebedev, A.A, Potterton, L., McNicholas, S., Long, F. & Murshudov, G.N. (2004). *Acta Cryst*. D**60**, 2184-2195.

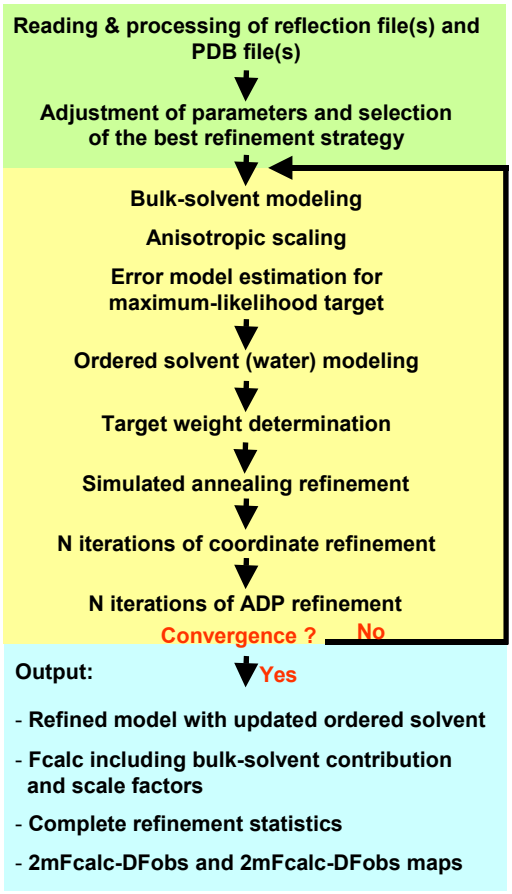

Figure 1. Phenix refinement protocol.

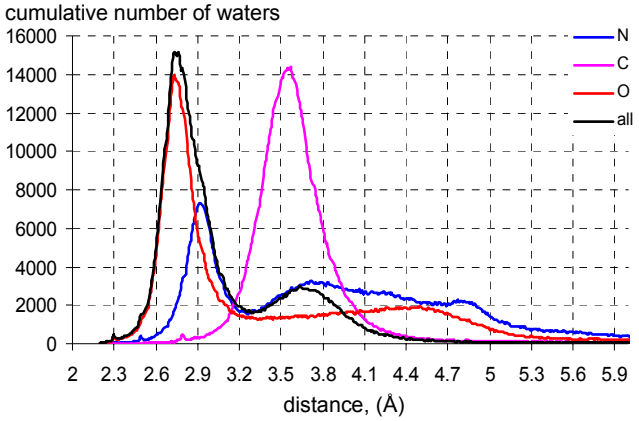

Figure 2. Statistics over high-resolution PDB models: distance distribution for water molecules; blue: water-protein N; magenta: water-protein C; red: water-protein O; black: sum of the three distributions.