Searching for Patterns of Amino Acids in 3D Protein Structures

Ruth V. Spriggs,[‡] Peter J. Artymiuk,[‡] and Peter Willett*,[†]

Krebs Institute for Biomolecular Research and Departments of Information Studies and of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Received September 2, 2002

This paper describes the program ASSAM, which has been developed to search for patterns of amino acid side-chains in the 3D structures in the Protein Data Bank. ASSAM represents an amino acid by a vector drawn from the main chain towards the functional part of the amino acid and then computes a graph representation of a protein in which the individual side-chain vectors are the nodes and the intervector distances are the edges. The presence of a query pattern in a Protein Data Bank structure can then be searched for by means of a subgraph isomorphism algorithm. Recent enhancements to ASSAM allow searches to include the following: the main-chain structure in addition to the side-chains; the secondary structure and solvent accessibility of side-chains; allowable distances from a known binding-site; disulfide bridges; and improved generic and wild-card queries. The effectiveness of these approaches is demonstrated by extensive searches of the Protein Data Bank for typical 3D query patterns.

INTRODUCTION

Tools for searching databases of protein sequences have for long played a central role in computational biology, but the much smaller numbers of protein structures (as against sequences) has meant that systems for structure-based searching have occasioned less interest. There is, however, an increasing need for such systems as a result of developments in protein NMR and in protein crystallography that are bringing about huge increases in the numbers of proteins for which a 3D structure is available.[1,2] A long-term collaboration between the Departments of Information Studies and of Molecular Biology and Biotechnology at the University of Sheffield[3−5] has investigated the use of methods from graph theory for the representation and searching of the structures in the Protein Data Bank (hereafter PDB).[6] Drawing on the subgraph isomorphism and maximum common subgraph isomorphism techniques that are used for structure matching in chemoinformatics applications such as substructure searching, pharmacophore mapping, and reaction indexing, we have developed techniques for structure-based access to proteins at both the secondary structure and tertiary structure levels.

Here, we describe the program ASSAM, which is designed to search for patterns of amino acid side-chains in PDB structures using a pseudo-atom representation of the structure of a side-chain and a modified version of the subgraph isomorphism algorithm of Ullmann (as described in the next section). ASSAM was first presented at the 1993 Noordwijkerhout conference[7] and has been in use in Sheffield since then. More recently, several other groups have described related programs, most notably the SPASM program of Kleywegt.[8] This adopts the idea of using pseudo-atoms, albeit using a rather different definition of their location within a residue, in conjunction with a depth-first tree search procedure, rather than the Ullmann algorithm; more recently, Kleywegt has described a postprocessing program, SAVANT, for the analysis of SPASM output.[9] In this paper, we report several major enhancements that have been made to ASSAM recently and that permit far more specific searches to be carried out. These enhancements allow searches to include the following: the main-chain structure in addition to the side-chains; the secondary structure and solvent accessibility of side-chains; allowable distances from a known binding-site; disulfide bridges; and improved generic and wild-card queries. The next section describes the basic program, together with these new facilities, and we then present the results of searches for typical query patterns with one of these, that for the serine protease catalytic triad, being described in some detail.

REPRESENTATION AND SEARCHING OF 3D PROTEIN STRUCTURES

The Basic Program. The program ASSAM has been developed for the representation and searching of patterns of amino acid side-chains in 3D space.[5,10] Proteins are represented by labeled graphs, and patterns are searched for in such graphs using a modification of the subgraph isomorphism algorithm due to Ullmann.[11]

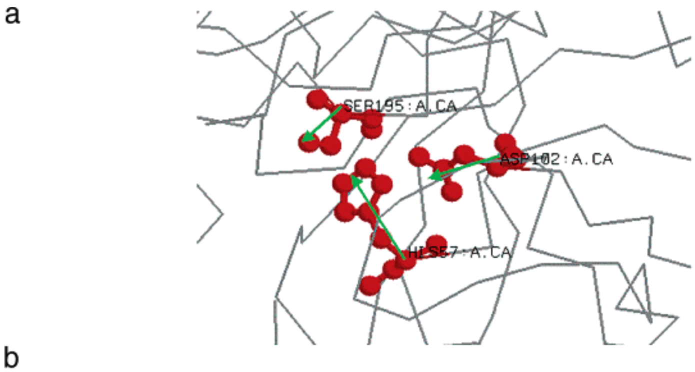
The nodes in the graph representation denote individual amino acid side-chains, and the edges denote the internode geometric relationships. Specifically, each node contains two *pseudo-atoms*, whose positions are chosen to emphasize the functional part of the side-chain corresponding to that node. The locations of the two pseudo-atoms are used to generate a vector, and each such vector corresponds to one of the nodes in a graph. The geometric relationships between pairs of residues are defined in terms of distances calculated between the corresponding vectors, and these relationships

* Corresponding author phone: +44-114-2222633; fax: +44-114-2780300; e-mail: p.willett@sheffield.ac.uk.

[†] Krebs Institute for Biomolecular Research and Department of Information Studies.

[‡] Krebs Institute for Biomolecular Research and Department of Molecular Biology and Biotechnology.

10.1021/ci0255984 CCC: $25.00 © 2003 American Chemical Society
Published on Web 01/01/2003



b

	HIS57						ASP102						SER195					
	SS	SE	ES	EE	MM	KK	SS	SE	ES	EE	MM	KK	SS	SE	ES	EE	MM	KK
HIS57	0	0	0	0	0	0	5.7	3.6	6.3	4.3	4.9	3.9	6.4	5.0	4.4	3.1	4.6	4.0
ASP102	5.7	6.3	3.6	4.3	4.9	3.9	0	0	0	0	0	0	9.3	9.0	7.7	7.2	8.2	7.2
SER195	6.4	4.4	5.0	3.1	4.6	4.0	9.3	7.7	9.0	7.2	8.2	7.2	0	0	0	0	0	0

Figure 1. (a) Serine protease catalytic triad query pattern from chymotrypsin (4CHA): triad residues highlighted together with arrows showing position and direction of vectors. (b) Pattern matrix for the catalytic triad of chymoptrypsin, taken from Chain A of 4CHA. Five numbers (the SS, SE, ES, MM, and KK distances described in the text, measured in Å) represent the relationship between the three side-chains.

correspond to the edges of a graph. Let S, M, and E denote the start, middle, and end, respectively, of a vector; then the graph edges contain five parts, these being the SS, SE, ES, EE, and MM distances (although only a subset of these five distances is normally used to specify a query pattern). A typical ASSAM query pattern, for the serine protease catalytic triad pattern discussed later in this paper, is shown in Figure 1(a). Queries can be created in two ways: by calculation of a matrix directly from the coordinates of a known structure using programs written for this purpose; and hypothetical patterns can be defined using another ancilliary program. The matrix for the catalytic triad is shown in Figure 1(b). The output from the program consists of the matching residues in the hit PDB files, and these can then be extracted automatically for display using the RASMOL program.[12]

The vectorial representation is clearly an extremely simple description of the relative orientations of the side-chains in a 3D protein structure. It does, however, have the advantage that it does not overdefine the orientations of ends of side-chains, as could occur if a more precise representation was to be used that was based directly on the individual atomic coordinates in the PDB. This ambiguity is in fact a useful feature for at least three reasons: in medium-resolution protein-crystallographic studies, it is often difficult to get the final torsion-angle value correct and so the fine details of the side-chain orientations may be in doubt; the identifications of the individual atoms in a residue can often be ambiguous; and side-chains can often move or twist, for example on binding substrates.

