Data mining the protein data bank: Residue interactions

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · DECEMBER 2002

Impact Factor: 2.63 · DOI: 10.1002/prot.10221 · Source: PubMed

1 AUTHOR:

Thomas J Oldfield
EMBL-EBI
30 PUBLICATIONS **874** CITATIONS

SEE PROFILE

PROTEINS: Structure, Function, and Genetics 49:510–528 (2002)

Data Mining the Protein Data Bank: Residue Interactions

T.J. Oldfield*
Accelrys Inc., Department of Chemistry, University of York, Heslington, York, Yorkshire, United Kingdom

ABSTRACT **The protein databank contains a vast wealth of structural and functional information. The analysis of this macromolecular information has been the subject of considerable work in order to advance knowledge beyond the collection of molecular coordinates. This article presents a method that determines local structural information within proteins using mathematical data mining techniques. The mine program described returns many known configurations of residues such as the catalytic triad, metal binding sites and the N-linked glycosylation site; as well as many other multiple residue interactions not previously categorized. Because mathematical constructs are used as targets, this method can identify new information not previously known, and also provide unbiased results of typical structure and their expected deviations. Because the results are defined mathematically, they cannot indicate the biological implications of the results. Therefore two support programs are described that provide insight into the biological context for the mine results. The first allows a weighted RMSD search between a template set of coordinates and a list of PDB files, and the second allows the labeling of a protein with the template results from mining to aid in the classification of this protein. Proteins 2002;49:510–528.**

Key words: mathematical data mining; active sites; binding sites; protein structure; templates; superposition; residue configurations

INTRODUCTION

The protein databank (PDB)[1,2] is rapidly increasing in size as protein structure determination methods become more automated. With the advent of structure genomics, it should be expected that the exponential grown in the number of protein structures within the PDB will be maintained or even surpassed.[3] It is necessary in order to collate information on protein structure to understand principles of fold and function within proteins.[4–11] The information derived from structural biology then has to be presented to scientists in the fields of genetics, biology, and chemistry who wish to focus on the details of this information. This requires the identification and subsequent presentation of the pertinent detail within a coordinate structure in a clear and concise way.

The region of protein structure that is of principal interest to the scientific community is generally the active site or binding site depending on the protein's biological role. There are other regions within a protein, such as the packing of the core, that are also of note, though usually these are foci only for protein scientists. The active site or binding site is often localized in space, although not necessarily in sequence. Recognition of these significant features has been integral to the structure determination process, but recently, with the advent of structure genomics, the aim has become structure solution without careful subsequent analysis. As more and more structures are solved in a high-throughput way, it becomes important to identify features in these proteins automatically. One problem is that novel proteins may have local structure features that are not described within the literature, so new information is missed. While the identification of known features within proteins is essential, the ability to identify new information, not yet assigned as biologically important, could be invaluable. Mathematical data mining, therefore, provides a method that can recognize local features common to a number of protein structures, which can then be correlated to biological function.

Methods that identify recurrent common interactions within proteins without prior knowledge of function are described elsewhere,[12] though these methods are limited to identifying the common feature within a homologous set of proteins due computational times. Template searching programs using a number of different algorithmic approaches have been described that are suitable for the analysis of a known residue interaction. These include the use of graph theory to look for isomorphous patterns of atoms,[13–15] geometry hashing to define templates[16] and other general methods of pattern identification.[17] These local structure-searching techniques are biased by the choice of homologous structures, or by the target itself. Although they can provide a mean coordinate template structure as a result, the original target definition is based on a chemical and biological description of what is required. Not only is the result limited to that of the original chemical definition, but the statistical distribution of it will be biased towards the origin. The description and analysis of local residue interactions is further complicated because proteins that have the same spatial relation of residues within an active site usually have the same functional activity but it is not necessarily true that these

*Correspondence to: T.J. Oldfield, EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
E-mail: oldfield@ebi.ac.uk

Received 18 December 2001; Accepted 13 June 2002

proteins will have the same structure.[18,19] These difficulties require that important spatial relationships of atoms must be identified regardless of their topological arrangement in a protein, and even independently of the amino acid type the atoms are part of. Even so, the number of residue interactions that have been studied in great detail is large. Metal binding motifs,[20,21] the catalytic triad,[16,22] and a number of ligand bind sites[23,24] have been studied specifically. The generation of a general database of 3D motifs is also being undertaken by systematic study of the PDB.[16]

This study describes a data mining method that can determine common atom/residue interactions within proteins. The method has been designed to be highly efficient, taking just seconds to analyze all the non-degenerate information within the PDB for interacting patterns of atoms. The analysis is not limited to amino acids as any chemical group can be defined including that of solvent. The output from the analysis is a huge database of common amino acid configurations found in the known protein structure space determined by mathematical targets. Data mining as implemented in SIDEMINE does have some disadvantages, especially when used on complex 3D data such as protein structure. Useful information can be swamped by noise and mining methods are sensitive to systematic error. Therefore, careful data selection and validation are required. The output is almost as large as the data itself, and so the data mining described here cannot reasonably be defined as a data reduction technique. Furthermore, since the targets are mathematical, no biological relevance is identified by the calculation. Therefore, other programs are required to collate the data (ANAMINE) and identify biological meaning (SITEMINE and TEMPLATE). The three analysis programs that collate and process the mine information use novel methods to handle the problems of chemical equivalence between amino acids and internal chemical symmetry within some amino acids. They also handle limited atom selections and variable atom weights and they are not prone to problems of instability of alignment of small number of atoms that are planar or linear. This study describes the algorithms used within the four programs as well as some preliminary results based on the analysis of protein fragments that have different sequence, although the methods are not limited to this type of analysis. The results are, therefore, based on the study of common analogous feature observations found in a set of proteins with partially dissimilar sequences. The results presented here aim to justify the methods and algorithms described. In this context, only results that should be expected to be produced by this type of analysis are shown as a number of lines of work are in progress to determine and justify some of the more novel details found; these will be presented elsewhere.

METHODS

A summary of the data flow and programs is shown in Figure 1. The mining procedure requires an initial data preparation stage, an N-body mine and a data collation step. The results consist of files containing coordinates and

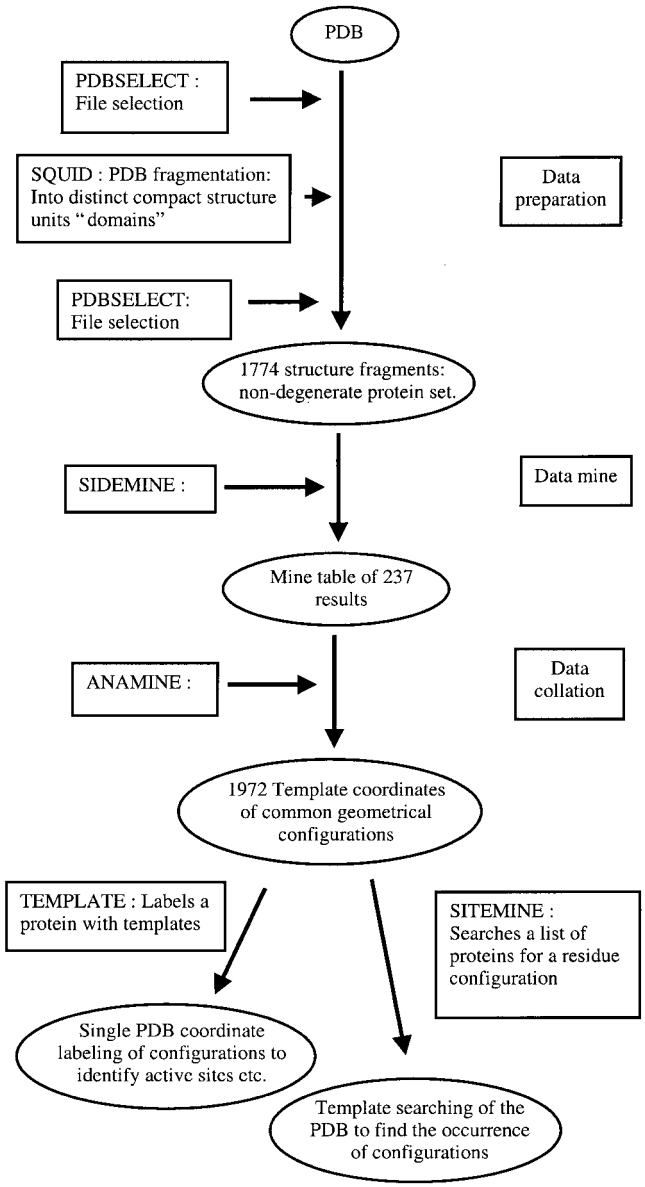

Fig. 1. Flow diagram to show data flow and programs used.

variance of common occurring geometric configurations of residues. The programs SITEMINE and TEMPLATE were written for general protein analysis and characterization and are designed to use these template coordinates.

Data Preparation

It is necessary to process the PDB before data mining because of the problems associated with degenerate structural entries. Therefore, the data was selected on the basis of sequence at a threshold of 80% exact residue optimal alignment and outlier removal. This will retain some homologous features within the protein data, and so some configuration results will reflect this. A second issue is that the PDB contains mainly crystallographic information, which represents the asymmetric unit of the crystal, and is not necessarily the biologically active protein. There are

two options for data preparation: the first is to generate these biologically active structures, and second, to fragment the proteins further so that the representative set of data is the non-degenerate set of protein fragments. The second approach has the advantage of removing homo-multimer structure from the representative set, which improves the signal-to-noise ratio, but prevents the analysis of inter-domain common sites of interest. The generation of the biologically active set of proteins is a major undertaking and is certainly not an automated process. This task is currently being undertaken elsewhere (personal communication K. Henrick, EBI). Therefore, the fragmentation method was used as this can be automated and thus allows the methods presented to be validated.

A program PDBSELECT (Oldfield, unpublished program) was used to provide a set of non-degenerate protein fragments defined by sequence similarity, with data outliers removed[25] (Appendix A). The 1301 sequence dissimilar proteins used for this study are listed in Appendix B. These protein structures were then fragmented with no user intervention by the program SQUID[26] using analysis of the two-dimensional finite difference of the Cα distance matrix. The definition of the fragmentation is entirely mathematically defined and details are presented elsewhere.[25] This was followed by another sequence similarity check to remove fragments of structure generated, for example, by the splitting of a homo-diamer proteins. In general, the proteins were divided up by this analysis into compact structural units that would normally be regarded as domains. The distribution of fragment sizes had the same mean (130 residues) and variance as estimated by Poisson curve fitting[26] to that of the CATH[9] domain data (data not shown). The resultant 1,774 fragments of protein were used in the subsequent analysis, where each protein fragment was given a seven-letter name. The first four letters originate from the original PDB, the fifth letter is an underscore, and the sixth and seven define a numerical fragment order within the original protein.

SIDEMINE: Mining for Atom Interactions

Local interaction mining was performed by the program SIDEMINE. This program carries out N-body correlation analysis of all interactions for a list of proteins simultaneously. The output from this program is a number of hits that gives information on the proteins, the residues and additional information to uniquely define each residue configuration. The design of a mine program must take into account the combinatorial problem of studying atomic interactions within three dimensions. Even the results presented here, which used a sub-set of 15 residues, required the correlation of more than 2.6 million three-dimensional (3D) coordinates. The aim of the analysis was to make as few biological/chemical assumptions as possible and define results using a mathematical metric. It is possible to carry out analysis with no prior expectation encoded at all but this yields few results because of the low signal-to-noise ratio (S/N) of the data.

Within the 20 amino acids found in proteins, there is some equivalence in chemistry between some atoms. For example, the residue TYR is aromatic and is therefore chemically similar to the PHE amino acid. It also has an oxygen atom with two lone pair electrons similar to that of the SER oxygen atom. A second important note is that five of the 20 amino acids have internal symmetry (PHE, TYR, ASP, GLU, and ARG), and this must be taken into account when searching for equivalent atomic positions in space. The similarity in chemical properties of amino acids allows the ability to improve the S/N of a search by defining equivalence between groups of atoms. Residue equivalence was made between the amide groups of GLN and ASN, between the acid grouping in ASP and GLU, the aromatic rings of PHE, TYR, and TRP, the oxygen atom in SER and THR, and the basic chemical grouping in LYS and ARG. These equivalence definitions can be varied by run time options of the SIDEMINE program. The SIDEMINE program handles internal symmetry by defining a single averaged point (over a number of atoms) that is representative of the side chain atoms. The single point definitions used are shown in Table I. All atomic interactions up to 8Å were considered when they contained more than five examples in a hash bin of bin width 0.5Å. Results were preserved between N-body calculations when they contained more than five members.

