Protein Science (1997), *6*:2308–2323. Cambridge University Press. Printed in the USA.

TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites

ANDREW C. WALLACE,[1] NEERA BORKAKOTI,[2] AND JANET M. THORNTON[1,3]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England
[2]Roche Discovery Wellwyn, 40 Broadwater Road, Welwyn Garden City, Hertfordshire AL7 3AY, England
[3]Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England

(RECEIVED March 5, 1997; ACCEPTED July 11, 1997)

Abstract

It is well established that sequence templates such as those in the PROSITE and PRINTS databases are powerful tools for predicting the biological function and tertiary structure for newly derived protein sequences. The number of X-ray and NMR protein structures is increasing rapidly and it is apparent that a 3D equivalent of the sequence templates is needed. Here, we describe an algorithm called TESS that automatically derives 3D templates from structures deposited in the Brookhaven Protein Data Bank. While a new sequence can be searched for sequence patterns, a new structure can be scanned against these 3D templates to identify functional sites. As examples, 3D templates are derived for enzymes with an O-His-O "catalytic triad" and for the ribonucleases and lysozymes. When these 3D templates are applied to a large data set of nonidentical proteins, several interesting hits are located. This suggests that the development of a 3D template database may help to identify the function of new protein structures, if unknown, as well as to design proteins with specific functions.

Keywords: catalytic triad; database; 3D structure; protein function; template

The detection of recurring structural motifs in proteins is already well documented; indeed, they exist in all levels of protein structure, from primary to tertiary. At the primary level, there are now comprehensive protein sequence databases such as SWISS-PROT (Bairoch & Boeckmann, 1994) and OWL (Bleasby et al., 1994), that have been analyzed, using both automatic and manual pattern-matching and sequence-alignment techniques, to produce databases of recurring sequence motifs or templates such as PROSITE (Bairoch & Bucher, 1994) and PRINTS (Attwood et al., 1994). At the tertiary level, protein structure analysis, using both automatic (Orengo et al., 1993) and manual techniques, has enabled the creation of classifications such as CATH, SCOP, (Murzin et al., 1995) and DALI (Holm & Sander, 1993), which cluster proteins with similar folds. These databases are useful for the identification of biological function and fold recognition (Lemer et al., 1995). Neighborhood relationships can also be identified, using structural similarity methods, as exemplified by MMDB (Ohkawa et al., 1997).

Reprint requests to: Janet M. Thornton, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England; e-mail: thornton@biochem.ucl.ac.uk.

Many investigations have taken place at the substructural level of proteins, for example, analysis of 3D topologies of metal-binding sites in proteins and small molecules (see reviews by Glusker, 1991; Jernigan et al., 1994). PROMOTIF (Hutchinson & Thornton, 1996) provides automatic assignment and analysis of structural motifs in proteins, such as secondary structure, β-and γ-turns, β-hairpins, and disulfide bridges. In addition, there are already various algorithms to detect similar 3D arrangements of secondary structure in proteins. Some are comparison techniques that require the linear order of the amino acid sequences to be conserved (Matthews & Rossmann, 1985); others allow some degree of insertion/deletion in the protein sequence (Alexandrov et al., 1992), whereas others match by secondary structure elements (Mitchell et al., 1990).

However, as yet, a database of recurring 3D templates or motifs in proteins does not exist; these can be thought of as the 3D equivalent of the 1D templates found in the PRINTS and PROSITE databases. The number of protein 3D structures being solved by X-ray crystallography and NMR spectroscopy techniques is increasing rapidly; there are expected to be more than 20,000 by the turn of the century. This suggests that the need for a 3D equivalent of PROSITE is also growing; this would enable us to suggest functions of proteins whose roles are unknown as well as allow us to locate automatically functional regions and catalytic residues

within the protein structure. Such databases could address many different substructural aspects of proteins, such as enzyme active sites, ligand binding sites, loop conformation, and metal binding sites.

Previously, Artymiuk et al. (1994) have used a graph-theoretic approach for searching 3D patterns of side chains in protein structures. For example, they constructed a search template from the side-chain atoms of the Ser 195-His 57-Asp 102 catalytic triad of chymotrypsin and, depending on the allowed interatomic distance tolerances, different numbers of catalytic triads were identified from their data set. The template also revealed the existence of a second Ser-His-Asp triad in trypsinogen and chymotrypsinogen. A different structural comparison of the serine proteinases, using a less specific technique, has been performed by Fischer et al. (1994). Their method, derived from geometric hashing methods first described by Lamdan et al. (1988) for use in computer vision research, treats all C^{α} atoms in a protein as points in space and compares proteins purely on the geometrical relationships between these points. It can detect recurring substructural 3D motifs and was able to identify the structural similarities of the active sites of the trypsin-like and subtilisin-like serine proteases based solely on the similarities of the C^{α} geometries of their constituent residues.

We have previously shown that a Ser-His-Asp 3D enzyme active site template can be defined that will identify all the serine proteinases and lipase active sites in a set of PDB structures with the exclusion of all other noncatalytic Ser, His, and Asp interactions (Wallace et al., 1996). However, this method is not applicable to the case where the search atoms lie far apart in the protein structure because the search procedure becomes combinatorially intense. Here, an algorithm called TESS (TEmplate Search and Superposition) is described that is also based on the geometric hashing paradigm. TESS searches through a data set of Protein Data Bank structures (PDB; Bernstein et al., 1977) for any user-defined combination of atoms in space, from single atoms to multiple residues. The method is faster yet just as accurate as the method implemented by Wallace et al. (1996).

The new method is described together with its application to deriving enzyme active site consensus templates. First, we investigate those enzymes with a His-based "catalytic triad," namely the serine proteinases (Blow et al., 1969; Wright et al., 1969), cysteine proteinases (Drenth et al., 1968), lipases (Brady et al., 1990), and α/β-hydrolase (Ollis et al., 1992) enzymes. In addition, the active sites of eukaryotic and prokaryotic lysozymes (Johnson & Phillips, 1965; Matthews & Remington, 1974) as well as ribonucleases A and T_1 (e.g., Beintema et al., 1990; Arni et al., 1992) are explored.

In addition, using the TESS program, we have created a database of enzyme active site templates called PROCAT that is available using the World Wide Web at: http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html.

Results

The TESS algorithm

The TESS algorithm must address the following problem. Given the coordinates of the 3D query template and a protein structure, is there a match between them and what is the transformation, if any, that will best superimpose the template onto the relevant part of the protein molecule? This means that the atoms, residue types, and 3D coordinates of the template need to overlap with those of the stored PDB structure within a user-defined distance cut-off. The method we have devised to solve this problem is summarized in Figure 1 and consists of two stages: pre-processing and search with query template.

Pre-processing

In the pre-processing stage, a set of hash tables is compiled, containing geometrical information about the atoms of all the structures in the PDB. This avoids having to recalculate the data each time TESS is run and speeds up the comparison process considerably.

The atoms comprising a given template need to be defined with respect to a reference frame. For example, in the Ser O^{γ}-His side-chain-Asp O^{δ} consensus template of the serine proteinases and lipases, the His side chain defines the reference frame (Wallace et al., 1996). In fact, only three atoms of the side chain are used ($C^{\delta 2}$, C^{γ}, and $N^{\delta 1}$). Table 1 shows the three reference atoms used for each of the 20 standard amino acid side chains. The method used to create a hash table is summarized in Figure 2, using a His reference frame residue as an example. For each of the His residues in the protein structure, all atoms within a user-defined distance are identified; we chose 18 Å. The transformation matrix is calculated that places the His residue reference frame C^{γ} atom at the origin, with the $C^{\delta 2}$ and $N^{\delta 1}$ either side of the positive x direction. The same transformation matrix is applied to the atoms surrounding each His residue. A grid of 1 Å separation is placed around the His and the surrounding atoms and the position within the grid that each atom occupies is noted. To enable rapid searching through the hash table with the query template, the atoms in each grid are reordered according to their grid position number and are assigned an atomic label according to the residue and atom type. This information, along with the PDB file atom numbers, is stored in the hash table.

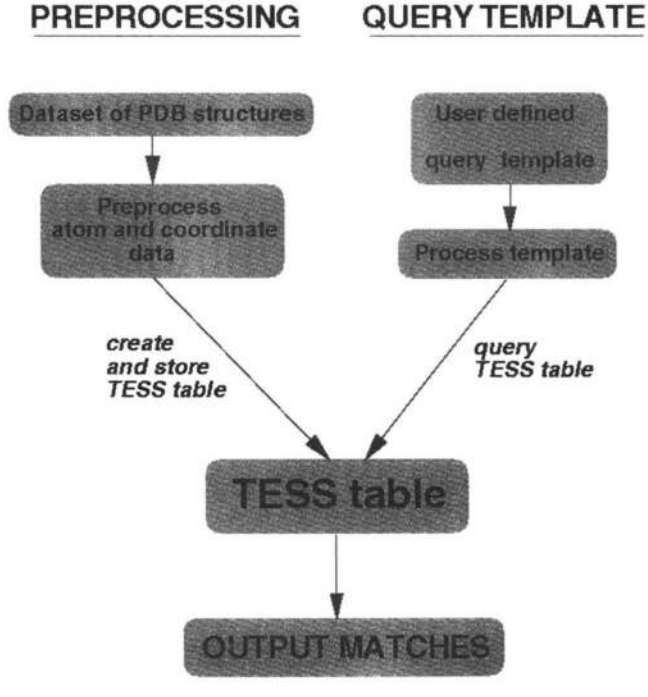

Fig. 1. Summary of the process involved when TESS searches through a data set of PDB structures for a user-defined 3D template.

