Article No. jmbi.1998.2393 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 285, 1887–1897

JMB

Recognition of Spatial Motifs in Protein Structures

Gerard J. Kleywegt

Department of Molecular Biology, Uppsala University Biomedical Centre, Box 590 SE-751 24 Uppsala, Sweden

As the structural database continues to expand, new methods are required to analyse and compare protein structures. Whereas the recognition, comparison, and classification of folds is now more or less a solved problem, tools for the study of constellations of small numbers of residues are few and far between. In this paper, two programs are described for the analysis of spatial motifs in protein structures. The first, SPASM, can be used to find the occurrence of a motif consisting of arbitrary main-chain and/or side-chains in a database of protein structures. The program also has a unique capability to carry out ''fuzzy pattern matching'' with relaxed requirements on the types of some or all of the matching residues. The second program, RIGOR, scans a single protein structure for the occurrence of any of a set of pre-defined motifs from a database. In one application, spatial motif recognition combined with profile analysis enabled the assignment of the structural and functional class of an uncharacterised hypothetical protein in the sequence database. In another application, the occurrence of short left-handed helical segments in protein structures was investigated, and such segments were found to be fairly common. Potential applications of the techniques presented here lie in the analysis of (newly determined) structures, in comparative structural analysis, in the design and engineering of novel functional sites, and in the prediction of structure and function of uncharacterised proteins.

Keywords: alanine racemase; lipid-binding protein; pattern recognition; protein structure; spatial motif

Introduction

In recent years, the number of experimentally determined three-dimensional (3D) protein structures deposited at the Protein Data Bank (PDB) has grown near-exponentially (Abola *et al.*, 1997). There is now such an overwhelming abundance of structural data, that new methods to employ and retrieve that data are needed (Thornton & Gardner, 1989). Such methods are required to aid structural biologists who, perhaps ten years ago, were able to memorise most of the relevant details of most of the previously solved structures. In addition, structural information is used increasingly by scientists from fields such as molecular biology, genetics, and medicinal chemistry, who are often unfamiliar with the intricacies of structural biology, and whose focus may well be at the level of individual residues rather than that of the fold.

The past few years have witnessed the development of ever more sophisticated methods for superpositioning and comparison of pairs or sets of protein structures, followed by methods for fold recognition and comparison (Holm & Sander, 1994), and finally by methods for fold classification, e.g., SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997). More recently, attention has slowly become focussed on structural similarities at a much ''lower'' level than that of the (domain) fold, namely on constellations of a limited number of residues (main-chain and/or side-chain atoms).

Abbreviations used: 2D, two-dimensional; 3D, three-dimensional; CBH, cellobiohydrolase; CRABP, cellular retinoic-acid-binding protein; NMR, Nuclear Magnetic Resonance; PDB, Protein Data Bank; RMSD, root-mean-square distance; SPASM, Spatial Arrangements of Side-chains and Main-chain.

E-mail address of the corresponding author: gerard@xray.bmc.uu.se

0022-2836/99/041887–11 $30.00/0

Whereas the overall fold is important as an enabling framework upon which protein function rests, the actual functional ''work'' that proteins do is usually carried out by a relatively small number of residues. Indeed, whereas many positions in a protein sequence are fairly tolerant to a variety of mutations, some are not, and it is usually these few exceptional residues that are crucial for protein function. Examples include enzyme active sites, metal-binding sites, anion-binding sites, and ligand-binding sites (where a ligand may be a substrate, substrate analogue, product, co-factor, or inhibitor).

Many specific types of site or motif have been studied in detail, for instance metal-binding sites (Glusker, 1991), the catalytic triad of the serine proteases (Fischer *et al*., 1994; Wallace *et al*., 1996), and binding sites for anions such as sulphate and phosphate (Chakrabarti, 1993; Copley & Barton, 1994). However, only a few methods have been published that enable more general investigations into small motifs in protein structures. The history of the computational problem of automatic matching of 3D patterns (sometimes called ''pharmacophoric pattern matching'', although this indicates too narrow a scope) goes back at least to the work by Lesk (1979). In the case of ''small molecules'', the problem is well known and has essentially been solved (Willett, 1987; Brint *et al*., 1988). However, macromolecules and macromolecular databases have been less amenable to similar analyses. A few years ago, Artymiuk *et al*. (1994, 1995) described a program called ASSAM which could be used to query the PDB using motifs consisting of side-chains of amino acid residues. Internally, a motif was represented by the distance matrix between pseudo-atoms (1, 2, or 3 per side-chain), and database proteins were represented similarly, which enabled the use of subgraph-isomorphism algorithms to find matches. The program appeared to function very well, but does not seem to have found wide application in the structural biology community. Wallace *et al*. (1997) described a geometric hashing algorithm, implemented in a computer program called TESS, that can be used to derive 3D co-ordinate templates for motifs. These templates can subsequently be used in a separate program to scan the structural database to find other occurrences of the motif, in essence providing a 3D counterpart to PROSITE searches that use sequence patterns (Bairoch & Bucher, 1994). More recently, Russell (1998) reported a method to detect such motifs automatically by pairwise comparison of protein structures. The method identifies cliques of side-chains (of residues with at least one polar atom in their side-chain) that form a similar spatial arrangement in both structures. In an all-against-all comparison of representative structures from the SCOP database, many motifs were identified that appear to be of functional significance.

In general, there are two extreme types of spatial motif recognition techniques: in one a motif is compared with a database of structures; in the other a database of motifs is scanned against a structure. Here, techniques are described to accomplish both tasks. SPASM (SPatial Arrangements of Side-chains and Main-chain) is a computer program that can be used to find matches in the structural database for any user-defined motif. RIGOR, on the other hand, is a program that can compare a database of predefined motifs against a perhaps newly determined structure. Major goals in the development of both programs have been to make the methods fast enough to enable interactive queries, to allow searches for motifs composed of arbitrary constellations of main-chain and/or side-chain atoms, to use intuitive input, and to provide interfaces to several other programs (e.g. for visualisation of matches, for least-squares superpositioning of matches to detect possible global similarities, and for profile analysis to detect other proteins in sequence databases that may be related in structure and/or function). In addition, a unique facility that allows for fuzzy pattern matching (*vide infra*) has been implemented in SPASM. Several applications of the method are discussed. In one (admittedly fortuitous) example, spatial motif recognition followed by profile analysis enabled assignment of the structural class and probable function of a hypothetical protein for which no structural or functional annotation was available in the sequence database.