The program SIDEMINE is designed to process the mass of information extremely fast (Table II) and return a vast number of approximate relationships together with some false positives. Since it is not known what the outcome of data mining should be, it is impossible to determine the false-positive rate. It, therefore, needs to be processed by the program ANAMINE.

TABLE I. Point Definitions for the Side Chain Mining Program SIDEMINE[†]

Amino acid	3 letter code	N-body definition
Alanine	ALA	Cβ
Arginine	ARG	NH1, NH2
Asparagine	ASN	Oδ1
Aspartic acid	ASP	Oδ1, Oδ2
Cystine	CYS	Sγ
Glutamic acid	GLU	Oε1, Oε2
Glutamine	GLN	Oε1
Glycine	GLY	Cα
Histidine	HIS	Nε2
Isoleucine	ILE	Cγ1, Cγ2, Cδ1
Leucine	LEU	Cγ, Cδ1, Cδ2
Lysine	LYS	Nζ
Methoinine	MET	Sδ
Phenylalanine	PHE	Cγ, Cδ1, Cδ2, Cε1, Cε2, Cζ
Proline	PRO	Cβ, Cγ, Cδ
Serine	SER	Oγ
Threonine	THR	Oγ1
Tryptophan	TRP	Cδ2, Cε2, Cε3, Cζ2, Cζ3, CH2
Tyrosine	TYR	Cγ, Cδ1, Cδ2, Cε1, Cε2, Cζ
Valine	VAL	Cγ1, Cγ2
Water	HOH	O

[†]The mean position of the list of atoms is used to define the search point in the N-body analysis. The crystallographic water names are all converted to that shown in the table during the triplet list generation for consistency.

TABLE II. Timing Statistics for the Side Chain Mining Program SIDEMINE

Calculation	Number of results	Time/seconds
550 Mbyte read	—	30
3 residue interaction	198	4
4 residue interactions	31	2
5 residue interactions	1	<1
6 residue interactions	6	<1
7 residue interactions	1	1
Total interactions	237	8

ANAMINE: Collating the Mine Results

The second program ANAMINE reads the output from SIDEMINE, and for each mine solution, collates all the information from each hit to provide specific configuration template files. This program uses weighted atomic RMSD, residue equivalence tables, and residue symmetry analysis to correctly handle the data and subsequently perform cluster analysis on each of the mine interaction hits. The output files are templates of common features in proteins and are classified only by their amino acid content. This collation program uses complete linkage analysis on the square symmetric matrix of atomic least squares (LSQ) pair-wise deviations between residue configurations. The results generated here are based on key atoms of a residue (Table III) within a least squares overlay of 1.5Å^2 that occurs five or more times. The program can either output files containing mean coordinate positions with standard deviations, or coordinates for all overlaid configurations in the same frame of reference. Any interacting ligands are also added to these result files to aid in classification. The biological relevance of these templates must be inferred from the associated information about ligand interaction, PDB header information, and surface accessible areas contained within them.

SITEMINE and TEMPLATE: Template Searching Programs

The two analysis programs SITEMINE and TEMPLATE were written to use the results from the mine analysis. The program SITEMINE can take a single template result (or any set of user-defined residues) and carry out an analysis against a list of PDB files for the presence of this template. The program handles internal residue symmetry and atomic weighting. The program TEMPLATE reads a single PDB structure, and determines from the list of mine result templates whether this PDB structure contains examples of any of the residue configurations. The program handles residue symmetry, standard deviations (SD), atomic weighting, absolute atomic deviation, and overall root mean square deviation (RMSD). These two programs are obviously very similar in their application but have been designed to concentrate on different aspects of protein structure analysis giving rise to the difference in options.

ALGORITHMS

SIDEMINE: Theoretical Issues

The basic unit of interaction between atomic coordinates is the distance between two atoms. Distance is a 2-body 1 Dimensional (1D) scalar property having no direction, and therefore is a sensible starting point for interaction searching in proteins. However, an attempt to use the 2-body concept was not successful due to the large number of non-bond interactions to combine by N-body analysis. This resulted in impractical computation times. It was, therefore, decided to use a 3-body interaction of atoms as the basic starting point of analysis. The algorithm described cannot, therefore, determine two atom (non-bonding) features.

The 3-body interaction (Fig. 2) can be defined by a number of metrics, the simplest of which is the separation of the three points in space, requiring three scalars. Interactions with more atoms can be defined by searching for N-body interactions directly, but it is more efficient to note that all N-body interactions are a union of two (N-1)-body interactions (Fig. 3). For example, to find a 4-body interaction it is necessary only to combine 3-body interactions, and store any common 4-body features that are found as defined by the scalar metrics shown in Figure 3. This analysis forms the basis of the SIDEMINE algorithm, and expansion continues until the set of N-body interactions is empty. It should be noted that if any (N-1)-body interaction has exactly the same members as an N-body interaction, then this (N-1) body interaction is degenerate by order and is not a useful result.

The next requirement is to define what constitutes the "body" for side chain atomic interaction. The external data file in Table I defines 21 template definitions. Symmetry can, therefore, be handled in a very simple way using this description as the average of equivalent atoms. The side chain point, which is the average of six atoms for a PHE amino acid, represents the principal point for the N-body definition. A main chain point, in this case just a Cα atom, represents a secondary classification that provides subgrouping information from the mine; it is not used for the mine computation.

SIDEMINE: Implementation

The first stage in the implementation of these concepts in SIDEMINE was a data transformation from the PDB format files to speed up the subsequent mine calculation. Reading and processing the PDB files represents the slowest part of the analysis and needs only to be carried out once for a set of proteins. The list of PDB files generated from the PDBSELECT program was searched for triplets of residues that were within a user-defined cut-off radius. The cut-off distance used in this analysis was 8Å between two out of three residues and 16Å for one of three interactions in a residue triplet. The triplet list was written as a binary file of data records that includes atomic coordinates, non-bond interactions to ligands (not "protein or water"), and the mean accessible surface area (ASA) of each residue. The triplet list was created by the SIDEMINE program in a "generate mode" and each inter-

TABLE III. Atom Match Table for an ASP Residue, Part of an External Definition File†

```
residue ASP
main CA
side OD1 OD2
all CA N C O CB CG OD1 OD2
match GLY CA:CA
match ALA CA:CA CB:CB
match VAL OD1:CG1 OD2:CG2       sym OD1:CG2 OD2:CG1
match ILE OD1:CG1 OD2:CG2       sym OD1:CG2 OD2:CG1
match LEU OD1:CD1 OD2:CD2       sym OD1:CD2 OD2:CD1
match TYR OD1:OH                sym OD2:OH
match PHE CG:CG
match TRP CG:CD2
match PRO CG:CA
match MET CG:CG
match CYS CG:CB OD1:SG          sym CG:CB OD2:SG
match SER CG:CB OD1:OG          sym CG:CB OD2:OG
match THR CG:CG OD1:OG1         sym CG:CG OD2:OG1
match HIS OD1:NE2               sym OD2:NE2
match ASN CG:CG OD1:OD1 OD2:ND2 sym CG:CG OD2:OD1 OD1:ND2
match GLN CG:CD OD1:OE1 OD2:NE2 sym CG:CD OD2:OE1 OD1:NE2
match ASP CG:CG OD1:OD1 OD2:OD2 sym CG:CG OD2:OD1 OD1:OD2
match GLU CG:CD OD1:OE1 OD2:OE2 sym CG:CD OD2:OE1 OD1:OE2
match LYS CB:CB CG:CG
match ARG OD1:NH1 OD2:NH2       sym OD2:NH1 OD1:NH2
end
```

†These entries define how the atom matches are set up for least squares fitting between one residue and other residues. The "sym" definition is used where there is an exact chemical match between atoms related by a rotational symmetry axis about the last χ angle.

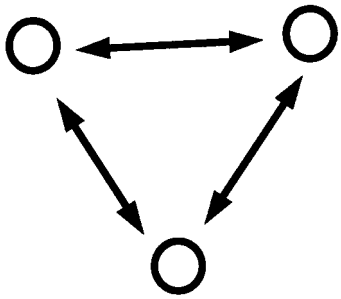

Fig. 2. Graphic to show the basis 3-body interaction with the scalar measurement of distance marked between each body.

action triplet could be sorted by sequence order or residue type order (Table I). The latter was the default and so subsequent analysis was independent of the residue sequence order within the protein. The triplet list was then read into memory and a geometrical hash table generated for each three-residue combination. Equation 1:

$$GH_{R[1][2][3]} = 3D \text{ hash-bin}[i][j][k]$$

Where:-

$i = \text{INT}(d_{12}/(\text{hash precision}))$

$j = \text{INT}(d_{23}/(\text{hash precision}))$

$k = \text{INT}(d_{31}/(\text{hash precision}))$

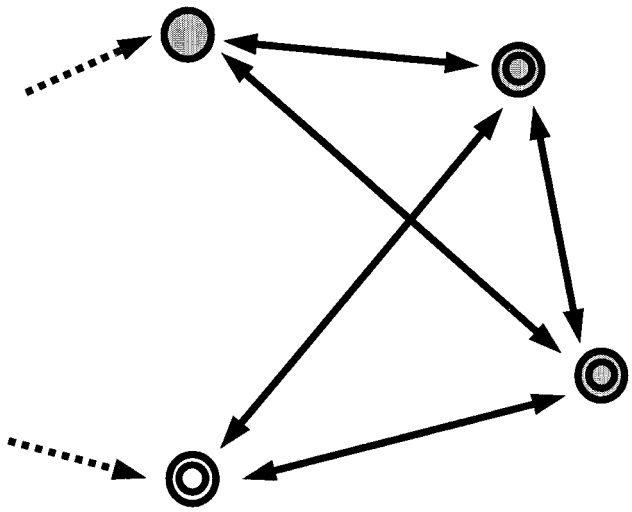

Fig. 3. The general N-body interaction, with 4 points explicitly shown. Two 3 body interactions (inset circles and shaded circles) are shown as subsets of a 4-body interaction.

$d_{12} = \text{dist}(R[1], R[2]) : d_{23} = \text{dist}(R[2], R[3]) :$

$d_{31} = \text{dist}(R[3], R[1])$ for all residue triplets (t)

where $(R[1], R[2], R[3]) \in t$

$\text{GH}_{\text{R[1][2][3}}$ = geometric hash table for residues types R[1], R[2] and R[3]. d_{12}, d_{23}, and d_{31} are the separation between

the 3 pairs of residue side chain points (Table I); the function INT specifies the nearest integer.

At this stage, if there were any equivalencies between residues then all residues within an equivalence were included in one hash. A hash table was set up using the three distance metrics, with a hash precision of 0.5Å. A peak search was carried out on the 3D-hash list to identify common interactions between residues where the peak was above a user-defined minimum (Equation 2). Note that shoulder points above the user-defined minimum were also included in subsequent analyses.

A 3-body interaction exits where: (Equation 2) (hash-bin[i][j][k] ≥ peak threshold)

A 3-body interactions consists of:

$$\left(\sum_{ii,jj,kk} \text{hash-bin}[ii][jj][kk]\right) : ii = i - r, i + r,$$
$$jj = j - r, j + r, kk = k - r, k + r$$

where r = peak range value.

Using each peak and shoulder as a starting point, a surrounding user-defined volume of data was selected from the hash list. If hash bin 10:10:10 (representing distances of 5Å) represents a peak, then all the values in bins 9 to 11 in each dimension were included for a user defined *peak range* value of "1." All the members of this peak, and the volume about it, represent a 3-body interaction, and were stored as a solution. This was repeated for all combinations of residues. The result of a hash list peak analysis for all possible combinations of residues (with equivalence) was a list of 3-body configurations where 1 body is the mean point for the atom lists in table I. Each 3-body configuration has multiple members that are the different occurrences of the common configuration within the data (equations 3a,b).