Table 1. *Side-chain atoms used to define the reference frames for each standard amino acid*[a]

Residue	1	2	3	Residue	1	2	3	Residue	1	2	3
Ala (A)	N	C^{α}	C^{β}	Gly (G)	N	C^{α}	C	Pro (P)	N	C^{α}	C^{β}
Arg (R)	$N^{\eta 1}$	N^{ϵ}	$N^{\eta 2}$	His (H)	$C^{\delta 2}$	C^{γ}	$N^{\delta 1}$	Ser (S)	C^{α}	C^{β}	O^{γ}
Asn (N)	$O^{\delta 1}$	C^{γ}	$N^{\delta 2}$	Ile (I)	$C^{\gamma 1}$	C^{β}	$C^{\gamma 2}$	Thr (T)	$O^{\gamma 1}$	C^{β}	$C^{\gamma 2}$
Asp (D)	$O^{\delta 1}$	C^{γ}	$O^{\delta 2}$	Leu (L)	$C^{\delta 1}$	C^{γ}	$C^{\delta 2}$	Trp (W)	$C^{\delta 1}$	C^{γ}	$C^{\delta 2}$
Cys (C)	C^{α}	C^{β}	S^{γ}	Lys (K)	C^{ϵ}	C^{δ}	N^{ζ}	Tyr (Y)	$C^{\delta 1}$	O^{η}	$C^{\delta 2}$
Gln (Q)	$O^{\epsilon 1}$	C^{δ}	$N^{\epsilon 2}$	Met (M)	C^{ϵ}	C^{γ}	S^{δ}	Val (V)	$C^{\gamma 1}$	C^{β}	$C^{\gamma 2}$
Glu (E)	$O^{\epsilon 2}$	C^{δ}	$O^{\epsilon 1}$	Phe (F)	$C^{\epsilon 1}$	C^{γ}	$C^{\epsilon 2}$				

[a]Atoms in column 2 are transformed to the origin with atoms in columns 1 and 3 either side of the positive x direction.

Searching with a 3D query template

A query template is defined in terms of a side-chain reference frame and the atoms or other side chains in its vicinity. The coordinate template file is in standard PDB format and each atom can be specified in terms of an atom type (e.g., C, N, S, O) or a particular side-chain atom (e.g., Asp O^{δ}).

An example of a typical query template, taken from the active site of α-lytic proteinase, *1lpr*, is shown in Table 2. The reference frame residue, in this case a His, is defined by putting a "−1" in the atom number column corresponding to $C^{\delta 2}$, C^{γ}, and $N^{\delta 1}$. This enables comparison of the relative positions of the atoms surrounding the query template with those of the protein structures stored in the hash tables. For the nonreference frame atoms in the query template, it is possible to define any combination of both atom and residue types that are to be searched for at that point. This is in contrast to the method of Fischer et al. (1994), which is limited to only C^{α} atom searches.

To define which residues are to be located at a given template atom position requires the one-letter amino acid code corresponding to that amino acid to be placed after the coordinates of the query template. The atom type to be searched for is defined by placing one of the numbers in listed in Table 3 in the atom number column of the query template. For example, there will be a search for any non-carbon side-chain atom of residue type Ser or Asp at the coordinate position in the first row of the query template in Table 2.

The relative position of the atoms around the query His residue are calculated in the same way as the pre-processing stage; a grid is placed around the 3D template atoms, with the His reference frame at the origin, and the grid positions of each of the atoms are calculated. First the grid positions and then the atomic labels (atom and residue type) of the query template are compared to those of the proteins stored in the hash table; a hit occurs when all the atoms of the query template have a corresponding match with atoms in the protein structure. When comparing the grid positions, the situation is slightly complicated because it is not possible to tell where the atoms are lying in the grid boxes; they may be very near to one of the sides. Therefore, we have to search all neighboring grid boxes stored in the hash table to the central query atom grid box. Of course, the larger the distance cut-off, the more layers of neighboring grid boxes that will have to be searched. For each match, the matching PDB file is opened and the relevant atomic coordinates are transformed to the same reference frame and the RMS distance of the match is calculated.

Once the search is completed, a consensus template may be calculated by taking the mean coordinates of all the hits located.

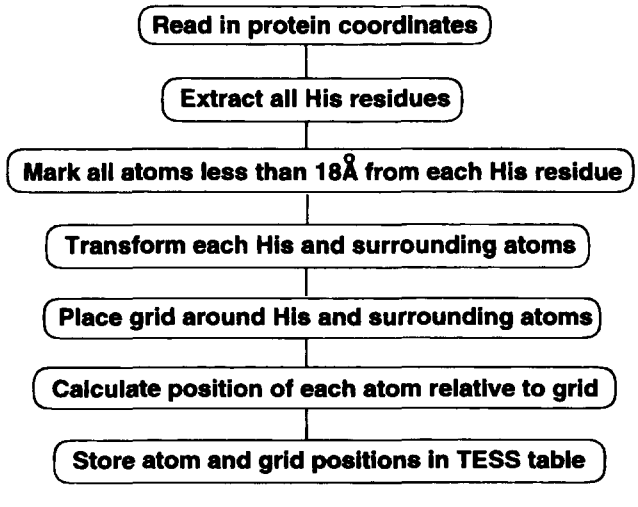

Fig. 2. Flow diagram showing the steps required in creating a hash table in the pre-processing stage, using the His reference frame residue as an example.

Performance of the TESS algorithm

TESS takes around 0.25 CPU seconds to locate a template match for a typical protein structure on a SGI R4400 Challenge. The run

Table 2. *An example of a typical query template taken from the active site of α-lytic proteinase, 1lpr*

Search option	Residue	Residue number	Atom	x	y	z	Residue search
1	Ser	195	O^{γ}	16.3	30.6	14.7	D
1	Asp	102	$O^{\delta 2}$	18.2	31.5	20.8	E
0	His	57	C^{β}	14.5	28.8	20.9	
−1	His	57	C^{γ}	15.0	29.3	19.5	
−1	His	57	$N^{\delta 1}$	16.2	30.0	19.3	
−1	His	57	$C^{\delta 2}$	14.3	29.1	18.3	
0	His	57	$C^{\epsilon 1}$	16.2	30.3	18.0	
0	His	57	$N^{\epsilon 2}$	15.1	29.8	17.4	

Table 3. *Search parameter numbers placed in the atom number column of the query PDB format file*[a]

Search option	Atom number
Template atom	−1
Search by atom type, i.e., $O^{\delta 1}$, $O^{\delta 2}$, O^{γ}	0
Search by non-carbon sidechain atom	1
Search by non-carbon atom	2
Search by specified atom, i.e., C, O, N	3
Search by non-carbon mainchain atom, i.e., O, N	4
Search by any mainchain atom	5
Search by any sidechain atom	6
Search by any atom type	7

[a]One of these numbers is placed against each of the atoms in the query template. This defines which atom types are to be searched for at the corresponding atom position. To search for different residue types at a given atom point requires the one-letter code of that amino acid to be placed after the coordinates in the query template file.

time is proportional to the order, $O(nh)$, where n is the number of atoms present in the template and h the number of hits. Each hash table uses around 70 megabytes of memory for the 3,019 structures found in the January 1995 PDB. The amount of memory used is proportional to l^3, where l is the length of the grid side, in our case, 36 Å.

The His-based catalytic triad

Template definition

The His-based catalytic triad was first identified in the serine proteinases (Blow et al., 1969; Wright et al., 1969) and comprises the residues Ser-His-Asp. Subsequently, other enzymes in the PDB, the α/β-hydrolase fold enzymes (Ollis et al., 1992), the esterase from *Streptomyces scabies* (Wei et al., 1995), and the cysteine proteinases (e.g., papain, Drenth et al., 1968) were shown to have a His-based catalytic triad. In fact, these triads can be generalized as Nu:-His-ELEC, where Nu: is the nucleophilic group and ELEC is the electrostatic group (equivalent to Asp in the Ser-His-Asp triad). Figures 3 and 4 show the heterogeneity in the orientation of the residues comprising these catalytic triads. We have divided these enzymes into five classes according to the residues that comprise their catalytic triads, which are Ser-His-Asp, Ser-His-Glu, Asp-His-Asp, Ser-His-Trp, and Cys-His-Asn. These are summarized in Table 4, which lists all PDB entries in the June 1996 release in each class. The α/β-hydrolase fold (Ollis et al., 1992) occurs in three of the five classes and, despite a low sequence identity, it suggests that the enzymes in this fold group have evolved from a common ancestor and have preserved the positions of the key catalytic residues.

Figure 5 is a 3D representation of the catalytic triads from classes 1, 2, 3, and 4. The His side chains have been superimposed, enabling us to compare the relative position of the nucleophilic and electrostatic side chains. Chymotrypsin 1*cho* (Fujinaga & James, 1987) represents class 1, although the dispositions of the side chains of the other class 1 members—namely subtilisin, serine-type carboxypeptidase, and lipase—are quite different to those of chymotrypsin and have been left out for clarity (Wallace et al., 1995). The catalytic triad of the cysteine proteinases (class 5) is

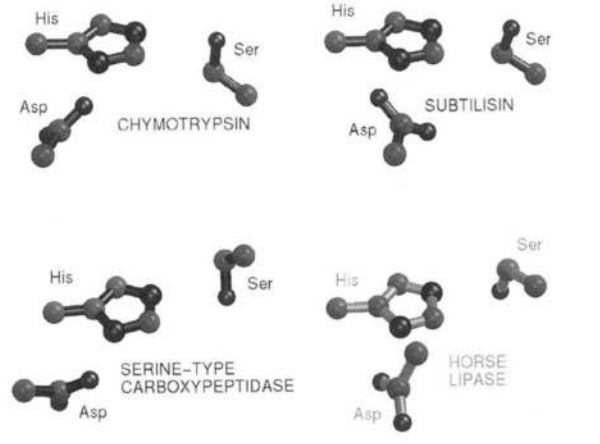

Fig. 3. Relative conformation of the catalytic triads from chymotrypsin 1*cho* (Fujinaga & James, 1987), subtilisin 2*sic* (Takeuchi et al., 1991), serine-type carboxypeptidase 3*sc*2 (Liao et al., 1992), and lipase 1*hpl* (Bourne et al., 1993). Diagram produced using MOLSCRIPT (Kraulis, 1991) and Raster3D (Bacon & Anderson, 1988)

different from those of classes 1, 2, 3, and 4 because the Cys 25 nucleophile interacts with the His $N^{\delta 1}$ rather than the $N^{\epsilon 2}$ (Fig. 4). The diagram shows that, although both side chains originate from different orientations in classes 1, 2, and 3, there is clustering of the functional atoms; the nucleophilic oxygens are all in the proximity of the acid/base catalyst His $N^{\epsilon 2}$ and the electrostatic residues are in a hydrogen bonding position with the His $N^{\delta 2}$. In contrast, the serine in the esterase (class 4) lies in a distinctly different position. We have shown previously that a Ser O^{γ}-His side-chain-Asp O^{δ} template can identify all serine proteinases and lipase catalytic triads of class 1 with the exclusion of all other interactions between Ser, His, and Asp side chains. Figure 5 suggests that this template might be extended to include classes 2 and 3, giving a class 1-2-3 consensus template. Sussman et al. (1991)

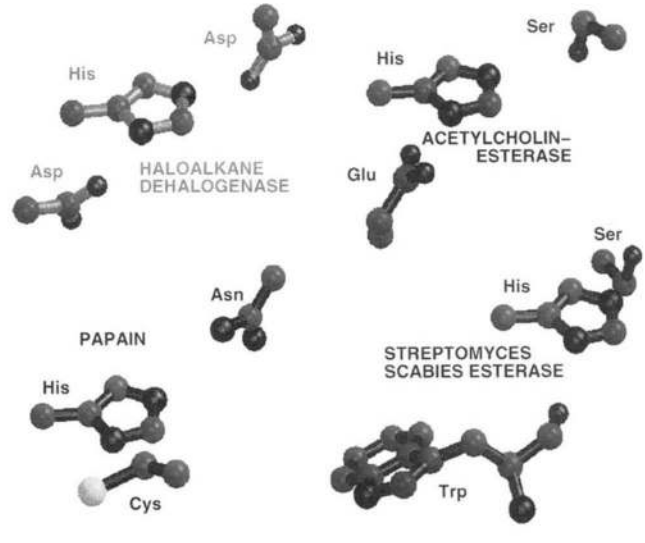

Fig. 4. Relative conformation of the catalytic triads from haloalkane dehalogenase (Verschueren et al., 1993), acetylcholinesterase (Sussman et al., 1991) and *Candida rugosa* lipase (Grochulski et al., 1993), papain (Drenth et al., 1968), and *S. scabies* esterase (Wei et al., 1995).

