Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

mrtailor: a tool for PDB-file preparation for the generation of external restraints

Tim Gruene

Department of Structural Chemistry, Georg-August-University Göttingen, Tammannstrasse 4, D-37077 Göttingen, Germany

Correspondence e-mail: tg@shelx.uni-ac.gwdg.de

Received 20 January 2013
Accepted 13 June 2013

Model building starting from, for example, a molecular-replacement solution with low sequence similarity introduces model bias, which can be difficult to detect, especially at low resolution. The program *mrtailor* removes low-similarity regions from a template PDB file according to sequence similarity between the target sequence and the template sequence and maps the target sequence onto the PDB file. The modified PDB file can be used to generate external restraints for low-resolution refinement with reduced model bias and can be used as a starting point for model building and refinement. The program can call *ProSMART* [Nicholls *et al.* (2012), *Acta Cryst.* D**68**, 404–417] directly in order to create external restraints suitable for *REFMAC*5 [Murshudov *et al.* (2011), *Acta Cryst.* D**67**, 355–367]. Both a command-line version and a GUI exist.

1. Introduction

Large structures are often composed of subunits of known structure. Most programs for macromolecular structure refinement allow the use of such structures as external geometric restraints, *e.g. BUSTER-TNT* (Blanc *et al.*, 2004), *CNS* (Brunger, 2007), *phenix.refine* (Adams *et al.*, 2010) and *REFMAC*5 (Murshudov *et al.*, 2011).

Especially at medium to low resolution, external restraints can lead to model bias, which may be difficult to detect and to correct for (Nicholls *et al.*, 2012). In order to reduce model bias, sequence alignment can be taken into account to remove regions in the template PDB file which are likely to be different from the target structure. The program *mrtailor* was developed to modify a PDB file based on sequence alignment prior to the generation of external restraints. It also provides an interface to the program *ProSMART* (Nicholls *et al.*, 2012).

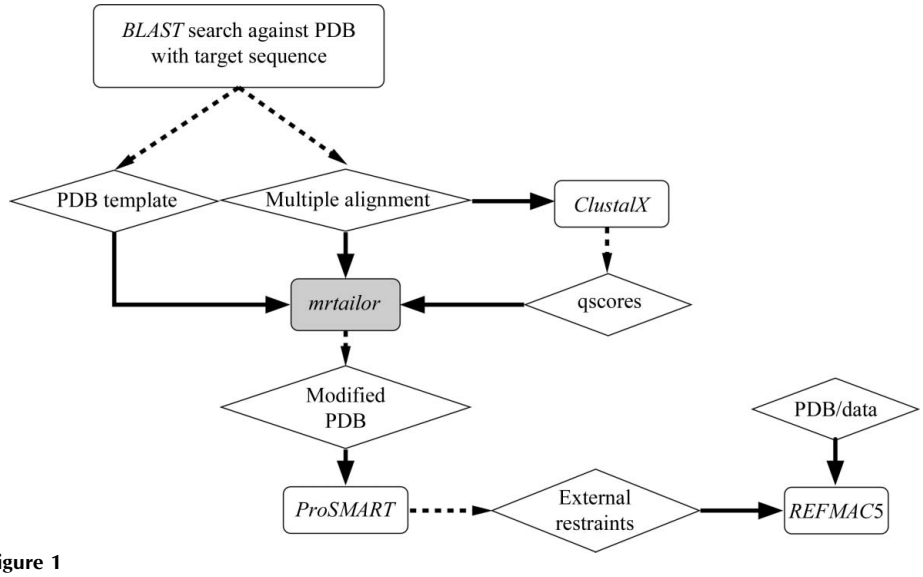

Figure 1
Workflow illustrating the use of *mrtailor*. Boxes enclose programs and diamonds enclose files; dashed arrows show program output and continuous arrows show program input. Considering only programs, the workflow is *BLAST*→*mrtailor*→*ProSMART*→*REFMAC*5. 'qscores' denotes the column scores file written by *ClustalX* for the target sequence.