Consider sets A and B, which are two N body interactions with "ma" and "mb" configurations present in the data:(equation 3)

$$A = \{1 \cdots N\} : \text{with configurations } a_{ie=1,ma} \in A$$

$$B = \{1 \cdots N\} : \text{with configurations } b_{je=1,mb} \in B$$

Where a and b are elements of the N-body interaction and are the individual configurations found within the protein data.

4-body solutions are now determined by combining two 3-body configurations. 3-body configurations (equations 3a,b) can be combined if there is equality in two out of three members from each triplet and where there is an equivalence in the member elements of these two that is above the peak threshold. For the general N-body interaction, if (N-1) bodies are equivalent over K elements then a larger (N+1) body is created (equation 4).

Then for the set C, which is a (N+1) body interaction: (equation 4)

$$C_{AB} = \{A \cup B\} \text{ iff } [(\#(A \cap B) = (N-1)) \wedge_{(a,b) \in A \otimes B} (\#(a \cap b) \geq K)]$$

Where iff : If and only if K is the peak threshold. # is the number elements in the set.

Elements "a" and "b" require their properties to be equal for an intersection. These properties are the residue type (including residue equivalence), protein (i.e., both 3-body interactions must be in the same protein) and hash position (interaction distance). If either of the N-body configurations has exactly the same elements as the new (N+1)-body configuration, then the N-body configuration is marked as not-interesting, but the data retained for other analyses. The practical implementation of equations 1 to 4 must be able to balance the consideration of the large size of the data, and the calculation time. The algorithm was implemented so that the N-body information was created when found and data deleted when no longer required. This design was implemented, instead of using pointers throughout, for the principal reason as to allow the program to carry out multiple passes of the interaction triplet data for each geometric hash table (equation 1). The 3-body interaction list is appended for each separate hash tables analysis. It is not required to load the entire interaction triplet information into memory at one time, although the data analyzed here was small enough to allow this. This design was considered important because the size of the protein databank is increasing and at the current time the triplet data for the 15 residue analysis presented here is already 550 Mbytes.

ANAMINE, SITEMINE, and TEMPLATE: Theoretical Issues

The three programs ANAMINE, SITEMINE, and TEMPLATE use atomic least squares (LSQ) to define similarities between sets of coordinates. Computational features common to these programs are described.

The first problem is that the mine program includes residue equivalencies; an example is the acidic residue pair ASP and GLU, which have a group of atoms that have the same chemistry. It is therefore necessary to define how each atom in a residue overlay with the atoms of other residue types. Table III shows the overlay table for the ASP amino acid and includes entries for all 20 amino acids. This information is stored in an external data file with all the other residue definitions.

The second problem with the atomic LSQ is that five residues have a 2-fold rotation symmetry axis concurrent with the last angle of the side chain (ASP, GLU, PHE, TYR, ARG). In the ASP residue, the atoms O1 and O2 are electronically equivalent, and nomenclature rules dictate that the absolute value of the torsion angle must be less than 90°. Since the root point definition of a torsion angle of a side chain is the main chain Cα atom, the nomenclature rule does not aid in side chain optimal overlap. Table III shows that "sym" definitions are included for the ASP residue. A "sym" definition is an alternative way for atoms to be overlaid, and each LSQ calculation performed by the ANAMINE, SITEMINE, and TEMPLATE programs tries all combinations of the "sym" definitions, where present, for all residues in the interaction. The RMSD returned

between residues with internal symmetry is the lowest value obtainable by combinatorial analysis.

The third issue for residue similarity measurement can be observed with the catalytic triad. There are five atoms in the HIS residue, one atom in the SER residue, and one atom in the ASP residue that have invariant positions important for catalysis. The five-atom match of a HIS residue will dominate the alignment over a single atom match of a SER and ASP residue, which may or may not be required. To overcome this, the LSQ method in ANAMINE, TEMPLATE, and SITEMINE uses a weighted atomic fit, so that some atoms can increase their dominance in the alignment. In the catalytic triad, it is possible to increase the weight of the SER-Oγ and the ASP-Oδ1 fivefold to balance the fit. The implementation of weighting is shown in Appendix C.

The LSQ routine used in the three alignment programs is based on the algorithm of Kabsch[27] but uses a 4×4 set of cross terms rather than the principal components of the atomic inertia tensor matrix (Appendix C). This mathematical description of shape is not sensitive to planar or linear sets of atoms, which can result in instability within an eigen routine during diagonalization. In particular, the important atoms that make up the catalytic triad occasionally produced spurious results when using the un-modified algorithm of Kabsch.

ANAMINE: Implementation

The ANAMINE program reads the output from the mine program SIDEMINE, and for each hit that represents a putative multiple residue interaction the ANAMINE program extracts the original atomic information from the PDB files. A square symmetry matrix of RMSD is generated by LSQ overlap for all against all based on the prescribed atom selection (Table III). The square symmetric matrix of RMSD values is analyzed by a complete linkage sub-structure analysis to determine the common configurations of residues (equation 5).

$$d(r, x) = \max(dist(x_{ri}, x_{rj})), i \in (1, n_r), j \in (1, n_S) < \text{cut-off}.$$

Where n_r is the number of configurations in cluster r and n_s is the number of objects in cluster s, and x_{ri} is the i^{th} object in cluster r.

For each group larger than the user-defined minimum, the mean, standard deviation, and occurrence of each atom within the residues are determined, and the template set of coordinates written out. Note that the atom occurrences will not necessarily be constant for a residue because there may be no equivalent atom defined for an equivalent residue match. An example of an average coordinate template is shown in Table IV. The template file name is defined by the amino acid names (i.e., CYSCYSCYSCYSPHE.PDB), and any multiple occurrence of a configuration is appended to this file within a MODEL and ENDMDL record pair.

SITEMINE: Implementation

The program SITEMINE is designed to use a single average template generated from the ANAMINE program (or any PDB file containing a number of residues) and search a list of PDB files for the occurrence of this template. The aim of the calculation is to determine the RMSD by least squares for all combinations of the template residues and protein residues of the same type. However, it is not practical to implement this as stated because of the computation time required, particularly with large templates. A number of filtering steps are used to optimize the calculation. The protein residue content is re-cast by generating link lists of the residues that are of the same type as those within the template, allowing rapid screening of the residues of interest. Each protein residue and template residue is reduced to a single point, by averaging all the atoms to be overlaid. The atoms to be overlaid are defined by the occupancy field of the search template and any value larger than 0.5 defines that that atom is to be used within the calculation. A difference distance matrix can then be progressively determined between the template residue average points and possible protein residue average points and, if any distance plus residue radius exceeds a threshold, then this combination of residues in the protein is rejected. The template distance matrix can be pre-calculated, and the multiple residue analysis is stack-based. If the difference distance matrix is completed with no significant error in any one value, then atomic LSQ is performed using the modified method of Kabsch with atom weights and cross terms vectors. It is also necessary to carry out a combinatorial analysis of conformations of those residues with internal symmetry. If the value of the RMSD is less than a user-defined value, then a "hit" is printed as shown in Table V.

TEMPLATE: Implementation

TEMPLATE is a program that takes a single protein coordinate PDB file and matches all the templates generated by SITEMINE against this file, printing out a list of solutions. The TEMPLATE program prints out the REMARK information stored in the templates on finding a solution (Table VI), and the residues associated with the interaction. The TEMPLATE program handles residue symmetry in the manner described earlier and uses progressive difference distance matrix filtering similar to that of SITEMINE to improve calculation speeds. There are a number of differences compared with the SITEMINE program. Each template solution file contains information from the original PDB files, the overall RMSD of the template and the RMSD values for each atom. The TEMPLATE program can screen the template file for limits on these two parameters using user-defined options. For each template alignment found in the PDB file, the program defines an overall RMSD for a LSQ hit, an absolute deviation limit for each atom, and a scaled limit based on the standard deviation (SD) for each atom. For example, it is possible to define an upper limit of 1Å deviation for all atoms, as well as a three SD limit on each atom. Those atoms in the template that have low variation in position will have a low SD, and so an alignment hit will be limited

TABLE IV. A Small Example Template File Generated by the Program SIDEMINE for the Interaction of Four CYS + PHE Residues Forming an Iron Binding Site

```
MODEL
REMARK   Number of examples = 6
REMARK   RMSD = 0.224
REMARK      6/ 6 have ligands
REMARK      6/ 6 have same ligand
REMARK   Ligand fields include (> 0.200000 occupancy)

REMARK          (1b2o_01)          HEADER          ELECTRON TRANSPORT                                30-NOV-98
REMARK          (1b71_01)          HEADER          ELECTRON TRANSPORT                                26-JAN-99
REMARK          (1bq8_01)          HEADER          IRON-SULFUR PROTEIN                               22-AUG-98
REMARK          (1rb9_01)          HEADER          IRON-SULFUR PROTEIN                               21-DEC-97
REMARK          (1rdg_01)          HEADER          ELECTRON TRANSFER (IRON-SULFUR PROTEIN)           17-MAR-88
REMARK          (6rxn_01)          HEADER          ELECTRON TRANSFER (IRON-SULFUR PROTEIN)           16-JAN-90
REMARK          (1b2o_01)          TITLE           CLOSTRIDIUM PASTEURIANUM RUBREDOXIN G10VG43A MUTANT
REMARK          (1b71_01)          TITLE           RUBRERYTHRIN
REMARK          (1bq8_01)          TITLE           RUBREDOXIN (METHIONINE MUTANT) FROM PYROCOCCUS FURIOSUS
REMARK          (1rb9_01)          TITLE           RUBREDOXIN FROM DESULFOVIBRIO VULGARIS REFINEDTITLE
REMARK          (1rdg_01)          TITLE           N/A
REMARK          (6rxn_01)          TITLE           N/A

REMARK      Molecule = 1b2o_01 :        6       39       49        9       42:       0      1      3      4
REMARK      Molecule = 1b71_01 :      158      174      184      161      177:       0      1      3      4
REMARK      Molecule = 1bq8_01 :        6       39       49        9       42:       0      1      3      4
REMARK      Molecule = 1rb9_01 :        6       39       49        9       42:       0      1      3      4
REMARK      Molecule = 1rdg_01 :        6       39       49        9       42:       0      1      3      4
REMARK      Molecule = 6rxn_01 :        6       32       42        9       35:       0      1      3      4

ATOM       1      CA       CYS       1      42.051      16.959      12.855      6.00      0.10
ATOM       2      N        CYS       1      42.218      18.112      11.974      6.00      0.13
ATOM       3      C        CYS       1      40.959      17.282      13.880      6.00      0.18
ATOM       4      O        CYS       1      41.087      18.258      14.609      6.00      0.31
ATOM       5      CB       CYS       1      43.358      16.650      13.556      6.00      0.10
ATOM       6      SG       CYS       1      43.212      15.212      14.656      6.00      0.08
ATOM       7      CA       CYS       2      47.757      12.973      12.487      6.00      0.23
ATOM       8      N        CYS       2      47.471      11.622      12.006      6.00      0.31
ATOM       9      C        CYS       2      49.124      12.983      13.137      6.00      0.21
ATOM      10      O        CYS       2      49.379      12.221      14.080      6.00      0.21
ATOM      11      CB       CYS       2      46.682      13.361      13.502      6.00      0.15
ATOM      12      SG       CYS       2      46.975      15.033      14.127      6.00      0.04
ATOM      13      CA       PHE       3      39.815      14.844       9.686      6.00      0.09
ATOM      14      N        PHE       3      39.583      13.520      10.242      6.00      0.20
ATOM      15      C        PHE       3      38.573      15.354       8.982      6.00      0.09
ATOM      16      O        PHE       3      37.799      14.572       8.413      6.00      0.13
ATOM      17      CB       PHE       3      40.985      14.801       8.686      6.00      0.16
ATOM      18      CG       PHE       3      42.318      14.849       9.395      6.00      0.09
ATOM      19      CD1      PHE       3      42.787      13.751      10.078      6.00      0.14
ATOM      20      CD2      PHE       3      43.070      16.009       9.363      6.00      0.11
ATOM      21      CE1      PHE       3      44.020      13.810      10.725      6.00      0.12
ATOM      22      CE2      PHE       3      44.298      16.074      10.004      6.00      0.12
ATOM      23      CZ       PHE       3      44.770      14.969      10.681      6.00      0.09
ATOM      24      CA       CYS       4      43.542      17.788      18.300      6.00      0.13
ATOM      25      N        CYS       4      42.546      16.920      17.708      6.00      0.15
ATOM      26      C        CYS       4      43.686      19.133      17.621      6.00      0.20
ATOM      27      O        CYS       4      44.193      19.831      17.910      6.00      0.95
ATOM      28      CB       CYS       4      44.906      17.082      18.358      6.00      0.08
ATOM      29      SG       CYS       4      45.783      17.090      16.759      6.00      0.10
ATOM      30      CA       CYS       5      47.762      13.067      18.072      6.00      0.13
ATOM      31      N        CYS       5      48.484      13.931      17.149      6.00      0.15
ATOM      32      C        CYS       5      47.658      11.613      17.646      6.00      0.16
ATOM      33      O        CYS       5      47.155      10.797      18.422      6.00      0.25
ATOM      34      CB       CYS       5      46.372      13.624      18.369      6.00      0.14
ATOM      35      SG       CYS       5      45.184      13.382      17.028      6.00      0.07
HETATM     3      E        FE        5      45.230      15.193      15.645      1.00     18.71      1b2o_01
HETATM     5      E        FE        7      45.330      15.185      15.661      1.00     19.98      1b71_01
HETATM     8      E        FE       15      45.281      15.204      15.690      1.00      2.91      1bq8_01
HETATM    11      E        FE2      20      45.264      15.180      15.668      1.00      4.98      1rb9_01
HETATM    14      E        FE       25      45.271      15.153      15.643      1.00      6.82      1rdg_01
HETATM    15      E        FE       26      45.253      15.150      15.635      1.00      6.88      6rxn_01
ENDMDL
```