Description of the Method

Representation of protein structure

SPASM currently only treats amino acid residues (recognised by the fact that they contain at least three main-chain atoms), and they are represented by the co-ordinates of their C^{α} atom and (for non-glycine residues) by a pseudo-atom located at the centre of gravity of their side-chain atoms. There are several advantages to such an abstracted representation. First, the database of structures will be considerably smaller than one adopting an all-atom representation. Second, operations on the database will be considerably faster, and speed is of the essence today, but even more in the future as the size of the structural database continues to increase. Third, the approach avoids many problems associated with ambiguities in the experimental identification of certain atoms (e.g. the side-chain oxygen and nitrogen atoms in asparagine and glutamine residues; McDonald & Thornton, 1995), ambiguities in nomenclature (e.g. equivalent oxygen atoms in aspartate and glutamate residues), as well as some possible model errors (e.g. erroneous peptide flips; Jones *et al*., 1991). The main drawback of the abstracted representation is that the searches may be less sensitive. On the other hand, if the program is used as a filter to quickly retrieve possible hits, which are subsequently scrutinised either by a full-atom least-squares program or by manual inspection on a graphics display, the severity of this drawback is much reduced.

Input and database

The input to the program consists mainly of the SPASM database file, a file containing the user's motif, and values for a number of program parameters. The motif is a small file in PDB format that contains only the residues that together constitute the motif, which was considered the most intuitive method from a user's point of view. The current database was derived from the June 1998 release of the PDB, using the June 1998 list of Hobohm & Sander (1994), including all protein chains whose mutual sequence identity is 95 % or less. This database comprises more than 450,000 residues from 2190 PDB entries.

Algorithm

In SPASM, a user-defined motif is compared with a database of protein models. A simple recursive, depth-first search algorithm is used (Kleywegt *et al*., 1989), with pruning of the search tree as early as possible (Kleywegt & Jones, 1997). Each residue in the user-defined motif is represented by its C^{α} atom and the centre of gravity of its side-chain atoms. The distances between all these (pseudo) atoms are calculated, and for each database protein the program will identify all sets of residues that are identical (or similar) in type, and display a similar spatial arrangement, as assessed by the extent to which their (pseudo) atoms have similar mutual distances to those in the user-defined motif. At the start, the user can select to have SPASM use all (pseudo) atoms in the search, or only the side-chain pseudo-atoms, or only the C^{α} atoms. In addition, the program contains a number of features that make it very flexible, and that also enable fuzzy pattern matching (i.e. relaxed requirements on the precise nature of matching amino acid residues).

Fuzzy pattern matching can be accomplished in a number of different ways. If a motif contains one or a few residues that are (expected to be) variable, their residue type in the PDB file can be replaced by ''XXX''. For any such residues, only the C^{α} atoms will be considered in the matching procedure, so that they can match any other type of residue in the database proteins. If all residues are deemed to be variable (e.g. in searches for similar main-chain conformations), the program can also be instructed to only consider the C^{α} atoms in the matching process, and to allow every residue in the motif to match any residue type in the database proteins (provided they are in the correct spatial arrangement, obviously). A third method involves the use of a few, hard-coded allowed substitutions (e.g. Asp/Glu, Phe/Tyr, etc.), which enables the retrieval of patterns with a similar spatial arrangement of conservatively substituted residues. A further method employs the BLOSUM-45 amino acid substitution matrix (Henikoff & Henikoff, 1992) and a user-defined cut-off level; all amino acid substitutions whose matrix element is not less than the cut-off value will be tolerated in the matching process. Finally, the user may supply a list of explicitly allowed substitutions for each residue type that occurs in the motif. For example, if the serine protease catalytic triad of trypsin (Ser195, His57, and Asp102) is used as the search motif, all expected hits (including several lipases) are found (results not shown). However, if the substitution Asp/Glu is specifically allowed, the program also retrieves the Ser-His-Glu catalytic triads in acetylcholinesterase (Ser200, His440, and Glu327; PDB code 2ACK) and in *Geotrichum candidum* lipase (Ser217, His463, and Glu354; PDB code 1THG).

The matching process can be made considerably more efficient if certain constraints are applied. Three constraints have been implemented in SPASM to date, all of which are optional. The first requires that sequence directionality be conserved, i.e. if residue I in the motif is matched to residue K in a database protein, then residue J, with $J > I$, in the motif can only be matched to a residue L in the database protein if $L > K$. This constraint makes the matching process faster (on average reducing the number of candidates for matching a residue J by a factor of 2), but it will often be undesirable to use it. The second constraint requires that neighbouring residues in the motif must also be neighbouring residues in the database protein, i.e. if residue I in the motif is matched to residue K in a database protein, then if residue $I + 1$ is also present in the motif, it can only be matched (if at all) to residue $K + 1$ in the database protein. This constraint is much more powerful than the previous one (reducing the number of candidates for matching residue $I + 1$ to only a single residue), and is used quite often in practice. The final constraint requires that gap sizes between matched residues are conserved, i.e. if residue I in the motif is matched to residue K in a database protein, then residue $I + J$, with $J > 0$, in the motif can only be matched to residue $K + J$ in the database protein. This constraint is equally powerful as the previous one, but rarely desired. It should be emphasised that any or all of the three constraints can be switched on or off by the user. For instance, if one searches the database for a motif consisting of two helices taken from a helix-turn-helix moiety, and if one wants to allow for variable length turns between them, only the first two constraints would be used.

For each database protein, the program collects all residues that (based on their type) could in principle be matched to any of the residues in the input motif (if one or more residues in the motif have no potential matches, the database protein is not further considered). Subsequently, it recursively generates all combinations of residues that potentially match the motif. While doing so, the distances between the (pseudo) atoms of the database residues and those of the motif are compared, and if any distance pair differs by more than a preset cut-off value, that combination is discarded (i.e. the corresponding branch of the search tree is pruned). If a complete set of residues is found that matches the

motif, the entire set of (pseudo) atoms is superimposed, and their RMSD is calculated. If this number is smaller than a preset cut-off value, the corresponding set of residues constitutes a match or hit. The algorithm is essentially the same as that used previously for the automatic assignment of 2D (two-dimensional) and 3D ^{1}H NMR spectra of proteins (where spin systems detected in the spectra must be matched to residues in the sequence; Kleywegt *et al*., 1989, 1991, 1993), and in the fold-recognition program DEJAVU (where secondary structure elements of one protein must be matched to those of another; Kleywegt & Jones, 1994, 1997). The major difference lies in the fact that SPASM does not attempt to identify a single, optimal solution. Since more than one copy of a motif may occur in a database protein, the program will enumerate all occurrences that satisfy the constraints.

