Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Creating structure features by data mining the PDB to use as molecular-replacement models

T. J. Oldfield

Accelrys Inc., Department of Chemistry, University of York, Heslington, York YO10 5DD, England

Correspondence e-mail: tom@ysbl.york.ac.uk

Received 9 May 2001
Accepted 18 July 2001

Mathematical data-mining techniques to generate a representative set of protein fragments are described. Protein fragments are used as search models within the macromolecular phasing method of molecular replacement to attempt to phase protein data without a homologous model correctly. Preliminary investigations using these fragments indicate that molecular replacement with *AMoRe* is not sensitive enough to phase myoglobin or insulin data sufficiently for successful refinement. The results suggest that more advanced molecular replacement techniques may be successful, though at present these are not computationally practical.

1. Introduction

As more model structures become available though protein crystallography and nuclear magnetic resonance (NMR), there have been a number of projects to generate a classification of protein structure [CATH (Orengo *et al.*, 1997); DALI (Holm & Sander, 1992, 1993, 1995); SCOP (Murzin *et al.*, 1995)]. This is important so as to move forward from the presentation of raw data to a greater understanding of protein structure and function. Classification has so far concentrated on analysing the domain structure of proteins to produce a hierarchy of structure definition (Orengo *et al.*, 1997; Sowdhamini *et al.*, 1996). However, it can be difficult to define what fragment of structure makes up a domain (Holm & Sander, 1996; Sowdhamini *et al.*, 1996; Sali & Blundell, 1990; Alexandrov & Go, 1992; Vriend & Sander, 1991; Fischer *et al.*, 1992; Grindley *et al.*, 1993; Sowdhamini *et al.*, 1996; Mizuguchi & Go, 1996; Rufino & Blundell, 1994; Boutonnet *et al.*, 1995) as it is based on the human interpretation of protein structure. From these domain definitions of protein structure, it has been proposed that the fold space is finite and therefore it is reasonable to assume that a representative set of protein fragments can be used as molecular-replacement (MR) models. A number of studies have approached this problem, but unfortunately the MR is very sensitive to dissimilarity between the model and target data.

The aim of the approach presented here is to generate a representative set of protein folds using mathematical targets to define these. This allows any definition to be predictable and reproducible even if it does not conform to current expectations. Since all parts of the calculation are deterministic, the representative set of protein folds can be reproduced in a trivial way. Finally, these recurring folds are used as molecular-replacement search models to determine whether such generic information can be used to provide any phase information.

1.1. Data mining

It is difficult to find a definition of data mining, as it is defined differently in different contexts. In this paper, the term is used to describe the extraction of information from data using targets defined only by mathematical descriptions of correlation, occurrence and deviation. Thus, any result of the data mining is only a mathematical construct of the data. This means that the use of templates to search a database is not data mining, as a template is a pre-defined target which must bias any subsequent conclusions. There is also a problem that the original data may be skewed; for example, fold analysis of the whole Protein Data Bank would indicate that the most common folding pattern in proteins is known as lysozyme, owing to the many examples of this structure.

There are problems associated with the use of data mining. It is necessary to re-cast the analysis in a way that does not require knowledge of the data; also, noise can swamp any useful information. The results also do not provide any explanation of the presence of a feature, though this does not matter when using the results for MR.

1.2. Data degeneracy

The Protein Data Bank contains significant amounts of structural degeneracy. This occurs because of multiple deposition of proteins by different scientific groups, species variants and studies of structure and function by site-directed mutagenesis. There is also structural similarity within single proteins and occasionally between parts of different structures such as the lysozyme–immunoglobulin complex and depositions of the single molecules of lysozyme and immunoglobulin. The problem is to decide where to draw the line between information that is to be discarded (homologous domains) because it is known to exist and detail that is to be determined by this analysis. It is unfortunate that this process requires that we apply some knowledge of protein structure, but the use of sequence alignment and mathematical fragmentation removes the human interpretation from this.