2. *mrtailor*: features

Fig. 1 places *mrtailor* into context between sequence alignment and the generation of external restraints for *REFMAC*5 using *ProSMART*.

2.1. Input

The input data to *mrtailor* are the following.

(i) A (multiple) sequence alignment, for example, created by *COBALT* (Papadopoulos & Agarwala, 2007) from a *BLAST* (McGinnis & Madden, 2004) search of the target sequence against the PDB.

(ii) The name of the target sequence in the alignment file.

(iii) The name of the template sequence in the alignment file.

(iv) The template structure in PDB format.

(v) Optionally, a column scores file from *ClustalX* (Larkin *et al.*, 2007) with respect to the target sequence.

(vi) Optionally, a cutoff value for the column scores file. The cutoff defaults to 20 in *mrtailor*.

2.2. Output

The example alignment in Fig. 2 illustrates how *mrtailor* modifies the template PDB file.

(i) Identical residue types: coordinates not modified.

(ii) Different residue types: only N, C^{α}, C^{β}, C, O copied from the template PDB file, residue type changed from template to target, *e.g.* from valine to threonine in Fig. 2. C^{β} is excluded for glycines.

(iii) Gap in target sequence: residue removed from the template PDB file, numbering of C-terminal residues shifted by −1.

(iv) Gap in template sequence: renumbering and renaming, *e.g.* F194–S195 to F397–I407 in Fig. 2.

(v) In cases where a *ClustalX* column scores file is provided, the coordinates of all residues below the cutoff are removed.

(vi) Chains which do not match the template sequence are copied unmodified.

The sequence numbering is always adjusted to the target sequence, and ANISOU and HETATM lines are always removed from the PDB file.

2.3. User interface

Both a command-line version called *mrtailor* and a graphical user interface (GUI) called *mrtailor-gui* exist. The command-line version prints usage instructions to the terminal when it is called without an argument; an example for using the GUI is available online (Gruene, 2013).

3. Two applications of *mrtailor*

The output PDB file from *mrtailor* is primarily intended as input to *ProSMART*, but as shown in the following section it can be beneficial to also use it for refinement.

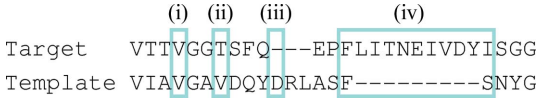

Figure 2
Example alignment to explain how *mrtailor* works. (i)–(iv) correspond to the items in §2.2.

3.1. PDB file used in refinement

In order not to decrease the phasing information by removing too many scatterers from the PDB file used for refinement, but still to correct the residue numbering and residue types, *mrtailor* should be run without using a *ClustalX* column scores file. Removing too much phasing information from the PDB file may make map interpretation after refinement unnecessarily difficult.

Fig. 3 shows an example of the reduction of model bias by *mrtailor*. Thr142–Val157 from the template PDB entry 1dbi (Smith *et al.*, 1999) match Gln338–Thr353 of the target PDB entry 3ee6 (Pal *et al.*, 2009), but the two structures deviate N-terminally. According to the sequence alignment in Fig. 3(*b*), *mrtailor* inserts a gap into the template PDB file. After refinement, the 16 common C^{α} atoms of the grey target and the green template have an r.m.s.d. of 1.1 Å (Kleywegt & Jones, 1997). The red trace in Fig. 3(*a*) was refined with the unmodified template PDB file and without external restraints and has an r.m.s.d. of 1.7 Å to the grey target trace. In the presence of external restraints in addition to the unmodified template PDB file for refinement, the r.m.s.d. of the orange trace is 2.0 Å, *i.e.* the gap between His140 and Thr142 introduced by *mrtailor* allows the helix to be refined towards the correct position.