The program can conveniently be run interactively. Depending on the search motif and the parameters and constraints used, typical runs require between 30 seconds and five minutes of real time on a Silicon Graphics Indigo^2 workstation equipped with an R10000 processor.

Results

Every match that the program finds is listed in the output, together with a comparison of the residue types and, optionally, the distance matrices of the (pseudo) atoms. In addition, files can be created that enable SPASM to be interfaced to several other programs. First, the program can create a macro file for the crystallographic modelling program O (Jones *et al*., 1991). When executed, this macro will read and draw the user's motif, and subsequently read, superimpose, and draw each of the matches that SPASM found in the database proteins. This allows for a very rapid visual inspection of the results of the SPASM run. Second, SPASM can create an input file for LSQMAN, our local least-squares structure superpositioning program (Kleywegt, 1996; Kleywegt & Jones, 1997). The purpose of this is to do a more sensitive least-squares superpositioning analysis of the motif and the matches, which sometimes reveals similarities that extend beyond the local ones embodied in the common motif. Third, the program can produce a partial multiple sequence alignment file of all matches, which can subsequently be used by the SBIN suite of programs (G.J.K., unpublished results) to generate structure-based sequence profiles. This option is most useful if one uses one or more stretches of consecutive main-chain as the motif. The aligned partial sequences can be analysed to see if they contain (possibly hidden) similarities by creating a profile (Gribskov *et al*., 1990; Gribskov & Veretnik, 1996) based upon them. This profile can subsequently be scanned against a sequence database, such as SWISS-PROT (Bairoch & Apweiler, 1997), to try and retrieve other proteins that contain a similar sequence/structure motif.

Inverse motif recognition

The motif recognition technique discussed above essentially compares a single motif to a database of protein structures, and enumerates all occurrences of that motif in these structures (subject to certain constraints). It is a small step to implement the inverse process, namely a technique that compares a database of motifs to a single structure. This idea has been implemented in a separate program, called RIGOR. In essence, RIGOR is a 3D cousin of PROSITE, a well-annotated collection of sequence motifs developed by Bairoch & Bucher (1994). Obviously, the utility of this approach depends critically on the quality of the motif database. Since generating a comprehensive, high-quality, well-annotated database of motifs is beyond the scope of this work (and, indeed, is a task more suitably co-ordinated by a large institute for structural bioinformatics), a less labour-intensive approach has been developed for the time being. Initially, the SITE records that are present in some PDB files were considered for forming the basis of a motif database, but closer inspection showed that these are unsuitable for this purpose. There are many PDB files without SITE records, most that do exist are not annotated, and either contain very few residues (which will lead to many false matches), or very many residues (which makes them too specific), reflecting widely differing views as to what constitutes a ''site''. The compromise method used here is based on automatic scrutiny of the proteins present in the SPASM database to identify motifs that are deemed ''interesting'' for one of several reasons. In the present implementation, a set of residues is considered interesting if it consists of N sequential identical residues (e.g. four sequential arginine residues). Alternatively, a set of residues that are in spatial proximity is considered interesting if it consists of only hydrophobic, only polar and charged, or mixed hydrophobic and polar/charged residues. Finally, a set of residues that all contact a single hetero-compound are also considered interesting. All motifs generated in the latter category are represented twice in the motif database, once as a residue-specific motif (all residue types must match exactly), and once as an ''engineerable'' motif (the residue types do not matter, only the spatial relationships of their C^α atoms). The rationale behind the latter category of motifs is that it might enable the identification of potential binding sites for metals, ions, and other ligands that could be engineered into proteins that do not contain them in their natural state. The automatically generated motif database is unsatisfactory in a number of respects (principally due to the fact that it is unclear what structural or functional importance, if any, should be attached to each motif), but serves to demonstrate the scope of the methodology. The present database contains $\sim$3400 regular motifs, and $\sim$1600 engineerable ones. Like SPASM, RIGOR is interfaced to O, allowing for rapid visual inspection of its results.

Several methods have been reported to automatically detect recurring motifs and patterns in protein sequences (for example, Karlin *et al*., 1990; Henikoff & Henikoff, 1992), as well as at the main-chain level of protein structures (for example Matsuo & Kanehisa, 1993; Han *et al*., 1997). Similarly, several studies have been carried out investigating specific types of motifs, clusters, and interactions at the residue and side-chain level (Warme & Morgan, 1978; Heringa & Argos, 1991; Chakrabarti, 1993; Copley & Barton, 1994). In order to be able to produce a high-quality motif database, manual curation, possibly supported by more intelligent methods for automatic motif discovery such as that described by Russell (1998), would appear to be the most promising approach.

Applications

In this section, three different types of application of spatial motif recognition using SPASM are discussed: main-chain recognition (using only C^{α} atoms), active-site recognition (using both C^{α} atoms and side-chain pseudo-atoms), and metal-binding site recognition (using only side-chain pseudo-atoms). The first three examples described below used the December 1997 database, whereas the fourth one used the June 1998 database.

Several ''real-life'' applications have been published in the literature. Harel *et al*. (1995), in their analysis of the interface of *Torpedo californica* acetylcholinesterase and the snake toxin fasciculin, encountered a seemingly unusual stacking interaction between a tryptophan ring and the S^{δ}-C^{ε} moiety of a methionine residue. Closer scrutiny with SPASM, however, revealed the existence of several similar, albeit not identical, interactions that occur in a variety of proteins, including the anionic site of acetylcholinesterase itself. Heikinheimo *et al*. (1996) used SPASM to investigate the metal-binding site of soluble inorganic pyrophosphatase, and found similarities to several proteins that interact with DNA, including rat DNA polymerase β. Matte *et al*. (1996) compared the main-chain conformation of eight residues in the phosphate-binding motif of phosphoenolpyruvate carboxykinase to the structural database using SPASM, and found striking similarities to the P-loop of adenylate kinase isozyme III, RecA protein, and $p21^{ras}$ complexed with GTP. Some additional applications of SPASM are discussed by Kleywegt & Jones (1998).