ASSAM has been in use in Sheffield for several years now. We have recently made several major enhancements to ASSAM, as described in the remainder of this section.

Query Formulation. The basic query formulation facilities have been augmented by the inclusion of Key-Key distances and by provision for generic residue types.

Key-Key, or KK, distances are measured between the key points of the residue side-chains, where the key point (one of S, M, or E) is the section of the side-chain that best describes the potential activity of that side-chain. For example, the key point of serine is the E pseudo-atom describing the position of the OH group which is the hydrogen bonding (and in some cases catalytically active) part of a serine; but the key point of a phenylalanine is the M pseudo-atom which best represents the side-chain's common function as a large hydrophobic moiety. In addition, it is now possible to define side-chains in generic terms, rather than using specific individual amino acid types. The generic residues currently available are as follows: ACI (acidic residues); ARO (aromatic residues, excluding histidine); BAS (basic residues); BAR (basic residues, excluding histidine); AMI (residues containing amides); PHO (specific hydrophobic residues: valine, leucine, or isoleucine); PHX (all hydrophobic residues); ASX (aspartic acid or asparagine); GLX (glutamic acid or glutamine); ACX (ASX or GLX); ROH (serine or threonine); RSH (serine, threonine, or cysteine); ROY (serine, threonine, or tyrosine); and RSY (serine, threonine, cysteine, or tyrosine). These classes have proved sufficient for the many searches that we have carried out; however, it would be entirely possible to enable a user to specify matching residue-types at search time.

Distance to Known Binding Site. The distance-to-known-binding-site (hereafter DBS) facility is based on distances that are calculated from the Key of a residue to the nearest nonwater, nonprotein atom. This information can be used to reduce the number of hits to only those where all matching residues are within the specified distance from a bound ligand. It is, of course, the case that not all proteins have their structure determined while their ligands are bound, and these hits could potentially be lost when a DBS is specified in a query. This problem is alleviated by the use of the PDB SITE information, which records those residues that are involved in sites that have been identified by the author of the structure. These sites can be active sites as well as binding sites and are given for 47% of the 9932 PDB files that were searched in our experiments. A DBS specification will hence find both bound heteroatoms and named SITEs.

Secondary Structure Information. The majority of PDB files include lines labeled HELIX and SHEET, detailing the residues involved in α-helices and β-sheets within the protein. This information is used to label each residue in the vector file with an 'h' to indicate it is positioned in an α-helix, an 's' to indicate it is involved in a β-sheet, or an 'x' for residues that are in neither of these; for example, an aspartic acid residue in an α-helix would be labeled as 'ASPh'. However, a minority of PDB files (about 7.5% in our search file) do not have secondary structure information provided. In these cases visual examination of the structure with RasMol was used to determine if secondary structure was present, and, if so, the appropriate helix and sheet labels were created using the ksdssp program in the MIDAS system,[13] which is based on the original DSSP program of Kabsch and Sander.[14] After this processing, there remained just 1.6% of the search files for which no secondary structure information was available.

The principal value of secondary structure information is the ability to increase the specificity of pattern searching. For example, this facility can be helpful in the case of searches for binding sites that may be positioned on the exposed ends of helices or on the loops between strands in a sheet, and this information is particularly valuable when main-chain vectors are used (see below), as queries involving the position of sections of main-chain will produce very large search-outputs unless some form of extra detail can be provided. However, secondary structure information can also be used to identify any trends in secondary structure for a particular tertiary structure motif. For example, the catalytic triad from chymotrypsin was used in searches with every possible combination of secondary structure: HISx ASPx SERx is the most common hit, suggesting that, in general, these three residues are more likely to be appropriately oriented if they are not involved in any secondary structure.

Disulfide Bridge Information. The cysteine residues of a protein are found in one of two forms: reduced (cysteine), with SH groups intact, or oxidized (cystine) to form disulfide bonds that link two sections of the same peptide, or different peptides, together. This information can be included in a query to further increase the specificity of searching. For example, disulfide bridges are most often found in extracellular proteins as they offer added stability to the protein structure (although there are exceptions to this), and a free cysteine may be part of an active site. For convenience, post-translationally modified cysteines (sulfenic acids, cysteines in Fe−S clusters) are currently regarded as cysteines.

PDB files for protein structures containing disulfide bridges have lines labeled SSBOND, which detail the pairs of cysteines involved in such bonds. This information is used to indicate the redox state of cysteines in the vector files searched by ASSAM. A reduced cysteine is labeled as CYH, and an oxidized one as CSS: a query pattern can specify either of these forms or the more general CYS (which matches both of them). Note that all cysteine residues in a database file to be searched can be labeled as either CSS or CYH because the absence of SSBOND lines indicates that any cysteine residues in that protein are not involved in disulfide bonds.

Solvent Accessibility Information. Specifying the solvent accessibility of the residues in a query enables the hits to be, for example, restricted to those sites that are more likely to be involved in binding substrates (i.e. sites on or near the surface of proteins). Knowledge of solvent accessibility may also enable the user to ask questions about the potential steric hindrance at the surface of a protein at a putative ligand binding site. The use of solvent accessibility hence represents a substantial improvement over the simple DBS approach mentioned previously, enabling areas of the protein that could potentially be binding sites, but are not already so identified within ASSAM, to be searched. Searches that do not specify solvent accessibility labels retrieve hits with all levels of solvent accessibility.

The accessibility of each residue in a protein is determined using the AREAIMOL program, which is available in the CCP4 suite[15] and which gives an absolute value in Å^2 for the accessible surface area for each atom in a protein using the algorithm of Lee and Richards.[16] The maximum solvent accessibility possible for each type of residue side-chain was determined using the exposed residue on the loop between two polyglycine strands of an antiparallel β-sheet, with the exposed residue replaced by each type of residue in turn. The resulting peptide was put through AREAIMOL to calculate the accessibility of the residue's side-chain. The accessibilities of the atoms making up the side-chain of each residue in the data set proteins are added together by the file-creation program, and the percentage of the side-chain exposed is then calculated using the maximum figures, together with an accessibility code B, S, or E, indicating buried, surface, and exposed respectively: an accessibility of less than 20.5% is coded as B, 67% to 100% inclusive as E, and those between as S. These codes can be specified in a query if required, as can the codes I and X: I (Internal) code corresponds to B and S, while X (eXternal) corresponds to S and E.

Various problems were encountered before the final accessibilities could be calculated, of which two were of particular importance. The first is that AREAIMOL cannot deal with very large PDB files. All files over 10 MB after generating the correct quaternary structure (see below) (which includes virus proteins, other large proteins and proteins with a very large number of subunits) were reduced in size by manually analyzing the file to determine which chain or chains of the protein made up the basic subunit of that protein; all residues greater than 25 Å from the residues of the basic subunit were then removed from the file before the solvent accessibility was computed. The postprocessing file was then further reduced to include only whole residues within 20 Å of the basic subunit, before the creation of the search file. The second problem is that a protein must be in the probable active quaternary structure form before the solvent accessibility of each of its atoms can be computed; this is discussed below.

Generating the Correct Quaternary Structure. The files in the PDB are deposited in such a way that the crystallographic coordinates given are the contents of the asymmetric unit of the unit cell of the crystal. For multisubunit proteins they may therefore comprise more or less than the true active biomolecule but contain at least the minimum needed to be able to recreate the entire active multimer by symmetry transformations that are often given in the PDB file as well. It is important when calculating the solvent accessible surface area of a protein to use the biologically relevant multimer of the protein to ensure that surfaces that, for example, are exposed in a monomer and buried in a dimer, are given the correct accessibility.