3.2. PDB file used as source of external restraints

The main purpose of *mrtailor* is the tailoring of a PDB file to yield reduced model bias from external restraints during refinement. Sequence alignment is used as criterion for the tailoring carried out by *mrtailor* according to §2.2. *mrtailor* improves the reliability of sequence alignment as source of structural similarity by taking scores

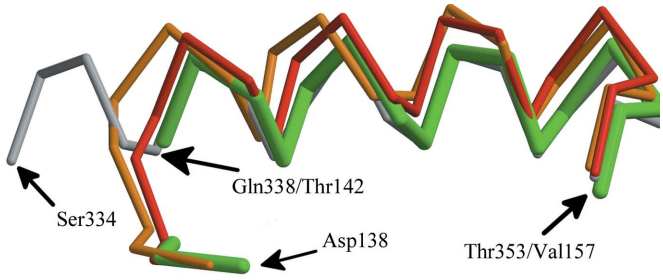

Figure 3
(*a*) Example of removal of model bias by the introduction of a gap in the template PDB file as explained in §3.1. (*b*) The corresponding sequence alignment with the gap highlighted in cyan. This figure was created with *MolScript* (Kraulis, 1991), *Raster3D* (Merritt & Bacon, 1997) and *GIMP* (v.2.8; http://www.gimp.org).

```
gi|6573500    1DBI   1:W  T  P  N  D  T  Y  Y  -:  8
gi|17943333           -  -  -  -  -  -  -  -  -
gi|51247507           T  L  D  R  A  H  V  A  A
gi|213424332          Q  K  C  H  S  V  I  T  Q
gi|215261288  3EE6 109:Q  K  C  H  S  V  I  T  Q:117
multiple alignment     0  0  0  0  0  0  0  0  0
pairwise alignment     4 43  8 34 35 28 16 11  2
```

Figure 4
Extract from the multiple sequence alignment of the sequence of PDB entry 3ee6. The distance between Trp2 from 1dbi and Lys110 from 3ee6 is greater than 50 Å. The benefit of using the scores from multiple sequence alignment over those from pairwise sequence alignment is explained in §3.2.

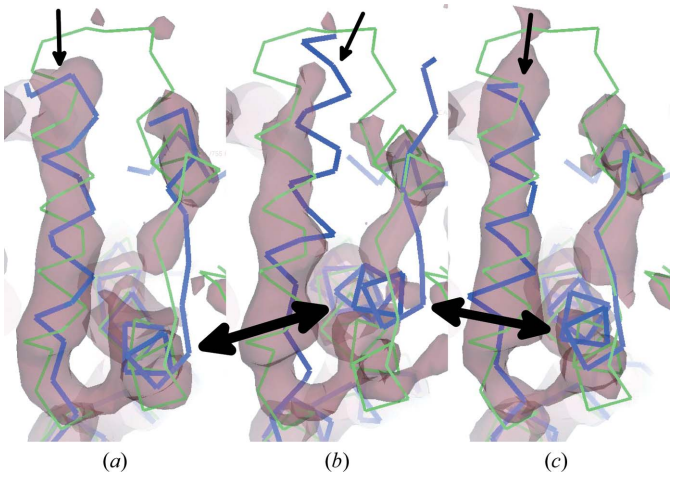

Figure 5
Illustration of how external restraints can hamper refinement. Green, thin C^{α} trace, target PDB entry 2y9y (Yamada *et al.*, 2011). Blue, bold C^{α} trace, PDB entry 1ofc (Grüne *et al.*, 2003) as a template refined against the 2y9y data. (*a*) Refinement without external restraints. (*b*) External restraints from the unmodified model 1ofc. (*c*) External restraints from 1ofc tailored by *mrtailor*. Both cases (*a*) and (*c*) allow the two helices marked with arrows to move into density during refinement, whereas the external restraints generated from the unmodified 1ofc coordinates in case (*b*) keep the model out of the density. The figures are screenshots from *Coot* (Emsley *et al.*, 2010) and were labelled with *GIMP* (v.2.8; http://www.gimp.org).

based on multiple sequence alignment into account. For example, after superposition the distance between Lys110 from PDB entry 3ee6 and Thr2 from PDB entry 1dbi is greater than 50 Å. Without scores, the sequence alignment in Fig. 4 would result in the truncation of residues Trp1–Tyr8 to alanines but not their entire removal. The *ClustalX* scores from multiple sequence alignment clearly lead to removal of the complete stretch. For comparison, the scores calculated from pairwise sequence alignment are much more ambiguous in deciding about a structural relationship between the template and target.

