Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Evaluating the potential of using fold-recognition models for molecular replacement

David T. Jones

Bioinformatics Unit, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, England

Correspondence e-mail: d.jones@cs.ucl.ac.uk

Received 3 May 2001
Accepted 8 August 2001

Here, the proposal is investigated that protein tertiary structure prediction methods and threading methods in particular might be applied to the problem of solving a protein structure by X-ray crystallography, thus reducing the need for the more traditional experimental intensity methods of data phasing, such as heavy-metal isomorphous replacement and anomalous scattering methods, and without reference to a very closely related protein of known structure. If this kind of approach were to become successful and reliable, this would represent a significant advance in protein structure determination, offering an easy and accessible method for the initial data phasing for proteins' crystal structures, utilizing the vast amount of structural data, deposited in the Brookhaven PDB, that has been accumulated over the past 30 years of crystallographic structural studies. In the light of the ongoing structural genomics initiatives, the successful development of this kind of approach would be of enormous benefit.

1. Introduction

Given that protein structure is much more highly conserved than protein sequence and that it is the tertiary structure of a protein which creates the means by which it functions, it is not surprising that many people believe that determining the three-dimensional structure of a protein can provide valuable information as to its function and mechanism. One result of this belief is the impetus to solve experimentally the structures of every protein encoded by a bacterial genome. Several such structural genomics initiatives are already under way, but as yet none of these projects have produced large numbers of new structures, as they remain in the pilot stage of development. Despite great improvements to the basic techniques of X-ray crystallography, particularly the use of synchrotron-radiation sources, the rate-limiting step in structure determination remains the expression, purification and crystallization of the target proteins. However, even here great strides are being taken towards automation.

Once improvements in purification and crystallization become available, attention will then focus on the problems of data phasing. In this stage of the crystallographic 'pipeline', a possible avenue for increasing the throughput of structure genomics initiatives would be to reduce the need for time-consuming traditional experimental intensity methods of data phasing, such as heavy-metal isomorphous replacement and anomalous scattering methods. Molecular replacement (MR) is a widely used method for bypassing these experimental methods, but here it is necessary for there to be a closely

related protein of known three-dimensional structure to act as a phasing model.

Although MR is widely used in cases where the target protein is closely related to the protein used as a phasing model, it is now well established that many proteins sharing no obvious sequence similarity can show remarkable similarities in their native folds (*e.g.* Orengo *et al.*, 1994). Examples of such proteins include the various TIM-barrel enzymes, interleukin 1b/soybean trypsin inhibitor and the globins/colicin A. It is currently estimated (C. Orengo, personal communication) that there is a 70% probability that a newly determined protein domain will have a native fold similar to one already solved. It is therefore possible that threading methods (Jones *et al.*, 1992) could provide suitable MR phasing models in cases where no closely related protein of known three-dimensional structure is available. This would of course greatly increase the scope of MR in protein structure determination.

In principle, the application of threading methods to molecular replacement (MR) is obvious. A threading method can produce a number of models for a protein being studied and each of these models can be used as a source of initial phasing data set (α_{cal}) for the experimentally collected diffraction data (F_{obs}). Once a computed optimal position and orientation has been determined between the experimentally collected diffraction data (F_{obs}) and those of the model (F_{cal}), an initial calculated phase data set (α_{cal}) can be associated with each of the observed reflections (with adequate weighting) and hence an initial electron-density map can be calculated (Fig. 1).

Despite the apparent simplicity of this approach, the use of prediction techniques in molecular replacement is not common. Of course, the usual assumption here is that prediction methods are not capable of generating sufficiently accurate models to allow a molecular-replacement solution to be found, but how true is this blind assumption?