2. Methods

The process to generate the fragments of protein structure for MR involves a number of steps. Firstly, it is necessary to select a representative set of data which will not skew any attempt at data analysis. Secondly, this selection of proteins is cut into fragments using a mathematical description to avoid recurrent structure. Next, the selection of protein fragments are optimally aligned using alignment length as a target for optimization and finally this square symmetric matrix of alignment lengths is analysed numerically for recurrent fold features. Although this process requires no human intervention, only the last stage, which involves studying data inflections, can be considered data mining.

2.1. Data selection

The program *PDBSELECT* was used for the initial selection of proteins (Oldfield, unpublished program). The results presented here pertain to the PDB from January 1999 (Bernstein *et al.*, 1977), which contains 11 208 protein structures.

NMR structures were not used, as they are determined using a different experimental target from that used to solve protein structures by crystallography and thus errors have a different distribution and meaning. In the former case experimental errors are defined on the basis of inconsistencies between measured atomic interactions (NOEs) and equivalent interactions within a model and in the latter case the difference between model structure factors and measured structure factors as well as geometry deviations. Since more proteins within the Protein Data Bank have been solved by crystallography, these represent the largest consistent set. Proteins solved before 1983 were rejected, as geometry restraints were not generally used within refinement before this date. DNA and RNA structures are rejected if they contain no protein part. In fact, proteins with less than ten C^{α} atoms were rejected from the analysis, as these could never align with another protein with more than ten C^{α} atoms. Structures are also rejected if any of the residues have names UNK or just C^{α} atoms, as these may be incomplete structures.

The next group to be deselected were geometric outliers. These are proteins with systematic errors that do not conform to expectation with respect to some geometrical property calculated from the coordinates. It would be more sensible to define the property based on the original data, but this is generally not available within the PDB. Geometric outliers are determined using the a Ramachandran energy surface described by Ramachandran & Sasisekharan (1969), calculated with *CHARMm* 22 (Brooks *et al.*, 1983), as well as distributions and standard deviations of torsion angles, C^{α} geometries, packing density and solvent analysis. Data resolution is used as a measure of the statistical error within the coordinates, owing to experimental limitations such as crystal quality. A limit of 2.5 Å is used and PDB structures that have a resolution field greater than this or no resolution information are rejected.

Finally, all proteins that satisfy the previous criteria were aligned by sequence. Where proteins were more than 80% identical by sequence (exact residue homology) only a single structure was retained. The structure with the higher resolution was retained unless ambiguous, in which case the structure with the better geometric criteria was selected. If still ambiguous then the protein with the latest submission date was selected.

The list of 2239 proteins generated using this rejection method contains protein structures solved by macromolecular crystallography to a resolution of 2.5 Å with good geometry and represents a non-homologous set.

2.2. Structure fragmentation

The proteins were fragmented into major structurally distinct units and generally, though not exclusively, this was based on the domain structure. This also removes a major part the non-transitive nature of the protein data found in structure

complexes. An example of this non-transitive problem occurs between the three structures lysozyme, immunoglobulin and the complex of immunglobulin and lysozyme. The complex aligns with both lysozyme and immunoglobulin, but lysozyme does not align with immunoglobulin. These three molecules therefore form an inconsistent triangle of structure alignment.

Fragmentation of the proteins was carried out by analysis of the two-dimensional finite difference matrix of the C^{α} distance matrix. The C^{α} distance matrix contains lines of discontinuity where the protein has a packing edge. To make this effect easier to analyse, the finite difference matrix was generated to create lines of peaks (Figs. 1 and 2). A packing edge in a protein can therefore be defined where

$$D_{ij} = |M_{ij} - M_{i(j+1)} + M_{(i+1)(j+1)} - M_{(i+1)j}| \text{ for } i, j = 1, (N-1)$$
$$\sum_{i=1,(N-1)} D_{ij} \geq a \times 3.8 \times N^{b} \text{ for } j = 1, (N-1),$$

where $[M]$ is the C^{α} distance matrix, $[D]$ is the two-dimensional finite difference matrix and D_{ij} is an element of the matrix and N is the number of C^{α} atoms in the protein. 3.8 is the mean separation between two consecutive C^{α} atoms in a protein. a and b are coefficients that define the sensitivity of edge finding, where a defines the sensitivity and b the slope for protein size.