The generation of correct multimers was achieved by various means. Some PDB files contain matrices under REMARK 350 that can be used to transform the chains given, into a number of other chains that together make up

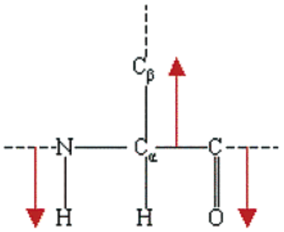


Figure 2. Main-chain vectors.

the known, or predicted, biologically relevant multimer for that protein. If no such transformation data are available in the PDB file, then one of two options can be used. If the file does not contain enough chains to produce the multimer, then a new file can be downloaded from the PQS (Protein Quaternary Structure) server, which is a database of the predicted and known, biologically active quaternary states of protein structures determined by X-ray crystallography.[17] Alternatively, if too many chains are present, then chains are removed, specified by the user, using a program written for this purpose. Taken together, these three procedures enable the production of the biologically active multimer for all those proteins in the data set for which such a multimer is known or predicted.

Addition of Main-Chain Vectors. Thus far, the vector representations used in ASSAM have described only the side-chains, but it is clearly desirable that main-chain vectors should also be included, so that searches can include backbone atoms, e.g., when they occur on the surface of the protein and are thus potentially involved in binding or activity. This, in turn, requires that we are able to identify those sections of the main-chain that are on the surface of a protein as, otherwise, very large numbers of (buried) hits may be retrieved in searches involving main-chain information.

The main-chain section of an amino acid is represented in ASSAM by three vectors that, taken together, describe the direction of the potential hydrogen bond donors and acceptors and the direction of the side-chain. The vectors, which are illustrated in Figure 2, comprise the following: one from the nitrogen (N) of the peptide bond to its hydrogen (H) (the position of which is computed from the positions of the C_{x-1} to N_x and Ca_x to N_x bonds); one from the carbon (C) of the peptide bond to its oxygen (O); and one from the α-carbon (Cα) to the β-carbon (Cβ) of the side-chain.

Each line of the main-chain vector file describes one residue of the protein and is composed of an 'm' to indicate main-chain, the three letter name of the residue, the secondary structure of the residue, and the chain label and residue number, followed by the coordinates of the pseudo-atom nodes N (S) and H (E), C (S) and O (E), and Cα (S) and Cβ (E). No M pseudo-atom coordinates are given, but these can be calculated at the time of searching if required and can be specified in a query. The three vectors have their individual accessibilities calculated by adding together the accessibilities of their constituent atoms, excluding the hydrogen of the NH vector which has no measured accessibility. As with the side-chain vector files, the accessibilities are converted into percentages using a maximum accessibility figure. For the main-chain the figures are taken from a glycine residue positioned at both ends of an isolated α-helix. The same maximum figures are used for all NH and CO vectors, regardless of residue type. The maximum figures for CαCβ vectors are taken from those used in the side-chain representation, and the CαCβ vectors are labeled with a solvent accessibility code that refers to the entire side-chain. The CαCβ solvent accessibilities do, therefore, depend on the specific residue they are a part of. The solvent accessibility of the CαCβ bond on its own would not be very meaningful in terms of matching like residues. Query patterns do not distinguish which residue a main-chain vector should be in. The accessibility codes are again defined as 0% to 20.5% for buried, 67% to 100% for exposed, and the remaining as surface.

EFFECTIVENESS OF RETRIEVAL

Having described the main components of the program, we now report the results obtained in a comprehensive set of test searches that was performed on 9932 PDB files downloaded from the Databank in October 2000; this test set comprised all of those structures then available after elimination of obvious duplicates, Cα-only files, and non-protein structures. The data set is updated periodically.

The searches test all the features that can be specified in an ASSAM query and are detailed in Table 1. In this table, and in the discussion below, the number of files with hits is the number of PDB files that matches are found in; there may be several occurrences of a particular motif in a particular file. The CPU time for each search are those obtained with the ASSAM search program, which is written in Fortran 77 and which runs under the Unix or Linux operating systems. The runs described here were carried out on a Silicon Graphics R12000 workstation with a single 300 MHz processor.

Query Patterns. Queries from known structures are used in searches 1−26 (denoted by S1−S26 in Table 1), and in S48−S65, while invented queries (i.e., residues and distance patterns specified by the user, rather than being extracted from a known protein structure) are used in S27−S47 and S66.

α-Chymotrypsin (PDB code 4CHA)[18] is used to create queries for S1−S18, S48−S51, and S63−S65. In S1−S14, the aspartic acid-histidine-serine active site triad (Figure 1) is used as the side-chain query motif, while in S48 and S49 the same active site triad is used, but with the NH main-chain groups of the three residues acting as the query vectors. In S15−S18 a random trio of residues from α-chymotrypsin, including a cysteine residue, is used as the query. In S50 and S51 the same three random residues are searched, but this time using one each of the CO, NH, and CαCβ main-chain vectors to represent them. S19 uses residues from two different chains of HIV-1 reverse transcriptase (1RTH).[19] Mastoparan-X, a 15-residue G protein-binding wasp venom (1A13),[20] is used in S20−S22. The three residues chosen are relatively highly exposed residues, due to the small size of the protein, and are hence used to test the S (surface) and E (exposed) solvent accessibility codes. Papain (1BP4)[21] is used to generate queries for S23−S26, using the catalytic triad found at its active site. Phosphate binding protein (1A40)[22] is used in S52−S62; the motif is composed of three of the clustered helix-end main-chain NH groups involved in the binding of the phosphate group.

Vector Type. Side-chain vectors are used to create queries for S1−S47 and S66, and main-chain vectors are used to

Table 1. Searches Designed To Test the Functionality of ASSAM[a]