2. Replacement models in the PDB

The first task in this case study is to look at 'prior art' in crystallographic molecular replacement in order to try to work out the minimum required levels of similarity between the phasing model and the target protein. As a first step, the question of sequence similarity was addressed by looking at structures deposited in the RCSB Protein Data Bank (Berman *et al.*, 2000). The PDB as of December 2000 was searched to find deposited structures which had been solved by molecular replacement and which consequently included the REPMOD record to denote the PDB entry which had been used as the phasing model. A total of 1349 structures were found in the data bank as of December 2000 which included a valid REPMOD record. Of course, this is an underestimate of the number of deposited structures solved by molecular replacement, as for earlier structures there was no standard way of specifying which phasing model had been used. It is worth noting that even where a phasing model has been specified, it is still rare to find any information in the PDB file which describes in detail which part of the phasing model was used (*e.g.* subdomains).

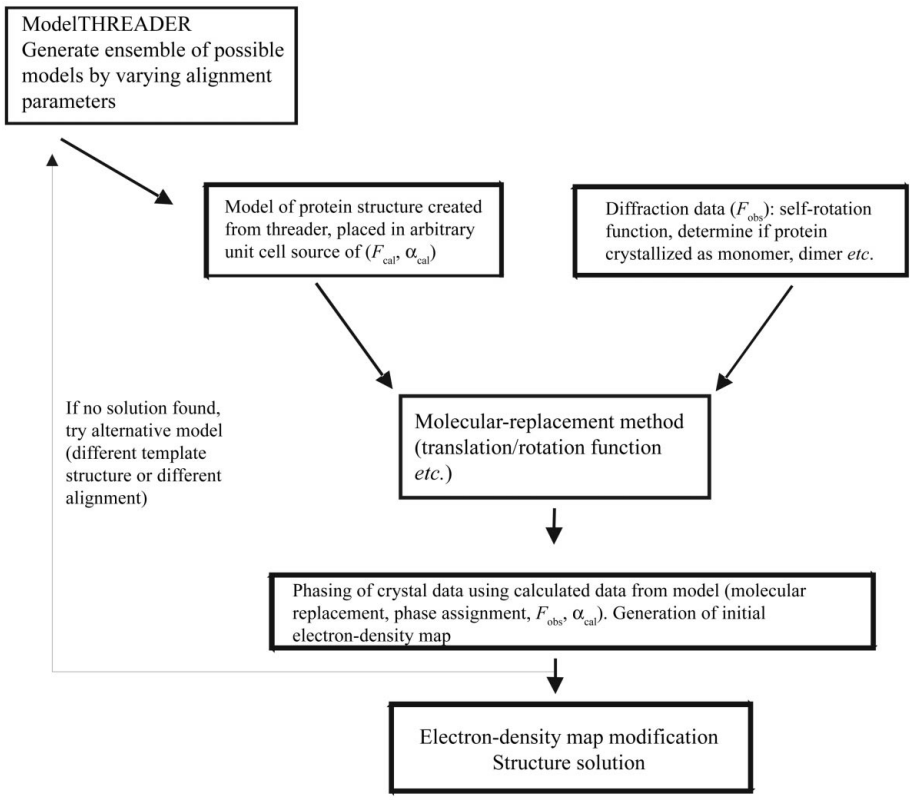

Figure 1
A flow chart illustrating the basic idea of using a threading method as a source of phasing models for molecular replacement.

Fig. 2 shows the distribution of sequence similarities observed for the 1349 MR structures extracted from PDB. As no information is usually given to specify which chain or domain was used as the phasing model, a local alignment was calculated between each chain of the target protein and each chain of the template structure. The sequence similarity reported for a particular target–template pair is therefore the highest percentage identity found across all target–template chain pairs. As expected, the vast majority of cases involve phasing models which are 100% identical to the target protein. These cases will include structures solved with different bound ligands, different crystal forms and so on. The next most highly populated group are those structures solved with phasing models in the range of 90–99%

identity to the target, which correspond to structures solved of close family members or natural or artificial mutants of the original protein. In terms of this analysis, both of these categories need to be considered special in that the problem at hand is probably not to solve a new protein structure *per se*, but to observe variations of an existing structure.