Table 4. *Five different classes of Nu:-His-ELEC catalytic triads found in the PDB (June 1996 release), where ELEC acts to perturb the pK_a of the acid/base His and Nu: is a nucleophilic group*

Class 1: Ser-His-Asp catalytic triad										
β-Sandwich trypsin-like fold										
Chymotrypsin E.C.3.4.21.1 31 chains										
1acb E	1cgi E	1cgj E	1chg	1cho E	1gcd	1gct A	1gha E	1ghb E	1gmc A	1gmd A	1gmh
2cga A	2cga B	2cha	2gch	2gct A	2gmt	3gch	3gct A	4cha A	4cha B	4gch	5cha A
5cha B	5gch	6cha A	6cha B	6gch	7gch	8gch					
Trypsin E.C.3.4.21.4 55 chains										
1bit	1bra	1brb E	1brc E	1btp	1btw A	1btx A	1bty	1btz A	1ept A	1ept B	1gbt
1mct A	1ppc E	1ppe E	1pph E	1sgt	1smf E	1tab E	1tbs	1tgb	1tgc	1tgn	1tgs Z
1tgt	1tld	1tng	1tnh	1tni	1tnj	1tnk	1tnl	1tpa E	1tpo	1tpp	1tps
1trm A	1trm B	1trn A	1trn B	1try	1tyn	2ptc E	2ptn	2tbs	2tga	2tgd	2tgp Z
2tgt	2trm	3ptb	3ptn	3tpi Z	4ptp	4tpi Z					
Thrombin E.C.3.4.21.5 58 chains										
1abi H	1abj H	1abj L	1aht H	1bbr H	1bbr E	1bbr K	1bbr N	1dwb H	1dwc H	1dwd H	1dwe H
1etr H	1ets H	1ett H	1fpc H	1fph H	1hag E	1hah H	1hai H	1hao H	1hap H	1hbt H	1hdt H
1hgt H	1hlt H	1hit K	1hrt H	1hut H	1ihs H	1iht H	1nrn H	1nro H	1nrp H	1nrq H	1nrr H
1nrs H	1ppb H	1thr H	1ths H	1tmh H	1tmt H	1tmu H	2hat H	2hgt H	2hgt L	2hnt C	2hnt E
2hnt F	2hpp H	2hpq H	2hpq P	2hpq L	2pf1	2pf2	2spt	3hat H	4htc H		
Plasmin E.C.3.4.21.7 9 chains										
1cea A	1cea B	1ceb A	1ceb B	1pk4	1pkr	1pmk A	1pmk B	2pk4			
α-lytic proteinase E.C.3.4.21.12 34 chains										
1gba A	1gbb A	1gbc A	1gbd A	1gbe A	1gbf A	1gbh A	1gbi A	1gbj A	1gbk A	1gbl A	1gbm A
1lpr A	1p01 A	1p02 A	1p03 A	1p04 A	1p05 A	1p06A	1p08 A	1p09 A	1p10 A	1p11 E	1p12 E
2alp	2lpr A	2p07	3lpr A	4lpr A	5lpr A	6lpr A	7lpr A	8lpr A	9lpr A		
Tissue kallikrein E.C.3.4.21.35 7 chains										
1ton	2kai A	2kai B	2pka A	2pka B	2pka X	2pka Y					
Pancreatic elastase E.C.3.4.21.36 26 chains										
1bma A	1eas	1eat	1eau	1eta A	1elb A	1elc A	1eld E	1ele E	1elf	1elg	1elt
1esa	1esb	1est	1inc	1jim	1nes E	2est E	3est	4est E	5est E	6est	7est E
8est E	9est										
Leukocyte elastase E.C.3.4.21.37 3 chains										
1hne E	1ppf E	1ppg E									
Complement factor D E.C.3.4.21.46 1 chains										
1dst											
Lysyl endopeptidase E.C.3.4.21.50 2 chains										
1arb	1arc										
Streptogrisin A E.C.3.4.21.80 5 chains										
1sgc	2sga	3sga E	4sga E	5sga E							
Streptogrisin B E.C.3.4.21.81 5 chains										
1sgp E	1sgq E	1sgr E	3sgb E	4sgb E							
Alternating α/β subtilisin-like fold										
Subtilisin family E.C.3.4.21.62 32 chains										
1cse E	1mee A	1mpt	1s01	1s02	1sbc	1sbh	1sbi	1sbn E	1sbt	1sca	1scb
1scd	1scn B	1sel A	1sel B	1sib E	1spb P	1spb S	1st2	1st3	1sub	1suc	1sud
1sup	2sbt	2sec E	2sic E	2sni E	2st1	3sic E	5sic E				
Endopeptidase K E.C.3.4.21.62 5 chains											
1pek E	1ptk	2pkc	2prk	3prk E							
Thermitase E.C.3.4.21.66 4 chains											
1tec E	1thm	2tec E	3tec E								
Thermitase E.C.3.4.21.68 10 chains											
1pk2	1pml A	1pml B	1pml C	1tpg	1tpk A	1tpk B	1tpk C	1tpm	1tpn		
Myeloblastin E.C.3.4.21.76 4 chains											
1fuj A	1fuj B	1fuj C	1fuj D								
α/β Hydrolase fold											
Serine-type carboxypeptidase E.C.3.4.16.5 6 chains											
1cpy	1whs A	1whs B	1wht A	1wht B	1yse						
Lipase E.C.3.1.1.3 10 chains											
1hpl A	1hpl B	1lpa A	1lpa B	1tah B	1tah A	1tah C	1tah D	3tgl	4tgl		
β-Barrel fold											
Coagulation factor XA E.C.3.4.21.6 4 chains											
1apo	1ccf	1hcg A	1hcg B								
U-plasminogen activator E.C.3.4.21.73 3 chains											
1kdu	1lmw B	1lmw D									
Glutamyl endopeptidase II E.C.3.4.21.82 1 chain											
1hpg A											
Class 2: Ser-His-Glu catalytic triad											
α/β Hydrolase fold											
Lipase E.C.3.1.1.3 10 chains											
1cle A	1cle B	1crl		1lpn	1lpp	1lps	1lpm	1lpo	1thg	1trh	
Acetylcholinesterase E.C.3.1.1.7 9 chains											
1ace	1acj	1ack	1acl	1amn	1fss A	1fss B	1mah A	1mah F			
Class 3: Asp-His-Asp catalytic triad											
α/β Hydrolase fold											
Haloalkane dehalogenase E.C.3.8.1.5 9 chains											
1edb	1edd	1ede	2dhc	2dhd	2dhe	2eda	2edc	2had			
Class 4: Ser-His-Trp catalytic triad											
α/β Fold											
Streptomyces scabies E.C.3.1.1.50 3 chains											
1esc	1esd	1ese									
Class 5: Cys-His-Asn catalytic triad											
$\alpha + \beta$ Cysteine proteinase fold											
Papain E.C.3.4.22.2 12 chains											
1pad	1pe6	1pip A	1pod A	1ppd	1ppn	1ppp	1stf E	2pad	4pad	5pad	6pad
Actinidin E.C.3.4.22.14 2 chains											
1aec	2act										
Caricain E.C.3.4.22.30 1 chain											
1ppo											

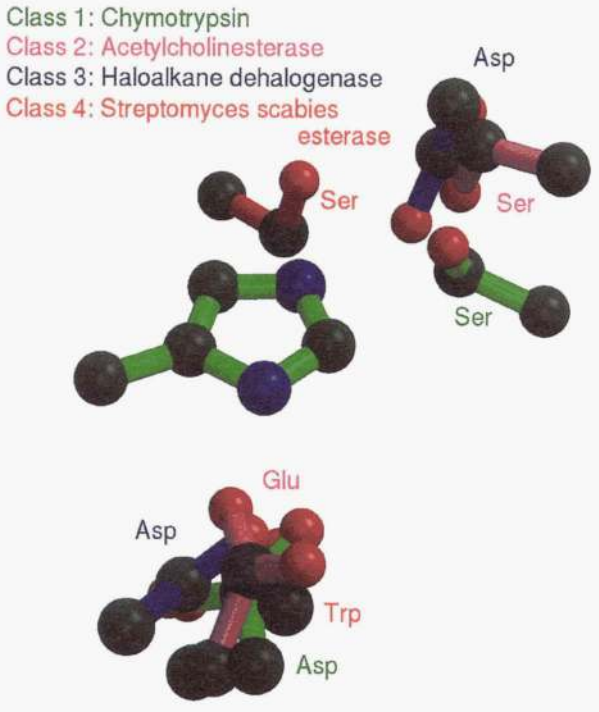

Fig. 5. Comparison of the catalytic triads from chymotrypsin 1*cho* class 1 (Fujinaga et al., 1987), acetylcholinesterase 1*ace* class 2 (Sussman et al., 1991), haloalkane dehalogenase 2*dhc* class 3 (Verschueren et al., 1993), and *S. scabies* esterase 1*esc* class 4 (Wei et al., 1995). The reference frame His residues are superimposed, allowing us to compare the relative conformations of the nucleophilic and electrostatic side chains.

noted that the main-chain atoms of both the Ser and His residues of the acetylcholinesterase triad originate from different sides when compared to the class 1 Ser-His-Asp catalytic triad. However, the difference in the handedness of these triads is not relevant to our consensus template because we are interested in the functional atoms and not the main-chain scaffold.