Main-chain recognition

Although the problem of recognising main-chain motifs was solved by Jones & Thirup (1986) more than a decade ago, the present implementation allows for increased flexibility. Moreover, as the following example shows, combining motif recognition with profile analysis yields a powerful tool for the prediction of structure and function. In this example, residues 2 to 14 were taken from the crystal structure of cellular retinoic acid-binding protein type II (Kleywegt *et al*., 1994; CRABP2; PDB code 1CBS). CRABP2 is a member of a large family of proteins with eight or ten-stranded β-barrel folds (the retinol-binding protein family, and the fatty acid-binding protein family, respectively; Banaszak *et al*., 1994). The N terminus forms a single helical turn, followed by a β-strand that is bent so as to participate in both of the two orthogonal β-sheets that make up the barrel. In addition, this part of the protein contains an almost absolutely conserved GXW sequence motif. Running SPASM with default parameters (maximum RMSD of 1.5 Å, and maximum distance mismatch of 2.0 Å) yields 11 matches (Figure 1(a)), all of which are proteins with the eight or ten-stranded β-barrel fold. The SPASM database contained a total of 16 such proteins. Further hits are found with more relaxed criteria, but then in addition other hits are found, e.g. in alcohol dehydrogenase. From a purely structural point of view, such additional matches are not ''false positives'', of course. Figure 1(b) shows the matching proteins after their global superpositioning operators were refined with the program LSQMAN. The aligned partial sequences of the 11 matching proteins were further used to generate a structure-based sequence profile (for the 13 residues alone), and this profile was scanned against release 35 of the SWISS-PROT database (Bairoch & Apweiler, 1997). Apart from the expected hits, and a few false ones, one hypothetical protein from *Caenorhabditis elegans* was retrieved in the scan. This hypothetical protein (Q09294 a.k.a., YQO3_CAEEL) looked interesting since the match to the profile occurred ~130 residues upstream of its C terminus, typical of the ten-stranded β-barrel protein family. The first ~60 residues of the hypothetical protein are very rich in charged residues, but the C-terminal ~135 residues can be aligned surprisingly well with other members of the family (Figure 2), revealing that it shares several key sequence features. Consequently, it seems reasonable to assume that the C-terminal domain of this protein will fold into a ten-stranded β-barrel, and that its function lies in the transport or storage of one or more types of carboxylate-containing hydrophobic compound. The hypothetical protein is not annotated with respect to possible structure or function in SWISS-PROT, probably because it does not fit the ''canonical'' PROSITE pattern (Bairoch & Bucher, 1994) for this family of proteins, and because both global and local sequence alignments with other members of the family are rather poor (results not shown). This serendipitous finding illustrates the potential of methods that investigate structural similarities at a more local level than that of the (domain) fold, and that use structural alignments to produce profiles that can be scanned against sequence databases (G.J.K., unpublished results).