Below 90% identity lie cases where we might assume that the primary goal is to determine the structure of a relatively distinct protein. The distribution in this region shows a slight bias to target–template pairs with 50% identity, but with a fairly even spread of examples down to 20% identity. It is interesting to speculate why we observe this slight bias towards structures in the range 30–60%. One possible explanation for this bias is that this region shows a compromise between the interest in the structure and the practicality of finding a suitable MR solution. It might well be felt that for structures showing sequence similarity >>60% that the structure is unlikely to provide enough novel insights to make the structure determination worthwhile. Certainly, in the case of structural genomics projects it would be assumed that structures in this region of the sequence similarity distribution could easily be modelled by homology. However, as the degree of sequence similarity falls below 60%, the likelihood of finding novel structural features increases and so presumably the impetus to solve these structures is that much higher. Another possibility is that this region of sequence similarity space relates to cross-species levels of sequence similarity. For example, in pharmacological crystallographic studies it is frequently of interest to have crystal structures of both the human protein and the rat homologue, for example.

Although it is interesting to note that it is fairly common to find MR solutions for template–target pairs which have <30% sequence identity, this still does not provide much information as to the degree of structural similarity required. To look more closely at the requirements for structural similarity, the 1349 MR structures were cross-referenced with the current release of the FSSP data bank (Holm & Sander, 1998). FSSP is a collection of structural alignments and optimal rigid-body superpositions for a non-redundant subset of the PDB, which is updated on a weekly basis. As entries are only included for structures which are non-redundant at the level of sequence similarity, only 329 of the 1349 MR structures could be found in FSSP. However, these 329 structures do provide a guide as to the degree of structural similarity required for a successful MR solution, as shown in Fig. 3. It is clear from this scatter plot that the vast majority of deposited MR structures involve template–target pairs which have very high degrees of structural similarity. Only 3% of the pairs have C^{α} RMSDs of more than 2.0 Å and two of these cases appear to be clear outliers which are apparently a consequence of domain shifts between the target and template proteins. Presumably, in these cases the domains in the phasing models were treated separately.

Another interesting aspect of Fig. 3 is the degree of structural overlap between the target and template protein structures. The overlap is defined as the fraction of the target protein which can be structurally aligned with the phasing model structure. Interestingly, in a few cases the phasing model only covers around half of the target protein; presumably, these cases correspond to domain-level similarities between the proteins. However, in the majority of cases the degree of structural overlap is very much higher, with more than 90% of the target protein chain equivalenced to the associated chain in the phasing model.

Given this very rudimentary analysis of MR models found in the PDB, it is possible to put some very broad boundaries on the acceptable levels of structural similarity that must be observed between the target protein and any potential phasing model. Looking at the data pessimistically, we can see that a typical successful phasing model covers more than 90% of the target protein and has a C^{α} RMSD of <2.0 Å. This is clearly a tough requirement for any viable structure-prediction method. For fold recognition, even ignoring the problem of

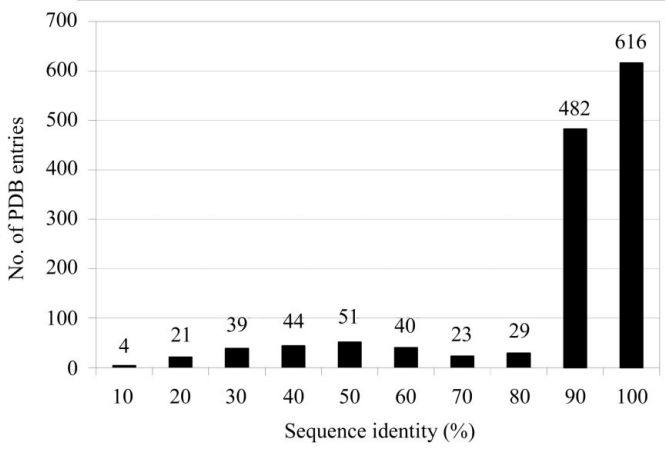

Figure 2
Bar graph showing the distribution of sequence similarities between phasing model and target for molecular-replacement structures deposited in the PDB.

