Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the potential of normal-mode analysis for solving difficult molecular-replacement problems

Karsten Suhre[a]* and Yves-Henri Sanejouand[b]

[a]Information Génomique et Structurale (UPR CNRS 2589), 31 Chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France, and [b]Laboratoire de Physique, Ecole Normale Supérieure, 46 Allées d'Italie, 69364 Lyon CEDEX 07, France

Correspondence e-mail: karsten.suhre@igs.cnrs-mrs.fr

Received 14 November 2003
Accepted 26 January 2004

Molecular replacement (MR) is the method of choice for X-ray crystallographic data phasing when structural data of suitable homologues are available. However, MR may fail even in cases of high sequence homology when conformational changes arising for example from ligand binding or different crystallogenic conditions come into play. In this work, the potential of normal-mode analysis as an extension to MR to allow recovery from such drawbacks is demonstrated. Three examples are presented in which screening for MR solutions with templates perturbed in the direction of one or two normal modes allows a valid MR solution to be found where MR using the original template failed to yield a model that could ultimately be refined. It has been shown recently that half of the known protein movements can be modelled by displacing the studied structure using at most two low-frequency normal modes. This suggests that normal-mode analysis has the potential to break tough MR problems in up to 50% of cases. Moreover, even in cases where an MR solution is available, this method can be used to further improve the starting model prior to refinement, eventually reducing the time spent on manual model construction (in particular for low-resolution data sets).

1. Introduction

Molecular replacement (MR) is the most cost-effective method for solving the three-dimensional structure of a new protein by X-ray crystallography. The success or failure of MR depends on a variety of factors, not least of which are the diffraction data quality in the mid- and low-resolution range and the number of molecules in the asymmetric unit cell that have to be placed. The decisive factor is the availability of a template of sufficiently high structural homology to the target. However, it is likely that most structural genomics projects have already encountered the frustrating situation where MR failed despite a 'good' template with high sequence similarity (or even identity) to the target. On the eventual phasing of the diffraction data by applying more time-consuming experimental methods such as multiple isomorphous replacement (MIR) and multiple-wavelength anomalous diffraction (MAD), it often turns out that the 'new' structure exhibits an important conformational change with respect to the original template, explaining *a posteriori* the failure of the MR attempt. *A priori* modelling of the most likely conformational changes that a given template might undergo would thus be of significant benefit and could eventually allow an increased number of crystal structures to be solved by MR.

Here, we propose normal-mode analysis (NMA) as a powerful tool for anticipating the most likely conformational changes of a given template and to screen its perturbed structures for possible MR solutions. The rationale behind this approach is as follows. It was noticed almost 20 years ago, using empirical force fields and a protein description at the atomic level, that one of the largest amplitude motions predicted by normal-mode theory for proteins, that is one of the lowest-frequency normal modes of motion, often compares well with their functional conformational change as observed by crystallographers upon ligand binding (Harrison, 1984; Brooks & Karplus, 1985). More recently, using much more simplified protein descriptions, namely elastic network models (Tirion, 1996; Bahar *et al.*, 1997; Hinsen, 1998), it was shown that of 3800 known protein motions more than half can be described well by perturbing the considered protein along the direction of at most two low-frequency normal modes (Krebs *et al.*, 2002), that is, by displacing the structure along two corresponding perpendicular directions of the configurational space. Moreover, when the collective character of the protein motion is obvious, a single low-frequency normal mode often proves to be sufficient and it is usually one of the three lowest-frequency modes (Tama & Sanejouand, 2001; Delarue & Sanejouand, 2002). Such results strongly suggest

Table 1
Overview of the crystallographic parameters, molecular-replacement results and refinement statistics.

	Maltodextrin-binding protein	HIV-1 protease	Glutamine-binding protein
Target	1omp	1ajx	1wdn
Template	1anf	1hhp	1ggg
No. residues	370	99	224
No. reflections†	5851	4280	3796
Completeness‡ (%)	97.9	78.9	97.1
Space group	$P1$	$P2_12_12_1$	$P2_12_12_1$
No. molecules	1	2	1
Best mode	Mode 7	Mode 11	Modes 7 + 8
Perturbation§	180	60	200 + 40
CC/*R* factor¶			
Target	86.0/20.9	73.2/31.3	72.1/29.9
Template	25.8/49.0	31.0/50.0	26.1/50.3
Best NMA	33.5/46.4	57.7/39.7	21.3/51.6
R_{work}/R_{free}††			
Target	16.5/23.3	35.4/38.6	25.9/35.9
Template	43.0/51.1	52.8/54.1	45.3/54.0
Best NMA	38.8/45.3	41.8/46.0	35.2/47.1