query	DBS	tolerance	distances	hits	truncated	time
S1: 4CHA, HIS57_B ASP102_B SER195_B	unspecified	1.5	SS SE ES EE MM KK	326	1	762
S2: 4CHA, HIS57x_ ASP102x_ SER195x_	unspecified	1.5	SS SE ES EE MM KK	101	1	774
S3: 4CHA, HIS57xB ASP102xB SER195xB	unspecified	1.5	SS SE ES EE MM KK	69	1	753
S4: 4CHA, HIS57xX ASP102xX SER195xX	unspecified	1.5	SS SE ES EE MM KK	0	1	773
S5: 4CHA, HIS57xI ASP102xI SER195xI	unspecified	1.5	SS SE ES EE MM KK	101	1	759
S6: 4CHA, HIS57__ ASP102__ SER195__	unspecified	1.0	SS EE	389	1	516
S7: 4CHA, HIS57x_ ASP102x_ SER195x_	unspecified	1.0	SS EE	84	1	509
S8: 4CHA, HIS57xB ASP102xB SER195xB	unspecified	1.0	MM	143	1	472
S9: 4CHA, HIS57xB ASP102xB SER195xB	4	1.0	MM	0	0	72
S10: 4CHA, HIS57xB ASP102xB SER195xI	unspecified	1.0	MM	176	1	450
S11: 4CHA, HIS57xB ASP102xB SER195xI	4	1.0	MM	0	0	72
S12: 4CHA, HIS57xB ASP102xB SER195xB	15	5.0	KK	159	0	81
S13: 4CHA, HIS57xB ASP102xB SER195xB	10	2.0	KK	17	0	76
S14: 4CHA, HIS57xB ASP102xB SER195xB	unspecified	1.0	KK	95	1	458
S15: 4CHA, PHE41__ CSS136__ ARG145__	unspecified	1.0	SS EE	16	0	455
S16: 4CHA, PHE41s_ CSS136s_ ARG145x_	unspecified	1.0	SS EE	3	0	433
S17: 4CHA, PHE41_B CSS136_B ARG145_S	unspecified	1.0	SS EE	6	0	435
S18: 4CHA, PHE41sB CSS136sB ARG145xS	unspecified	1.0	SS EE	1	0	431
S19: 1RTH, ALAA33h_ THRA69x_ LEUB92x_	unspecified	1.0	SS EE	48	9	757
S20: 1A13, MET9_S ALA10_E LEU14_E	4	1.0	SS EE	0	0	72
S21: 1A13, MET9_S ALA10_E LEU14_E	10	1.0	SS EE	0	0	78
S22: 1A13, MET9_S ALA10_E LEU14_E	unspecified	1.0	SS EE	1	7	629
S23: 1BP4, CYS25__ HIS159__ ASN175__	unspecified	2.0	SS EE	116	0	429
S24: 1BP4, CSS25__ HIS159__ ASN175__	unspecified	2.0	SS EE	6	0	415
S25: 1BP4, CYH25__ HIS159__ ASN175__	unspecified	2.0	SS EE	111	0	427
S26: 1BP4, CYH25__ HIS159__ ASN175__	4	2.0	SS EE	1	0	73
S27: TRP__ TRP__ TRP__ (all MM 3.4 Å)	unspecified	0.5	MM	0	0	386
S28: TRP__ TRP__ TRP__ (all MM 3.4 Å)	unspecified	2.0	MM	4	0	387
S29: TRP__ TRP__ TRP__ (all KK 3.4 Å)	unspecified	1.0	KK	1	0	390
S30: ARO__ ARO__ ARO__ (all MM 3.4 Å)	4	2.0	MM	5	0	126
S31: BAS__ BAS__ BAS__ (all MM 3.4 Å)	4	2.0	MM	11	0	127
S32: BAR__ BAR__ BAR__ (all MM 3.4 Å)	4	2.0	MM	2	0	126
S33: BAR__ BAR__ BAR__ (all MM 3.4 Å)	4	5.0	MM	110	0	130
S34: ACI__ ACI__ ACI__ (all MM 3.4 Å)	4	2.0	MM	46	0	127
S35: AMI__ AMI__ AMI__ (all MM 3.4 Å)	4	2.0	MM	2	0	126
S36: AMI__ AMI__ AMI__(all MM 3.4 Å)	4	5.0	MM	60	0	128
S37: PHO__ PHO__ PHO__ (all MM 3.4 Å)	4	2.0	MM	1	0	127
S38: PHO__ PHO__ PHO__ (all MM 3.4 Å)	4	5.0	MM	278	0	132
S39: PHX__ PHX__ PHX__ (all MM 3.4 Å)	4	2.0	MM	90	0	147
S40: ASX__ ASX__ ASX__ (all MM 3.4 Å)	4	2.0	MM	38	0	127
S41: GLX__ GLX__ GLX__ (all MM 3.4 Å)	4	2.0	MM	7	0	126
S42: ACX__ ACX__ ACX__ (all MM 3.4 Å)	4	2.0	MM	104	0	130
S43: ROH__ ROH__ ROH__ (all MM 3.4 Å)	4	2.0	MM	44	0	127
S44: RSH__ RSH__ RSH__ (all MM 3.4 Å)	unspecified	5.0	MM	119	0	130
S45: ROY__ ROY__ ROY__ (all MM 3.4 Å)	4	2.0	MM	58	0	127
S46: RSY__ RSY__ RSY__ (all MM 3.4 Å)	4	2.0	MM	133	0	131
S47: wil__ wil__ wil__ (all MM 3.4 Å)	4	2.0	MM	414	0	146
S48: 4CHA, nh57xE nh102xE nh195xE	unspecified	1.5	SS SE ES EE MM KK			Halted
S49: 4CHA, nh57xB nh102xB nh195xI	unspecified	1.5	SS SE ES EE MM KK			Halted
S50: 4CHA, co41_S nh136_B ab145_B	unspecified	1.0	SS EE			Halted
S51: 4CHA, co41_S nh136_B ab145_B	4	1.0	SS EE	9	0	199
S52: 1A40, nh38h_ nh140h_ nh141h_	unspecified	1.0	SS EE	3289	1419	4150
S53: 1A40, nh38h_ nh140h_ nh141h_	unspecified	0.5	SS EE	152	1419	3897
S54: 1A40, nh38h_ nh140h_ nh141h_	5	1.0	SS EE	24	0	193
S55: 1A40, nh38h_ nh140h_ nh141h_	4	1.0	SS EE	11	0	162
S56: 1A40, nh38h_ nh140h_ nh141h_	2	1.0	SS EE	0	0	166
S57: 1A40, nh38_B nh140_B nh141_B	4	1.0	SS EE	78	0	163
S58: 1A40, nh38__ nh140__ nh141__	4	1.0	SS EE	79	0	164
S59: 1A40, nh38hB nh140hB nh141hB	unspecified	1.0	SS EE	3283	1419	4191
S60: 1A40, nh38hB nh140hB nh141hB	unspecified	0.5	SS EE	152	1419	3916
S61: 1A40, nh38hB nh140hB nh141hB	4	1.0	SS EE	11	0	165
S62: 1A40, nh38hB nh140hB nh141hB	5	1.0	MM	336	0	206
S63: 4CHA, cab57x_ ASP102x_ SER195x_	unspecified	1.0	SS EE	218	1326	6440
S64: 4CHA, cab57xB ASP102xB SER195xB	unspecified	1.0	SS EE	128	1326	6148
S65: 4CHA, cab57xB cab102xB SER195xB	unspecified	1.0	SS EE	2074	1269	12508
S66: PHE__ PHE__ PHE__ (all MM 3.4 Å)	4	2.0	MM	2	0	126

[a] The ‘Query’ column gives the query number, the PDB code, and the residues involved in the query pattern. This includes the name and number of the residue/vector, the secondary structure code, and the solvent accessibility code. If the residues come from more than one chain, then the chain label is also given, before the residue number. ‘DBS’ is the distance to a known binding site, as specified in the query, in Å (if specified). The ‘Tolerance’ is that placed (±) on the specified distances, in Å. The distances used are those specified in the query and hence matched during the search (S_1S_2, S_1E_2, E_1S_2, E_1E_2, M_1M_2, and K_1K_2). The ‘Hits’ column gives the number of PDB files retrieved in the search, and ‘Truncated’ is the number of files that were truncated by ASSAM, denoting that only a portion of the file will have been searched for that query. ‘Time’ is in CPU seconds, with ‘Halted’ denoting a search that was stopped after 24 h.

create queries for the remaining searches. Side-chain searches are generally faster than main-chain searches, primarily because matching main-chains require the consideration of all residues in a protein, whereas side-chains require that only matching residue types be considered; indeed, three of the main-chain searches were aborted after running for 24 h. Main-chain queries are especially time-consuming when all intervector distances are used in the matching and when a mixture of all three vector types is used in one search (requiring that the program looks at all residues three times to find matching motifs). In these cases, the new facilities described here—such as distance-to-known-binding-site or solvent accessibility—proved highly effective in reducing the search space, thereby obtaining acceptable run times and more relevant hits.