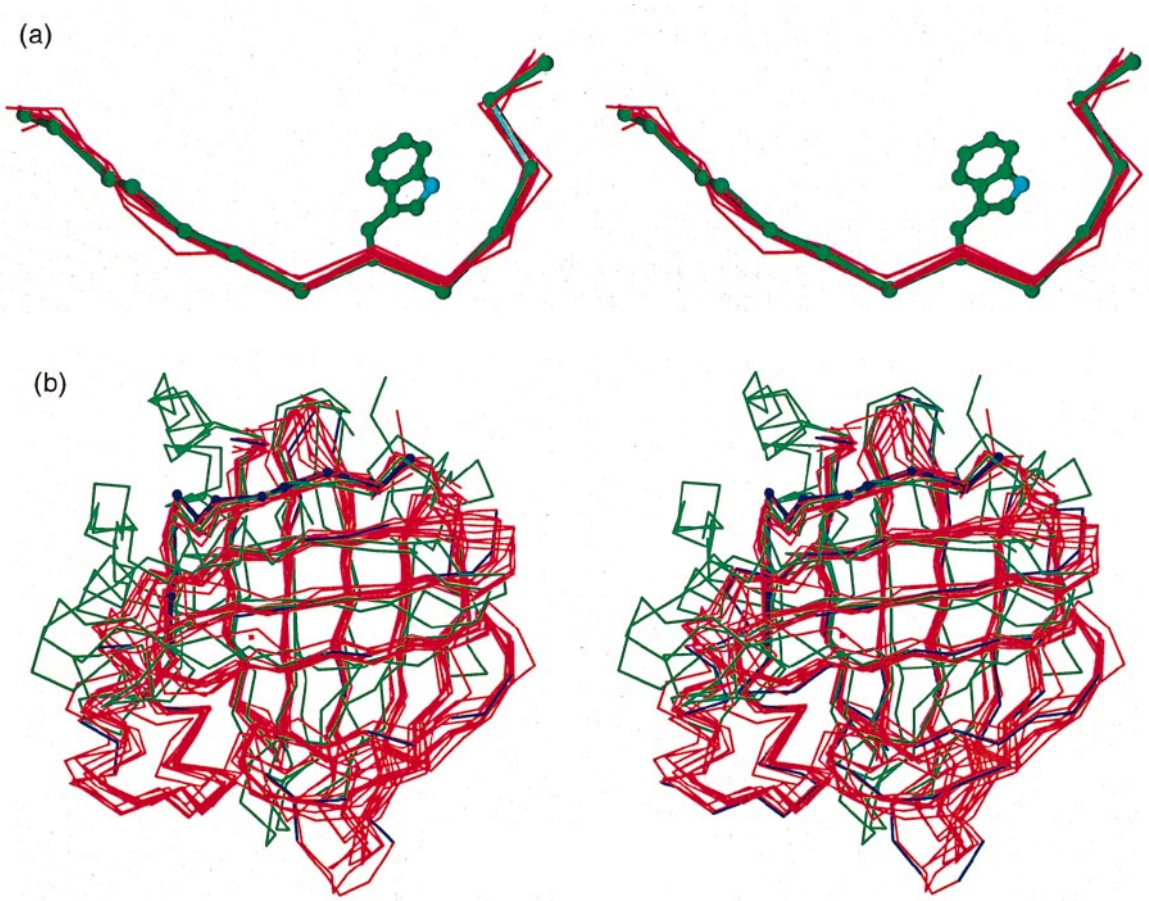

Figure 1. (a) Results of a SPASM search using residues 2-14 of CRABP2 (1CBS; green ball-and-stick C^{α} trace; the side-chain of Trp7 is included for reference). A total of 11 matches is found (red traces), nine of which belong to the ten-stranded β-barrel family (1CBS, RMSD = 0.0 Å for 13 C^{α} atoms; 1OPB, 0.31 Å; 1CBI, 0.35 Å; 1CRB, 0.36 Å; 1HMT, 0.44 Å; 1LID, 0.48 Å; 1PMP, 0.58 Å; 1ICN, 0.78 Å; 1EAL, 0.98 Å), and two to the eight-stranded β-barrel family (1MUP, 0.72 Å; 1EPA, 0.76 Å). (b) Results of automatic superpositioning with LSQMAN of the 11 matching proteins shown in (a) and the entire structure of CRABP2 (in blue, with the 13-residue motif shown as a ball-and-stick trace). The members of the ten-stranded β-barrel family are shown in red (between 115 and 137 residues can be aligned, with RMSDs varying from 0.0 to 1.62 Å), and those of the eight-stranded β-barrel family in green (44 aligned residues for 1MUP with an RMSD of 1.68 Å; 52 aligned residues for 1EPA, with an RMSD of 1.64 Å).

Active-site recognition

In the second example, the three acidic residues (Glu212, Asp214, and Glu217) that make up the active site of the cellulose-degrading enzyme cellobiohydrolase I from *Trichoderma reesei* (Divne *et al.*, 1994; CBH I; PDB code 1CEL) were used. For the enzymes that were expected to contain a similar set of residues (CBH I, endoglucanase I, and bacterial β-glucanase), four PDB entries were present in the SPASM database (PDB codes 1CEL, 2AYH, 1GBG, 1MAC; other entries were either not available when the database was created, or they had more than 95% sequence identity to any of these four proteins). The C^{α} and the side-chain pseudo-atoms were used in the search (RMSD cutoff 1.0 Å, maximum distance mismatch 1.5 Å), and all four expected proteins were retrieved by SPASM, with no further matches (Figure 3).

Metal-binding site recognition

As a direct comparison with the method by Artymiuk *et al.* (1994, 1995), a motif was used composed of the zinc-binding side-chains of thermolysin (His142, His146, Glu166; PDB code 4TMN). Artymiuk *et al.* (1994) found matches with zinc and iron-binding sites in five distinct proteins, namely thermolysin itself, carboxypeptidase, haemoglobin, hemerythrin, and the photosynthetic reaction centre. Reflecting the rapid growth of the structural database, SPASM finds 125 hits in 58 proteins (results not shown) with similar criteria as used by Artymiuk *et al.* (1994): 0.5 Å RMSD cutoff, and a maximum distance mismatch of 1.0 Å.

```
Q09294          EAPIQILTAMIGKWKLASSENLQEYFTLEKFPEITQM----AWEHGITCYKM-NGNQLHV
P55054          -----MIEPFLGTWKLVSSENFENYVRELGVECEPRK--VACLIKPSVSISF-NGERMDI
P51673          -------PNFSGNWKIIRSENFEEMLKALGVNMMMRKIAVAAASKPAVEIKQENDDTFYI
P06768          ------TKDQNGTWEMESNENFEGYMKALDIDFATRK--IAVRLTQTKIIVQ-DGDNFKT
P02689          ------SNKFLGTWKLVSSENFDDYMKALGVGLATRK--LGNLAKPTVIISK-KGDIITI
P29498          ------MSSFLGKWKLSESHNFDAVMSKLGVSWATRQ--IGNTVTPTVTFTM-DGDKMTM
                           1 1      *

Q09294          HTDLLGKS---LIPTIFEFDKPIAR-DDNAVSTHAEGNMMS-TICKRIADGSIV--WKVE
P55054          QAGSACRNTEISFKLGEEFEETTA--DNRKVKSLITFEGGS-MIQIQRWLGKQ---TTIK
P51673          KTSTTVRTTEINFKIGEEFEEQTV--DGRPCKSLVKWESENKMVCEQRLLKGEGPKTSWS
P06768          KTNSTFRNYDLDFTVGVEFDEHTKGLDGRNVKTLVTWEGNT-LVCVQKGEKEN---RGWK
P02689          RTESTFKNTEISFKLGQEFEETTA--DNRKTKSIVTLQRGS-LNQVQRWDGKE---TTIK
P29498          LTESTFKNLSCTFKFGEEFDEKTS--DGRNVKSVVEKNSESKLTQTQVDPKNT---TVIV
                      ~          **~       *      ~

Q09294          RLIKN-GNLVVFNSRGNFRCKRVYKRVN
P55054          RRIVD-GRMVVECTMNNVVSTRTYERV-
P51673          RELTNDGELILTMTADDVVCTRVYVRE-
P06768          QWVEG-DKLYLELTCGDQVCRQVFKKK-
P02689          RKLVN-GKMVAECKMKGVVCTRIYEKV-
P29498          REVDG-DTMKTTVTVGDVTAIRNYKRLS
                3       ~          ~ 3 2 2
```

Figure 2. Sequence alignment of residues 65-194 of the hypothetical *C. elegans* protein YQO3_CAEEL (Q09294) with the complete sequences of several representative members of the ten-stranded β-barrel family: P55054, TLBP_RAT, rat testis lipid-binding protein; P51673, RET4_RAT, rat cellular retinoic acid-binding protein II; P06768, RET2_RAT, rat cellular retinol-binding protein II; P02689, MYP2_HUMAN, human myelin P2 protein; P29498, FABP_SCHMA, 14 kDa fatty acid-binding protein from *Schistosoma mansoni*. Several residues characteristic of this protein family are conserved in the hypothetical protein, including the GXW motif near the start (indicated by ''1''), the Y/F-X-K/R motif near the C terminus (indicated by ''2''), and several others (indicated by ''*'' for absolutely conserved, and '' ~ '' for moderately conserved residues). It is well known (Banaszak *et al.*, 1994) that the two residues indicated by ''3'' determine the specificity for ligands that contain an alcohol moiety (such as retinol-binding proteins, where they are glutamine residues), or for ligands that contain a carboxylate moiety (such as fatty acid and retinoic acid-binding proteins, in which case they are arginine residues). Since the hypothetical protein contains an arginine residue in both positions, it is extremely likely that this protein is involved in transport or storage of carboxylate-containing hydrophobic compounds.