We derived such a template by averaging the functional consensus templates created for classes 1, 2, and 3. This class 1-2-3 template was compared with the consensus templates derived for each of the five classes and the RMS differences are summarized in Table 5. Figure 6 shows the distribution of individual RMS deviations (RMSDs) for each of the triads in all class 1, 2, 3, and 4 proteins from the class 1-2-3 template. The majority of triads in classes 1–3 have an RMSD between 0.3 Å and 1.3 Å—triads above this value are the class 1 triads with inhibitors bound to their active sites. All three triads in class 4 cluster around 2.0 Å RMS distance despite the fact that two of the structures are crystallized with inhibitors. Figure 5 indicates that, in class 4, the *S. scabies* triad Ser 14 residue is in a different position with respect to the His residue and this is reflected in the high RMSD. Given the different tertiary fold, it would seem likely that this different conformation is a consequence of convergent evolution. Whether the catalytic mechanism differs in this enzyme compared to classes 1, 2, and 3 is a matter for speculation (Wei et al., 1995).

Class 1-2-3 template search through the PDB

It is interesting to see if the arrangement of the atoms in the class 1-2-3 consensus template, in terms of the functional oxygens surrounding the His side-chain reference frame, occurs elsewhere in the PDB. We searched through a data set of 639 representative protein structures in the January 1995 PDB. Some of the structures in classes 1, 2, and 3 were present in this data set, others having been excluded on the basis of having greater than 95% sequence identity. A triad located in this nonidentical data set was considered interesting if its RMSD was less than 2.0 Å from the class 1-2-3 consensus template and there was an Asp or Ser side-chain atom in the position equivalent to Nu:, the nucleophilic group, and an Asp or Glu side-chain atom in the position equivalent to ELEC, the electrostatic group. We found a number of proteins, not members of classes 1, 2, and 3, having the characteristic Nu:-His-ELEC triad conformation. Two of the triads found resemble the Ser-His-Asp catalytic triad of the class 1 enzymes; cyclophilin 2*cpl* and immunoglobulin 2*ig*2. These triads have already been discussed in detail

Table 5. *Coordinates of the functional templates for the class 1-2-3 template and those of the five separate classes*[a]

Residue	Atom	*x*	*y*	*z*
Class 1-2-3 functional consensus template				
Ser/Asp	O^{γ}/O^{δ_1}	4.9	0.9	−0.3
Glu/Asp	$O^{\epsilon_1}/O^{\delta_1}$	−0.4	−3.8	0.1
Class 1: Ser-His-Asp RMS from class 1-2-3 template 0.6 Å				
Ser	O^{γ}	4.9	0.8	−0.3
Asp	O^{δ_2}	0.4	−3.7	0.1
Class 2: Ser-His-Glu RMS from class 1-2-3 template 0.6 Å				
Ser	O^{γ}	4.9	1.3	0.3
Glu	O^{ϵ_1}	−0.6	3.7	0.3
Class 3: Asp-His-Asp RMS from class 1-2-3 template 0.2 Å				
Asp	O^{δ_1}	4.8	0.8	−0.3
Asp	O^{δ_2}	−0.5	−3.6	0.2
Class 4: Ser-His-Trp RMS from class 1-2-3 template 2.1 Å				
Ser	O^{γ}	3.2	3.1	−0.4
Trp	O	−0.4	−3.4	−0.9
Class 5: Cys-His-Asn RMS from class 1-2-3 template 7.9 Å				
Cys	S^{γ}	0.5	−4.0	2.1
Asp	O^{δ_2}	4.4	2.3	0.5
His template residue				
His	C^{β}	−1.4	−0.1	−0.0
His	C^{γ}	0.0	0.0	0.0
His	N^{δ_1}	0.8	−1.1	0.0
His	C^{δ_2}	0.8	1.1	0.0
His	C^{ϵ_1}	2.1	−0.7	−0.0
His	N^{ϵ_2}	2.1	0.6	−0.0

[a]Each 3D template is superimposed onto the same His residue. The RMS distances of the individual templates from the class 1-2-3 template are also given.

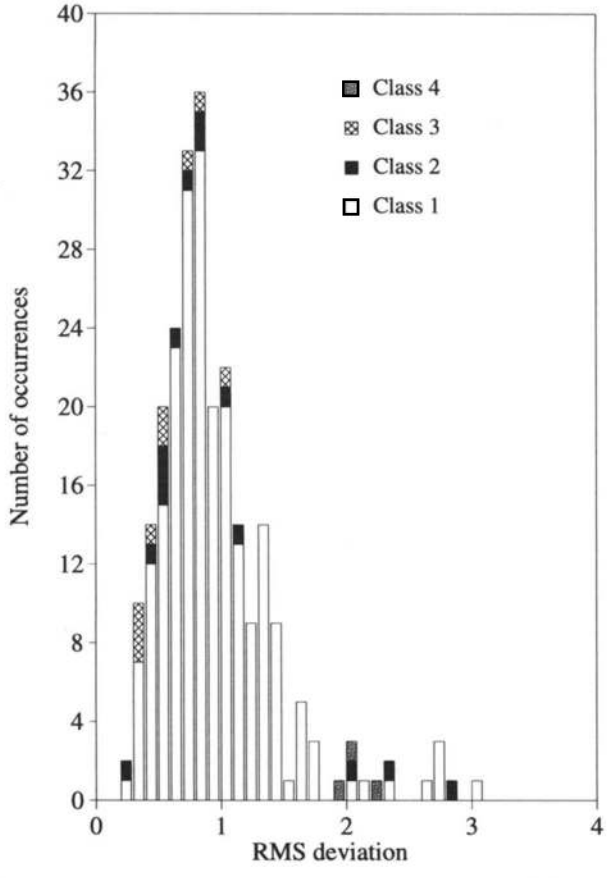

Fig. 6. Distribution of RMSDs of observed catalytic triads in all the structures in classes 1, 2, 3, and 4 from the class 1-2-3 template in the June 1996 PDB.

previously (Wallace et al., 1996). A further six proteins have triads resembling the Asp-His-Asp triads of the class 3 enzymes. These are listed in Table 6 in order of increasing RMSD from the template. We take the first of these, the nitrogenase molybdenum–iron protein, as an example.

Nitrogenase molybdenum–iron protein E.C.1.18.6.1

The MoFe protein from *Azotobacter vinelandii* is an α_2/β_2 tetramer and the X-ray structure has been determined to 2.7 Å resolution, PDB code 1*min* (Jongsun & Rees, 1992). It is part of the nitrogenase enzyme system, which consists of two metalloproteins, the molybdenum–iron (MoFe) protein and the iron Fe-protein. There is another cofactor, the P-cluster pair, which is thought to transfer electrons between the 4Fe:4S cluster Fe-protein and the MoFe cofactor. The precise catalytic mechanism is unknown, but the N_2 substrate is proposed to bind directly to the MoFe cofactor. The MoFe cofactor is bound to homocitrate and is surrounded by water molecules; it may be a source of protons for the formation of the NH_3 product.

A close-up view of the Asp-His-Asp triad and the MoFe and P-cluster cofactors is shown in Figure 7. The Asp 160-His 90-Asp 116 triad and the P-cluster are located within about 8 Å of each other. The triad has not been mentioned or implicated in the reaction course and whether it is involved in proton or electron transferral is unknown. Closer inspection reveals that the Asp 116 is, in fact, accessible to the surface of the protein, indicating that it could have access to solvent or ligands.

We checked the sequence of the MoFe protein (PDB code 1*min*) against the SWISS-PROT (Bairoch & Boeckmann, 1994) database using the automatic sequence alignment program BLAST (Altschul et al., 1990) and obtained several MoFe protein sequences, all derived from nitrogen-fixing bacterium, although the sequence identity with 1*min* may be as low as 25%. The His residue is conserved in all but one case, and there are several instances of a Ser or a Thr in the position of Asp 116. The Asp at position 160 is conserved in all cases.

Table 6. *List of the potential catalytic triads found when the 95% nonidentical data set of PDB structures was searched with the class 1-2-3 catalytic triad template*

Putative Asp-His-Asp catalytic triads					
Residue	Res. number	Atom	*x*	*y*	*z*
		Nitrogenase molybdenum–iron Protein RMSD 0.42 1*min*(D)			
Asp	160	$O^{\delta 2}$	5.0	1.4	−0.1
Asp	116	$O^{\delta 2}$	−0.5	−3.4	0.6
His	90	Side chain	—	—	—
		Pyruvate oxidase E.C.1.2.3.3 RMSD 0.58 1*pox*(D)			
Asp	69	$O^{\delta 1}$	4.7	1.8	−0.4
Asp	27	$O^{\delta 2}$	−0.4	−3.6	0.4
His	28	Side chain	—	—	—
		Macromycin RMSD 1.08 2*mem*			
Asp	100	$O^{\delta 2}$	5.8	1.9	−0.6
Asp	53	$O^{\delta 1}$	0.2	−3.5	−0.3
His	32	Side chain	—	—	—
		Protein R2 of ribonucleotide reductase E.C.1.17.4.1 RMSD 1.37 1*rib*(A)			
Asp	237	$O^{\delta 2}$	4.3	2.1	−0.2
Asp	84	$O^{\delta 1}$	−1.4	−4.8	0.7
His	118	Side chain	—	—	—
		Superoxide dismutase E.C.1.15.1.1 RMSD 1.51 1*sos*(D)			
Asp	124	$O^{\delta 1}$	4.1	2.1	−0.8
Asp	83	$O^{\delta 1}$	1.6	−3.7	0.8
His	71	Side chain	—	—	—
		D-glyceraldehyde-3-phosphate dehydrogenase E.C.1.2.1.12 RMSD 1.72 1*gdl*(O)			
Asp	312	$O^{\delta 2}$	6.2	1.8	−1.0
Asp	47	$O^{\delta 1}$	−1.3	−4.7	0.3
His	50	Side chain	—	—	—

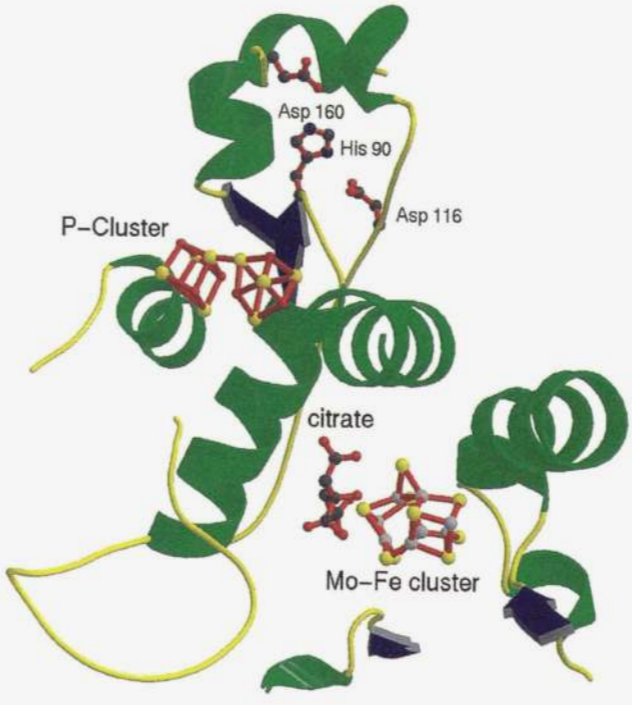

Fig. 7. A view of the MoFe protein from *Azobacter vinelandii* showing the positional relationship of the P-cluster, the Asp-His-Asp triad and the MoFe cofactor associated with the nitrogenase enzyme (Jongsun & Rees, 1992).