The current version of ASSAM truncates hit-files that have too much potentially matching structure, the number of such occurrences increasing greatly when main-chain vectors are used; in such cases, retrieved hits will be displayed for the section of protein analyzed before the ASSAM search was halted. In S50 and S51 the three vectors of the query are the three different main-chain vectors (NH, CO, and CαCβ), and a DBS value needs to be specified if the query is to run in a reasonable length of time. The one truncated file seen in nine of the searches S1−S14 involving the catalytic triad from 4CHA is the large file 1C8N (tobacco necrosis virus coat protein).

The S51 query used the same residues as S15−S18, but with side-chain vectors replaced by different main-chain vectors. None of the nine hits (1A7A (S-adenosylhomocysteine hydrolase), 1AQC (peptide binding protein), 1EG5 (aminotransferase), 1GAH (glucoamylase), 1GAI (glucoamylase), 1QR6 (NAD(P)-dependent malic enzyme), 1X11 (peptide binding protein), 2HDH (L-3-hydroxyacyl coA dehydrogenase), and 7RSA (ribonuclease A)) retrieved in S51 are retrieved in S15−S18, indicating that searching for the main-chain of a query can retrieve very different hits to searching for the side-chains. S51 did not find 4CHA itself because a DBS of 4.0 Å is specified.

The use of queries composed of a mixture of side-chain and main-chain vectors is tested in S63−S65, using the catalytic triad from 4CHA. In S63 and S64, residue 57 is represented by the CαCβ vector and residues 102 and 195 are represented by their side-chains; while in S65, residue 102 is also converted into the CαCβ vector. Not surprisingly, the number of hits retrieved increases when one of the side-chains (residue 57) is replaced by the more general CαCβ main-chain vector (compared to S7) and increases further when two side-chains (residues 57 and 102) are replaced by the CαCβ vector. There is also a big increase in the CPU time as the two side-chains are replaced by the main-chain vectors. Surprisingly, the number of truncated files decreases slightly, from 1326 to 1269, when two main-chain vectors are used instead of one. These searches show that ASSAM is capable of finding matches to queries that involve mixtures of side-chain and main-chain vectors as well as queries that involve just one of these vector types.

Secondary Structure. Queries specifying: residues involved in α-helical secondary structure (h) are used in S19, S52−S56, and S59−62; residues involved in β-sheet (s) are specified in S16 and S18; that the secondary structure of residues is neither α-helix nor β-sheet (x) are used in S2−S5, S7−S14, S16, S18, S19, S48, S49, and S63−65. In addition, secondary structure is left unspecified in S1, S6, S15, S17, and S20−S47.

Searches that do not specify secondary structure retrieve hits with a mixture of secondary structure designations, for example, S1 retrieves hits that include many secondary structure patterns, such as hxx (in 1A0J, trypsin) and hss (in 1A46, α-thrombin). Secondary structure is seen to reduce the number of hits retrieved in a search, compared to the same search with unspecified secondary structure. The addition of secondary structure also slightly decreases search times, as this reduces the number of nodes in each file that could match the query pattern.

The only variable that changes in S1−S5 is the addition of various combinations of secondary structure and solvent accessibility information. The number of hits retrieved is reduced to a greater extent when both types of information are used, as compared to just one type, and adding secondary structure alone reduces the number of hits to more of an extent than adding solvent accessibility alone.

The only variable altered in S6 and S7 is the secondary structure designation. When the secondary structure is specified as x (neither α-helix nor β-sheet), the number of retrieved hits decreases from 389 to 84, and the search time decreases by seven seconds. In S15−S18 the use of the cysteine residue CSS is tested, but secondary structure and solvent accessibility are also used. The number of hits is reduced as secondary structure and solvent accessibility are added, and, as is also seen above, secondary structure has more of an effect. Finally, adding the secondary structure and solvent accessibility detail at the same time reduces the number of hits to just one, 4CHA itself.

The use of secondary structure with main-chain queries is tested in S52−S62. At a tolerance of 1.0 Å, using distances S_1S_2 and E_1E_2, and a DBS of 4.0 Å, the same search is run with various levels of secondary structure detail. In S58 no secondary structure or solvent accessibility information is given and there are 79 hits. Adding native secondary structure (all-helix, S55) reduces the hits to 11. S57 with native solvent accessibility (all buried) finds 78 hits, and S61 with both secondary structure and solvent accessibility specified also finds 11 hits. These four searches show that, for this data set, all the hits except one have matching motifs that are buried and that all hits in helices are buried.

Disulfide Bridges. In queries involving cysteine residues, disulfide bridges are specified in S15−S18 and S24, and free cysteines are specified in S25 and S26. In S15−S18 a random trio of residues from α-chymotrypsin, including a cysteine residue, is used as the query. These searches are designed purely to ensure that the correct form of cysteine can be retrieved, in this case CSS; and all cysteine residues in retrieved hits are indeed in disulfide bridges. S23 labels the cysteine residue as CYS, i.e., either form of cysteine is acceptable in a retrieved hit. The hits retrieved from this search contain disulfide bridges and free cysteines, as expected.

The difference in hits retrieved using the various forms of the cysteine residue is explored in S23−S26. In S23, 116 files are retrieved when the more general CYS is specified, while six files are retrieved with CSS and 111 are retrieved with CYH. These 111 hits are reduced to one when a DBS of 4.0 Å is required. The number of hits quoted here is the number of PDB files that contain matches to the query motif: in this case, one of the files contains both a CSS and a CYH match, explaining the 116 hits from the CYS search (S23) compared to the 117 hits from the CSS search and the CYH search combined (S25 and S26).

The use of the more specific CSS or CYH labels reduces the search times very slightly. However, there are usually only a few cysteine residues in a protein, compared to other residue types, so this distinction does not affect search speeds to any extent.

Solvent Accessibility. Solvent accessibility is set to the following: to buried (B) in S1, S3, S8−S14, S17, S18, S49−S51, S57, S59−S62, S64, and S65; to surface (S) in S17, S18, S20−S22, S50 and S51; and to exposed (E) in S20−S22 and S48. In addition, ‘I’ is specified in S5, S10, S11, and S49 and ‘X’ in S4. The use of solvent accessibility very slightly reduces search times. The addition of solvent accessibility information decreases the number of hits retrieved; for example, S63 (with no solvent accessibility labels) retrieves 101 hit files, whereas S63 (with B labels) retrieves 69 hit files.

In S1−S5, the only variable that changes is the addition of various combinations of secondary structure and solvent accessibility. The exchange of B for I increases the number of hits from 69 to 101, in S3 and S5, respectively; the exchange of B for X (for eXternal) unsurprisingly produces no hits. With the same search pattern in S9−S11, the replacement of one B with an I has no effect until the DBS restriction is lifted.

The use of the residue CSS is tested in S15−S18, but secondary structure and solvent accessibility are also used. The number of hits decreases as secondary structure and solvent accessibility are added; this is also the case for S52−S62, which test the use of solvent accessibility with main-chain queries. The serine in serine protease active triads is usually described as a surface residue.[23] To attempt to include serine residues that are more exposed than buried, the solvent accessibility code I (Internal: buried or surface) is used in S10, this increasing the number of files with hits from 143 to 176.