A number of parameters have been used, though the results presented here are for a = 2.2 and a b of unity, though better results have been obtained recently where $b \neq 1.0$.

Separate files were generated for each fragment part where a packing edge was found, as long as the different fragments had different sequence. This analysis was carried out with the program *SQUID* (Oldfield, 1992) using the list of proteins generated from the program *PDBSELECT* and resulted in a set of coordinates used for the alignment analysis. A copy of the water structure and any HETAM were also written with each of the structure fragments though not used in the analysis of protein features.

2.3. Structure alignment

The optimal alignment of proteins has been described by a number of authors (Taylor & Orengo, 1989; Holm & Sander, 1993; Mizuguchi & Go, 1995) and these involve using motif (helix/strand) alignment or C^{α} distance matrices. The use of alignment by vectors had been tried previously using an algorithm within the program *SQUID*, but was difficult to analyse owing to difficulty in using the number of equivalent structural elements in subsequent analysis. The optimal alignment of two proteins using atom positions used here is theoretically simple to describe; one only need to try aligning every part of one protein with every part of another protein. Unfortunately, this results in computational times that are large.

A program *CAMINE* (Oldfield, unpublished program) was written to carry out optimal structure alignment. This program carries out optimal alignment of the C^{α} atoms of a pair of protein structures and takes approximately 0.1 s to find all the possible ways of aligning a pair of proteins above a threshold length of alignment. The program is designed to use a user-defined RMSD and maximize the length of alignment to give the largest set of C^{α} atoms that align. The algorithm trims back any deviating sections of protein at the end of superposed structure, so the returned RMSD is not necessarily equal to

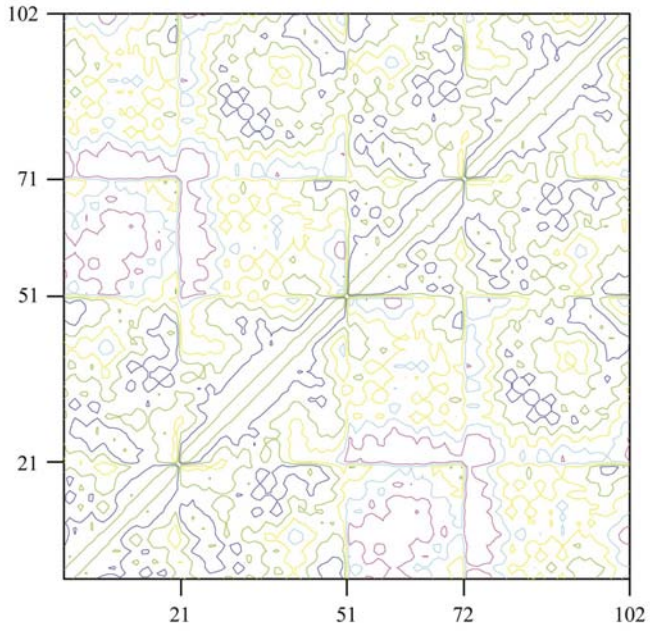

Figure 1
The C^{α} distance matrix for the protein insulin (3ins). The matrix is contoured at 10 (blue), 15 (green), 20 (yellow), 25 (cyan) and 30 Å (purple). The x and y ordinates are for the residue order in the protein and only the sequence position for the C-termini of each chain are labelled.