These results show this to be a region worthy of further investigation, which may give insight into the catalytic mechanism of MoFe proteins.

The ribonucleases

Ribonucleases are found in both prokaryotes and eukaryotes (Beintema et al., 1990); they catalyze the hydrolysis of phosphodiester bonds in RNA chains. Structurally, the best understood ribonucleases are bovine pancreatic ribonuclease (RNase A, E.C.3.4.27.5) and ribonuclease T_1 (RNase T_1, E.C.3.4.27.3) from the fungus *Aspergillus oryzae*. Both enzymes have an $\alpha+\beta$ roll architecture, but adopt a different topology.

The crystal structures of ribonuclease H and barnase have also been solved, but they will not be considered here.

Ribonuclease A

Ribonuclease A has 124 amino acids and is a member of a large superfamily of homologous bovine RNases (Beintema et al., 1988). RNase A is a monomer, with an $\alpha+\beta$ fold and four intrachain disulfides. Ribonuclease S (RNase S, e.g., 1*rbc*; Varadarajan & Richards, 1992) is a product of the cleavage of RNase A by subtilisin between residues 20 and 21.

RNase A is specific for the pyrimidines uridine and cytidine; this specificity is achieved by hydrogen bonding from the backbone NH and side-chain -OH atoms of the sequentially conserved residue Thr 45. Chemical modification studies (Crestfield et al., 1963) have shown that the residues His 12 (acid/base catalyst), His 119 (acid/base catalyst), and Lys 41 (stabilizes transition state) are involved in the catalytic mechanism (Blackburn & Moore, 1982).

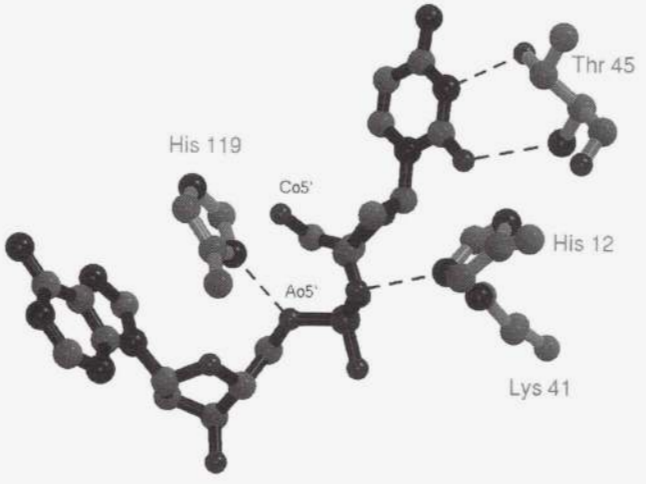

Fig. 8. 3D representation of the active site of ribonuclease A complexed with D(CPA) (Zegers et al., 1992). The catalytic residues are His 119, His 12, and Lys 41.

Figure 8 is a 3D representation of the inhibitor deoxycytidyl-3′,5′-deoxyadenosine, D(CPA) bound to ribonuclease A 1*rpg* (Zegers et al., 1992). The His 12 interacts with the 2′-oxygen (equivalent to CO5′ in the inhibitor in Fig. 8) and Lys 41 is in the vicinity.

Borkakoti et al. (1982) noticed that in a phosphate complex of RNase A, there are two distinct conformations of His 119 (A and B) that are related by a 180° rotation about the His 119 C^{β}-C^{γ} (χ_2) bond. There is still controversy as to which of these conformations is catalytically active (Borkakoti et al., 1982; deMel et al., 1992; Zegers et al., 1992).

Two consensus templates have been created describing both the A and B conformations (Table 7). The seed atoms used were the His 12 side chain, His 119 $N^{\delta 1}$, and Lys 41 N^{ζ} taken from the structure 3*rn*3 (Borkakoti et al., 1982), which has coordinates for both the A and B His 119 conformations. Figure 9 is a 3D representation of the distribution of the His 119 $N^{\delta 1}$ atoms relative to His 12 for all structures in the ribonuclease A and ribonuclease S entries in the PDB. There are two distinct clusters of His 119 $N^{\delta 1}$ atoms (in blue) representing the A and B conformations; Table 8 lists the PDB structures responsible for these clusters. Figure 10 is a histogram of the number of hits versus RMSD from the B conformation of the RNase template for all the ribonuclease A and S PDB structures (Table 9); it confirms that the A and B conformation occupy distinct positions for the His 119 $N^{\delta 1}$ atom.

Template search through the PDB

The 95% by sequence nonidentical protein data set was searched for other proteins with residues and atoms in a similar conformation to the two RNase A consensus templates. An RMS distance cut-off of 2 Å was used when defining matches to either template.

All the hits located for the A conformation, other than the RNase structures in this data set, have RMSDs greater than 2 Å; they can therefore be discounted as unlikely to have RNase A catalytic activity. The B conformation consensus template located His 285, His 321, and Lys 191 from the A and B chains of the same enzyme RUBISCO, ribulose-1,5-bisphosphate carboxylase, E.C.4.1.1.39 (5*rub*, Schneider et al., 1990) with RMSDs from the consensus templates of 1.51 Å and 1.86 Å, respectively. Figure 11 is a 3D

Table 7. *Coordinates of the consensus templates that describe the two conformers, A and B, for the active site of ribonuclease A as well as the coordinates of the consensus template for ribonuclease T*[a]

Residue	Number	Atom	*x*	*y*	*z*
		Ribonuclease A Template conformer A coordinates			
Lys	41	N^{ζ}	5.3	−1.1	−2.8
His	119	$N^{\delta 1}$	6.2	3.2	3.8
		Ribonuclease A Template conformer B coordinates			
Lys	41	N^{ζ}	5.1	−1.4	−2.9
His	119	$N^{\delta 1}$	5.8	5.8	2.0
		Ribonuclease T			
His	92	N^{ϵ}_2	9.7	2.5	−2.1
Glu	58	C^{δ}	3.4	2.8	−2.6
Glu	58	$O^{\epsilon 2}$	3.6	3.3	−1.5
Glu	58	$O^{\epsilon 1}$	4.3	2.3	−3.2
		Reference frame atoms			
His	12/40	C^{γ}	0.0	0.0	0.0
His	12/40	$N^{\delta 1}$	0.8	−1.1	0.0
His	12/40	$C^{\delta 2}$	0.8	1.1	0.0
His	12/40	$C^{\epsilon 1}$	2.1	−0.7	0.0
His	12/40	$N^{\epsilon 2}$	2.1	0.6	0.0

[a]The two templates share the same reference frame His 12/40 coordinates.

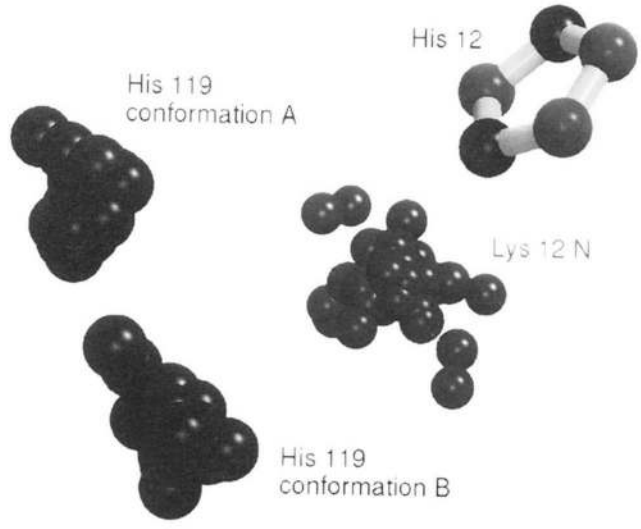

Fig. 9. 3D representation of the distribution of the His 119 $N^{\delta 1}$ active site atom conformations A and B for all the RNase A and RNase S structures in the PDB. Also shown is the side chain of His 12 and the distribution of the Lys 41 N^{ζ} atoms.

representation of the active site of RUBISCO from 5*rub* (Lundqvist & Schneider, 1989) with substrate ribulose-1,5-bisphosphate. Also shown is the active site Mg that is coordinated to Asp 193 (yellow bonds) and the substrate. The residues located, His 285, His 321, and Lys 191 (red bonds), are found in the large L chain and are also part of the active site. In addition, they are all conserved in the RUBISCO sequence (Schneider et al., 1990). Lys 191 is the site of carbamylation during activation, and His 321 is involved in binding the phosphate group of the substrate (Lundqvist & Schneider, 1989). There is no clear functional role assigned to His 285. When compared to ribonuclease A, the position of the active site residues with respect to the substrate is different and the equivalent residues also have different roles. However, it is interesting that the RNase template has identified the catalytic residues in RUBISCO even though they have different mechanisms, different primary E.C. numbers, and a totally different function.

Ribonuclease T_1

The bacterial RNase T_1 is isolated from *A. oryzae*. The enzyme has two isoforms containing Lys or Gln at position 25 of the polypeptide chain, denoted Lys 25-RNase T1 and Gln 25-RNase T1.