Since the motif contains two histidine residues, and only three pseudo-atoms are used in the comparison (that, in addition, are arranged as a near-equilateral triangle), many of the hits occur twice, once with swapped histidine residues. The 58 proteins include the five that were found by Artymiuk *et al*. (1994). Many of the matches are indeed zinc-binding sites (PDB entries 1ATL, 1CGL, 1FBL, 1DMX, 1EZM, 1FUA, 1IAG, 1JAP, 1KAP, 1KCW, 1KUH, 1MMQ, 1NPC, 1OBR, 1PCA, 1PMI, 1SAT, 1SLM, 2CTC, 2HMZ, and 8TLN). Several others are (part of) iron-binding sites: 2,3-dihydroxybiphenyl 1,2-dioxygenase (1DHY, 1HAN), hexose-1-phosphate uridylyltransferase (1HXP), the photosynthetic reaction centre (1PCR, 1PRC), hemerythrin (2HMZ), and myohemerythrin (2MHR). In one case, the site contains nickel (1IAE, astacin with zinc replaced by nickel), and in one case it contains copper (1KCW, ceruloplasmin). In the other hits that are not permutations of any of the above, either nothing is bound or modelled (e.g. in cytochrome *C* oxidase, 1OCC), or a water molecule is present (e.g. in cyclin H, 1KXU).

Left-handed helices

This application demonstrates how the present method can be used to rapidly provide an answer to questions such as ''do left-handed helices occur in natural proteins?'' To address this particular question, an ideal seven-residue poly-Ala right-handed α-helix (i.e. two turns) was used to generate a left-handed helix by negating the *x*-co-ordinates of all atoms (i.e. a mirroring operation in the plane $x = 0$). The resulting left-handed (poly-D-Ala) helix was used in a SPASM search (using only the C^{α} atoms) with default parameters. Eight matches were found of which one was essentially perfect (RMSD = 0.3 Å), namely residues Lys39 to His45 in the 1.9 Å crystal structure of alanine racemase (Shaw *et al*., 1997; PDB code 1SFT, chain A). This enzyme, interestingly enough, catalyses the

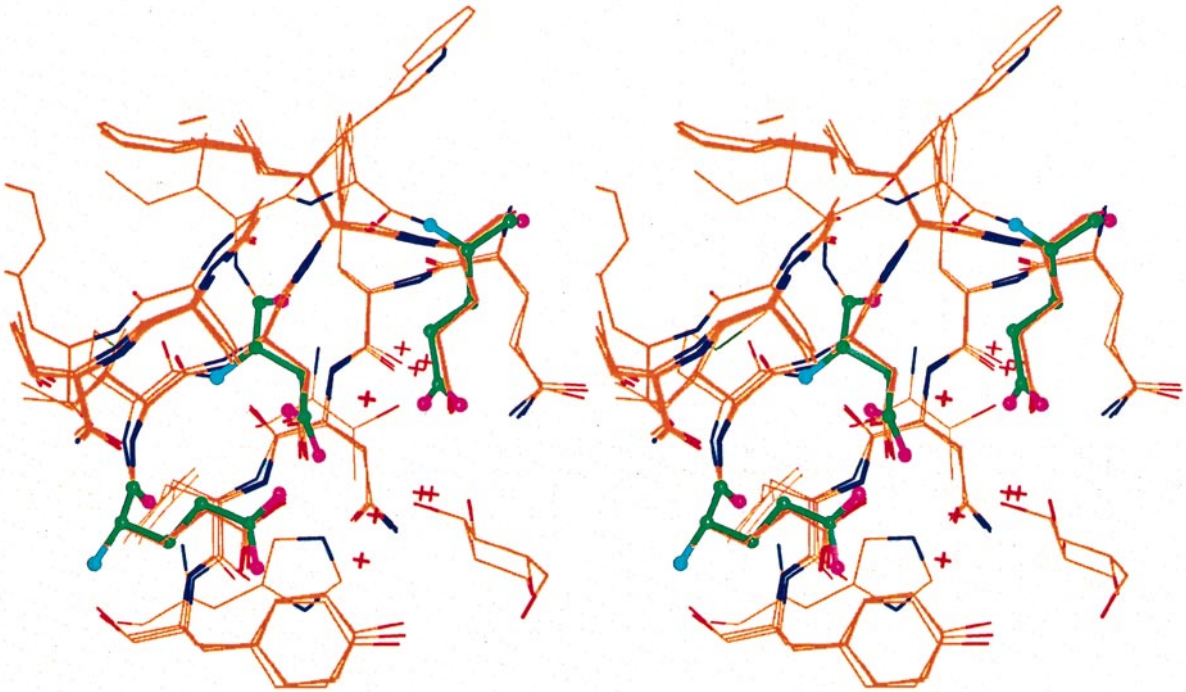

Figure 3. Results of a SPASM search using the catalytic residues of CBH I (1CEL), GluA212, AspA214, and GluA217 (ball-and-stick model). All four expected hits are found (thin stick models), namely CBH I itself (1CEL; RMSD = 0.0 Å), and three entries related to bacterial β-glucanase: 1GBG (residues 105, 107, and 109; RMSD = 0.25 Å), 1MAC (residues 103, 105, and 107; RMSD = 0.27 Å), and 2AYH (residues 105, 107, and 109; RMSD-0.24 Å). Note how well the side-chains of the three residues superimpose, despite the fact that there are different numbers of residues separating them in CBHI and the β-glucanases.

pyridoxal-dependent conversion of L-alanine into D-alanine. Residues 40 to 43 all lie in the left-handed helical region of the Ramachandran plot (ϕ and ψ both near $+50^\circ$), and Gly44 has positive ϕ and ψ values. The two turns of left-handed helix form the connection between the first strand of the α/β barrel and the subsequent (regular) α-helix (*vide infra*). Lys39 provides the covalent attachment site for the enzyme's co-factor, pyridoxal 5′-phosphate, and Tyr43 forms a hydrogen bond to one of the phosphate oxygen atoms of the co-factor. Interestingly, the left-handed helix forms the core of the (highly conserved) PROSITE (Bairoch & Bucher, 1994) sequence motif characteristic of alanine racemases (V-x-K-A-(DN)-(GA)-Y-G-H-G), which maps to residues Val37 to Gly46 as revealed by PDBsum (Laskowski *et al.*, 1997).