TABLE V. Output From the Program SITEMINE†

```
Mol :343 : Solution /y/database/brookhaven/pdb/1ajy.pdb : 0.268652
 → (B) 53 (B) 37 (A) 53 (A) 50
HEADER    TRANSCRIPTION REGULATION              12-MAY-97   1AJY
TITLE     STRUCTURE AND MOBILITY OF THE PUT3 DIMER: A DNA PINCER,
TITLE    2 NMR, 13 STRUCTURES
```

†The output shows a single hit result using the template in Table IV to search the protein databank files. The output is shown with verbose level 3 so that header cards and title cards are include for each hit.

TABLE VI. The Output File From the Program TEMPLATE That Shows a Hit Between One of the 1,972 Templates and the File 1ajy.pdb†

```
There are 4 residues in hit
 RMSD = 0.813843 for 8 atoms masked
  Residues in hit : CYS CYS CYS CYS
 The residues in the PDB file are . . .
 Seg. num., Seg. name, Res. num., Res. Inum., Res. name
    1         (A)         50         21         CYS
    1         (A)         44         15         CYS
    1         (A)         34          5         CYS
    1         (A)         37          8         CYS
 The template hit contains the following . . .
 There are 6 ligands at this site
  and 6 are bound in the same way
 Site RMSD is 0.224000 over 6 examples
 The sites are found in the following proteins / Iresidue
  protein 1b2o_01 : Iresidues >       6      39       9      42
  protein 1b71_01 : Iresidues >     158     174     161     177
  protein 1bq8_01 : Iresidues >       6      39       9      42
  protein 1rb9_01 : Iresidues >       6      39       9      42
  protein 1rdg_01 : Iresidues >       6      39       9      42
  protein 6rxn_01 : Iresidues >       6      32       9      35
```

†The file contains the RMSD of the hit, the residues aligned, the residue information within the file, and then the details of the template information.

by the SD value, whereas atoms with a high SD will be subject to the 1Å absolute limit.

RESULTS

A number of different analyses have been performed with SIDEMINE to focus in detail on different properties of amino acids such as donor/acceptor interactions, water interactions, hydrophobic and aromatic interactions. The results presented here are based on an analysis of the protein databank as of March 2001 for common configurations of 15 amino acid types (not GLY, ALA, LEU, ILE, VAL). The mine of the database resulted in 237 different configurations where deviations of the single point definitions were less than 1.5Å. Subsequent use of the program ANAMINE generated 1,972 templates containing between 3 and 7 amino acids, many with ligand interactions identifying these as functional sites within the proteins. There was an expansion in the number of results obtained from ANAMINE because the mine program uses a simplified interaction definition that does not classify the results as a function of orientation whereas ANAMINE uses atomic LSQ. The results presented show examples from six classes of interaction found by this mine analysis. An important feature of data mining with mathematical targets is that if the method is correct it will reproduce known results, as well as potential new information.

Data Content of the Resultant Template Files

The output from the SIDEMINE program was extensive and contained 11,393 lines of at least 15 columns of data for the 15-residue analysis presented here. This amount of tabulated information is not suitable for direct analysis, but required collating. The output of ANAMINE (Table IV) represents the usable output from data mining. These files have a standard PDB format, where separate mine results with the same residue content are written to the same file. Different configurations of these same residues are enclosed in MODEL and ENDMDL records within each file. Table IV was therefore cut from a file of name CYSCYSCYSCYSPHE.PDB of averaged coordinates. Most graphical programs can, therefore, view the configurations stored in this format. All the traceback information is included in REMARK records to make it transparent to general display programs. The REMARK information is essentially a keyword record of the traceback details that can be parsed to indicate the parent protein and ligand association.

The first record contains the number of proteins in which the configuration is found, and in Table IV there

were six present in the fragmented PDB files. The second record indicates the overall RMSD of the template atoms defined by the matching information in part shown within Table III. The following three records correspond to information on ligand content and are present even if no ligand information was associated with the amino acid configuration. The first of these ligand records in the file indicates how many parent proteins have an associated ligand, while the second indicates how each ligand is associated with the residue configuration. This field defines the non-bonding pattern of the residue configuration with respect to the ligand, and is the maximum number of times the same non-bond pattern occurs between the ligand and residue configuration. In this example, the metal ions all interacted with all the four CYS residues in the same way and therefore the field has value six. The third ligand record is a threshold value used to determine whether the actual ligand coordinates are included in the file. There then follows HEADER and TITLE records from the parent PDB files, and only the first one is included if more than one is present. A number of older PDB files do not contain TITLE cards, so these are marked N/A if no information is available. The Molecule records provide the exact traceback details to the parent PDB, including the file name, the residue offset from the beginning of the file, and ligand association fields. The ligand association fields are present only if an amino acid in the configuration is non-bonded to the ligand, and for Table IV, all CYS amino acids were non-bonded to the ligand and the PHE not non-bonded.

The ATOM records that follow are in standard PDB format. For the averaged coordinate output, the file contains atom records for each residue within the configuration that are mean values over all example proteins, and the occupancy field gives the number of proteins that contain each atom. The value of the occupancy field does not necessarily equal the number of proteins containing the configuration for two reasons. There is a possibility that the original PDB file did not contain a coordinate for a particular atom, although this is unlikely. The main reason for reduced values of occurrence is equivalence between different residues. An average template file only contains atoms for the first residue of an equivalence definition, so if the definition is ASP-GLU, then only record entries exist for the atoms of an ASP residue. The match information (Table III) then define which atoms to use from the GLU residue (i.e., Oε1 for Oε2), and the rest are discarded reducing their occupancy. The B-value field indicates the atomic deviation of each atom. The all atom output contains coordinates for all configuration examples overlaid to the same frame of reference. All equivalenced residues are written out in full.

Finally the file contains HETATM records only if the number of ligand associations is greater than the occupancy threshold value defined in the 5th record. All ligands in the parent PDB files that interact with any the residues in the configuration are included. All HETATM records are individual coordinates of ligands and not average coordinates because of the problem of attempting to match different ligand atoms.

Summary of Results From Mining

The SIDEMINE calculation generated a total of 237 side chain interactions between 15 amino acids from the degenerate list of protein fragments. The calculation time was eight seconds (Table I). The ANAMINE program took a further 12 min and 52 sec. The result of this analysis was 83 different files (the largest of which contained seven residues) consisting of a total of 1,972 different geometrical configurations. Many of the configurations were similar; for example, there were 93 examples of three SER residues with different orientations. The results generated from the SIDEMINE and ANAMINE programs can be divided into six broad classifications.

1. The metal binding sites: There were a number of residue configurations that interacted with metal ions: zinc, calcium, magnesium, cobalt, cadmium, iron, and copper. The metal binding sites were highly conserved in their configuration and show little variation in key atom position due to the high energy of the coordination bonds. These interactions are well characterized in the literature.
2. Various binding sites were found with associated ligands such as sugars, nucleic acids, sulfates, peptides, and phorphyrins. These interactions had moderate conservation of residue conformation and orientation.
3. The catalytic triad: These included the hydrolases, lipase's, thio-esterases, and all combination of acid (though not exclusively)-HIS-nucleophile. The interaction was characterized by a highly restricted configuration with all key atom positions highly conserved in relative position. No other configuration that had associated ligands showed such a level of atom position conservation.
4. The salt bridge: There were a number of three residue salt bridges of the form acid-base-acid and base-acid-base. There was a large amount of variability in the residue positions as this was a non-specific electrostatic interaction. There was some ligand association.
5. The general polar residue interaction: These were interactions between the hydrophilic, non-charged residues and many involved the residue SER. Some of these included associated ligands.
6. The aromatic interaction: The programs identified a number of aromatic ring interactions, most of the type edge to ring plane. The amino acids generally had zero surface accessible area as the configuration was found within the core of proteins.

Six Specific Results From Mining

Figure 4 shows an interaction of CYS-CYS-CYS-CYS -PHE with iron metal ions at the center of the 4-CYS cluster. Figure 4 is the graphic for Table IV and examples of this configuration were found in the proteins: 1b2o, 1b71, 1bq8, 1rb9, 1rdg, and 6rxn. The six proteins are all classified as electron transport proteins or similar, and contain a common structural motif important for metal binding with four CYS residues coordinated to Iron ions. The PHE residue was picked up as a residue close (< 6.5Å)

TABLE VII. A Sequence Comparison of the Metal Bind Motif That Contains the Five Residues Shown in Table IV and Figure 4[†]

```
6rxn MQKYVCNVCGYEYDPAEHd.......NVPFDQLPDDWCCPVCGVSKDQFSPA..
     :    *  *|| :     :          :   : :: *| *   |  *
1b2o MKKYTCTVCVYIYNPEDGdpdngvnpGTDFKDIPDDWVCPLCAVGKDQFEEVee
     :    *  *|| :     :          :   : :: *| *   |  *
1bq8 MAKWVCKICGYIYDEDAGdpdngispGTKFEELPDDWVCPICGAPKSEFEKLed
     :    *  *|| :     :          :   : :: *| *   |  *
1rb9 MKKYVCTVCGYEYDPAEGdpdngvkpGTSFDDLPADWVCPVCGAPKSEFEAA
     :    *  *|| :     :          :   : :: *| *   |  *
1rdg MDIYVCTVCGYEYDPAKGdpdsgikpGTKFEDLPDDWACPVCGASKDAFEKQ
          *  *||                          *|   * |  *
1b71 ATKWRCRNCGYVHEGtgap.................ELCPACAHPKAHFELLginw
```

[†]The sequence alignment was generated by optimal *structure* alignment of the motif using the program CAMINE (Oldfield, forthcoming). The sequence homology value is based on exact residue typing; the 4 CYS + PHE are marked with a "*," residues that are invariant in all sequences are marked with a "|," and residues in the first 4 sequence that are invariant are marked with a ":." The uppercase letters indicate the common coordinate structure identified by structure alignment, and lowercase letters do not have common structure.

to a CYS residue and present in all six proteins. The PHE residue does not have a direct role for the metal binding, but is presumably important within an evolutionary context, so is invariant and therefore found as part of this configuration during mining. The sequence alignment resultant from an optimal structure alignment over the six proteins (CAMINE: Oldfield, forthcoming) is shown in Table VII. The data mine found a common configuration that was part of a structural motif. The SITEMINE program was used to search the protein databank for the metal 4-CYS residue site (excluding the PHE residue) defined in Table IV. Table V shows a small part of the results of searching the June 2001 PDB (15,705 files) against the 4-CYS metal binding site using the SITEMINE program. Three hundred twenty-two hits within 74 different proteins were found at two different RMSD clusters around 0.16Å^2 and 0.82Å when all atoms of the four CYS residues were used for the search. A total of 6,129 hits in 169 different proteins were found with two different RMSD clusters around 0.28Å^2 and 0.34Å^2 when the Sγ atom and Cβ atom of the CYS residues were used; 55,140 hits in 276 different proteins were found with a single mean cluster of 0.17Å when just using the Sγ atoms. The very high number of solutions, particularly when using just the Sγ atoms, is because the interaction has tetrahedral symmetry and so multiple hits occur against each site. The mean values were calculated by fitting gaussian curves to the distributions of hits using the average of the degenerate alignment solutions over each molecule within the range of 0.0 Å^2 and 1.0Å^2. The distribution of hits was bifurcated for the Cβ-Sγ and all atom searches indicating two different classes of metal binding sites with different structural motifs (data not shown).