Distance-to-Known-Binding-Site. A required distance-to-known-binding-site (DBS) is specified in S9, S11−S13, S20, S21, S26, S31−43, S45−S47, S51, S54−S58, S61, and S62. These searches retrieve hits that are all within the specified distance from a known bound heteroatom or from an author-designated site. This feature is useful to reduce the hits to only those at known binding sites but will fail to find hits that are at previously unlabeled sites. Searches that do not specify a DBS find matches at any point in a protein structure. The use of a DBS specification greatly reduces search times, especially when used in conjunction with main-chain queries. As would be expected, the smaller the DBS that is set, the greater the reduction in search time, and the smaller the number of retrieved hits.

In S1−S14, the catalytic triad from 4CHA is used as the query motif. There are no retrieved hits (S9) when native secondary structure and solvent accessibility are combined with a 4.0 Å DBS and a 1.0 Å tolerance; there are 143 hits when the DBS constraint is lifted.

The difference in the number of hits retrieved using the various forms of cysteine residues is explored in S23−S26. As described above, 111 files are retrieved with CYH, and these 111 hits are reduced to one when a DBS of 4.0 Å is added. The 4.0 Å DBS criterion reduces the hits to zero in S9 (from 143 in S8) and also reduces the hits to zero in S11 (from 176 in S10). S12 and S13, using the same query as S8, show that increasing the tolerance, increasing the DBS, and using the K_1K_2 distance, all increase the number of hits retrieved. The specification of a DBS can sometimes be too precise, especially as this information is not available for every protein; thus, in S20−S22, the single hit is not retrieved until all DBS requirements are removed.

Many of the invented generic residue queries are run using a 4.0 Å DBS specification, so as to reduce the time taken and the number of hits retrieved. In S50 and S51 it can be seen that this query involving the three different main-chain vectors takes over 24 h until a 4.0 Å DBS is used; then, nine hits are retrieved in 199 s. Similarly, in the NH vector queries in S52−S62, the addition of more specific DBS values reduces the number of hits retrieved.

Other Search Facilities. The set of 66 searches also allowed us to analyze the effects of variations in the distance tolerances allowed for a match, of variations in the types of intervector distances specified in a query, and validated the effectiveness of the generic residue classes discussed previously. These additional analyses are described by Spriggs.[23]

Searches for the Serine Protease Catalytic Triad. We now describe in some detail searches for the serine protease catalytic triad, which is a common active site motif that consists of aspartic acid, histidine, and serine[24] and that is found in several families of enzyme. The catalytic triad query from 4CHA (Figure 1) was used again here: i.e, the HIS57, ASP102, and SER195 residues with S_1S_2, S_1E_2, E_1S_2, and E_1E_2 intervector distance measurements, at four different tolerance levels (0.5 Å, 1.0 Å, 1.5 Å, and 2.0 Å). 159 matching motifs in 149 PDB files were retrieved at 0.5 Å, 406 matching motifs in 348 PDB files at 1.0 Å, 559 matching motifs in 458 PDB files at 1.5 Å, and 767 matching motifs in 549 PDB files at 2.0 Å.

These results can be compared with those that would have been predicted from other information available in the PDB and in the literature. Specifically, the Unix grep command was used to search the entire contents of the 9932 PDB files for words and phrases indicating that a protein might be expected to contain an ASP−HIS−SER catalytic triad. Thus the following strings were searched for: ‘SERINE PROTEASE’, ‘TRYPSIN’, ‘CHYMOTRYPSIN’, ‘ELASTASE’, ‘KALLIKREIN’, ‘RAT MAST CELL’, ‘STREPTOMYCES GRISEUS’, ‘ALPHA LYTIC PROTEINASE’, ‘ALPHA-LYTIC PROTEINASE’, ‘SINDBIS VIRUS’, ‘SUBTILISIN’, ‘THERMITASE’, ‘MESENTERICOPEPTIDASE’, ‘FUNGAL LIPASE’, and ‘THROMBIN’. The retrieved files were checked individually before being labeled as predicted hits, giving a total of 450 structures that were expected to contain the ASP−HIS−SER catalytic triad; in fact, hits were found in 360 of them. Of the 90 files that did not produce hits, 37 were found to have been predicted incorrectly on further inspection, due to, for example, the exclusion of the catalytic domain from the structure. Of the remaining 53 files, the lack of a hit in 48 of them can be explained by looking at mutation, covalent modification, etc. at the active site, meaning that we achieved an overall recall (the fraction of the predicted structures that were retrieved) of 0.986, i.e., 360/(450−37−48). There were just five structures that were

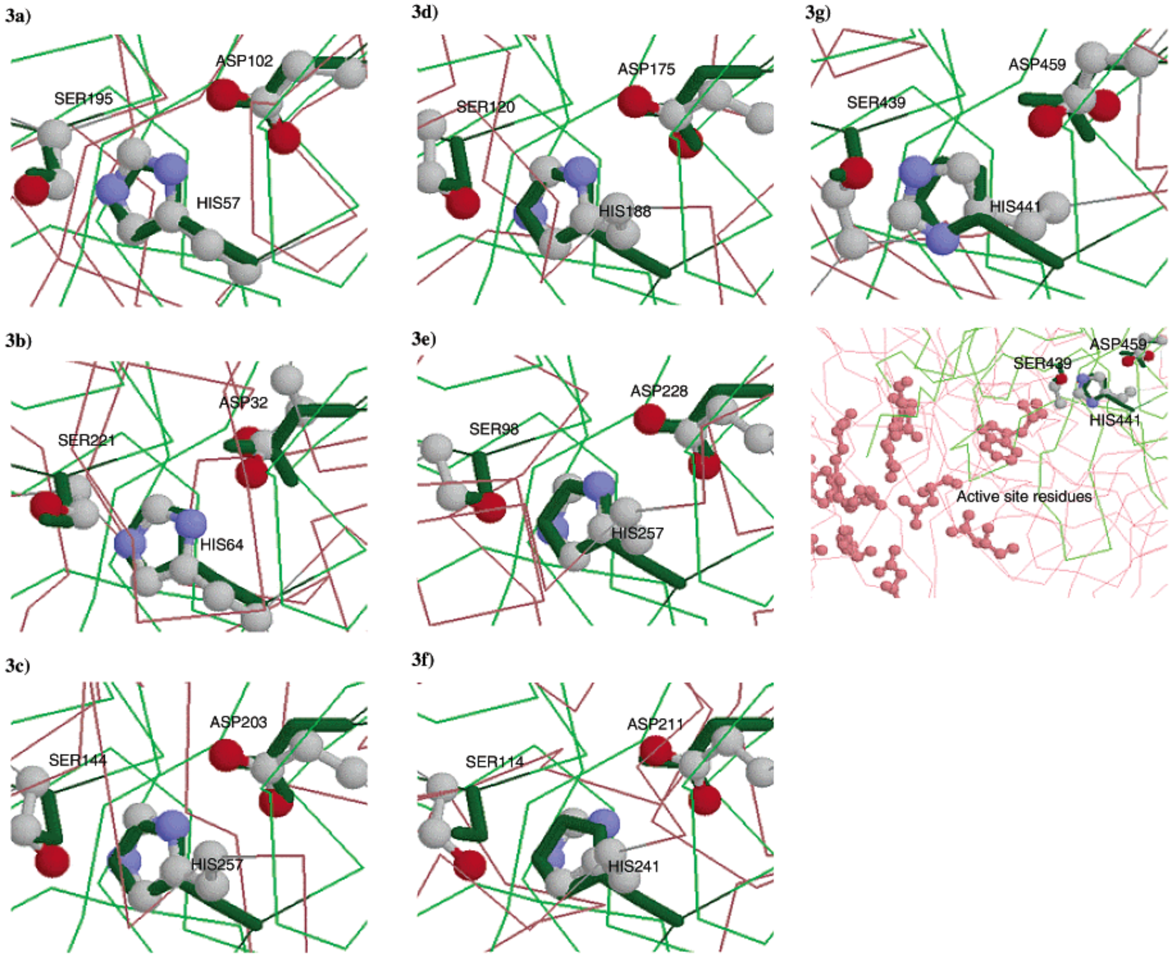


Figure 3. Retrieved motif in (a) trypsin (1AQ7), (b) subtilisin (1AF4), (c) fungal lipase (3TGL), (d) serine esterase (1AGY), (e) chloroperoxidase (1A7U), (f) thioesterase (1THT), and (g) neuraminidase (1KIT). The query pattern is dark green with the rest of the query protein in light green; matched side-chains in the hit protein are colored with C atoms gray, N atoms blue, and O atoms red. The lower figure in part (g) shows the matched side-chains in relation to the known active site residues for this hit protein (pink, ball-and-stick).