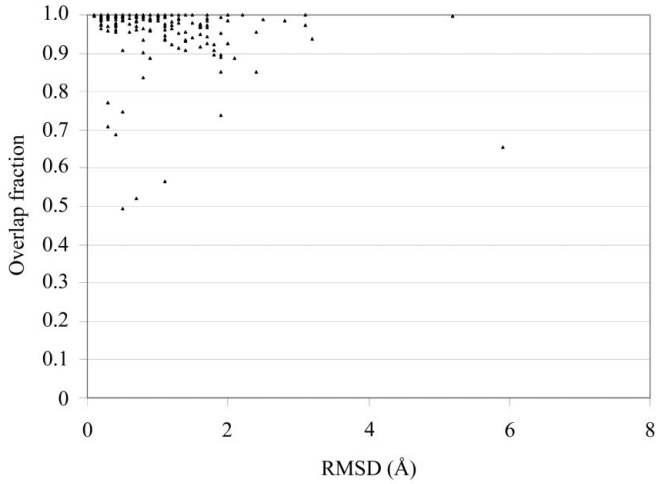

Figure 3
Scatter plot showing the structural similarities between phasing model and target for molecular-replacement structures deposited in the PDB. The *x* axis gives the C^{α} root-mean-square deviation and the *y* axis the fraction of the target protein chain which can be superposed onto the phasing model.

actually finding the alignment and building the model, it is unlikely that any suitable templates will even exist in the current PDB. However, it is possible to look more optimistically at the data. Ignoring the two obvious outliers with RMSDs > 5.0 Å, we can see that several structures have been solved where the C^{α} RMSD is 2.5–3.0 Å and the degree of overlap is as little as 50%. Of course, there does appear to be a correlation between these variables in that the structures solved with RMSDs >2.5 Å had overlaps of >90% and structures solved with structural overlaps of <60% had RMSDs of around 1.0 A, but nonetheless a highly optimistic view suggests that as an absolute minimum, an RMSD of 3.0 Å and an overlap of >50% to allow a successful MR experiment. In defence of this optimistic view is the fact that the MR structures in PDB do not represent a systematic study of the limits of MR techniques. It is reasonable to suppose that many more structures could be solved with highly divergent template–target pairs, but that crystallographers will have typically assumed that MR is impossible in the absence of very high degrees of structural or sequence similarity. These much more difficult MR experiments will therefore either not be carried out at all, or not carried out with much effort owing to the low expectations of success.

3. The state of the art in fold recognition

Given that we have some rough guidelines as to how accurate a phasing model needs to be to give any chance of success in MR, the next part of the question is to ask how often models of this accuracy can be generated with the best available prediction methods. In this brief survey, I will focus on fold-recognition methods, as although comparative modelling is still the most reliable available prediction method, it is already apparent that closely related protein structures can be used to provide phasing models (see, for example, Fig. 2). I will also not look at methods for *ab initio* prediction, as although these methods have developed significantly in recent years it is still clear that they are severely limited in terms of the size of proteins which can be modelled and the accuracy of the resulting models.

Rather than focus on a single method or methods from a particular group, here I will make use of the publicly available data from the recent 4th CASP (Critical Assessment of Methods for Structure Prediction of Proteins) experiment. Although the CASP experiments are limited in terms of the numbers of target proteins involved, the fact that almost all groups are able to apply their methods to a single set of test cases means that we can infer more about the general state of the whole protein structure-prediction field rather than the abilities of one or two groups.

This review will cover some general observations from the most recent CASP meeting which was again held at Asilomar in California in December 2000. Full details of the experiment will again be published in a forthcoming special issue of the journal *Proteins*, along the same lines as the previous special issues covering the first three CASP experiments (Moult *et al.*, 1999). The raw data which is used in this evaluation is available from the URL http://predictioncenter.llnl.gov/casp4.