† Number of reflection theoretically available to a resolution of 3.2 Å. ‡ Completeness to a resolution of 3.2 Å. § Arbitrary units. Using Tirion's elastic network model, normal-mode frequencies as well as the corresponding unit for the displacements along a normal mode are defined through a scaling free factor, which was set to $k = 10$ in the present study. ¶ The CC and *R* factor of the best translation/rotation solution(s) found by *AMoRe* (Navaza, 1994) when using data to 3.2 Å resolution. †† Final *R* factor for the working and the test set after *CNS* (Brünger *et al.*, 1998) refinement using standard parameters to 3.2 Å resolution.

that protein movements between open and closed forms (*e.g.* without and with ligand) may actually be under selective pressure to follow mainly one or a few low-frequency normal modes of the protein.

This suggests a screening approach in which a given template is perturbed with varying amplitudes in the direction of one or two of its low-frequency normal modes. One would then look for minima in the resulting *R* values after standard MR, optionally followed by a simulated-annealing and/or an energy-minimization refinement step. To evaluate the potential of this idea, which has been considered once before in the case of the determination of the low-resolution structure of F-actin (Tirion *et al.*, 1995), we selected proteins for which the structures are known in two different conformations (Echols *et al.*, 2003) and for which structure factors have been deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000). Proteins with two distinct domains connected by a single linker peptide (such as the diphtheria toxin or immunoglobin) were excluded, as the standard MR procedure would be to attempt a two-body search in these cases. Also omitted are proteins with only small conformational changes as these would most likely be of no challenge in MR. The idea is then to use one of the two protein models as an MR template in order to solve the structure of the other protein. Here, we present results for three representative cases in which the original template failed to give a proper solution but the perturbed model did: maltodextrin-binding protein (Zanotti *et al.*, 1992), HIV-1 protease (Backbro *et al.*, 1997) and glutamine-binding protein (Hsiao *et al.*, 1996). Although we limit this analysis to cases with 100% sequence homology for the sake of simplicity, the ideas presented in this paper can readily be applied to the more general situation where the sequences of the template and target are different. In that case a number of different protocols can be applied, including using all-alanine or homology-based models in different stages of the procedure.

2. Methods

Protein models (targets and templates) and the corresponding structure factors were downloaded from the Protein Data Bank. Targets were orientated onto the templates using rigid-body superposition. All atomic *B* factors were set to a value of 10 Å^2 prior to MR. The MR search was performed using the stand-alone version of *AMoRe* in fully automatic mode and data to a resolution of 3.2 Å (Navaza, 1994). Based on the best rotation/translation function(s) found by *AMoRe*, the models were refined to 3.2 Å resolution using the standard protocol implemented in *CNS*: reorientation with the 'realspace_transform.inp' script followed by initial *B*-factor and bulk-solvent corrections and two cycles of simulated annealing (only applied for the glutamine-binding protein, as discussed below) and coordinate and individual *B*-factor minimization (constrained torsion-angle dynamics) as implemented in the 'refine.inp' script. All parameters of these script were kept to their default values (Brünger *et al.*, 1998). Root-mean-square deviations (r.m.s.d.s) between the C$^\alpha$ atoms of two proteins and the fraction of C$^\alpha$ atoms that are closer than 3 Å were computed using *LSQMAN* from the *DEJAVU* package (Kleywegt, 1996). Normal modes were computed based on ideas from Tirion (1996) as implemented by Tama & Sanejouand (2001) using the RTB approach (Durand *et al.*, 1994; Tama *et al.*, 2000). All calculations were performed in all-atom mode applying a 5 Å cutoff in the definition of the elastic interactions. A web interface to these tools for automatic computation of NMA-perturbed structures for MR (including the cases presented here as pre-computed examples) is available at http://igs-server.cnrs-mrs.fr/elnemo/index.html. The Fortran source code of the software is available at http://ecole.modelization.free.fr/modes.html. Normal modes are numbered by increasing frequency values, the first six zero-frequency modes being the three trivial rigid-body translational and rotational modes. The lowest non-trivial normal mode is thus mode 7.