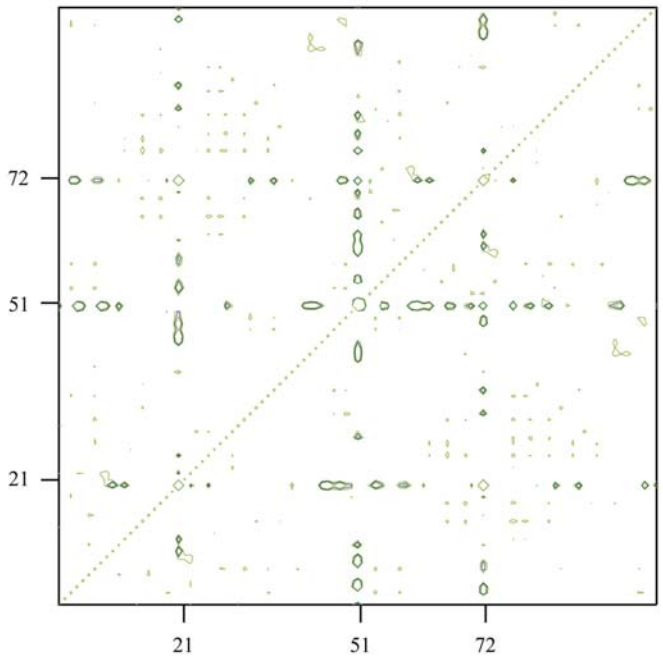

Figure 2
The two-dimensional finite difference matrix of Fig. 1 contoured at 10 (green) and 20 Å (blue).

the threshold value provided, but can be smaller. The program will also return multiple solutions to structure alignment; for example, myoglobin can be aligned with haemoglobin in four different ways and therefore the program actually produces four results. Since multiple alignment solutions could not be handled in a simple way in this analysis, the longest alignment of residues was taken as the only solution. The result of the alignment is a square symmetric matrix of alignments for all fragments with all other fragments.

A database was generated for each protein fragment that contains the coordinates of all fragments that aligns with this protein fragment. Some databases were small and contain two aligned fragments and some databases were large, containing more than 1000 members. Since the matrix of results defines only the length of alignment and not position, each database may contain information on multiple folds or no consistent fold information. It is therefore necessary to analyse each database to determine a consistent set of atom positions that occurs with a significant frequency.

For each of the databases, the set of aligned coordinates was analysed to find the longest common feature that occurs in the most fragments. A program *ANAFRAG* (Oldfield, unpublished program) reads in all the aligned coordinates and searches for the data inflections as a function of length of alignment and number of superposed structure motifs. An inflection (a rapid change in a target value) occurs as more C^{α} atoms are added to the common feature until the number of fragments within that feature suddenly reduces. The features of interest have a length just smaller than this inflection point. Each feature is written as a multiple superposed set of coordinates and these represent the information used for the molecular replacement.

2.4. Molecular replacement

The aim of the molecular-replacement work carried out to date is to determine what is possible with this type of analysis. Firstly, what size of fragment of protein structure can be used to obtain an MR signal that can be correctly identified? Since there are a large number of results here, it is possible to identify small signals from the MR results. Secondly, what is the best type of model coordinates to use within MR? This could be a single average structure or multiple overlaid coordinates, C^{α} atoms or polyalanine. Finally, what is the best metric over the large number of calculations that give the best signal-to-noise ratio? It is necessary to define the number of answers that are correctly identified, the number of correct answers not identified (false negatives), the number of wrong answers that are incorrectly signalled as correct answers (false positives) and the number of results that are correctly defined as wrong. A method of highlighting the correctness of solution needs to minimize the false negatives and false positives. To date, only the program *AMoRe* (Navaza, 1994) has been used, as this is the quickest though certainly not the most sensitive program.

The program *AMoRe* was used to carry out molecular replacement between the structural features found by data mining and the 2.0 Å data of porcine myoglobin (Smerdon *et al.*, 1990) and porcine insulin (Whittingham *et al.*, 1995). The resolution range used was 3–12 Å, though other ranges were tried with no obvious improvement in the results. C-shell scripts were written so that for each structural feature a cross-rotation function was performed followed by a translation function and finally a rigid-body refinement. The best 20 solutions were taken at each stage in the molecular replacement. The result of this analysis was 20 molecular-replacement solutions for each structural feature. That is, 20 solutions from the molecular replacement were generated for each of the structure features generated by data mining. Various metrics of the results were correlated with results from the least-squares structure alignment between each of the features and the myoglobin/insulin final coordinates.