A total of 16 other four-CYS interactions were found by the data mine calculation, the largest containing 15 members and the smallest containing eight members. A significant proportion of these configurations is associated with double disulphide interactions and an example is shown in Figure 5. This double disulphide has a parallel hydrophobic stacking interaction and is observed 15 times within the non-degenerate list of protein fragments.

Figure 6(a,b) shows the catalytic triad with an additional SER residue found within nine protein fragments used for the study: 1ppf, 1acb, 1agj, 1avw, 1ct0, 1dan, 1hpg, 1jrt, and 1au8. Figure 6(a) is the mean coordinate view and Figure 6(b) is an overlaid graphic of the nine results. The triad of three residues (HIS, ASP, SER) is found in many proteins and catalyses nucleophilic attack on a peptide bond.[13,22,28–36] In this example, a second SER residue was found to give a 4-residue configuration and this feature has been described in the literature.[16,32] Figure 7 shows distributions generated from a search of the PDB using the four-residue and three-residue catalytic site with zero weights on all atoms except the HIS ring atoms (weight = 1), the SER Oγ atom (weight = 5) and on the ASP Oδ1 atom (weight = 5). The three-residue site search result is plotted with positive occurrence and the four-residue search result is plotted with negative occurrence. It should be noted that the alignment tried both symmetry conformations of the acid oxygen atoms to minimize the residual. Both of the distributions in the graph have no zero RMSD values because the template represents a mean structure. The form of the distributions is similar to that described by Wallace et al.[16] although a different metric is used to define the RMSD. The details of the interactions within a catalytic triad are comprehensively described in Wallace et al.[16], although the mining described here provides additional information that will be presented elsewhere.

Figure 8 shows an acid-base-acid interaction with 14 examples: 1ffr, 1a70, 1q0a, 1aye, 2pgd, 1fce, 1a3a, 1qu1, 1bva, 1cru(2), 1dmr, 1bam, and 1nsy. The main chain atoms and Cβ are not shown for clarity because the interaction was very diffuse, although common. If the ANAMINE program is used with a RMSD cut-off of 2.0Å then this configuration expands to 65 members. The mean coordinate description of this configuration is probably not that useful due to the variability. In this example, the base

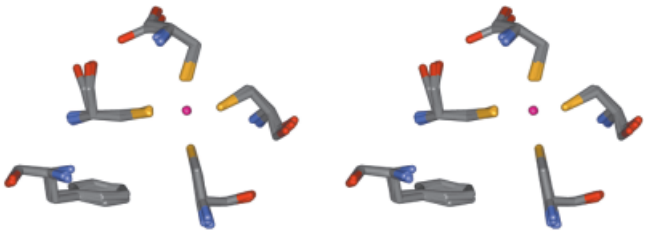

Fig. 4. Stereo graphic of a metal binding site with four-CYS residues and a PHE. Carbon atoms are shown in black, nitrogen in blue, oxygen in red, and sulfur in yellow. The central atom of the 4-sulphur set is the six superposed metal atoms.

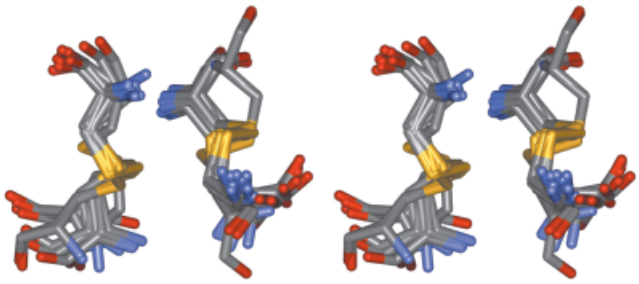

Fig. 5. Stereo figure of a double disulphide overlaid template result. Carbon atoms are shown in black, nitrogen in blue, oxygen in red, and sulfur in yellow.

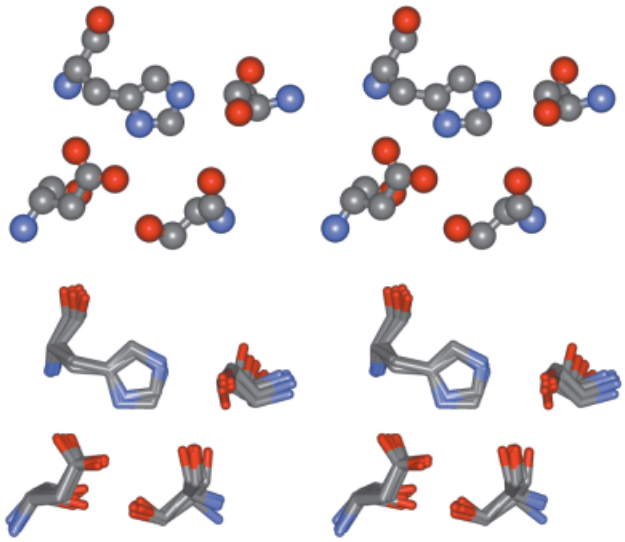

Fig. 6. Stereo figure of the four residue mine result which is a catalytic triad + helper SER found in a number of proteins. **a:** The averaged coordinates. **b:** The overlaid coordinates. Carbon atoms are shown in black, nitrogen in blue, and oxygen in red. The nucleophilic SER residue is at the top, the ASP residue is at the bottom.

was either LYS (8) or ARG (6), and the acid group was either ASP (15) or GLU (13) where the number in brackets indicates the occurrence.

Figure 9 shows a configuration of three hydrophilic residues ASN, SER/THR, and SER with associated ligands in the same frame of reference. There are 24 examples of this configuration: 1a53, 1avb, 1b1c, 1skf, 1ayx, 1cf3, 1onr, 2myr, 1ceg, 1the, 1ct0(2), 1chm, 1clx, 1be9, 1bza, 1cle, 1ql0, 1a5v, 1pme, 1clx, 1dhk, 1bue, and 2shp. The ASN and (SER/THR) pair of residues (closest to the ligands) are described in the literature as the N-linked glycosolation site.[37] This is defined on the basis of the sequence signature[38,39] (ASN)-X-(SER/THR) where a sugar is covalently attached to the ASN. The six ligands associated with the 24 configurations consisted of five copies of N-acetyl-D-glucosamine, which were covalently attached to the ASN residue and a single indole-3-glycerol phosphate that was not covalently linked to the protein 1a53. This configuration is interesting because there is a second common SER residue that is hydrogen bonded in each example to the carbonyl oxygen of glycosolation site SER. This is residue X within the sequence signature. The second result of note is that the protein 1a53 is an indol-3-glycerolphosphate synthase molecule, which as the name suggests, is an enzyme. The ligand is within the active site of this molecule,[40] which means the active site has the same configuration of three key residues as found in N-linked glycosolation site. The importance of this correlation is being investigated further.

Figure 10 shows a configuration of three aromatic rings, with 16 examples in the list of protein fragments used for this analysis: 1lar, 1wdc, 6rxn, 1mpb, 1tn4, 4icb, 1mr8, 1yge, 1a7x, 1btc, 1qd2, 1dmr, 1dm1, 1cwy, 1bq8, and 1b2o. Two of the ring sites were predominately occupied by PHE, though there was one TYR example in each. The third ring site contains examples of TRP, PHE, and TYR and was more variable in its orientation.

Template Search With SITEMINE

The template search program SITEMINE was used to carry out a number of analyses to reproduce the results found by SPASM[17] and ASSM.[13] The first was an analysis of the PDB and the non-degenerate fragment list for the triplet of acid residues (GLU212, ASP214, GLU217) taken from the protein 1CEL.[41] A second search was carried out using the zinc binding motif (HIS142, HIS146, GLU166) taken from the protein 4TMN. In both studies, the analysis was carried out using an all atom search and a key atom search (COO^-/ring). The acid residue triplet search of the PDB resulted in the return of 93 hits within an RMSD of 1.0Å and 34 within 0.5Å of the search template residues from 1CEL. The four proteins found by Kleywegt (1cel, 2ayh, 1gbg, 1mac) were all identified within the PDB. The search of the 1,774 protein fragments resulted in three hits within the proteins (1ajk, 1cel, 2ayh). Since there were only three examples of this interaction within the list of structure fragments, the data mine did not identify it as a common feature because the numerical target for mining was five occurrences. Using the key atoms within the side chain (COO^-) results in no further close hits, indicating that this site feature is fold specific and the whole residue configuration is invariant. The search for the zinc binding site (HIS142, HIS146, GLU166) taken from the protein 4TMN demonstrates the advantage a full atomic RMSD target has over a single point description of amino acids. There are a number of residue configurations that coordinate to zinc ions found in pro-

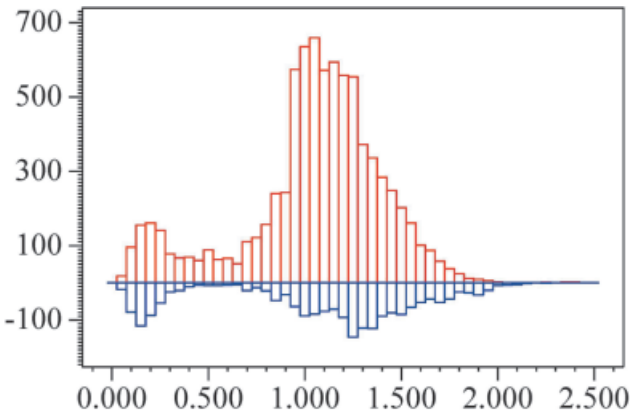

Fig. 7. The distribution of residual from the alignment of the four residue site [Fig. 6(a)] as negative occurrence values, and the catalytic triad as positive occurrence values.

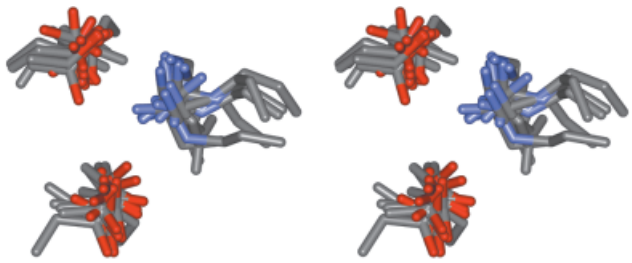

Fig. 8. Stereo figure of an acid-base-acid interaction. Carbon is shown in black, nitrogen is shown in blue, and oxygen red. The main chain atoms and Cβ are not shown for clarity.

Fig. 9. A three-residue interaction of SER SER/THR and ASN that is an N-linked glycosolation site. A number of ligands are included that associate with this three residue interaction.