3. Results

The maltodextrin-binding protein is a prime example for the application of NMA to MR (see Table 1 for details). It is constituted of two domains of about equal size. Upon ligand binding, the maltodextrin-binding protein undergoes a hinge movement from its open towards its closed conformation. We assume the open form to be our MR target as structure factors are only available for this model. Rigid-body superposition between template and target allows fitting of the larger of the two domains, which corresponds to 59% of the residues. The r.m.s.d. between the C$^\alpha$ atoms belonging to this domain is 0.98 Å. An MR search using the closed form as a template allowed the identification of the correct rotation/translation function, but even after a refinement step (energy minimization) the final free *R* factor of the resulting model remained at the quite high value of $R = 51.1\%$. This is because of the unmatched second domain of the protein. By perturbing the structure of the protein, following the lowest-frequency mode (mode 7), we find that this movement captures the conformational change quite well. Rigid-body superposition of the target and the perturbed template shows that 91% of the residues of the perturbed template have a r.m.s.d. of 1.2 Å with respect to the C$^\alpha$ atoms of the target. The total r.m.s.d. for all C$^\alpha$ atoms is 1.4 Å, compared with 3.8 Å in the case of the unperturbed template. Application of standard MR and refinement with the optimal

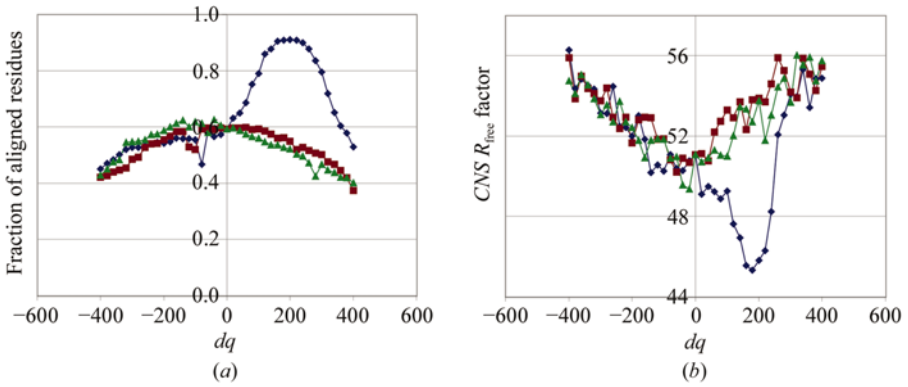

Figure 1
Maltodextrin-binding protein. Fraction of aligned residues, using *LSQMAN* and a cutoff distance of 3 Å (*a*) and *CNS* final free *R* factor (*b*), as a function of the perturbation *dq* (in arbitrary units) of the template along modes 7 (blue circles), 8 (red squares) and 9 (green triangles).

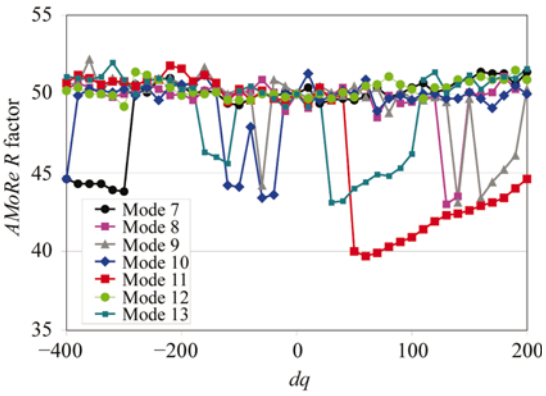

Figure 2
HIV-1 protease. *AMoRe R* factor as a function of the perturbation (in arbitrary units) of the template along one of its seven lowest-frequency normal modes.

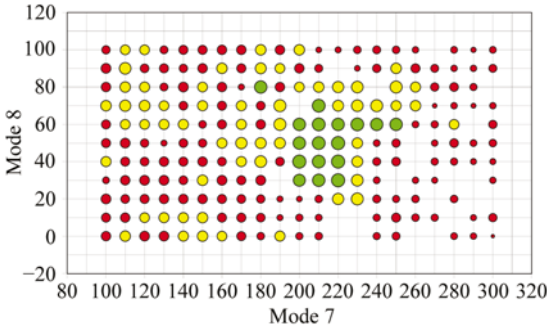

Figure 3
Glutamine-binding protein. *CNS* final free *R* factor as a function of the perturbation along modes 7 and 8; circle size is proportional to 65 − *R* factor; colour code: *R* factor better than 50, green; better than 55, yellow; other, red; missing circles in the screened range (100–300 arbitrary units for mode 7; 0–100 for mode 8; step size 10) indicate that *CNS* failed to refine the corresponding model.

perturbed model yields a final free *R* factor of 45.3%. An animated view of the predicted protein movement following its lowest-frequency mode is presented online (Suhre, 2004). A full analysis of the observed conformational changes between its opened and closed state is available *via* the molmovdb website (http://molmovdb.org).