4. Comparative modelling

Although it is not within the scope of the current study, it is worth taking a few notes from the comparative-modelling section of the CASP4 experiment. The comparative-modelling process can be divided into five basic steps: alignment of the target sequence with the sequence of a protein of known three-dimensional structure, building of a framework structure based on the alignment, loop building, addition and optimization of side chains and finally model refinement. In the first three CASP experiments, there was a general feeling of disappointment in the limited degree of technical development that has been apparent in comparative modelling, but nonetheless it remains the best means of obtaining an accurate protein structural model by theoretical means. In recent years, there has been a definite advance in the accuracy of sequence alignments for target–template pairs which are only distantly related. This has come from the common usage of sensitive sequence-profile alignment methods such as PSI-BLAST (Altschul *et al.*, 1997) or one of the several methods based on Hidden Markov Models (Eddy, 1996). Despite the evident improvements in automatic alignment accuracy in CASP4, there is still a lot to criticise in the comparative-modelling field, at least as viewed in the CASP experiment. Although alignment accuracy has certainly improved since the first CASP experiment, it is still fair to say that apart from cases where the target has a very close homologue of known structure, the vast majority of comparative models entered into CASP4 still display quite serious errors in alignments. As a result of these alignment errors, it becomes very difficult to make any reasonable attempt at loop fitting or side-chain building because the basic backbone structures are too inaccurate. As a result of this difficulty, it has been proposed that for CASP5, a second deadline in the comparative-modelling section will be set. After the first deadline has passed, comparative-modelling groups will be provided with a reference alignment and asked to fit loops and build side chains based on this alignment. This simple step should allow more meaningful comparisons to be carried out between different methods.

Probably the most disappointing aspect in comparative modelling continues to be the fact that final models are still no closer to the experimental structures than the original template protein. This clearly indicates that none of the molecular-mechanics refinement procedures are actually managing to move the unrefined structures towards the correct structures. The failure of molecular-mechanics methods to refine structures remains a fundamental barrier in comparative modelling and ultimately places a limit on the accuracy that one can expect even from the best models. Note that this is quite contrary to the situation in X-ray crystallography or NMR studies, where molecular-mechanics-based refinement methods are of great benefit in producing improved structures. Of course, in these cases large amounts

of experimental data are incorporated into the refinement objective function and so the refinement process does tend to converge on a structure closer to the 'correct' native conformation of the protein.

5. Fold recognition

Classically, fold-recognition or threading methods have been applied in cases where no suitable homologous template structure can be found to permit the building of a model by comparative modelling. The earliest fold-recognition approaches (*i.e.* threading methods) were designed specifically to recognize folds in the absence of sequence similarity and, indeed, the sequence of the template protein was usually not taken into account at all. However, the boundaries between comparative modelling and fold recognition are now becoming increasingly indistinct, as it can be increasingly expected, with the growth of both sequence and structure data banks, that for targets which belong to a currently known superfamily of known structure, sensitive sequence-comparison methods will equal or even surpass the abilities of true fold-recognition methods. This should not be surprising, as in cases of even very distant homology sequence conservation provides a great deal of information with which to produce an accurate sequence-structure alignment. Ignoring this information, as with many fold-recognition methods, will almost certainly therefore produce less accurate alignments. In view of this, of course, many developers of fold-recognition methods have been attempting to combine sequence-profile alignment methods with fold recognition. This should in principle produce an alignment method which can produce accurate alignments both where the target and template proteins are in the same superfamily and when they are not.

A total of 34 domains were considered during the main fold-recognition assessment at CASP4. Of these 34 domains, 11 belonged to a superfamily of known three-dimensional structure (homologous structures; FR/H category), 11 had a known fold but were most probably analogues rather than homologues (FR/A category) and 12 were arguably new folds though with some weak similarities with known folds (FR/NF category). Considering all three categories together, of the 34 target domains the fold for 26 was recognized by at least one group. Of the remaining eight target domains, at least some structural similarity was identified by one group for at least four of them. The best groups managed to predict ten or 11 folds correctly out of the 34 (22 if the almost new folds are discounted). This success rate of around 33% is somewhat lower than for previous CASP experiments, where the better groups typically achieved success rates of 50–60%. This is of course mainly a consequence of the inclusion of the 'almost new fold' targets in the assessment, which should probably be best left out of consideration for the purposes of this evaluation.