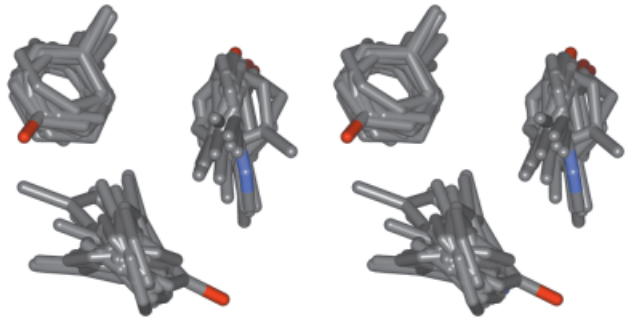

Fig. 10. A three-residue interaction of aromatic residues. Carbon is shown in black, nitrogen is shown in blue, and oxygen red. The main chain atoms are not shown for clarity.

teins. 4tmn is an example of a tetrahedral coordination with HIS 142, GLU 143, HIS 146, and a phosphate group from a modified PHE of an inhibitor. The GLU and phosphate coordination bonds are both bidendate. It is noted by Artymiuk et al.[14] that a search using single point descriptions of residues results in a number of false positives. Therefore, it was not possible to reproduce the results of these two studies using SITEMINE where the analysis is based on a multiple atom RMSD and thus very discriminatory. A search of the non-degenerate list of 1,774 protein fragments yields four hits (1bqb, 1ezm, 1lne, 2fua), while a full search of the PDB yields 29 hits, none of which represent false positives. A HIS-HIS-HIS metal tetrahedral coordination site with an additional GLU residue (conserved but not involved with metal coordination) occurs twice within the data mine template results. The first is a zinc site (e.g., 1bsw) that occurs 11 times, the second is an iron/manganese site that occur six times. The members of this different interaction were picked up by the single point search methods, but not by SITEMINE, as equivalent to that in 4tmn. The SITEMINE program is, therefore, shown to be able to discriminate between similar but unrelated residue interactions and offers a considerable advantage over single point approximate methods.

Protein Labeling With TEMPLATE

The TEMPLATE program used the list of 1,972 templates to search the protein "1ajy" for any occurrence of a configuration. Table VI shows one hit from the analysis for the protein structure that took approximately one second to complete. Table VI shows details of the alignment to the protein as well as information that was stored in the template file.

DISCUSSION

Mathematical data mining has been described and used to determine local residue interactions within proteins. Since it does not produce results using biological targets, the output cannot easily be defined as correct or not correct in a biological context. The definition of the false-positive and false-negative rate cannot be known when the complete set of local interactions expected in proteins is not known. The results are justified on the basis of the reproduction of known protein interactions, and the understanding of the effects of data noise. Noise in the data consists of any background rate of interaction within the triplet list, and this is dictated by the precision of the hash table and various run time parameters. When parameters are set so that the sensitivity is too high, then the N-body interaction list increases exponentially with "N." Data saturation can be observed when one set of parameters gives few results, but a change in a parameter to increase the sensitivity generates an explosion of configurations, though this is generally observed as termination of the mine calculation due to time and memory constraints.

Parameter Dependence

Tables VIIIa–c show the effect of changing three parameters: the peak height for geometric hash detection, the

TABLE VIII. Mine Results From SIDEMINE as a Function of Different Parameters[†]

a. Calculation statistics as a function of the peak volume

Number in peak	Number of results	Time/sec.
4	377/44/–	N/A
5*	198/31/1/6/1	8
6	98/12/2/2/1	8
7	50/71/0/1	8
8	34/5	8
9	25/3	8

b. Calculation statistics as a function of peak range value

Range	Number of results	Time/seconds
0	22/5	5
1*	198/31/1/6/1	8
2	286/63/–	N/A

c. Calculation statistics as a function of residue separation cutoff

Separation/Å	Number of results	Time/seconds
3.5	3	6
4.5	6/2	6
5.5	36/9/0/2	6
6.5*	198/31/1/6/1	8
7.5	531/64/–	N/A

[†]The parameters used for the results presented in this paper are marked with a "*," and calculations that fail to complete are marked as N/A. The "Number of Results" column gives the result numbers for 3-residue configurations/4-residue configurations/etc.

peak range value for the geometric hash detection, and the residue distance separation limit. These three parameters have the most effect on the outcome of the mine calculation. Other parameters that control the amount of output (Minimum number in a printed peak, the peak size saved between N-order analysis, suppression of printed outliers, and topological equivalence of the interaction) do not effect the propagation of the information in the same way and are not shown. Table VIIIa indicates that if the threshold of peak height is set too low; then noise is included within the results. Table VIIIa shows that a peak height of less than five results in propagation of data that is certainly the result of structural homology. Table VIIIb shows results as a function of the hash bin range value. This parameter defines the number of hash bins that are used around a solution peak, and values within these adjacent hash bins are included in subsequent analysis. Increasing the range value results in configurations with larger variance that can then merge during N-body analysis giving a large number of disparate results. Spread values greater than one prevent completion of the mine given the value of other parameters. The separation value (Table VIIIc) is used to define the distance limit between two key atoms in the analysis, and larger values produce more results. The value of 6.5Å was used as greater separation limits include too many distant interactions and prevents the completion of a mine. A successful mine can, therefore, be defined as an analysis that completes but where any increased parameter sensitivity prevents normal termination of the program.

Systematic error analysis, therefore, represents an important part of the process of data mining. It is necessary to run a mine a number of times to judge where a parameter has a critical value for a data set. Parameters must be optimized to improve the signal detection without losing all useful information or producing an explosion of the analysis. Since the data mining is usually complete within seconds, this not difficult as any calculation taking more than 1 minute is likely to be due to a run-away correlation indicative of systematic error.

Additional Parameters for Mining

There are a number of additional features of the SIDEMINE program that aid the analysis of specific questions by data screening. A residue sequence restraint can be used that prevents output of configurations where residues are also neighbors in sequence. Results that have a high correlation of spatial and sequential information are interesting and appeared to form about half of the mine results when the restraint was not used. A limitation that requires more than a certain percentage of the solution members to have a different sequence topology can be used to screen out configurations of residues that result from a common local protein motif. This run time parameter defines the cut-off percentage of residue configurations allowed with a particular numerical variance in sequence separation. It is possible to supply residue type and distance restraints to the mine analysis. In this way, a particular interaction type can be studied at parameter settings that would normally swamp results with background noise. In particular, this has aided in the analysis of those residues that can act as a nucleophile within the catalytic triad using a multiple equivalence statement (forthcoming). The residues that form the triplet data can be sorted by the order they appear in the PDB file, or they can be sorted by the numerical order of the residue type in Table I. If the sorting is by the sequential order within PDB file, then configuration HIS-ASP-SER is not equivalent to SER-ASP-HIS and will be returned as two separate mine results. Sorting by residue type order will the remove the sequence dependence, and so all different sequence orders will be returned as a single mine solution improving the S/N. Sorting by residue type increases the background noise as the results are compressed into a smaller number of valid residue combinations. The two sorting options can, therefore, be seen as complimentary because some interesting configurations may or may not be sequence dependent. It can be seen that the SIDEMINE analysis time is small in comparison with the triplet read time (Table II). It is, therefore, possible to load and leave resident within computer memory the triplet list. Different parameter values can then be tried to obtain results very quickly. Finally, the initial triplet generation renames the crystallographic waters to a single residue name, allowing the analysis of water structure about proteins.

The SIDEMINE program has been optimized so that it is possible to run data mine calculations on typical departmental computers. The only requirement is a large amount of real memory with one Gbtye necessary for the calcula-

TABLE IX. Calculation Time for the Program SIDEMINE on Four Different Computers Using the Same Parameters[†]

Computer	Processor	Real memory/Gb	Time/s
Compaq Alpha	ES40	4	8
Compaq Alpha	DS10	0.5	805
SGI origin	R10000	1.2	110
PC (linux)	P*III*450	1.0	105

[†]Identical mine results were obtained in each case.

tions described. In the near future, it is likely that many desktop PCs will have this amount of memory as standard. Table IX shows the calculation time on four different computers and indicates that the program is certainly memory access time dependent. The Compaq DS10 has a real memory smaller than the data size and hence the long calculation time due to page swapping.

It can be theorized that allowing the program to identify configurations that propagate between a set of structurally homologous proteins would be an ideal method of generating optimal structure alignments. Although this appears to be a useful feature, in fact the SIDEMINE program starts alignment at every residue, and N-body combination results in an exponential increase of (N+1) interactions. Even the alignment of two small proteins such as myoglobin species variants cannot be completed due to computational resource limits. This explains why the algorithm described appears ideal for fold analysis, but cannot be used for this without modification.

Review of Results

The aim of protocols and methods discussed here was to remove human intervention within the data preparation, target description, and much of the analysis. This could not produce ideal results, but it generated information that could not be obtained by a more interventionist method. The starting set of sequentially dissimilar protein fragments was generated using a number of rules. The selection of proteins to determine significant local structure was a problem. How do you remove structural equivalence that is not interesting while retaining structural equivalence that is interesting? In particular, selection by sequence dissimilarity did not exclude the common fold motif of the electron transport proteins that make up the configuration in Table IV. The six proteins in this configuration have high structure similarity even though they have sequence homology less the 80% (Table VII). An all-pairs structural alignment of the protein fragments[25] can provide additional information as to whether the selected data is non-degenerate by overall structure, but it is a problem to select data with the same metric (structure alignment RMSD) as used to determine the results.

The use of the data mine program SIDEMINE was not simple and the results obtained represent a starting point of further characterization. This article represents a description of an analytical method and results to justify the implementation. The main development issue was that of improving and characterizing the sensitivity of the analysis. It was only possible using data mining to observe residue interactions that occur more often than a background rate. Careful selection of data was also important. Even so, it was found that the general analysis of all aspects of interactions within proteins could not be studied in one single calculation. For example, a study of water structure and its interaction with proteins cannot be performed with the same parameter sensitivity. It should also be noted that the PDB currently contains approximately 15,000 proteins, of which about 2,000 have unique folds. Any results generated from this small subset of possible proteins may not be representative of protein structure space.

The SITEMINE program offers a considerable advantage over the methods that use single point approximations to amino acids as it does not produce false positives. In fact, it offers a powerful discrimination between different variants of metal binding sites and so represents a useful tool for protein characterization. It is not as rapid, taking 111 seconds to analyze the 1,774 non-degenerate proteins, but at 16 proteins per second it does represent a practical solution for database analysis.

SUMMARY

Four different programs have been described in this article, all with common methods of handling a number of issues associated with data mining and template searching the protein databank. A method of mathematical analysis of 3D data has been presented that minimizes prior expectation in the generation of common configurations of amino acid atoms. Algorithmic details of handling atomic weights and instability within the least square routine itself are addressed within the programs. The issue of internal chemical symmetry within five amino acids has been handled, as well as the ability to match different amino acids against each other. Calculation speed has been optimized in all the programs enabling the analysis of all sequence dissimilar protein fragments within the PDB now and for the near future. A number of ongoing lines of research are in progress to analyze the results using different parameters in the calculations. These include the analysis of protein solvation and patterns of hydrophobic residues in proteins. The programs can also be used to provide details of specific biological questions such as the many facets of the catalytic triad interaction. Its main use though is to study new features of interest within proteins that cannot be identified by amino acid template searching and, therefore, provides starting points for new lines of research. It also provides results that are un-biased representative configurations of known residue interactions. Subsequent statistical analysis with these template configurations results in clarity of information. The number of configurations of residues generated by mining is huge and these provide the starting point for detailed analysis rather than a definitive solution to the determination of active sites and binding sites.

The programs SQUID and PDBSELECT are available free of charge to academic users from http://www.ysbl.york.ac.uk/~oldfield. The programs are available from Accelrys as part of the Discovery Studio protein analysis

suite of tools. Information is available from the URL "http://www.accelrys.com/contact/" or by E-mail: "solutions@accelrys.com" or Telephone: +1 858 458 9990. The results described in this study can be found at the URL: http://www.ysbl.york.ac.uk/~tom

ACKNOWLEDGMENTS

I thank James Todd for writing the web pages for the database and Barry Grant and Ana Rodrigues for discussions associated with the mining work.

