electronic reprint

Journal of
Applied Crystallography
ISSN 0021-8898
Editor: **Gernot Kostorz**

Molecular replacement with multiple different models

Nicholas M. Glykos and Michael Kokkinidis

Journal of
Applied Crystallography
ISSN 0021-8898

Received 1 November 2003
Accepted 28 November 2003

Molecular replacement with multiple different models

Nicholas M. Glykos[a]*‡ and Michael Kokkinidis[a,b]

[a]IMBB, FORTH, PO Box 1527, 71110 Heraklion, Crete, Greece, and [b]Department of Biology, University of Crete, PO Box 2208, 71409 Heraklion, Crete, Greece. Correspondence e-mail: glykos@imbb.forth.gr

Classical molecular replacement methods and the newer six-dimensional searches treat molecular replacement as a succession of sub-problems of reduced dimensionality. Due to their 'divide-and-conquer' approach, these methods necessarily ignore (at least during their early stages) the very knowledge that a target crystal structure may comprise, for example, more than one copy of a search model, or several models of different types. An algorithm for a stochastic multi-dimensional molecular replacement search has been described previously and shown to locate solutions successfully, even in cases as complex as a 23-dimensional 4-body search. The original description of the method only dealt with a special case of molecular replacement, namely with the problem of placing n copies of only one search model in the asymmetric unit of a target crystal structure. Here a natural generalization of this algorithm is presented to deal with the full molecular replacement problem, that is, with the problem of determining the orientations and positions of a total of n copies of m different models (with $n \geq m$) which are assumed to be present in the asymmetric unit of a target crystal structure. The generality of this approach is illustrated through its successful application to a 17-dimensional 3-model problem involving one DNA and two protein molecules.

1. Introduction

Traditional molecular replacement methods (Rossmann, 1990; Navaza, 1994; Navaza & Saludjian, 1997) and the newer six-dimensional searches (Chang & Lewis, 1997; Kissinger *et al.*, 1999, 2001; Jamrog *et al.*, 2003) use a 'divide-and-conquer' approach to tackle molecular replacement problems. Nevertheless, and as the success of these methods convincingly demonstrates, the heuristic lying behind those approaches (namely, the properties of the Patterson function and its equivalence to the measured crystallographic data) is so powerful that treating the molecular replacement problem in its full dimensionality is more often than not a waste of computational resources. It is only when the assumptions behind those methods break, that a full multi-dimensional molecular replacement search could provide a useful alternative. We have previously described an algorithm (and its implementation in the form of a computer program) for a stochastic multi-dimensional molecular replacement search (Glykos & Kokkinidis, 2000; Glykos & Kokkinidis, 2001; hereinafter referred to as GK0 and GK1, respectively) and have recently applied it to the determination of a previously unknown structure *via* a 23-dimensional 4-body search (Glykos & Kokkinidis, 2003). Here we report an extension of this algorithm (and of its corresponding computer program) to tackle the problem of simultaneously and independently placing an arbitrary number of different search models in the asymmetric unit of a target crystal structure.

‡ Present address: Democritus University of Thrace, Department of Molecular Biology and Genetics, Dimitras 19, Alexandroupolis, 68100, Greece.

2. Formulation of the problem

The formulation of the problem follows closely its original derivation (GK0, GK1) and is as follows. If the asymmetric unit of a target crystal structure with s crystallographic symmetry operators comprises m different search models, each of which is present in c_m copies, with each model (m) consisting of i_m atoms, then the structure factor corresponding to the Bragg reflection $\mathbf{h}$ (and omitting for brevity the required orthogonalization and de-orthogonalization matrices) is

$$\mathbf{F}_{c,\mathbf{h}} = \sum_s \sum_m \sum_{c_m} \sum_{i_m} f_{m,i_m,\mathbf{h}} \exp[-B_m(\sin\theta_{\mathbf{h}}/\lambda)^2] \times \exp\{2\pi i \mathbf{h}[\mathcal{R}_s(\mathcal{R}_{m,c_m}\mathbf{x}_{i_m} + \mathcal{T}_{m,c_m}) + \mathcal{T}_s]\},$$

where, f_{m,i_m} is the atomic scattering factor of the atom i_m of the model m, B_m is an overall temperature factor assigned to model m, $\theta_{\mathbf{h}}$ is the Bragg angle corresponding to the reflection $\mathbf{h}$, $\mathcal{R}_s$ and $\mathcal{T}_s$ are the rotation and translation matrices corresponding to the crystallographic symmetry operator s, $\mathbf{x}_{i_m}$ are the orthonormal coordinates of the atom i_m of the model m in a standard (fixed) molecular reference frame, and finally, $\mathcal{R}_{m,c_m}$ and $\mathcal{T}_{m,c_m}$ are the rotation and translation matrices describing the orientation and position of the c_m copy of the m model in the orthonormal frame.