expected to contain the triad but that were not retrieved: 1EAP, a catalytic antibody/serine protease; 1EZM, an elastase; 1SAT, a serine protease; 1BT7, a serine protease; and 1DSU, a serine protease. However, all of these missing files were retrieved when the distance tolerance was increased: 1SAT, 1BT7, and 1DSU were found at a tolerance of 3.0 Å and the two remaining files at a tolerance of 5.0 Å.

In addition to the predicted hits, we also found unpredicted hits in 189 files. In retrospect, we should have been able to predict the presence of the triad in three of these files: 1QRX (‘ALPHA-LYTIC PROTEASE’, rather than the searched-for phrase ‘ALPHA-LYTIC PROTEINASE’), and 1BQY and 2PKC (‘SERINE PROTEINASE’, rather than the searched-for ‘SERINE PROTEASE’). The remaining 186 unpredicted hits are spread across 78 types of protein: types that were retrieved containing more than five hit files included 13 lipases, 10 oligo-peptide binding proteins, eight ribonucleoside reductases and serine esterases, seven haloperoxidases and transferases, and six carboxylesterases and G proteins (though it should be remembered, as always with the PDB, that some of the repeats will be from near-identical proteins because the PDB contains multiple examples of many proteins).

Examples of some of the hits from the catalytic triad searches are shown in Figure 3. Figure 3a shows the retrieved motif in trypsin (1AQ7[25]). This motif was retrieved at all tolerance levels and is a very close match to the chymotrypsin catalytic triad. This was a predicted hit as trypsin and chymotrypsin are homologous proteins, and the active site of trypsin is known to contain the same serine protease catalytic triad as the active site of chymotrypsin. Trypsin and chymotrypsin are likely to have come about through divergent evolution from a common ancestor; trypsin has a different specificity to chymotrypsin, cleaving peptide bonds on the carboxyl side of arginine and lysine residues.

Another predicted motif was found, again at all tolerance levels, in subtilisin (1AF4[26]), as shown in Figure 3b. In this case, the protein is a nonhomologous serine protease, and the bacterial subtilisin active site triad is the same as that in chymotrypsin through convergent evolution. Subtilisin is a much less specific serine protease in terms of which peptide bonds it will cleave.

A third predicted motif was found in fungal lipase (3TGL[27]), shown in Figure 3c. Fungal lipases are triacylglycerol lipases that cleave the ester bonds in triacylglycerol to produce glycerol and fatty acids. This motif was not retrieved until 1.0 Å tolerance. The retrieved motif in cutinase (1AGY,[28] Figure 3d) was not predicted but coincides with the labeled active site on the structure. Cutinase is a lipolytic serine esterase and so, by its similarity in function to the fungal lipases, it could perhaps have been predicted to contain a similar active site. Many proteolytic enzymes can also catalyze the unrelated hydrolysis of an ester bond,[24] and it is therefore not unusual that similarities are found between proteases and esterases at the active site.

An unpredicted hit was also found in chloroperoxidase (1A7U;[29] Figure 3e). The motif is found at 1.0 Å tolerance, and the three residues that are matched are the three residues of the labeled active site. Other haloperoxidases are also retrieved. Haloperoxidases catalyze the halogenation of organic compounds using halide ions and peroxides. The first step of the reaction mechanism has been shown to be analogous to the reaction catalyzed by serine proteases and esterases. This observation suggests that haloperoxidases have both esterase and haloperoxidase activity.

Another unpredicted, but active, motif was retrieved in thioesterase (1THT[30]), as shown in Figure 3f. Again, the retrieved motif is the labeled active site for the enzyme. This first thioesterase structure revealed a lipase-like catalytic triad for this subclass of hydrolases. Thioesterases are common hydrolytic enzymes involved in many biochemical pathways, for example, the biosynthesis of fatty acids. The same triad of residues has thus been retrieved in serine proteases, lipases, esterases, haloperoxidases, and thioesterases; in all cases the triad coincides with the labeled active site and so is involved in the catalytic activity of the enzyme.

The last example, Figure 3g, shows an unpredicted motif retrieved from neuraminidase (1KIT[31]). This is a hydrolase, with a labeled active site (highlighted in pink in Figure 3g), but the three residues of the retrieved motif are in no way involved in the known activity of this enzyme. In this case, the three residues happen to be near each other in the body of the protein structure, but the site is not active.

Finally here, a search was carried out that included the added detail of the 'native' (4CHA) secondary structure and solvent accessibility information; specifically, the query specified that all three residues are in neither α-helix nor β-sheet secondary structure, and all three residues have less than 20.5% of their surface area exposed to the solvent. This resulted in the retrieval of 73 hits, of which 42 (58%) were predicted to contain the query pattern, whereas the equivalent search without native structure detail discussed above retrieved 549 files, of which 360 (66%) were predicted to contain a hit. Thus, while the additional information has allowed a more specific search to be executed, it has not biased the search in favor of predicted hits, in this case.

Further details of all of the searches reported here are provided by Spriggs in her doctoral thesis.[23]

CONCLUSIONS

This paper describes the program ASSAM, which has been developed for searching for patterns of amino acid side-chains in the 3D structures in the PDB using techniques drawn from graph theory. Recent enhancements to ASSAM enable it to carry out searches that describe the main-chain structure, the secondary structure and solvent accessibility of residues, allowable distances from a known binding-site, disulfide bridges; and improved generic and wild-card queries. We have carried out an extensive series of searches that demonstrates the effectiveness of the new program and are now designing a Web interface to provide remote access to it.

ASSAM is based on the use of a subgraph isomorphism algorithm, which retrieves just those files that satisfy the structural constraints specified in the query. Alternatively, it would be possible to use a maximum common subgraph isomorphism algorithm to find those files that had a substructure of some minimal size, in terms of geometrically equivalent amino acids, in common with a query pattern, this being either an entire protein or some part of it such as a binding pocket. We have recently developed such a program, called ASPROTE, that allows similarity queries to be searched against the PDB, thus complementing the substructural queries that are searched for by ASSAM. This work will be reported shortly.