Fig. 5 shows an example in which external restraints from an unmodified PDB file result in two helices not being refined into their electron density (Fig. 5*b*), whereas both refinement without external restraints (Fig. 5*a*) and refinement with external restraints after using *mrtailor* (Fig. 5*c*) allow the desired shift. Note the difference from Fig. 3: the shifts described here are not owing to gaps in the PDB file used for refinement but are owing to a lack of improper external restraints.

4. Conclusions

The program *mrtailor* modifies a PDB file in a similar fashion as, for example, the programs *CHAINSAW* (Stein, 2008) or *Sculptor* (Bunkóczi & Read, 2011) do. These latter two programs are intended for the preparation of PDB files for molecular replacement, whereas *mrtailor* is intended as an improvement step before the generation of external restraints or for bias removal from a PDB file before refinement. With these two purposes in mind, *mrtailor* can be used repeatedly in the presence of a heteromeric multi-subunit PDB file because it does not remove chains that do not match the template sequence. In addition to a simple comparison between two sequences, *mrtailor* takes scores from a multiple sequence alignment into account. The work presented here shows that this feature can enhance the reliability at which structural similarity is deduced from sequence similarity and that a template PDB file tailored using *mrtailor* can lead to reduced model bias.

The importance of external restraints increases as the data-set resolution and the model quality decrease. At the same time, it becomes impossible to predict which method will produce the best results and the user is strongly advised to test several possible paths, *i.e.* using or not using *mrtailor*, using *Sculptor* instead of *mrtailor* or using or not using external restraints.

5. Availability

mrtailor and *mrtailor-gui* are available for download subject to the GNU General Public License (v.3; http://www.gnu.org/licenses/gpl.html) from TG's homepage (Gruene, 2013).

TG appreciates useful support from and constructive discussion with Rob Nicholls. Moritz Wette tested the online tutorial. TG was supported by the Volkswagen Stiftung *via* the Niedersachsen-professur awarded to Professor G. M. Sheldrick.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 2210–2221.
Brunger, A. T. (2007). *Nature Protoc.* **2**, 2728–2733.
Bunkóczi, G. & Read, R. J. (2011). *Acta Cryst.* D**67**, 303–312.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Gruene, T. (2013). *Online Tutorial for mrtailor*. http://shelx.uni-ac.gwdg.de/~tg/research/programs/mrtailor/tutorial.
Grüne, T., Brzeski, J., Eberharter, A., Clapier, C. R., Corona, D. F. V., Becker, P. B. & Müller, C. W. (2003). *Mol. Cell*, **12**, 449–460.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 525–545.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. (2007). *Bioinformatics*, **23**, 2947–2948.
McGinnis, S. & Madden, T. L. (2004). *Nucleic Acids Res.* **32**, W20–W25.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Nicholls, R. A., Long, F. & Murshudov, G. N. (2012). *Acta Cryst.* D**68**, 404–417.
Pal, A., Kraetzner, R., Gruene, T., Grapp, M., Schreiber, K., Grønborg, M., Urlaub, H., Becker, S., Asif, A. R., Sheldrick, G. M. & Steinfeld, R. (2009). *J. Biol. Chem.* **284**, 3976–3984.
Papadopoulos, J. S. & Agarwala, R. (2007). *Bioinformatics*, **23**, 1073–1079.
Smith, C. A., Toogood, H. S., Baker, H. M., Daniel, R. M. & Baker, E. N. (1999). *J. Mol. Biol.* **294**, 1027–1040.
Stein, N. (2008). *J. Appl. Cryst.* **41**, 641–643.
Yamada, K., Frouws, T. D., Angst, B., Fitzgerald, D. J., DeLuca, C., Schimmele, K., Sargent, D. F. & Richmond, T. J. (2011). *Nature (London)*, **472**, 448–453.