Given a set of trial $\mathcal{R}_{m,c_m}$ and $\mathcal{T}_{m,c_m}$ (and possibly after correcting F_c values with an overall scale and temperature factor), we can calculate the value of a target function (in this case the linear correlation coefficient between the observed and calculated structure-factor amplitudes) as follows:

Table 1
Final statistics for the five independent *Queen of Spades* minimizations.

The values shown correspond to those reported by the program after completion of ten thousand steps of Monte Carlo rigid-body minimization of the best solutions encountered during each of the five minimizations.

Minimization	$1.0 - \mathrm{Corr}(F_o, F_c)$	Free value
1	0.5215	0.5061
2	0.5641	0.5639
3	0.5215	0.5092
4	0.5600	0.5707
5	0.5210	0.5190

$$\mathcal{C}(\mathcal{R}_{m,c_m}, \mathcal{T}_{m,c_m}) = \frac{\sum_{\mathbf{h}}(F_{o,\mathbf{h}} - \bar{F}_o)(F_{c,\mathbf{h}} - \bar{F}_c)}{\left[\sum_{\mathbf{h}}(F_{o,\mathbf{h}} - \bar{F}_o)^2\right]^{1/2}\left[\sum_{\mathbf{h}}(F_{c,\mathbf{h}} - \bar{F}_c)^2\right]^{1/2}},$$

where $F_{o,\mathbf{h}}$ and $F_{c,\mathbf{h}}$ are the observed and calculated structure-factor amplitudes of the reflection $\mathbf{h}$, $\bar{F}_o$ and $\bar{F}_c$ are their respective means, and the sums are taken over all Miller indices ($\mathbf{h}$).

The aim of our algorithm is the unconditional global optimization of $\mathcal{C}$ as a function of $\mathcal{R}_{m,c_m}$ and $\mathcal{T}_{m,c_m}$. The matrices $\mathcal{R}_{m,c_m}$ and $\mathcal{T}_{m,c_m}$ for which the value of the target function is optimized are the solution of our method.

3. Method of solution and implementation

The unconditional global optimization of the target function is based on a modified reverse Monte Carlo algorithm (McGreevy & Pusztai, 1988; Keen & McGreevy, 1990). A complete account of the method has been presented previously (see §2.1 of GK0 and §§2.2–2.3 of GK1) and will not be repeated here.

The algorithm was implemented by modifying the program *Queen of Spades* (see §6 for program availability details). The annealing schedules, move-size control and other implementation details have been extensively discussed in §§2 and 3 of GK1. Here we will only discuss features and limitations of the program that arose as a result of implementing support for multiple different models. As previously, we traded physical memory requirements for speed of execution by calculating the molecular transforms of all models on the same grid (see §2.2 of GK1). The implication of this choice is that for models of considerably different size (or proportions) one or more of the molecular transforms may by unnecessarily over-sampled. The advantage is that the inner loop of calculating the structure factors of the current trial structure (which is where the program spends more than 85% of its time) can proceed *via* one and the same interpolation function. A second limitation arising for similar reasons is that in the multi-model mode the program will calculate a molecular transform for each and every of the search models, even if these correspond to the same molecule. In other words, if there is a total of n copies of m different models, the current implementation of the program will calculate and store in memory n (and not m) molecular transforms. For $n \gg m$ this choice can become rather wasteful with respect to physical memory resources. One other important difference from previous versions of the program concerns the annealing schedule and target function selected by default. This version uses a Boltzmann annealing schedule with a fixed (resolution-dependent) move-size, a fixed starting temperature, and a target function corresponding to the unweighted linear correlation coefficient between the observed and calculated structure-factor amplitudes. These choices bring the program more in-line with the current thinking in the field, reflect the strong theoretical foundation of Boltzmann annealing and of the correlation-based targets, and, finally, are identical to the choices made in all successful high-dimensionality applications of previous versions of the program.