Specificity and catalytic mechanism

RNase T_1 is specific for the purine nucleotide guanosine (as opposed to pyrimidines in RNase A) and is strictly limited to hydrolysis at 3′-phosphate groups in RNA. The reason for this specificity is not fully understood.

The catalytic residues are Glu 58, which takes a proton from H_2O in the first step, and His 92, which donates a proton to the leaving group and activates the water molecule used in hydrolysis. The guanosine-specific recognition occurs in the loop Tyr 42-Asn 43-Asn 44-Tyr 45-Glu 48.

The consensus template

The atoms used to generate a consensus template were the side chain of His 40, His 92 $N^{\epsilon 2}$, and Glu 58 C^{gd}, $O^{\epsilon 1}$, $O^{\epsilon 2}$ taken from the RNase T_1 X-ray crystal structure 1*rn*1 (Arni et al., 1992); the distance cut-off was set to 3.0 Å. Table 9 gives the data set of RNase T_1 PDB codes and their RMSDs from the resultant consensus template.

There are two structures in the ribonuclease T_1 data set, 2*aae* (Zegers et al., 1992) and 5*rnt* (Lenz et al., 1991), whose active site residues are not identified by the consensus template. 2*aae* has its His 40 mutated to a lysine and 5*rnt*'s structure is only refined to 3.2 Å and has the inhibitor guanosine-3′,5′-bisphosphate bound to its active site.

No potential ribonuclease T_1 active sites were located when the 95% nonidentical-by-sequence data set of PDB structures was searched.

Comparison of ribonuclease A and T_1 active sites

It is interesting to compare the three different templates, RNase A conformations A and B and RNase T_1. Figure 12 is a 3D representation of these templates with the atoms superimposed according to their proposed role in the catalytic mechanism. As Zegers et al. (1992) have shown, the general base catalysts, His 12 and Glu 58, the electrostatic stabilizing groups, Lys 41 and His 40, and the general acid catalysts, His 119 and His 92, of RNase A and RNase T_1, respectively, all superimpose, indicating convergent evolution of these enzyme active sites.

Table 8. *Summary of the ribonuclease PDB structures and their RMSDs from their respective consensus templates that adopt either the A or B conformation of their active site His 119 residue*[a]

Template conformer A: Ribonuclease E.C.3.1.27.5						
Ribonuclease A						
1bsr B 0.45	**3rn3** 1.16	1ras 1.50	1rat 0.90	2rat 0.74	3rat 0.92	4rat 0.79
5rat 0.31	6rat 0.43	7rat 0.48	8rat 0.44	9rat 0.41	1rbn 0.62	1rcn E 0.71
1rnc 0.35	1rnd 0.46	1rar 1.03	1rob 0.49	1rpg 0.33	1rph 0.53	1rtb 1.56
5rsa 0.34	6rsa 0.53	7rsa 0.48	1rtb 1.56			
Template conformer B: Ribonuclease E.C.3.1.27.5						
Ribonuclease A						
1bsr B 0.64	1rbn 0.97	**3rn3** 1.05	1rpf 1.29	1rph 0.78	9rsa B 0.71	1srn A 0.65
3srn A 0.49	4srn A 0.32	1ssa A 0.78	1ssb A 0.50			
Ribonuclease S						
1rbc S 1.00	1rbd S 0.30	1rbe S 0.74	1rbf S 1.50	1rbg S 0.49	1rbh S 0.44	1rbi S 0.40
2rln S 0.29	1rnu 1.36	1rnv 0.54	2rns 1.38			

[a]PDB codes in bold have coordinates describing both conformations.

Lysozyme

Lysozyme kills certain bacteria by cleaving the polysaccharide component, N-acetylglucosamine (NAG) and N-acetylmuramate (NAM), of their cell wall. There are two catalytic residues, Glu 35 and Asp 52, in the eukaryotic lysozyme (Blake et al., 1965; Johnson & Phillips, 1965) and Glu 11 and Asp 20 in the prokaryotic lysozyme from bacteriophage T4 (Matthews & Remington, 1974).

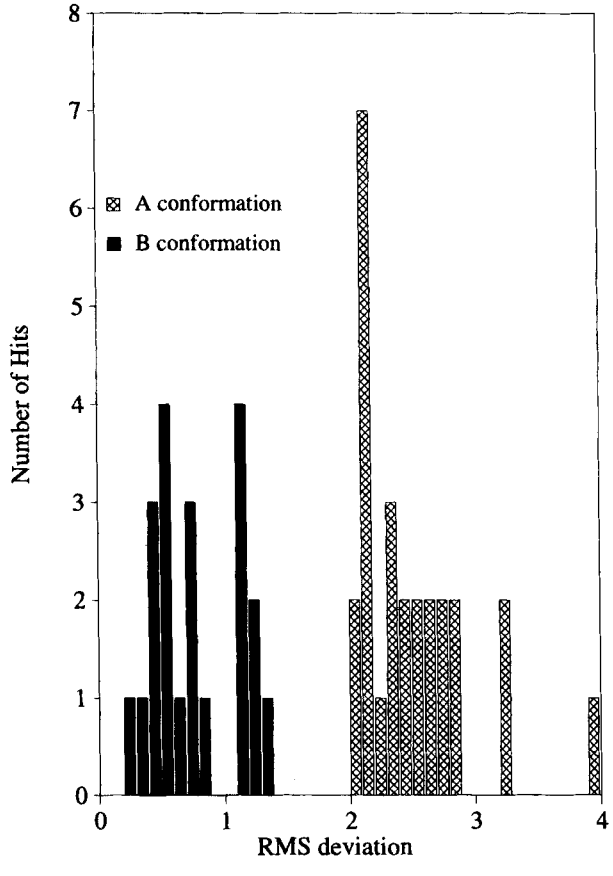

Fig. 10. Histogram of the number of hits against RMS distance from the B conformation of the RNase A consensus template. It shows that the A and B conformations of the His 119 residues are in distinct positions.

The prokaryotic and eukaryotic lysozyme structures have low sequence identity, but have a similar $\alpha+\beta$ tertiary fold (Remington & Matthews, 1978) and may have evolved from a common precursor. They can, however, be divided into eukaryotic and prokaryotic groups on the basis of the geometry of their catalytic residues.

Two other lysozymes, characterized by a low sequence identity and different tertiary fold to those mentioned above, are found in the goose (154*l*, Weaver et al., 1995) and black swan (1*gbs*, Rao et al., 1995). In fact, they also lack a catalytic Asp and for this reason will not be considered further.

Eukaryotic: Mammalian and avian lysozyme

There are 76 eukaryotic lysozyme structures in the PDB from several species, although many of these originate from mutational studies of the same structure. Figure 12 shows 3D representations of the conformation of the Glu 35 and Asp 52 catalytic residues in the mammalian and avian lysozymes, together with the corresponding catalytic residues in the bacteriophage T4 lysozymes. In all cases, the Glu residues have been superimposed so that the relative conformation of the Asp side chain can be compared. Although the conformations of the Asp side chains are similar, in all the lysozymes there is no obvious clustering of the functional Asp O^{δ} atoms. This is in contrast to the situation in the Ser-His-Asp catalytic triads, where we have shown that, although the relative position and orientation of the catalytic Asp shows great variability, the location of the functional Asp O^{δ} is highly conserved. In the lysozyme catalytic mechanism, the negative charge on Asp 52 stabilizes the formation of the positive charge on the sugar in the D-site (Phillips, 1966). This suggests that the Asp 52 side-chain carboxyl group only needs to be in the vicinity of the D-site sugar and its precise position is not so important.

A consensus template was constructed for the mammalian and avian data set using the seed template atoms of the Glu 35 side chain and Asp 52 C^{γ}, $O^{\delta 1}$, and $O^{\delta 2}$ from the PDB structure 135*l* (Harata et al., 1993) of human lysozyme with a 3.0 Å cut-off; the

Table 9. *Summary of the ribonuclease T_1 PDB structures and their RMSDs from the ribonuclease T_1 consensus template*[a]

1fus 1.08	1fut 1.28	1rga 1.28	1rgeB 1.40	1rgcA 1.92	1rgk 1.78	1rgl 0.66	1rls 1.71	1rms 1.56
1rnlA 1.53	1rnlB 1.53	1rnlC 0.62	1rnt 0.99	2rnt 0.47	3rnt 0.76	6rnt 0.41	7rnt 0.90	8rnt 1.78
9rnt 1.77	1trpA 0.85	1trpB 1.00	1trqA 0.87	1trqB 0.83	**2aae**	**5rnt**		

[a]PDB codes in bold are missed by the RNase T_1 template using a 3.0-Å distance cut-off.

coordinates are given in Table 10. Nine hen-egg lysozyme structures, which are different refinement models based on the original Blake et al. (1965) lysozyme structure (Diamond, 1975), were removed. Although they gave extraordinary insight into lysozyme structure and function at the time, they are poorly refined compare to the lysozyme structures solved in the last 15 years. Table 11 summarizes the number of catalytic Glu 35–Asp 52 diads located using the template for each species.

There are three hen-egg white lysozyme structures whose Glu 35-Asp 52 catalytic residues are not located. Two of these, 1*lym* (Rao et al., 1983) and 3*lyt* (Young et al., 1993), are only refined to 2.5 Å, and the third, 2*hfl* (Sheriff et al., 1987), has been crystallized in a complex with an immunoglobulin molecule. This also applies to three structures of mouse lysozyme (e.g., Fischmann et al., 1991), which are not identified. The catalytic residues in horse lysozyme (Tsuge et al., 1992) have an RMSD of 2.52 Å from the consensus template; Figure 12 shows that the relative position of the Asp 52 is shifted with respect to the other mammalian Asp residues.

Prokaryotic: Bacteriophage T4 lysozyme

T4 lysozyme is produced late in the infection of *Escherichia coli* by T4 bacteriophage. The structure was first determined by Matthews and Remington (1974). This enzyme has been used as a model to study the effects of mutations on protein stability and function and there are 160 different mutant forms deposited in the PDB (e.g., Alber et al., 1987; Weaver & Matthews, 1987).

A consensus template was constructed from the wild-type T4 structure 2*lzm* (Alber et al., 1987) using the same atoms as for the mammalian template; the distance cut-off was set to 3.0 Å. The coordinates of the resultant template are given in Table 10.