Rather than look at the results of a single group, for this study it is informative to look at the whole set of results for the best 20 groups and to ask what is the best result for each target domain from any of the 20 groups. For molecular-replacement studies, it is common to use not just one model but a set of different models in the search for a phase solution. In the case of fold recognition it is reasonable to consider a scheme where a set of different models could be generated by a number of different approaches (perhaps by means of automated submission to various fold-recognition web servers). It would then be reasonable to ask whether or not a useful model exists at all within this set of models. Clearly, if an accurate model does exist within this ensemble of alternative models it is possible (and hopefully probable) that it will be found in the course of the molecular-replacement study.

Table 1
Summary of fold-recognition results at CASP4, calculated as a consensus of the top 20 groups' submissions.

Results are presented for cases where 50% of the model is correct and where 60% of the model is correct (see text).

	CM	CM/FR	FR/H	FR/A
Total targets	8	6	10	11
Correct fold	8	6	10	9
50% model	8	6	8	3
60% model	8	5	7	0
	100%	83%	70%	0%

Table 1 shows the result of using this ensemble approach to evaluating the CASP4 predictions based on the results from the 20 best groups (ranked in terms of how many folds were correctly identified). The results are broken down into four categories: CM (easy comparative-modelling targets), CM/FR (difficult comparative-modelling targets), FR/H (fold-recognition targets belonging to an existing superfamily of known three-dimensional structure) and FR/A (fold-recognition targets which are analogues of known structures rather than homologues). The raw data used to compile Table 1 are the GDT-4 values available from the Prediction Center website. These values denote the number of residues in the model which after superposition with the experimental structures are found equivalenced to within 4 Å. These numbers are then simply calculated as a percentage of the experimental structure domain length. For example, a model of length 100 for which the C^{α} atoms of 60 residues could be superposed to within 4 Å of equivalent C^{α} atoms in the experimental structures would be considered 60% 'correct'. Using these values, we can ask whether any of the models submitted by the top 20 groups fall within particular thresholds of accuracy. Based on Fig. 2, two thresholds are defined: a very optimistic threshold of 50% and a somewhat less optimistic threshold of 60%. The basic assumption here is that if a model can be found with at least 50% (60%) of the C^{α} positions correctly modelled then there is at least a chance of success in a molecular-replacement experiment.

From Table 1 it is clear that finding models for which the majority of residues are correctly modelled is relatively trivial for easy comparative-modelling targets and becomes progressively more difficult as the cases move into the hardest fold-recognition category. Indeed, within the FR/A category none of the 20 top groups submitted predictions where 60% or

more of residues were correctly modelled, although for three of the FR/A targets borderline predictions (50% correct) were submitted. Of course, these results are based on a very small sample of just 34 domains and so it is not reasonable to try to extrapolate too much from this analysis. Nonetheless, this does give at least a sketch of the current abilities in the fold-recognition field in general.

6. Applications to structural genomics

Despite the limitations already noted, Table 1 does give a rough idea as to the likelihood of finding a useful model from a set of good fold-recognition methods for targets of varying levels of difficulty. Given these observations, what can be inferred about structural genomics? To map the results from Table 1 to structural genomics, we have to investigate the distribution of target difficulties within a single genome. Fig. 4 shows the estimated distribution of targets within the 470 ORFs from the *Mycoplasma genitalium* genome. The CM category is defined as those ORFs for which a homologue of known three-dimensional structure can easily be found using a standard BLAST search. The CM/FR category is defined as those ORFs which match known three-dimensional structures using an iterative PSI-BLAST search. The FR/H category is defined using the program *GenTHREADER* (Jones, 1999) which is designed to recognize superfamily matches to known three-dimensional structures using a combination of sequence profiles and threading potentials. The FR/A category is estimated based on the observation that currently 70% of newly solved structures have significant similarity to an existing protein of known three-dimensional structure. This leaves a category labelled 'unknown' which will be a mixture of proteins with as yet novel folds and proteins which are non-globular (*e.g.* transmembrane proteins).