3. Results

3.1. Data selection

The data selection produces 2239 structures from the original set of 11 208 proteins within the Protein Data Bank. The rejection statistics in Table 1 result from the rejection criteria described.

The splitting of the 2239 proteins into fragments of structure using the two-dimensional finite difference matrix of C^{α} distances results in a total of 3331 fragments of protein.

3.2. Alignment

The program *CAMINE* was provided with the list of protein fragments and it carried out the simultaneous pairwise alignment of all member of the list of proteins. The calculation using the list of 3331 fragments took 148 h on a Pentium II 450 PC running Red Hat Linux and required 5.5 million pairwise structural alignments. The program provides a sequence alignment based on the structure alignment, a transformation matrix that orientates one of a fragment pair to the other, an

Table 1
The rejection criteria defines a summary of each rejection method described in the text.

The limit column in the table defines the numerical limit of the rejection where this applies, or text keyword (as defined by IUPAC rules) recognized from the PDB file.

Rejection criterion	Limit	No. rejected
NMR structures	\$NMR/MOL†	31
Date limit	After 1983	152
DNA + RNA structures	\$DNA/\$RNA‡	96
Too few residues	<10 residues	352
Too many C^{α}	$>0.25 \times N_{\rm atom}$	31
UNK sequence entries		4
Bad Ramachandran	>10% OUB§	1
Resolution limit/No info	>2.5/No data¶	3792
Homology rejection	>80%	4510

† The keyword \$NMR should be declared within a PDB file that is solved by NMR, but in most cases this is not present. Therefore, a NMR structure is also detected by the presence of the MOL and ENDMOL cards used to delimit the multiple solutions within such files. ‡ The \$DNA and \$RNA should be used in a PDB file to indicate that the structure only contains nucleic acid residues. § OUB: out of bounds. ¶ A PDB file is rejected if it contains no resolution information.

RMSD, length of alignment and a set of translated atom coordinates.

The longest aligned fragment pair was 690 residues and the shortest aligned fragment pair was 20 residues, the minimum recorded alignment length. With the minimum recorded alignment length it is found that all except 142 proteins aligned with some other protein in this analysis. The number of unique fragments is zero for ongoing analysis. The distribution of alignment lengths on a $\log_e$ scale indicates that only alignments below about 70 residues are significant (Oldfield, in preparation). Thus, data mining will not find significant common features longer than 70 residues.

One problem to contend with is that the helix in a protein represents the most compact method of packing a sequential list of residues, while a strand represents the most inefficient method of packing a sequential set of residues. This means that β-sheets contain relatively few residues to form a significant fold pattern, while the packing of just two helices consists of 20–30 residues. Hence, the recognition of β-strand/sheet information within the protein alignment is difficult as significant β-strand information lies within the noise of the helical alignment.

The pairwise alignment of the 3331 protein fragments generates a matrix of aligned metrics, where each element is the number of residues that align between a pair of proteins. This matrix is mostly complete, with non-zero (large) values on the leading diagonal. Elements for pairs of proteins that have no alignment solution contain zero.

3.3. Common feature analysis

Each database of aligned fragments is read by the program *ANAFRAG* and fold analysis carried out by inflection detection. A total of 358 features were generated with a distribution of length as shown in Fig. 3. This graph shows that the most common feature size occurs with a length of 28 residues and subpeaks occur at 36 and 45 residues. The significance of these peaks in this graph results from two-, three- and four-helix packing, where the average length of a helix is about 12 residues. It should be noted that while helical structure is highly conserved and β-strand structure is very variable, the packing of helices is very variable and the packing of β-strands is highly conserved. Therefore, most of the features of structure result from various packing occurrences of helices. Some example folds are shown in Fig. 4(*a*)–4(*d*), although most of the smaller features consist of two/three helices with different packing angles.