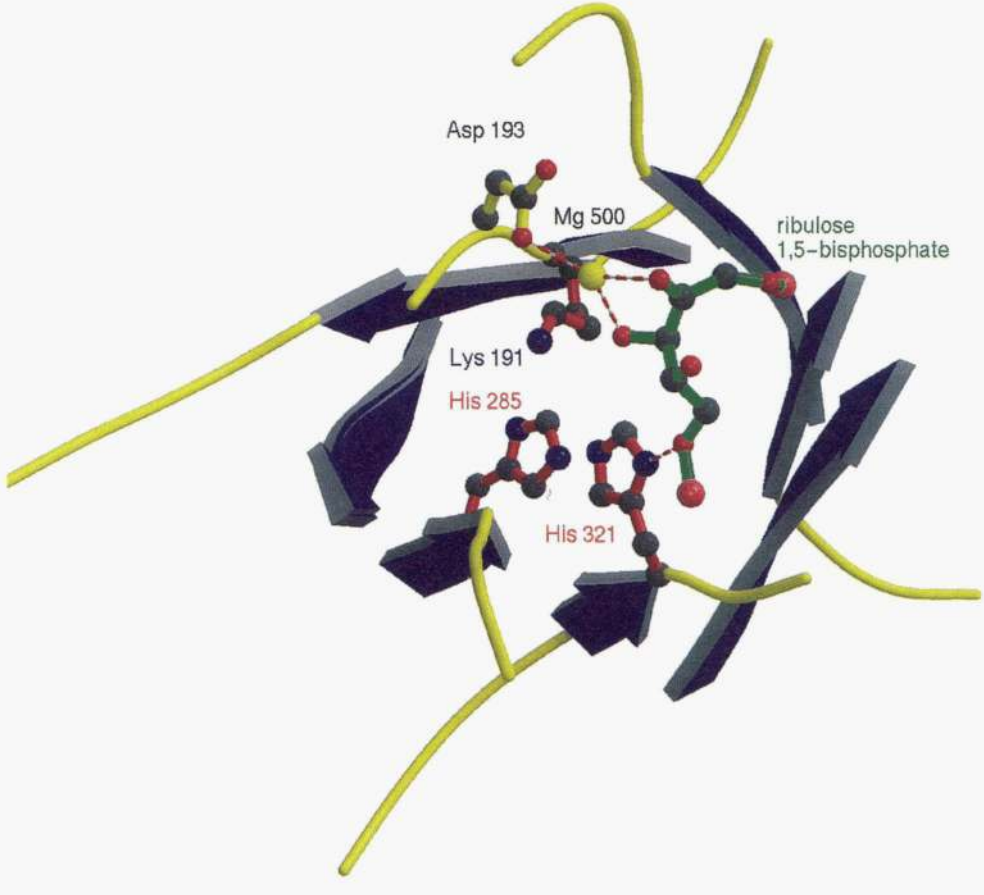

Fig. 11. 3D representation of the active site of RUBISCO (Lundqvist & Schneider, 1989) showing the three residues His 285, His 321, and Lys 191 (red bonds) that have the same conformation as the active site residues of ribonuclease A.

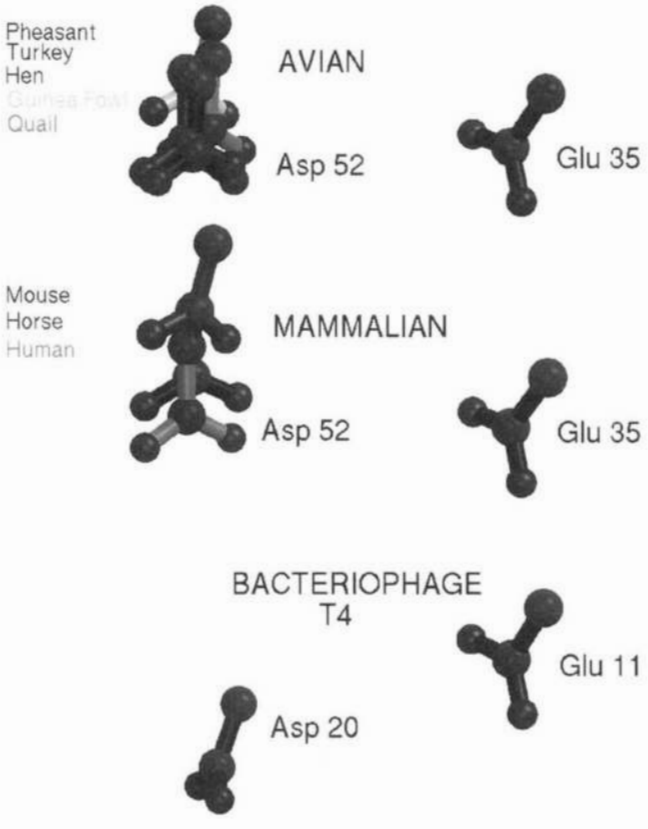

Fig. 12. Relative conformations of the catalytic Glu 35-Asp 52 residues from avian and mammalian lysozymes as well as the bacteriophage T4 catalytic Glu 11-Asp 20.

Of the 160 T4 structures in the PDB, 7 are not located by the consensus template. All "missed" structures have mutations of the active site residues or of regions around the active site, perturbing the conformation of the catalytic residues.

Comparison of prokaryotic and eukaryotic lysozymes

Due to the similarity of the tertiary fold of both the prokaryotic and eukaryotic lysozyme structures, and because they both have a Glu and Asp residue as their catalytic residues, it would not be surprising if the conformation of the catalytic residues is also the same. In fact, as Figure 12 indicates, when the Glu residues of the two templates are superimposed, the Asp residues lie around 4.5 Å apart. It is proposed that the general mechanism of catalysis of these two lysozyme is the same. As for the mammalian and avian lysozymes, Asp 20 acts as an electrostatic stabilizer of the positively charged sugar in the D-site, so it is only required to be in the vicinity of the D-site.

When the prokaryotic and eukaryotic lysozyme consensus templates are used to search through the representative structures of the PDB, there are about 100 hits located for each template. This large number of matches is not surprising given that the template consists of only two residues. This shows that, although the lysozyme templates are able to extract the majority of catalytic Glu-Asp diads, there are many false hits involving noncatalytic interactions, and so the template has a low specificity.

Table 10. *Coordinates of the consensus template describing the active site of the prokaryotic T4 as well as mammalian and avian lysozymes present in the PDB*

Residue	Residue number	Atom	*x*	*y*	*z*
		Mammalian and avian Asp 52			
Asp	52	C^{γ}	4.0	−4.3	4.1
Asp	52	$O^{\delta 1}$	4.4	−3.8	5.1
Asp	52	$O^{\delta 2}$	4.4	−4.3	3.1
		Phage t4 Asp 20			
Asp	20	C^{γ}	7.5	−3.6	1.5
Asp	20	$O^{\delta 1}$	4.4	−3.8	5.1
Asp	20	$O^{\delta 2}$	4.4	−4.3	3.1
		Template residue			
Glu	11/35	C^{γ}	−1.5	0.1	−0.0
Glu	11/35	C^{δ}	0.0	0.0	0.0
Glu	11/35	$O^{\epsilon 1}$	0.6	−1.0	0.0
Glu	11/35	$O^{\epsilon 2}$	0.6	1.1	0.0

Discussion

We have developed an algorithm called TESS that is based on the geometric hashing paradigm. The program enables us to search through a data set of 3D PDB structures for any user-defined, sequence order-independent 3D template. The seed for the 3D template consists of atoms or residues extracted directly from a PDB file, and it is possible to define explicitly for each coordinate which atoms or residue types should be included. This work enables us to create a collection of 3D enzyme active site or functional templates, that, by analogy to the 1D templates present in secondary protein sequence databases such as PROSITE or PRINTS, may allow the assignment of the biological function and evolutionary origins of a new protein structure. Of course, TESS could

Table 11. *Numbers of Asp and Glu catalytic residues located for lysozymes originating from different species*

Name	Total no. PDB structures	Hits
	Avian lysozymes	
Guinea fowl	1	1
Hen	25	22
Pheasant	1	1
Quail	1	1
Turkey	4	4
	Mammalian lysozymes	
Mouse	4	1
Horse	1	1
Human	17	17