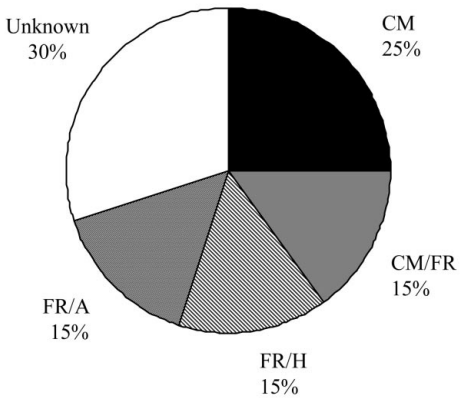

Figure 4
Pie chart showing the distribution of different prediction-target difficulties in the *M. genitalium* genome (470 open reading frames). Key: CM (easy comparative-modelling targets), FR/CM (difficult comparative-modelling targets), FR/H (fold-recognition targets in the same superfamily), FR/A (fold-recognition targets not in the same superfamily). The remaining 'unknown' category includes ORFs with novel folds, and non-globular proteins.

Based on the results in both Table 1 and Fig. 4, we can infer that, in theory, MR solutions should be possible for a total of 48% of the ORFs in the *M. genitalium* genome. This assumes than none of the FR/A cases would be predicted within the boundaries set by Fig. 3, but it is possible that with the slightly more lenient cutoff of 50% as many as 25% of the FR/A targets might be modelled within acceptable limits of accuracy. In this case, as many as 52% of the structures for these ORFs might be solved by a combination of structure prediction, modelling and molecular replacement.

7. Conclusions

The above feasibility study is admittedly somewhat biased towards an optimistic view of the possibilities of combining fold-recognition techniques with MR techniques. Certainly, many assumptions have been made in all the calculations. For one thing, I have not considered the well known adage that every error in the phasing model contributes to the noise. This would suggest that even where the majority of the phasing model is correctly built, the regions which are even slightly mis-modelled would contribute to the background noise and therefore make it hard if not impossible to find the correct MR solution. Investigations are currently under way in my own laboratory to investigate this and related issues. In particular, we feel the following questions are critical to the successful combination of FR (fold-recognition) and MR (molecular-replacement) techniques.

(i) How accurate does the initial model have to be to provide a useful source of initial phase information?

(ii) What effect does the resolution, completeness and quality of the collected diffraction data have on the success of MR?

(iii) How much detail is required in the model to phase the collected diffraction data?

(iv) How should loops be treated – should they be modelled or deleted altogether?

(v) Can a small substructure fragment (domain) library be used to phase data?

(vi) Can we automatically extract the correctly modelled regions of a predicted structure?

Given the difficulties highlighted within this paper, why should we continue to be optimistic about the development of hybrid FR/MR techniques? Perhaps the simplest answer is that protein structure-prediction results can be obtained for almost no cost with respect to both money and manpower. Computer time is now virtually free in most cases, with cheap desktop PCs offering all of the required computational power required to run MR experiments. Given that there are now many automated protein structure-prediction servers available, and even now 'meta servers' (*e.g.* http://bioinfo.pl) which can obtain results from many other servers, it is now relatively easy to obtain a reasonable ensemble of possible models for any given target protein. Given these two facts, it is reasonable to envisage a scenario where systematic MR experiments can be carried out almost entirely automatically according to the

system outlined in Fig. 1. This systematic search would require only a dedicated computer on which it could run; almost no human intervention would be required. Even if the overall success rate turned out to be as low as say 5%, this would still mean that four additional X-ray structures could be solved for the FR/H category ORFs in *M. genitalium* with no real human effort beyond the initial crystallization and native data collection. For larger organisms with many more ORFs, this number would be proportionately larger. With suitable answers to the open questions posed above, it is likely that the overall success rate of these methods would in fact be much greater than 5%. In this case, many hundreds or even thousands of new X-ray structures could potentially be solved relatively quickly from native diffraction data alone.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

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.

Eddy, S. R. (1996). *Curr. Opin. Struct. Biol.* **6**, 361–365.

Holm, L. & Sander, C. (1998). *Nucleic Acids Res.* **26**, 316–319.

Jones, D. T. (1999). *J. Mol. Biol.* **287**, 797–815.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). *Nature (London)*, **358**, 86–89.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). *Proteins, Suppl.* **3**, 2–6.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). *Nature (London)*, **372**, 631–634.