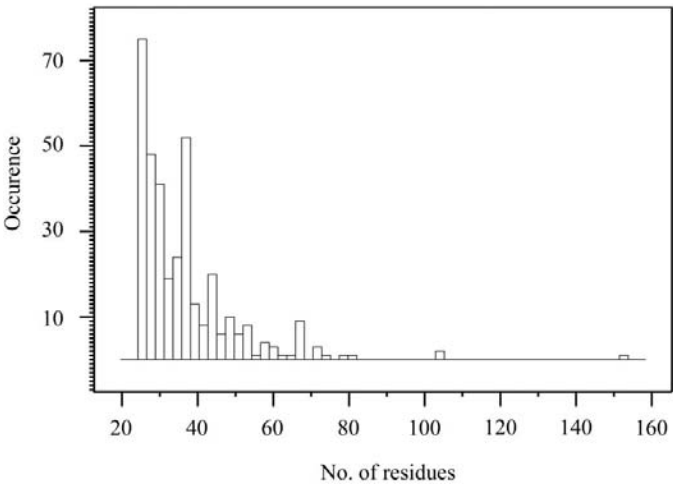

Figure 3
The distribution of feature sizes generated by mining the protein fragments.

3.4. Molecular replacement

To test whether a small feature of protein structure could be used as a molecular-replacement model, a single model feature was used. A 76-residue feature generated by this analysis was used to test the principle of using parts of a structure for MR. The feature is a four-helix packing of the helices *E*, *F*, *G* and *H* from myoglobin and is shown in Fig. 5 as a multiple overlaid C^{α} trace. The MR was carried out with just C^{α} atoms and a polyalanine structure using (i) an ensemble of overlaid structures, (ii) an average structure of the ensemble and (iii) a single feature element of the ensemble. A summary of the results of this analysis is shown in Table 2.

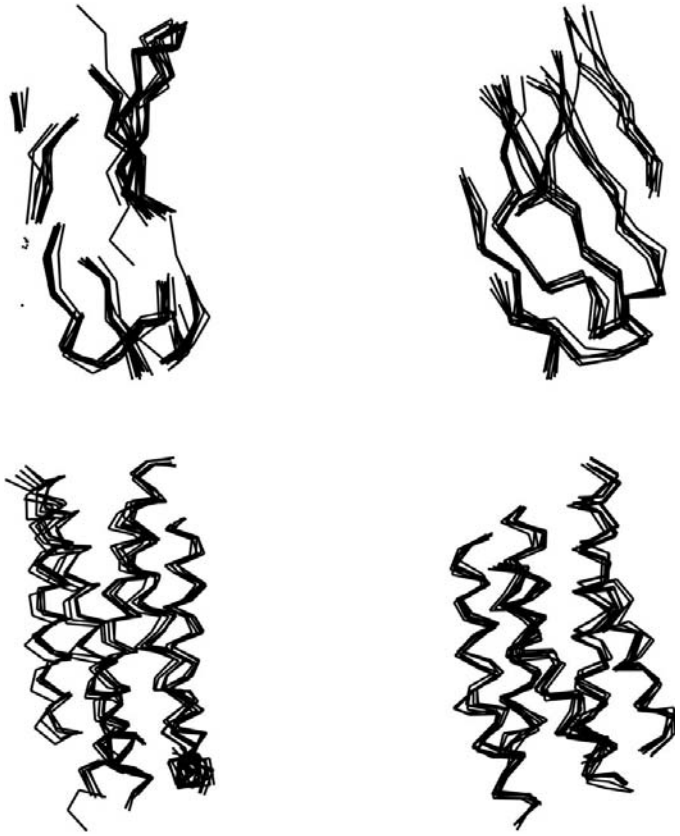

Figure 4
Four example features generated by mining the protein structure coordinates. Each figure consists of multiple superposed C^{α} traces from a number of proteins that contain the same feature.

Table 2
Molecular replacement.