This corresponds to a gain of almost 6% in the free *R* factor. Visual inspection of the corresponding electron-density map and further refinement to higher resolution shows that here a solution for the entire protein was found, whereas only 60% of the protein could be modelled using the original template. In Fig. 1, the result of the search for MR solutions is shown in which one of the three lowest normal modes (modes 7, 8 and 9) is used to apply the perturbation. In this case, the free *R* factor depends continuously on the perturbation and the optimum can already be identified after the MR step. Note that modes 8 and 9 do not yield any MR solution better than the original template.

Our second example, HIV-1 protease, pictures a case where two molecules per asymmetric unit have to be placed by MR. This is a relatively small protein with less than 100 residues and the conformational change between the two forms mainly concerns a single antiparallel β-sheet (see Suhre, 2004). Using one molecule of the original template, corresponding to the open non-liganded form of the protein, the correct position of the second molecule, with both molecules of the target being in the closed form, could not be found. Unsurprisingly, the final free *R* factor obtained with such a model is quite high, *R* = 55.1%. Application of an appropriate normal-mode perturbation along mode 11 eventually allows the identification of a suitable rotation/translation function which results in a sudden decrease of the *R* factor, the best model having a final free *R* factor of *R* = 46.0% (Fig. 2 and Table 1). Visual inspection of the corresponding electron-density maps shows that in this case NMA clearly yields a solution, whereas this is not the case when the original template is used. Note that some of the other modes also yield 'good' solutions.

As a third example, we chose glutamine-binding protein. This protein undergoes extensive conformational changes (hinge movement) between its open and closed forms. Rigid-body superposition of both models allows the fitting of a large domain that contains about 60% of the residues (134 of 224) with an r.m.s.d. of 0.8 Å, while the r.m.s.d. between all C^{α} atoms is as large as 5.3 Å. An MR attempt using the open form as a template to solve the structure of the closed form failed with a free *R* factor after *CNS* refinement (simulated annealing and energy minimization) of 54.0%. Similarly, screening for an MR solution using NMA perturbation along only one low-frequency mode also failed. Eventually, a bimodal scan yielded a solution when combining perturbations in the direction of the two lowest normal modes, the best model having a final free *R* factor of 47.1% (Fig. 3 and Table 1). The r.m.s.d. between all C^{α} atoms of this highest-scoring perturbed template and the target is 2.1 Å and more than 90% (202) of the residues have an r.m.s.d. of 1.6 Å to the target. However, in this case a simulated-annealing step was necessary for convergence of the refinement, because no optimal model was found at the level of the MR step, which is at variance with what was observed with our two previous examples. This particular bimodal screen took about 13 h of CPU time to complete on an Athlon 2400+ processor (MR with *AMoRe* and *CNS* refinement of 231 models).

Finally, we should also briefly mention two cases where the use of NMA for MR is not that efficient. In order to obtain better results with a perturbed model than with the original template to solve the structure of the closed form of citrate synthase (437 residues; PDB code 6csc) with its open form (PDB code 5csc) as a template, we had to screen three low-frequency normal modes; this yielded only a slightly improved free *R* factor of 44.5% compared with 45.2% obtained with the original template. However, as the original template already yields a 'good' solution, this may not be such a surprising result. Attempts to apply the same approach to the quite large lactoferrin protein (691 residues; PDB code 1cb6, open form, used as target; 1lcf, closed form, used as template) failed. In fact, successive

application of the optimal perturbation (computed by projecting the difference vector between the open and closed forms onto the normal modes) along the first ten normal modes steadily decreases the r.m.s.d. between all C^{α} atoms, but in the process the final free R factor goes through a maximum of R = 55.3% when the fourth mode is added, reaching a value of R = 45.4% only when the eleventh mode is also taken into account. If we compare this to the free R factor of R = 47.0% obtained with the unperturbed template, we conclude that lactoferrin falls into the class of cases where more than two low-frequency modes are needed in order to anticipate its conformational change.