ACKNOWLEDGMENT

We thank Eleanor Gardiner, Helen Grindley, Andrew Poirrette, and David Rice for their assistance with the development of the ASSAM program and the referees for comments on the initial draft of this manuscript. We also thank the Engineering and Physical Sciences Research Council for funding and the Royal Society, Tripos Inc. and the Wolfson Foundation for hardware, software, and laboratory support. The Krebs Institute for Biomolecular Research is a designated center of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Christendat, D.; Yee, A.; Dharamsi, A.; Kluger, Y.; Savchenko, A.; Cort, J. R.; Booth, V.; Mackereth, C. D.; Saridakis, V.; Ekiel, I.; Kozlov, G.; Maxwell, K. L.; Wu, N.; McIntosh, L. P.; Gehring, K.; Kennedy, M. A.; Davidson, A. R.; Pai, E. F.; Gerstein, M.; Edwards, A. M.; Arrowsmith, C. H. Structural Proteomics of an Archaeon. *Nature Struct. Biol.* **2000**, *7*, 903−909.

(2) Blundell, T. L.; Jhoti, H.; Abell, C. High-Throughput Crystallography for Lead Discovery in Drug Design. *Nature Reviews Drug Discov.* **2002**, *1*, 45−54.

(3) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins. *J. Mol. Biol.* **1990**, *212*, 151−166.

(4) Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. *J. Mol. Biol.* **1993**, *229*, 707−721.

(5) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A Graph-Theoretic Approach to the Identification of Three-Dimensional Patterns of Amino Acid Side-chains in Protein Structures. *J. Mol. Biol.* **1994**, *243*, 327−344.

(6) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Blum, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr.* **2002**, *D58*, 899−907.

(7) Artymiuk, P. J.; Grindley, H. M.; Poirrette, A. R.; Rice, D. W.; Ujah, E. C.; Willett, P. Identification of β-Sheet Motifs, of φ-Loops and of Patterns of Amino Acid Residues in Three-Dimensional Protein Structures Using a Subgraph-Isomorphism Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 54−62.

(8) Kleywegt, G. J. Recognition of Spatial Motifs in Protein Structures. *J. Mol. Biol.* **1999**, *285*, 1887−1897.

(9) Madsen, D.; Kleywegt, G. J. Interactive Motif and Fold Recognition in Protein Structures. *J. Appl. Crystallogr.* **2002**, *35*, 137−139.

(10) Poirrette, A. R.; Artymiuk, P. J.; Grindley, H. M.; Rice, D. W.; Willett, P. Structural Similarity between Binding Sites in Influenza Sialidase and Isocitrate Dehydrogenase: Implications for an Alternative Approach to Rational Drug Design. *Prot. Sci.* **1994**, *3*, 1128−1130.

(11) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *16*, 31−42.

(12) Sayle, R. A.; Milner-White, E. J. RasMol − Biomolecular Graphics for All. *Trends Biochem. Sci.* **1995**, *20*, 374−376. The RasMol home page is at URL http://www.umass.edu/microbio/rasmol/index2.htm.

(13) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. The MIDAS Database System. *J. Mol. Graph.* **1988**, *6*, 2−12. Ferrin, T. E.; Huang,

C. C.; Jarvis, L. E.; Langridge, R. The MIDAS Display System. *J. Mol. Graph.* **1988**, *6*, 13−27.

(14) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure − Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.

(15) Bailey, S. The CCP4 Suite − Programs for Protein Crystallography. *Acta Crystallogr.* **1994**, *D50*, 760−763. The AREAIMOL home page is at URL http://www.ccp4.ac.uk/dist/html/areaimol.html.

(16) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(17) Henrick, K.; Thornton, J. PQS: a Protein Quaternary Structure File Server. *Trends Biochem. Sci.* **1998**, *23*, 358−361. The PQS home page is at URL http://pqs.ebi.ac.uk/.

(18) Tsukada, H.; Blow, D. M. Structure of Alpha-Chymotrypsin Refined at 1.68 Å Resolution. *J. Mol. Biol*. **1985**, *184*, 703−711.

(19) Ren, J. S.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High-Resolution Structures of HIV-1 RT from Four RT-Inhibitor Complexes. *Nature Struct. Biol.* **1995**, *2*, 293−302.

(20) Kusunoki, H.; Wakamatsu, K.; Sato, K.; Miyazawa, T.; Kohno, T. G Protein-Bound Conformation of Mastoparan-X: Heteronuclear Multidimensional Transferred Nuclear Overhauser Effect Analysis of Peptide Uniformly Enriched with 13C and 15N. *Biochemistry* **1998**, *37*, 4782−4790.

(21) Lalonde, J. M.; Zhao, B.; Smith, W. W.; Janson, C. A.; Desjarlais, R. L.; Tomaszek, T. A.; Carr, T. J.; Thompson, S. K.; Oh, H. J.; Yamashita, D. S.; Veber, D. F.; Abdel-Meguid, S. S. Use of Papain as a Model for the Structure-Based Design of Cathepsin K Inhibitors: Crystal Structures of Two Papain-Inhibitor Complexes Demonstrate Binding to S′-Subsites. *J. Med. Chem*. **1998**, *41*, 4567−4576.

(22) Ledvina, P. S.; Tsai, A. L.; Wang, Z.; Koehl, E.; Quiocho, F. A. Dominant Role of Local Dipolar Interactions in Phosphate Binding to a Receptor Cleft with an Electronegative Charge Surface: Equilibrium, Kinetic, and Crystallographic Studies. *Protein Sci*. **1998**, *7*, 2550−2559.

(23) Spriggs, R. V. *Development of the ASSAM and ASPROTE Programs for Protein Tertiary-Structure Searching*. Ph.D. Thesis, University of Sheffield, in preparation.

(24) Stryer, L. *Biochemistry*, 4th ed.; W. H. Freeman and Company: New York, 1995.

(25) Sandler, B.; Murakami, M.; Clardy, J. Atomic Structure of the Trypsin-Aeruginosin 98-B Complex. *J. Am. Chem. Soc.* **1998**, *120*, 595−596.

(26) Schmitke, J. L.; Stern, L. J.; Klibanov, A. M. The Crystal Structure of Subtilisin Carlsberg in Anhydrous Dioxane and its Comparison with those in Water and Acetonitrile. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 4250−4255.

(27) Brzozowski, A. M.; Derewenda, Z. S.; Dodson, E. J.; Dodson, G. G.; Turkenburg, J. P. Structure and Molecular-Model Refinement of Rhizomucor-miehei Triacylglyceride Lipase − A Case-Study of the Use of Simulated Annealing in Partial Model Refinement. *Acta Crystallogr. B* **1992**, *48*, 307−319.

(28) Longhi, S.; Czjzek, M.; Lamzin, V.; Nicolas, A.; Cambillau, C. Atomic Resolution (1.0 Å) Crystal Structure of *Fusarium solani* Cutinase: Stereochemical Analysis. *J. Mol. Biol.* **1997**, *268*, 779−799.

(29) Hofmann, B.; Tolzer, S.; Pelletier, I.; Altenbuchner, J.; Van Pee, K. H.; Hecht, H. J. Structural Investigation of the Cofactor-Free Chloroperoxidases. *J. Mol. Biol.* **1998**, *279*, 889−900.

(30) Lawson, D. M.; Derewenda, U.; Serre, L.; Ferri, S.; Szittner, R.; Wei, Y.; Meighen, E. A.; Derewenda, Z. S. Structure of a Myristoyl-ACP−Specific Thioesterase from *Vibrio harveyi*. *Biochemistry* **1994**, *33*, 9382−9388.

(31) Crennell, S.; Garman, E.; Laver, G.; Vimr, E.; Taylor, G. Crystal Structure of *Vibrio cholerae* Neuraminidase Reveals Dual Lectin-Like Domains in Addition to the Catalytic Domain. *Structure* **1994**, *2*, 535−544.

CI0255984