Search model data	Noise (CC) (%)	Solution (CC) (%)
76 alanine residues of myoglobin	33.9	39.0
76 C^{α} atoms of myoglobin	21.0	No solution
C^{α} trace, ensemble	32.4	34.9
C^{α} trace, average structure	22.0	No solution
C^{α} trace, representative structure	22.7	No solution
Alanine, ensemble	45.3	48.1
Alanine, average structure	36.6	39.4
Alanine, representative structure	36.3	No solution

The myoglobin data has two molecules in the asymmetric unit and thus a total of 306 amino acids, as well as two porphyrin ligands. Two solutions were determined by molecular replacement and found to overlay myoglobin with a correlation coefficient of 48% for both molecules in the asymmetric unit when using a polyalanine ensemble. This peak was 3% larger than the mean of the remaining 18 solutions. It was interesting to note that the correct solution was found when just using the C^{α} ensemble atoms of the feature, though the solution was not so clear at 2.5% above noise. From this analysis, it was observed that the best signal could be obtained using the ensemble of alanine coordinates, though an ensemble of C^{α} traces was almost as useful.

A C-shell script was written to use each feature as a model to search myoglobin and insulin data with *AMoRe*. The results from the MR (correlation coefficient) were correlated with the structural alignment of each of the 358 feature fragments with myoglobin and insulin. Figs. 6(*a*) and 6(*b*) show these alignment results plotted against the product of maximum correlation coefficient and feature size.

The largest alignment between the fragments and myoglobin is that from the four-helix packing found about the haem group. This does not give the largest correlation coefficient, even when scaled by the size of this feature, and in fact there are a number of non-aligned fragments that result in

Figure 5
The myoglobin core feature (helices *E*, *F*, *G*, *H*) shown as multiple superposed C^{α} coordinates.

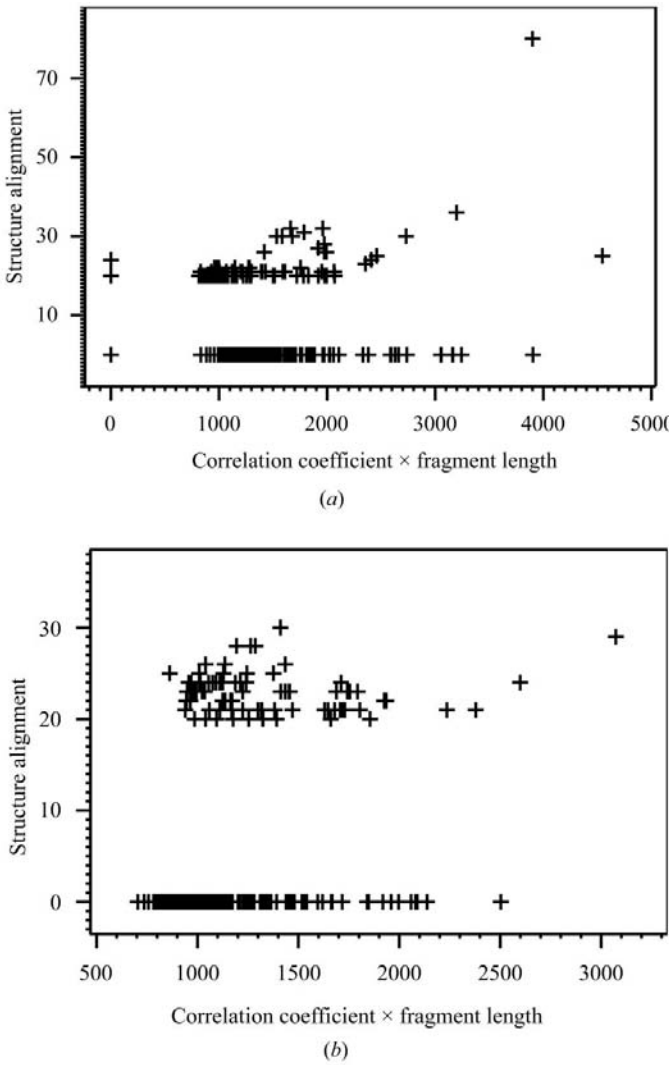

Figure 6
Graphs of alignment length between each of the feature fragments and the proteins myoglobin (*a*) and insulin (*b*) plotted as a function of the product of MR correlation coefficient and fragment size. The discontinuity observed between the alignment length of zero and 20 is the result of recording alignment lengths only above 20 residues. Hence, if a fragment aligns with a length of less than 20 residues then it lies at $y = 0$. The fragments longer than 80 residues (see graph in Fig. 3) are not shown for clarity, but have an x ordinate that is less than 30.