4. Discussion

Molecular replacement is the most cost-effective method for solving the three-dimensional structure of a new protein by X-ray crystallography and it is thus the method of choice for structure determination. However, MR may fail even in cases of high sequence homology when conformational changes, *e.g.* arising from ligand binding or different crystallogenetic conditions, come into play (by failure of the MR approach we mean that no refinable model can be found and that additional experimental techniques are required to achieve phasing of the diffraction data). Here, we demonstrate the potential of normal-mode analysis as an extension to MR that allows recovery from such drawbacks. We have provided three examples where application of standard MR protocol did not allow solution of the structure of a protein in one conformation (open or closed), whereas using a template perturbed following one or two low-frequency modes allowed lowering of the final free R factor below the 'noise' limit of about 50% (also confirmed by visual inspection of the resulting electron-density maps and further refinement to higher resolution). Although we limited our analysis to cases where template and target have 100% sequence identity, this approach should also be applicable to templates with much lower sequence similarity. Ideally, one would then start by building a homology model, *e.g.* using programs such as *MODELLER* (Sali & Blundell, 1993). Such a protocol is already implemented in *CCP*4 (Collaborative Computational Project, Number 4, 1994), so that an extension of *CCP*4 including normal-mode perturbation between the homology-modelling step and MR can be envisaged. Note also that NMA has already been used in order to refine temperature factors (Diamond, 1990; Kidera *et al.*, 1992) based on the fact that atomic fluctuations computed by NMA are often found to be well correlated with crystallographic B factors (Bahar *et al.*, 1997). Therefore, adding B-factor values predicted by NMA to the perturbed templates could also prove useful.

Overall, our approach can be viewed as perturbing the original template structure in different (but still physically meaningful) directions until one of the new models comes close enough to the searched structure to identify an MR solution. Even in cases where an MR can be found, this method can be of interest to further improve the starting model for refinement, eventually reducing the time spent on manual construction. This should be particularly true when working with low-resolution data sets. Jones (2001) has discussed the potential of such fold-recognition models for MR, but without referring explicitly to normal-mode perturbations, as presented here. As previously mentioned, an increase in MR success rates would be especially valuable in the context of the ongoing high-throughput structural genomics projects (Rupp *et al.*, 2002).

It has been shown recently (Krebs *et al.*, 2002) that half of the known protein movements can be modelled by displacing the studied structure using at most two low-frequency normal modes. A screening procedure following ideas presented in this paper would be highly efficient and parallelizable on a cluster of Linux PCs. Thus, NMA analysis may prove able to break tough MR problems in up to 50% of cases.

Protein models and structure factors were obtained from the Protein Data Bank (PDB). The authors wish to thank J. Navaza for access to the latest version of his *AMoRe* program. We acknowledge free access to the *CNS* (Brünger *et al.*, 1998) and *DEJAVU* (Kleywegt, 1996) software packages. We are grateful to C. Abergel for her interest in this work and helpful discussions.

References

Backbro, K., Lowgren, S., Osterlund, K., Atepo, J., Unge, T., Hulten, J., Bonham, N. M., Schaal, W., Karlen, A. & Hallberg, A. (1997). *J. Med. Chem.* **40**, 898–902.
Bahar, I., Atilgan, A. R. & Erman, B. (1997). *Fold Des.* **2**, 173–181.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brooks, B. & Karplus, M. (1985). *Proc. Natl Acad. Sci. USA*, **82**, 4995–4999.
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., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Delarue, M. & Sanejouand, Y. H. (2002). *J. Mol. Biol.* **320**, 1011–1024.
Diamond, R. (1990). *Acta Cryst.* A**46**, 425–435.
Durand, P., Trinquier, G. & Sanejouand, Y. H. (1994). *Biopolymers*, **34**, 759–771.
Echols, N., Milburn, D. & Gerstein, M. (2003). *Nucleic Acids Res.* **31**, 478–482.
Harrison, R. W. (1984). *Biopolymers*, **23**, 2943–2949.
Hinsen, K. (1998). *Proteins*, **33**, 417–429.
Hsiao, C. D., Sun, Y. J., Rose, J. & Wang, B.-C. (1996). *J. Mol. Biol.* **262**, 225–242.
Jones, D. T. (2001). *Acta Cryst.* D**57**, 1428–1434.
Kidera, A., Inaka, K., Matsushima, M. & Go, N. (1992). *J. Mol. Biol.* **225**, 477–486.
Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.
Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H. & Gerstein, M. (2002). *Proteins*, **48**, 682–695.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Rupp, B., Segelke, B. W., Krupka, H. I., Lekin, T., Schafer, J., Zemla, A., Toppani, D., Snell, G. & Earnest, T. (2002). *Acta Cryst.* D**58**, 1514–1518.
Sali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.
Suhre, K. (2004). *ElNémo Examples.* http://igs-server.cns-mrs.fr/elnemo/examples.html.
Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. (2000). *Proteins*, **41**, 1–7.
Tama, F. & Sanejouand, Y. H. (2001). *Protein Eng.* **14**, 1–6.
Tirion, M. (1996). *Phys. Rev. Lett.* **77**, 1905–1908.
Tirion, M., ben-Avraham, D., Lorenz, M. & Holmes, K. C. (1995). *Biophys. J.* **68**, 5–12.
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.