REFERENCES

1. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archive for macromolecular structures. J Mol Biol 1977;112:535–542.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
3. Abola EE, Sussman JL, Prilusky J, & Manning NO. Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol 1997;277:556–571.
4. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
5. Hubbard TJP, Ailey B, Brenner SE, Murzin AG Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 1999;27:254–256.
6. Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res 2000;28:257–259.
7. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton, JM. CATH: a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
8. Salem GM, Hutchinson EG, Orengo CA Thornton JM. Correlation of observed fold frequency with the occurrence of local structural motifs. J Mol Biol 1999;287:969–981.
9. Bray JE, Todd AE, Pearl FMG, Thornton JM. Orengo CA. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Prot Eng 2000;13:153–165.
10. Kinoshita K, Mizuguchi K, Go, N. Classification of protein 3D structures. Prog Biophys Mol Biol 1996;65:203–203 .
11. Sanchez R, Sali A. MODBASE: A database of comparative protein structure models. Bioinformatics 1999;15:1060–1061.
12. Russel RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
13. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW. Willett P. A graph-theoretic approach to the identification of 3-dimensional patterns of amino-acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
14. Artymiuk PJ, Poirrette AR, Rice DW. Willett P. Comparison of protein folds and sidechain clusters using algorithms from graph theory. In: Mailey S, Hubbard R, Waller D. editor. From first map to final model. SERC Daresbury Laboratory, Daresbury, UK 1995 p 71–81.
15. Grindley HM, Artymiuk PJ, Rice DW, Willett P. Identification of tertiary structure resemblance in protein using a maximal common subgraph isomorphism algorithm. J Mol Biol 1993;229:707–721.
16. Wallace AC, Borkakoti N, Thornton JM. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
17. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;265:1887–1897.
18. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
19. Russell RB, Saqi MAS, Bates PA, Sayle RA, Sternberg MJE. Recognition of analogous and homologous protein folds: assessment of prediction success and associated alignment accuracy using empirical substitution matrices. Protein Eng 1998;11:1–9.
20. Glusker JP. Structural aspects of metal liganding to functional groups in proteins. Adv Protein Chem 1991;42:1–76.
21. Jernigan R, Raghunathan G, Barhar I. Characterization of interactions and metal-ion binding sites in proteins. Curr Opin Struct Biol 1994;4:256–263.
22. Fischer D, Wolfson H, Shuo LL, Nussinov R. Three-dimensional, sequence order-independent comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding. Protein Sci 1994;3:769–778.
23. Chakrabarit P. Anion binding sites in protein structures J Mol Biol 1993;234:463–482.
24. Copley RR, Barton GJ. A structural analysis of phosphate and sulphate finding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. J Mol Biol 1994;242:321–329.
25. Oldfield TJ. Creating structure features by data mining the PDB to use as Molecular Replacement models. Acta Cryst. 2001;57:1421–1427.
26. Oldfield TJ. SQUID : A program for the analysis and display of data from crystallography and molecular dynamics J Mol Graphics 1992;10:247–252.
27. Kabsch W. A solution for the best rotation to retate two sets of vectors. Acta Cryst 1978;A32:922–923.
28. Blow DM. Structure and mechanism of chymotrypsin. Acc Chem Res 1976;9:145–152
29. Blow DM. More of the catalytic triad. Nature 1990;221:337–340.
30. Perona J, Craik C. Structural basis of substrate specificity in the serine protease. Protein Sci 1995;4:337–360.
31. Barth A, Wahab M, Brandt W, Frost K. Classification of serine protease derived from steric comparisons of their active sites. Drug Design Discovery 1993;10:535–542.
32. Barth A, Frost K, Wahab M, Schlader HD. Classification of serine protease derived from steric comparisons of their active site geometry, part II. Drug Design Discovery 1994;12:89–111.
33. Brady L, Brzozowski AM, Derewenda ZS, Dodson E, Dodson G. Tolley S, Turkenberg JP, Christianson L., Huge Jenson B, Norskov L, Thrim L, Menge U. A serine protease triad forms the catalytic centre of tri-acylglycerol liplase. Nature 1990;343:767–770.
34. Brzozowski AM, Derewenda U, Derewenda ZS, Dodson GG, Lawson DM, Turkenberg JP, Bjorkling F, Huge Jenson B, Patkar SA, Thim L. A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. Nature. 1991;351:491–497.
35. Corey DR, McGrath ME, Vasquez JR, Fletterick RJ, Craik CS. An alternative geometry for the catalytic triad of serine proteases. J Am Chem Soc 1992;114:4905–4907.
36. Wallace AC, Laskowski RA, Thornton JM. Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His_Asp catalytic triads in the serine proteinase and lipases. Protein Sci 1996;5:1001–1013.
37. Imberty A, Perez S. Stereochemistry of the n-glycosylation sites in glycoproteins. Protein Eng 1995;8:699–709.
38. Bairoch A. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 1992;:2013–2018.
39. Attwood TK, Beck ME, Bleasby AJ, Parrysmith DJ. PRINTS: a database of protein motif fingerprints. Nucleic Acids Res 1994;22:3590–3596.
40. Hennig M, Darimont B, Sterner R, Kirschner K, Jansonius JN. 2.0 angstrom structure of indole-3-glycerol phosphate synthase from the hyperthermophile Sulfolobus solfataricus: Possible determinants of protein stability. Structure 1995;3:1295–1306.
41. Divne C, Stahlberg J, Reinikainen T, Ruohonen L, Pettersson G, Knowles JKC, Teeri TT, Jones TA. The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma resser. Science 1994;265:524–528.

APPENDIX A

Proteins for this analysis were selected by the program PDBSELECT (http://www.ysbl.york.ac.uk/~oldfield) on the basis of nine rejection criteria.

1. The field $NMR should be declared within a PDB file that is solved by NMR, but in most cases this is not

	Rejection criterion	Limit
1	NMR structures	$NMR/MOL
2	Date limit	After 1983
3	DNA + RNA structures	$DNA/$RNA
4	Too few residues	<10 amino acids
5	Too many Cα	>0.25 * Natom
6	UNK sequence entries	
7	Bad Ramachandran	>10% OUB
8	Resolution limit/No info	>2.5/no data
9	Sequence rejection	>80%

present. Therefore, a NMR structure was also detected by the presence of the MOL and ENDMDL cards used to delimit the multiple solutions within such files. These files were rejected as they have different experimental variance and so would effect the mathematical targets.

2. Proteins solved after 1983 were used as these were solved by restrained least squares or maximum likelihood methods not available before this date.
3. Structures that contains only nucleic acids should contain $RNA or $DNA fields. Where this field was not present then the structures were rejected by rule 4.
4. Small structure size. Small peptides are unlikely to contain compact residue interactions
5. The Cα atom proportion limit was used to detect proteins where only the Cα or backbone atoms were present. Structures that do not contain side chain atoms would not contribute to the analysis.
6. Proteins were rejected where UNK sequence entries were present.
7. Geometrical structure quality was defined by the number of non-glycine outlier/torsion angles with respect to a Ramachandran energy surface calculated by CHARMm 22.
8. The resolution limit was used to define an experimental quality metric; structures where no resolution limit was provided were also rejected.
9. Sequence similarity pruning of the PDB was based on an 80% exact optimal sequence alignment (Oldfield, unpublished algorithm). If a pair of proteins align in sequence above the 80% threshold, and the difference in the number of residues within the two proteins deviated less than 10%, then the latest deposited structure with the best resolution that was not a mutant structure was selected. If the pair of proteins differ in length by more than 10%, then the largest structure was retained.

APPENDIX B

List of 1,301 proteins selected on the basis of sequence homology, geometry quality, resolution, submission date, and submission type. The proteins were split into 1,774 fragments with more than 20% difference in sequence.

104m	107l	11bg	134l	13gs	154l	16pk	19hc	1a00	1a05	1a0g	1a0j	1a12	1a1j
1a28	1a2j	1a2z	1a30	1a34	1a3a	1a3b	1a3c	1a3d	1a3h	1a44	1a48	1a4i	1a4m
1a4u	1a4v	1a4y	1a53	1a54	1a58	1a5m	1a5v	1a62	1a68	1a6q	1a6v	1a70	1a73
1a75	1a77	1a78	1a7c	1a7d	1a7s	1a7u	1a7w	1a7x	1a7y	1a88	1a8h	1a81	1a8o
1a8p	1a8q	1a8s	1a8v	1a92	1a96	1a9s	1aa2	1aac	1aal	1aap	1aaz	1ab1	1abf
1abo	1acb	1acf	1ad2	1ade	1ado	1ae2	1ae7	1ae9	1aec	1aew	1af7	1afb	1afw
1ag6	1ag9	1agi	1agj	1agq	1ah7	1aho	1ahx	1ai3	1aik	1ail	1aim	1aiu	1aiz
1aj0	1aj8	1ajk	1ak0	1ak2	1akd	1ake	1ako	1akr	1aky	1akz	1al3	1alo	1alu
1alv	1aly	1aml	1amf	1amk	1amp	1amu	1amx	1aoc	1aoe	1aoh	1aok	1aol	1aom
1aoz	1ap6	1aph	1apm	1apt	1apy	1aq0	1aq6	1aqm	1aqy	1aqz	1ar5	1arb	1arp
1as7	1ast	1at0	1atg	1atz	1au8	1aug	1aun	1auo	1ava	1avb	1avw	1avy	1aw7
1awp	1awr	1axn	1axy	1ay7	1aye	1ayf	1ayi	1ayl	1ayo	1ayx	1az5	1az9	1azo
1azq	1b00	1b0b	1b0n	1b0u	1b0w	1b0x	1b0y	1b12	1b1c	1b25	1b2k	1b2n	1b2o
1b2p	1b2r	1b2v	1b31	1b3a	1b31	1b3m	1b43	1b4e	1b4f	1b4k	1b4l	1b57	1b59
1b5e	1b5f	1b5q	1b63	1b65	1b66	1b67	1b6q	1b6t	1b71	1b7d	1b7v	1b80	1b8a
1b8c	1b8d	1b8j	1b8p	1b8z	1b93	1b94	1b9d	1b9h	1b9k	1b9m	1b9w	1ba0	1bam
1bav	1baz	1bbl	1bbh	1bbp	1bbz	1bc2	1bc8	1bck	1bcx	1bd3	1bd8	1bdb	1bdm
1bdo	1be0	1be9	1bea	1beb	1bec	1ben	1bf2	1bf6	1bfd	1bfg	1bft	1bg0	1bg6
1bg7	1bgc	1bgf	1bgp	1bgv	1bhd	1bhe	1bhh	1bhp	1bht	1bif	1bio	1bir	1bis
1bit	1bj7	1bjw	1bk1	1bk7	1bk9	1bkb	1bkc	1bkj	1bkp	1bkz	1b13	1bli	1blx
1blz	1bm7	1bm8	1bm9	1bma	1bmb	1bn7	1bn8	1bnf	1bol	1bpi	1bpq	1bq8	1bqb
1bqc	1bqk	1bqk	1bqu	1brs	1bs0	1bs4	1bs9	1bsg	1bsl	1bsw	1bt0	1bt5	1btc
1bte	1btk	1btn	1bto	1bu7	1bu8	1bue	1buo	1bur	1buu	1bv1	1bv4	1bva	1bvq
1bvw	1bw9	1bx0	1bx4	1bx7	1bxe	1bxk	1bxt	1bxu	1bxy	1by2	1by7	1byf	1byo
1byp	1byr	1byz	1bz4	1bza	1bzs	1bzy	1c02	1c1k	1c1l	1c1y	1c2a	1c2p	1c39
1c3c	1c3d	1c3j	1c3k	1c3p	1c3q	1c3w	1c52	1c5e	1c6o	1c8x	1c8z	1c97	1c9o
1c9s	1cal	1cai	1cb0	1cb6	1cb7	1cb8	1cbj	1cbs	1cc3	1cc7	1ccr	1ccz	1cdc
1cdm	1cdw	1cdy	1ce9	1ceg	1cel	1cem	1ceo	1ceq	1cew	1cf3	1cf9	1cfb	1cfm
1cfw	1cg5	1cgk	1cgo	1cgt	1chd	1chh	1chm	1chp	1ci1	1ci3	1cjb	1cjc	1cjd
1cjw	1cka	1ckq	1cl1	1cl6	1clc	1cle	1clw	1clx	1cm2	1cm4	1cmc	1cns	1cnv
1cnz	1co6	1coj	1cot	1coz	1cp2	1cp7	1cpo	1cpq	1cpu	1cq3	1cq4	1cqq	1cqx
1cqy	1crm	1cru	1cs1	1cs6	1cs8	1csn	1ct0	1ct5	1ct9	1ctf	1ctj	1ctq	1cua