good correlation coefficients. From this analysis only two results from a cluster of seven, with scaled correlation coefficients greater than 3000, are correct solutions within the myoglobin example. The insulin example has five results spread out to higher correlation values and the two largest are correct solutions, but both from the same feature. It should be noted that owing to the method of generation some feature fragments are small parts (subsets) of other fragments that occur with higher frequency. When phasing by MR is difficult, analysis of ten solutions would be considered a trivial problem; however, automated methods using multiple solutions at different sites in the molecule require more conclusive results. It is felt that these correlations do not represent conclusive results, but do suggest that more sensitive MR methods could be productive.

4. Conclusions

The aim of this work was to generate a series of structural motifs by looking for common folds within a non-degenerate set of proteins and use these with molecular replacement. Although the generation of the features was successful and is being used in a number of lines of research, the MR results have only been partially successful with the myoglobin and insulin data. In particular, it can be seen that the size of the features generated by data mining is restricted to small values. The PDB data is currently being reduced again with a number of algorithmic limitations adjusted to produce a better feature set. Statistical analysis of the results indicates that the recurrent features will still be limited in size, less than 70 residues. The algorithm within *AMoRe* is probably not sensitive enough to produce MR solutions with moderate and large proteins using features less than 70 residues in length. These examples are probably at the limit of sensitivity of this technique and worked only because the packing of the *EFGH* helices was present as a feature in myoglobin and because insulin is small. As more powerful MR methods become available that are more sensitive, this methodology described may yield useful results for small proteins.

5. Program availability

The programs *SQUID* and *PDBSELECT* are available to academic institutes from http://www.ysbl.york.ac.uk/~oldfield. The program *PDBSELECT* available from this site has a rather slow sequence-alignment algorithm based on dynamic programming; a later version using very high-thoroughput sequence alignment is not available.

The programs *CAMINE* and *ANAFRAG* are not available, although publications on the methods used are in preparation.

I would like to thank Eleanor Dodson for help with the molecular replacement and Leo Caves for constructive discussions on structure alignment and data analysis. This work was entirely funded by Accelrys Inc.

References

Alexandrov, N. N. & Go, N. (1992). *J. Mol. Biol.* **225**, 5–9.
Bernstein, F. C, Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Boutonnet, N. S., Rooman, M. J., Ochagavia, M. E., Richelle, J. & Wodak, S. J. (1995). *Protein Eng.* **8**, 647–662.
Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187–217.
Fischer, D., Bachar, O., Nussinov, R. & Wolfson, H. (1992). *J. Biomol. Struct. Dyn.* **9**, 769–789.
Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1993). *J. Mol. Biol.* **229**, 707–721.
Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.
Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.
Holm, L. & Sander, C. (1996). *Science*, **273**, 595–602.
Mizuguchi, K. & Go, N. (1995). *Protein Eng.* **8**, 353–362.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Smerdon, S. J., Oldfield, T. J., Dodson, E. J., Dodson, G. G., Hubbard, R. E. & Wilkinson, A. J. (1990). *Acta Cryst.* B**46**, 370–377.
Oldfield, T. J. (1992). *J. Mol. Graph.* **10**, 247–252.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.
Ramachandran, G. N. & Sasisekharan, V. (1969). *Adv. Protein Chem.* **23**, 283–437.
Rufino, S. D. & Blundell, T. L. (1994). *J. Comput. Aided Mol. Des.* **8**, 5–27.
Sali, A. & Blundell, T. L. (1990). *J. Mol. Biol.* **212**, 403–428.
Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *Folding Des.* **1**, 209–220.
Taylor, W. R. & Orengo, C. A. (1989). *J. Mol. Biol.* **140**, 77–199.
Vriend, G. & Sander, C. (1991). *Proteins*, **11**, 52–58.
Whittingham, J. L., Chaudhuri, S., Dodson, E. J., Moody, P. C. E. & Dodson, G. G. (1995). *Biochemistry*, **34**, 15553–15563.