A further seven matches for the seven-residue left-handed helix were found, but these were not as good as the one encountered in alanine racemase, mainly because only four to six residues superimposed well. The additional matches were found in P1 endonuclease (PDB code 1AK0, RMSD = 0.6 Å, residues Ala129 to Asn135), matrix porin outer membrane protein F (2OMF, RMSD = 0.8 Å, Asn141 to Leu147), acyl carrier protein (1ACP, RMSD = 0.9 Å, Lys9 to Leu15), activated protein C (1AUT, RMSD = 1.1 Å, SerL99 to CysL105), thermolysin (8TLN, RMSD = 1.1 Å, ThrE224 to ValE230), neutral protease (1NPC, RMSD = 1.1 Å, Ser225 to Val231), and FD major coat protein (1FDM, RMSD = 1.4 Å, Lys43 to Ala49). When comparing the sequences of the eight matches, no obvious patterns emerge, although there is an abundance of residues with small side-chains.

In their paper describing the crystal structure of alanine racemase, Shaw *et al.* (1997) regarded residues 39 to 44 as an ordinary α-helix that is part of the α/β barrel, whereas the subsequent helix was depicted as an extra inserted secondary structure element. It would seem more appropriate to regard the left-handed helix as an unusual element, and the subsequent right-handed helix as a regular component of the barrel. The reason why the handedness of the helix went unnoticed may well lie in the fact that programs that assign secondary structure classes based on distances and angles (but not torsion angles) cannot discriminate between left and right-handed helices (since distances and angles are invariant under a mirroring operation; torsion angles, on the other hand, undergo a sign change). Hence, left-handed helices may easily escape attention. It might be worthwhile to modify such programs so that they determine the handedness of helices, since short left-handed helical segments appear to be quite common. A search of the SPASM

database (containing 2190 proteins) using a five-residue left-handed helical fragment yields 1745 hits in 943 proteins (43 %) with an RMSD below 1.0 Å, of which 266 hits in 230 proteins (11 %) have an RMSD below 0.5 Å. A similar search with a four-residue fragment yields 1132 hits in 738 proteins (34 %) with an RMSD below 0.25 Å.

Conclusion

The scope of applications of spatial motif recognition techniques is more extensive than can be covered here. Potential applications include: (1) analysis of newly determined protein structures. Often in a newly solved structure, one observes a local main-chain conformation, or an arrangement of side-chains that may seem odd or unusual. Spatial motif recognition can be used to rapidly answer questions such as ''is this a unique loop conformation?'', ''in what other structures does a similar constellation of residues occur?'', etc. (2) Comparative structural analysis. Scientists interested in aspects of protein structure between the levels of domain folds and individual residues will be interested in identifying all proteins that contain a certain arrangement of helices, strands, turns, and loops, or in all proteins that contain a certain constellation of residues or side-chains (e.g. metal-binding sites, anion-binding sites, left-handed helices). Spatial motif recognition methods can be employed to quickly enumerate all such occurrences. (3) Design and engineering. A search of the database with a motif consisting of certain metal-binding residues, with relaxed criteria on the exact nature of one or more of the residues, may identify other proteins in which one or a few mutations might suffice to introduce a novel binding site. Similar searches could be carried out with residues that bind anions, co-factors, ligands, or drugs. Some dedicated programs for this task exist, e.g. for the case of tetrahedral metal-binding sites (Clarke & Yuan, 1995), but the approach presented here is completely general, and therefore has wider applicability, although it may be less sensitive. (4) Prediction of structure and function. As the example involving CRABP2 demonstrated, there is scope for the use of small structural fragments in the identification of proteins of unknown structure and function. In this case, quite unexpectedly, a hypothetical protein of unknown structure and function was shown to belong to the fatty acid-binding protein family, based on structural and profile analysis of a stretch of only 13 residues.

Clearly, spatial motif recognition has the potential to become a powerful tool in structural analysis, just as sequence motif recognition is important in the analysis of sequence databases (Bork & Gibson, 1996). SPASM is a rapid, intuitive, and user-friendly tool to quickly identify proteins that display structural similarities at the level of residues, and that perhaps have (or once had) unsuspected functional similarities. RIGOR, if provided with a high-quality motif database, has the potential to become a very useful tool for structural biologists who want to make the most of their models, and to protein engineers who want to understand, influence, or alter protein function.

Availability

The SPASM software package (including all program executables, manuals, and databases) is available free of charge to academic users (see http://alpha2.bmc.uu.se/usf/for details). Other users may contact G.J.K. for licensing details (mailto:gerard@xray.bmc.uu.se). The software package includes the programs SPASM, RIGOR, MKSPAZ (to generate custom SPASM databases), and MAKRIG (to generate custom RIGOR motifs and databases).

Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research (SSF), and its Structural Biology Network (SBNet). The work on SPASM was inspired by a lecture by Peter Artymiuk at the 1995 CCP4 Study Weekend (Artymiuk *et al.*, 1995).

References

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* **277**, 556-571.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. & Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **243**, 327-344.

Artymiuk, P. J., Poirrette, A. R., Rice, D. W. & Willett, P. (1995). Comparison of protein folds and sidechain clusters using algorithms from graph theory. In *From First Map to Final Model* (Bailey, S., Hubbard, R. & Waller, D. A., eds), pp. 71-81, SERC Daresbury Laboratory, Daresbury, UK.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucl. Acids Res.* **25**, 31-36.

Bairoch, A. & Bucher, P. (1994). PROSITE: recent developments. *Nucl. Acids Res.* **22**, 3583-3589.

Banaszak, L., Winter, N., Xu, Z., Bernlohr, D. A., Cowan, S. & Jones, T. A. (1994). Lipid-binding proteins: a family of fatty acid and retinoid transport proteins. *Advan. Protein Chem.* **45**, 89-151.

Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162-184.

Brint, A. T., Mitchell, E. & Willett, P. (1988). Substructure searching in files of three-dimensional chemical structures. In *Chemical Structures. The International Language of Chemistry* (Warr, W. A., ed.), pp. 131-144, Springer-Verlag, Berlin.

Chakrabarti, P. (1993). Anion binding sites in protein structures. *J. Mol. Biol.* **234**, 463-482.

Clarke, N. D. & Yuan, S. M. (1995). Metal search: a computer program that helps design tetrahedral

metal-binding sites. *Proteins: Struct. Funct. Genet.* **23**, 256-263.
Copley, R. R. & Barton, G. J. (1994). A structural analysis of phosphate and sulphate binding sites in proteins. Estimation of propensities for binding and conservation of phosphate binding sites. *J. Mol. Biol.* **242**, 321-329.
Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science,* **265**, 524-528.
Fischer, D., Wolfson, H., Lin, S. L. & Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* **3**, 769-778.
Glusker, J. P. (1991). Structural aspects of metal liganding to functional groups in proteins. *Advan. Protein Chem.* **42**, 1-76.
Gribskov, M. & Veretnik, S. (1996). Identification of sequence patterns with profile analysis. *Methods Enzymol.* **266**, 198-212.
Gribskov, M., Lüthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146-159.
Han, K. F., Bystroff, C. & Baker, D. (1997). Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* **6**, 1587-1590.
Harel, M., Kleywegt, G. J., Ravelli, R. B. G., Silman, I. & Sussman, J. L. (1995). Crystal structure of an acetylcholinesterase-fasciculin complex: interaction of a three-fingered toxin from snake venom with its target. *Structure,* **3**, 1355-1366.
Heikinheimo, P., Lehtonen, J., Baykov, A., Lahti, R., Cooperman, B. S. & Goldman, A. (1996). The structural basis for pyrophosphate catalysis. *Structure,* **4**, 1491-1508.
Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.
Heringa, J. & Argos, P. (1991). Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **220**, 151-171.
Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.
Holm, L. & Sander, C. (1994). Searching protein structure databases has come of age. *Proteins: Struct. Funct. Genet.* **19**, 165-173.
Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819-822.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A,* **47**, 110-119.
Karlin, S., Blaisdell, B. E. & Brendel, V. (1990). Identification of significant sequence patterns in proteins. *Methods Enzymol.* **183**, 388-402.
Kleywegt, G. J. (1996). Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallog. sect. D,* **52**, 842-857.
Kleywegt, G. J. & Jones, T. A. (1994). Halloween . . . masks and bones. In *From First Map to Final Model* (Bailey, S., Hubbard, R. & Waller, D. A., eds), pp. 59-66, Daresbury, UK, SERC Daresbury Laboratory.
Kleywegt, G. J. & Jones, T. A. (1997). Detecting folding motifs and similarities in protein structures. *Methods Enzymol.* **277**, 525-545.
Kleywegt, G. J. & Jones, T. A. (1998). Databases in protein crystallography. *Acta Crystallog. sect. D,* **54**, 1119-1131.
Kleywegt, G. J., Lamerichs, R. M. J. N., Boelens, R. & Kaptein, R. (1989). Toward automatic assignment of protein ^{1}H NMR spectra. *J. Magn. Reson.* **85**, 186-197.
Kleywegt, G. J., Boelens, R., Cox, M., Llinás, M. & Kaptein, R. (1991). Computer-assisted assignment of 2D ^{1}H NMR spectra of proteins: basic algorithms and application to phoratoxin B. *J. Biomol. NMR,* **1**, 23-47.
Kleywegt, G. J., Vuister, G. W., Padilla, A., Knegtel, R. M. A., Boelens, R. & Kaptein, R. (1993). Computer-assisted assignment of homonuclear 3D NMR spectra of proteins. Application to pike parvalbumin III. *J. Magn. Reson. sect. B,* **102**, 166-176.
Kleywegt, G. J., Bergfors, T., Senn, H., Le, Motte P., Gsell, B., Shudo, K. & Jones, T. A. (1994). Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-*trans*-retinoic acid and a synthetic retinoid. *Structure,* **2**, 1241-1258.
Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488-490.
Lesk, A. M. (1979). Detection of three-dimensional patterns of atoms in chemical structures. *Commun. ACM,* **22**, 219-224.
Matsuo, Y. & Kanehisa, M. (1993). An approach to systematic detection of protein structural motifs. *Comput. Applic. Biosci.* **9**, 153-159.
Matte, A., Goldie, H., Sweet, R. M. & Delbaere, L. T. J. (1996). Crystal structure of *Escherichia coli* phosphoenolpyruvate carboxykinase: a new structural family with the P-loop nucleoside triphosphate hydrolase fold. *J. Mol. Biol.* **256**, 126-143.
McDonald, I. K. & Thornton, J. M. (1995). The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. *Protein Eng.* **8**, 217-224.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.
Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211-1227.
Shaw, J. P., Petsko, G. A. & Ringe, D. (1997). Determination of the structure of alanine racemase from *Bacillus stearothermophilus* at 1.9-Å resolution. *Biochemistry,* **36**, 1329-1342.
Thornton, J. M. & Gardner, S. P. (1989). Protein motifs and data-base searching. *Trends Biochem. Sci.* **14**, 300-304.
Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteases and lipases. *Protein Sci.* **5**, 1001-1013.

Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308-2323.

Warme, P. K. & Morgan, R. S. (1978). A survey of amino acid side-chain interactions in 21 proteins. *J. Mol. Biol.* **118**, 289-304.

Willett, P. (1987). A review of chemical structure retrieval systems. *J. Chemometrics,* **1**, 139-155.

Edited by J. Thornton

(Received 25 March 1998; received in revised form 2 November 1998; accepted 4 November 1998)

Note added in proof: Recently, Brodersen *et al.* (1998) *Biochemistry*, in the press, have used SPASM to locate and identify structurally similar motifs for a new zinc-binding site discovered in psoriasin (S100A7), a protein belonging to the S100-class of calcium-binding EF-hand proteins. They found that the new zinc-binding site is reminiscent of the pattern seen for the active site in certain metalloproteases. This finding has important consequences for the understanding of the regulation of the S100-class of proteins.