APPENDIX B. (Continued)

1cuk	1cun	1cv8	1cvl	1cvr	1cw2	1cwy	1cxc	1cxp	1cxy	1cy5	1cy9	1cyd	1cyi
1cyn	1cyo	1cz1	1cza	1czf	1czk	1czq	1czs	1czy	1d02	1d06	1d0b	1d01	1d0q
1d0v	1d0z	1d1q	1d1z	1d2e	1d2i	1d2m	1d2n	1d2s	1d2z	1d3b	1d3v	1d3y	1d4a
1d4o	1d4t	1d4w	1d5n	1d5z	1d6j	1d7e	1d7f	1d7o	1d7p	1d7u	1d8c	1d8d	1d8h
1d9b	1d9c	1dae	1dan	1dbl	1dbf	1dbg	1dbi	1dbu	1dce	1dci	1dck	1dcs	1dd9
1ddt	1ddv	1dek	1deo	1deu	1df4	1df7	1dfa	1dfn	1dfu	1dfx	1dg3	1dg6	1dg9
1dgh	1dgy	1dhj	1dhk	1dhn	1di2	1di6	1din	1diw	1dj0	1dj7	1dj8	1dja	1djo
1dki	1dkx	1dl2	1dlg	1dlj	1dlt	1dm1	1dm5	1dm9	1dmm	1dmr	1dmw	1dnl	1do6
1dob	1doi	1dok	1dor	1doz	1dp7	1dpe	1dpf	1dpg	1dpj	1dpo	1dps	1dpt	1dqa
1dqs	1dth	1dun	1dup	1dut	1dvf	1dw9	1dxy	1eay	1ecf	1ecp	1edg	1edh	1edm
1eer	1egp	1eif	1elp	1elt	1emg	1enx	1epn	1esf	1esl	1eso	1euh	1eus	1evh
1ext	1ezm	1f3z	1fas	1fbn	1fce	1fdd	1fdn	1fdr	1fds	1feh	1fen	1fgk	1fia
1fil	1fit	1fjl	1fld	1flm	1flt	1fmb	1fmc	1fmt	1fna	1fon	1frd	1frp	1frr
1ftr	1fuq	1fus	1fxd	1g3p	1gai	1gar	1gca	1gce	1gcm	1gdl	1gdi	1gdo	1ger
1gfl	1gic	1gky	1gnd	1gnk	1gog	1got	1gox	1gp1	1gpe	1gpr	1gr2	1grg	1gsa
1gse	1gso	1gsu	1guq	1gux	1h2r	1han	1hav	1hbg	1hck	1hcr	1hcv	1hcz	1hfc
1hfe	1hgx	1hil	1hka	1hmd	1hmk	1hms	1hna	1hoe	1hpg	1hpi	1hrc	1hsb	1hsh
1hsl	1htr	1hur	1huu	1huw	1hxn	1hy1	1hyp	1iak	1ibe	1icf	1icm	1idk	1ids
1ift	1igd	1ihb	1iib	1ilk	1iow	1ir3	1isa	1isu	1iuz	1jdd	1jdw	1jer	1jfr
1jhg	1jkm	1jlm	1jpc	1jrt	1jug	1kao	1kce	1kdb	1kdi	1kdn	1kel	1kid	1klt
1koe	1kop	1kp6	1kpa	1kpt	1krn	1kuh	1kvc	1kvd	1kvo	1kwa	1lam	1lar	1lat
1lbu	1lci	1lcj	1lcl	1lfa	1lhs	1lid	1lis	1lit	1lkf	1lki	1lld	1llo	1lmb
1lml	1lne	1loc	1loe	1lop	1lou	1lst	1lt5	1lts	1lve	1mai	1mbl	1mdc	1meg
1mem	1mfa	1mff	1mfm	1mgt	1mhl	1mho	1mhy	1mjc	1mjh	1mjw	1mka	1mla	1mld
1mml	1mmq	1mns	1mof	1mol	1moq	1mpb	1mpg	1mpp	1mr8	1mro	1mrp	1msc	1msk
1mty	1muc	1mud	1mug	1mwe	1mwp	1myt	1mzl	1nar	1nba	1nbc	1ncg	1nci	1nco
1nec	1neu	1nfp	1nhk	1nhq	1nic	1nkr	1nlr	1nox	1np4	1nps	1nsd	1nsj	1nsy
1ntn	1nue	1nul	1nwo	1nzy	1oaa	1oac	1obp	1onc	1one	1onr	1opa	1opc	1ops
1orc	1osp	1otf	1oth	1ova	1oyc	1pbv	1pbw	1pcf	1pch	1pda	1pdo	1pfz	1phm
1phn	1php	1pht	1pi2	1pii	1pin	1pjc	1plc	1pme	1pmi	1pmy	1pnf	1pnk	1poc
1pot	1ppa	1ppd	1ppf	1ppr	1prx	1psc	1psr	1psz	1ptf	1ptq	1pty	1puc	1pud
1pva	1pvb	1pym	1qaa	1qac	1qad	1qau	1qav	1qaz	1qb0	1qb7	1qba	1qbz	1qc5
1qci	1qco	1qcs	1qcz	1qd1	1qd2	1qd9	1qde	1qe3	1qf8	1qfl	1qfo	1qfs	1qft
1qg6	1qgi	1qgj	1qgs	1qgv	1qgw	1qgx	1qh3	1qh4	1qh6	1qh8	1qhf	1qhi	1qi7
1qin	1qiz	1qj4	1qj5	1qj8	1qja	1qjd	1qk3	1qk5	1qk8	1q10	1q13	1qlm	1qlp
1qlw	1qm5	1qmg	1qmp	1qnf	1qnn	1qnt	1qo7	1qoa	1qoi	1qou	1qow	1qoy	1qpe
1qqp	1qqy	1qr0	1qrg	1qrp	1qrr	1qrz	1qs1	1qsa	1qsr	1qto	1qtw	1qu1	1qu9
1qup	1r69	1rb9	1rcd	1rcy	1rdg	1rds	1rec	1reg	1req	1rfs	1rgp	1rhs	1rie
1rkd	1rl6	1rmg	1rnl	1rpj	1rro	1rss	1rst	1rsy	1rtp	1rtu	1sac	1sat	1sbp
1sct	1sei	1sel	1sem	1sfp	1sft	1sgt	1shf	1shk	1skf	1sky	1skz	1slg	1slm
1slt	1slw	1sml	1smr	1snl	1snb	1sox	1spb	1spg	1sph	1sra	1srd	1srr	1srv
1st3	1stm	1sur	1svb	1svf	1svp	1svy	1swf	1szj	1taf	1tcl	1tca	1ten	1tf4
1tfa	1tfe	1tgj	1tgx	1the	1thm	1thv	1thx	1tib	1tif	1tig	1tki	1t12	1tml
1tmy	1tn3	1tn4	1toa	1tol	1ton	1tph	1trb	1trk	1try	1tud	1tux	1tvd	1tvx
1tx4	1u9a	1uae	1ubi	1uch	1uda	1udh	1ukz	1uok	1uox	1upl	1urn	1uro	1ush
1ute	1utg	1uxy	1v39	1vap	1vcc	1vfr	1vfy	1vhb	1vhh	1vid	1vif	1vin	1vjw
1vls	1vns	1vpe	1vpn	1vpp	1vsc	1vsr	1wab	1wad	1wba	1wdc	1wdn	1wej	1wer
1wfa	1wgj	1wgt	1whi	1who	1wht	1wpo	1wwc	1xer	1xgs	1xid	1xik	1xyf	1xyn
1xyz	1yac	1yag	1yai	1yal	1yge	1ygh	1yjc	1yna	1ypb	1zfj	1zin	1zpd	1zrm
256b	2a0b	2aac	2abk	2acr	2acy	2ae2	2afg	2ahj	2ak3	2aop	2apr	2aps	2atj
2ayh	2bam	2bbk	2bce	2bmi	2bop	2bos	2c2c	2cav	2cb5	2cbp	2ccy	2cd0	2cd2
2cdv	2ci2	2clr	2cmd	2cmk	2cpg	2cts	2cua	2cy3	2dhq	2dnj	2dpm	2dtr	2e2c
2ebn	2ebo	2end	2era	2erl	2exo	2fcb	2fcr	2fha	2fiv	2fua	2gnk	2gpn	2gst
2hbi	2hdd	2hft	2hrv	2hts	2ilb	2imn	2jia	2kin	2knt	2lbd	2lhb	2mcg	2mcm
2mlt	2myr	2nad	2nap	2nll	2oat	2pgd	2pia	2pii	2plc	2por	2prd	2prk	2psp
2ptd	2pth	2pvi	2rhe	2rmc	2rsp	2sak	2scp	2sfa	2shp	2sim	2sn3	2spc	2sqc
2src	2tdt	2tgi	2tir	2tnf	2tod	2tps	3bdp	3cla	3cms	3cox	3csu	3cyr	3eip
3eng	3era	3ert	3ezm	3gcb	3gct	3had	3il8	3kvt	3lyn	3mat	3mds	3nul	3pcc
3prn	3pro	3rab	3rp2	3rub	3seb	3sli	3tgi	3tgl	3thi	3tmk	3vub	4aig	4bcl
4gcr	4icb	4lip	4mt2	4nse	4tmk	4tsv	4uag	4ubp	5csm	5eug	5hpg	5pal	5rub
5ukd	6ldh	6paz	6rlx	6rxn	6std	7ahl	7odc	7taa	830c	8dfr	9gaf	9ldt	

APPENDIX C

The algorithm for least square was adapted from that of Kabsch[27] so as to use a 4 × 4 cross term definition for overlay analysis. The following defines these cross terms and the subsequent rotation matrix extraction. This implementation results in an algorithm much more stable to ill-defined conditions, such a planar atom sets, or superposition with a small number of atoms.

$$T_{i,j} = \sum_{i=1,3;\, j=1,3;\, n=1,N} (w_n * R_{n,i}) * (w_n * W_{n,j})$$

$$V_{00} = T_{00} - T_{11} - T_{22} \qquad V_{11} = T_{11} - T_{00} - T_{22}$$
$$V_{22} = T_{22} - T_{00} - T_{11} \qquad V_{33} = T_{00} + T_{11} + T_{22}$$
$$V_{10} = T_{10} + T_{01} \qquad V_{20} = T_{20} + T_{02}$$
$$V_{30} = T_{21} - T_{12}$$
$$V_{21} = T_{21} + T_{12} \qquad V_{31} = T_{02} - T_{20}$$
$$V_{32} = T_{10} - T_{01} \qquad V_{ij} = V_{ji}: \quad \text{For } i < j$$

VV = Diagonalise (V)

$$R_{00} = VV_{00} - VV_{11} - VV_{22} + VV_{33}$$
$$R_{11} = -VV_{00} + VV_{11} - VV_{22} + VV_{33}$$
$$R_{22} = -VV_{00} - VV_{11} + VV_{22} + VV_{33}$$
$$R_{10} = 2*(VV_{10} + VV_{32}) \qquad R_{01} = 2*(VV_{10} - VV_{32})$$
$$R_{20} = 2*(VV_{20} - VV_{31}) \qquad R_{02} = 2*(VV_{20} + VV_{31})$$
$$R_{21} = 2*(VV_{21} + VV_{30}) \qquad R_{12} = 2*(VV_{21} - VV_{30})$$

T = Tensor of atomic coordinates
R = Reference atoms
W = Working set of atoms
N = number of overlay atoms.
V = cross term vector matrix
R = rotation matrix.
w = weight for atom n.