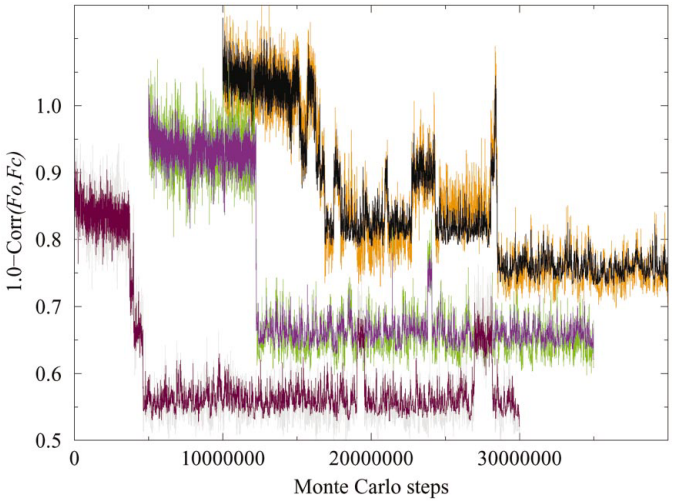

Figure 1
Evolution of the average values of the target function $[1.0 - \mathrm{Corr}(F_o, F_c)]$ *versus* Monte Carlo moves for the three successful *Queen of Spades* minimizations. To minimize overlap, successive graphs have been translated by 0.1 unit along y, and by five million Monte Carlo steps along x. Each graph includes both the target function (foreground, dark colour) and its cross-validated value (background, light colour).

4. Application to a 17-dimensional 3-model problem

The structure chosen to illustrate the application of the program is the complex of NFAT, Fos and Jun with DNA, corresponding to the PDB entry 1A02 (Chen *et al.*, 1998). To reduce the dimensionality of the problem we have treated Fos and Jun as one (rigid) search model, leaving us with three models (NFAT, Fos-Jun and the DNA molecule). We used real data deposited with the PDB (entry r1a02sf.ent), and to make the example more realistic we modified the deposited coordinates by subjecting (without experimental restraints) the starting models to energy minimization with *CNS* (Brünger *et al.*, 1998) for 500 cycles. The resulting models deviated from the deposited structures with r.m.s. deviations of 1.1, 1.5 and 2.2 Å for the NFAT, Fos-Jun and DNA, respectively, and were used as search models in the ensuing molecular replacement calculations. The *Queen of Spades* run was performed in its fully automatic mode (simply by entering `Qs -auto 3`), requiring only the PDB files containing the search models and a free-format ASCII file containing the observed data that the program should use for the calculation (all data between 19.5 and 4 Å resolution for this example). In this default mode, the program performed five independent minimizations, each lasting 30 million Monte Carlo moves and taking approximately 73 h of CPU time on a personal computer equipped with a 1.8 GHz Pentium IV processor, 1 GByte of random access memory and a proper operating system (GNU/Linux, RedHat distribution v.7.3). The total physical memory requirements of the program amounted to 196 MBytes (a significant proportion of which corresponds to high resolution volumes of the molecular transforms and, thus, does not have to be resident in memory). All minimizations used the strongest 70% of all reflections with $F/\sigma(F) > 2.0$ (4943 reflections in total), with 10% of these being reserved for statistical cross-validation (Brünger, 1997). The target function for the minimization was $[1.0 - \mathrm{Corr}(F_o, F_c)]$, where $\mathrm{Corr}(F_o, F_c)$ is the linear correlation coefficient between the observed and calculated struc-

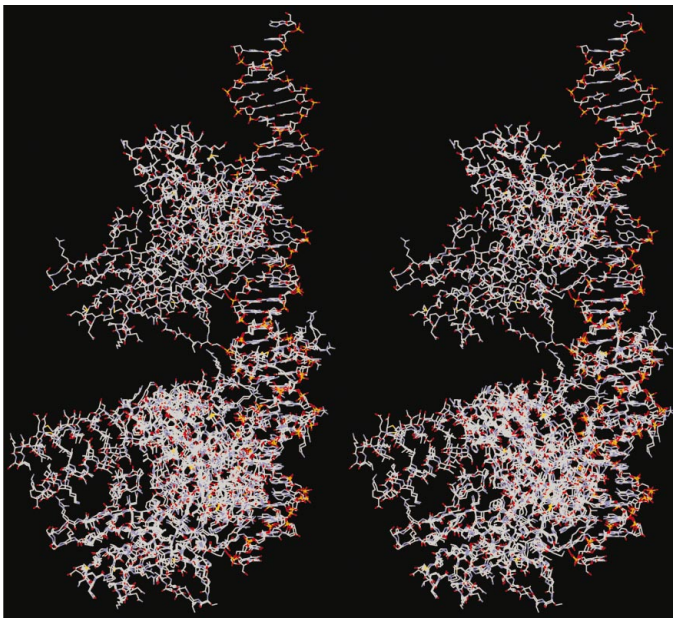

Figure 2
Stereodiagram of a small portion from the packing arrangement corresponding to the first *Queen of Spades* minimization showing the characteristic stacking of the DNA molecules to form a continuous helix, and the relative placement of the protein molecules. For comparison, the lower half of the figure also shows superimposed a wireframe representation of the deposited (1a02.pdb) coordinates. Figure prepared with the program *RasMol*.

ture-factor amplitudes. A Boltzmann annealing schedule was used with the temperature T at step k given by $T = T_o/\log(k)$, where T_o is the starting temperature for the minimization [set to 0.070 (arbitrary units) in the default program mode].

Table 1 shows the final statistics for each of the five minimizations, and Fig. 1 shows the evolution of the average values of the target function (and its cross-validated counterpart) *versus* Monte Carlo moves for three minimizations. All five minimizations resulted in closely related structures, with three of them (first, third and fifth minimization, Fig. 1) converging to the correct crystal structure. This is illustrated in Fig. 2, which compares the results obtained from the first minimization with the deposited coordinates. As Fig. 1 shows, all minimizations located correct or partially correct structures early during the simulations, with the first minimization converging to the correct crystal structure after only five million moves. This behaviour (and the high success rate of the algorithm for such a high-dimensionality problem) is probably the result of the (atypically) accurate search models used for the calculation.

5. Discussion

We have described what, to our knowledge, is the most general formulation and method of solution of the molecular replacement problem. The method uses all information available at hand for a given problem (both with respect to the number and type of the search models, but also with respect to the measured crystallographic data) without resorting to Patterson-function-based heuristics. The algorithm only demands that the true crystal structure is the one for which the agreement between the observed and calculated structure-factor amplitudes is maximized. As usually is the case, this generality of treatment comes at a very high computational cost. The calculations described in section §4 could have been performed equally well with classical molecular replacement methods at a computational cost three orders of magnitude lower than the one reported here. To reiterate what has already been said, it is only when the assumptions behind the traditional methods break, that this multi-dimensional multi-model molecular replacement search could provide a useful alternative.

6. Program and data availability

The program *Queen of Spades* is free open-source software which is immediately available for download *via* http://origin.imbb.forth.gr/~glykos/ or from the various mirrors generously provided by the Collaborative Computational Project, Number 14. All files from the application of the program described in section §4 can be obtained *via* http://origin.imbb.forth.gr/~glykos/Qs_1A02.tar.gz.

We should like to thank Professor Jorge Navaza for his useful comments on the manuscript and for suggesting that it would be a worthwhile exercise writing the code for a multi-model mode of *Qs*.

References

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Sinonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.
Chen, L., Glover, J. N. M., Hogan, P., Rao, A. & Harrison, S. (1998). *Nature (London)*, **392**, 42–48.
Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.
Glykos, N. M. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462–1473.
Glykos, N. M. & Kokkinidis, M. (2003). *Acta Cryst.* D**59**, 709–718.
Jamrog, D. C., Phillips, G. N. J. & Zhang, Y. (2003). *Acta Cryst.* D**59**, 304–314.
Keen, D. A. & McGreevy, R. L. (1990). *Nature (London)*, **344**, 423–425.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kissinger, C. R., Gehlhaar, D. K., Smith, B. A. & Bouzida, D. (2001). *Acta Cryst.* D**57**, 1474–1479.
McGreevy, R. L. & Pusztai, L. (1988). *Mol. Simul.* **1**, 359–367.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Navaza, J. & Saludjian, P. (1997). *Methods Enzymol.* **276**, 581–593.
Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.