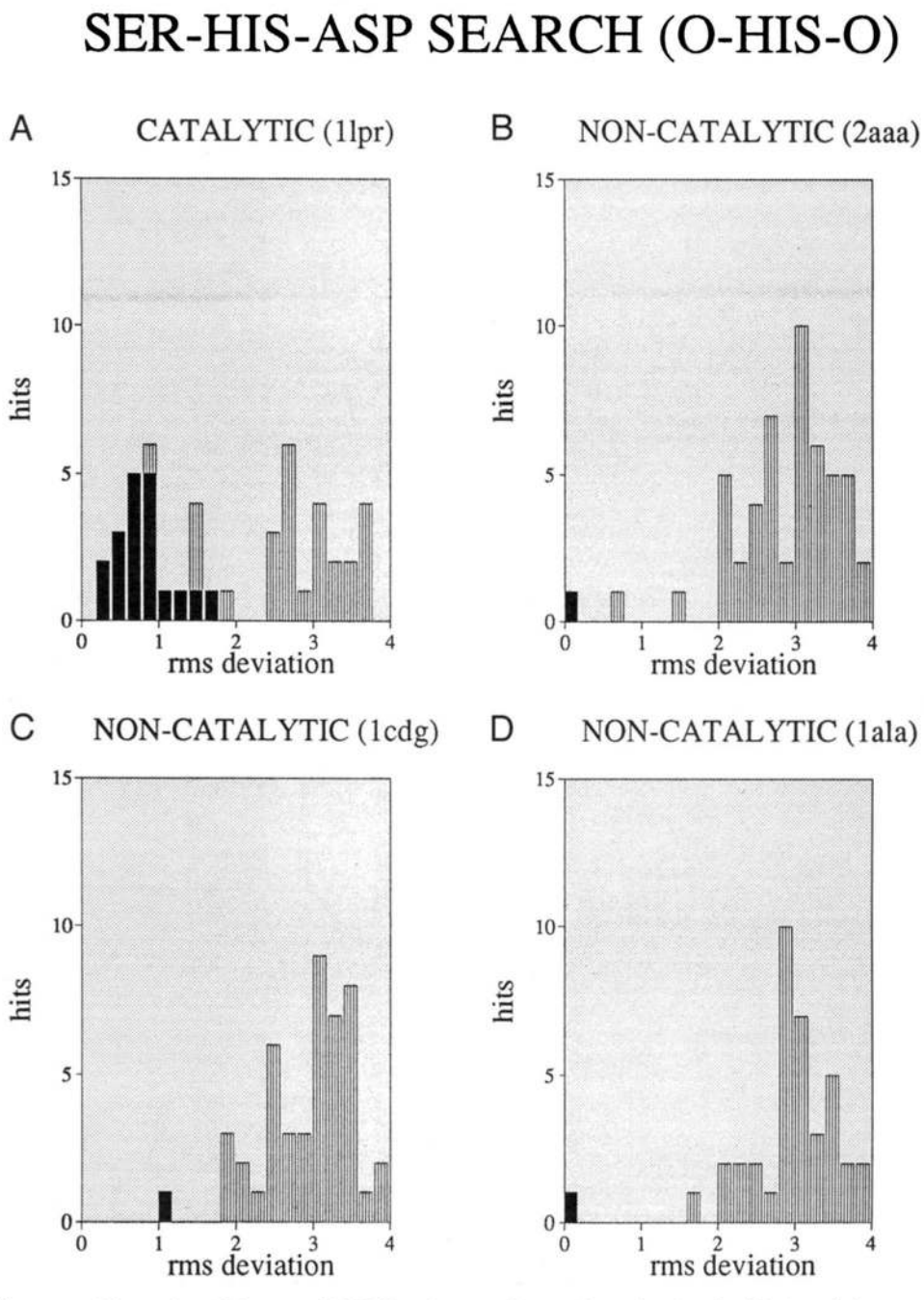

Fig. 13. Histograms of the number of hits versus RMSD from the respective templates when four Ser-His-Asp triads are used to search through the nonidentical data set. The *1lpr* triad is the catalytic consensus template for the serine proteinases and lipases. The other three are randomly chosen noncatalytic triads. Black bars are diads originating from the same protein by function as the template diad.

be used to produce databases of other recurring 3D templates, such as metal or nonenzyme ligand binding sites. In addition, the templates are useful as a purely practical method to identify automatically the catalytic triad in, for example, a serine proteinase. It also allows us to orient all serine proteinases into the same reference frame, to facilitate simple visual comparisons.

Having studied all the enzymes in the PDB with the catalytic residues of type Nu:-His-ELEC, the conservation of these triads is striking. Although the residue types of the Nu: and ELEC groups can vary according to enzyme type, we find that the positions of the functional atoms in the triads are conserved. Indeed, one template is able to describe the active site of all the serine proteinases, acetylcholinesterase, and haloalkane dehalogenase proteins. This suggests that convergent evolution has drawn the functional atoms into optimal catalytic positions. In addition, these triads are confined to the active sites of enzymes, although we did find several examples of other unidentified Nu:-His-ELEC triads that may have biological relevance. This suggests

that, as the number of protein structures deposited in the PDB increases, other common 3D templates, functional in more than one enzyme type, will occur.

Conversely, the Asp-Glu catalytic diad found in all lysozyme structures in the PDB does not have such a conserved structure. The Asp is thought to act as an electrostatic stabilizer of the transition state and therefore its precise position in the active site is not so critical. In addition, when the lysozyme Glu 35–Asp 52 consensus templates are used to search through the representative structures of the PDB, there are about 100 hits for each of the templates. This larger number of hits occurs simply because there is greater chance of two rather than three residues being located randomly in the same conformation as the catalytic pair in a given data set of protein structures.

To investigate this, we have taken three randomly picked noncatalytic Ser, His, and Asp interactions and compared the number of hits located when these triads are used to search our representative nonidentical data set of PDB structures with those for a catalytic triad. Secondly, we took three randomly picked noncatalytic Asp and Glu residues and compared the number of hits located in the protein data set with the catalytic Asp and Glu from lysozyme. The representative nonidentical data set is different from

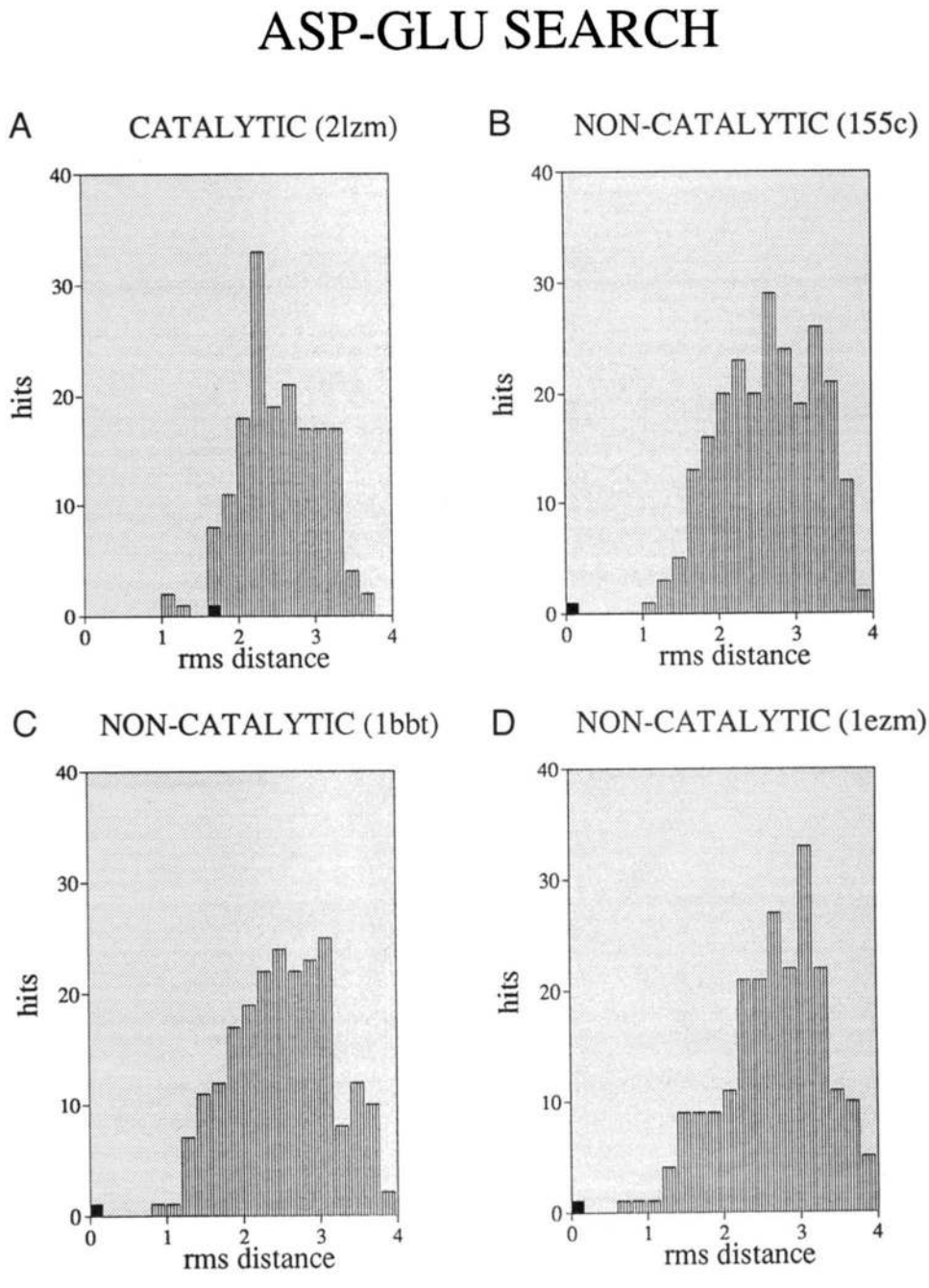

Fig. 14. Histograms of RMSDs from the four different Glu-Asp triad templates of hits found in the data set of nonidentical protein structures. Black bars represent diads located from the same protein by function as the template triad.

the 95% nonidentical data set used previously and comprises protein structures that all have an SSAP score (Orengo et al., 1993) less than 80 when compared and less than 25% sequence identity (i.e., it excludes homologous proteins). This was extracted from the June 1996 release of the PDB.

The histogram in Figure 13 (top left) represents the search with the catalytic triad consensus template derived from the Ser 195 O^{γ}-His 195 side-chain-Asp $102^{\delta 2}$ catalytic triad from *1lpr* (Bone et al., 1991). The other three histograms are searches using noncatalytic Ser-His-Asp interactions that have an RMSD between 3 Å and 6 Å from the *1lpr* consensus template; these are Ser 202 O^{γ}-His 205 side-chain-Asp 172 $O^{\delta 2}$ from chicken annexin *1ala* (Bewley et al., 1993), Ser 13 O^{γ}-His 503 side-chain-Asp 518 $O^{\delta 2}$ from cyclodextrin glycosyltransferase *1cdg* (Lawson et al., 1994), and Ser 104 O^{γ}-His 108 side-chain-Asp 201 $O^{\delta 1}$ from α-amylase *2aaa* (Boel et al., 1990). For the catalytic triad search, 20 other catalytic triads are located from the data set of structurally nonidentical proteins, compared to only one for each of the noncatalytic Ser, His, and Asp interactions. In addition, for the catalytic triad, there are only a few hits below 2.0 Å RMS distance that are noncatalytic. although other factors, such as accessibility to substrate, are important, this does suggest that when Ser, His, and Asp are found in this conformation in a protein structure, we can be confident that they play a catalytic role.

A similar test was performed on a template consisting of only two residues; here we compared the nonidentical data set search of the T4-lysozyme catalytic Asp 11-Glu 20 diad with three other randomly chosen noncatalytic Glu-Asp interactions. These diads were the side chains of Glu 11-Asp 109 from cytochrome C550 *155c* (Timkovich & Dickerson, 1976), Glu 128-Asp 169 from foot-and-mouth virus *1bbt* (Parry et al., 1990) and Glu 172-Asp 189 from the elastase structure *1ezm* (Thayer et al., 1991). Figure 14 shows the number of hits located for searches with these templates. There are around the same number of hits located for the catalytic as the noncatalytic diads.

This proves that, as expected, the number of hits located when searching a data set of protein structures with a consensus template depends on the number of atoms and residues in that template. If hits are located below the chosen RMS cut-off, it does not necessarily mean that they are functionally significant, but merely provides a possible starting point for further experimental investigation. Indeed, when such a hit is located, other factors should be considered, such as locality with respect to potential ligand binding site or accessibility to the protein surface.

These results are borne out by the searches using the ribonuclease and lysozyme templates. Although there were several hits below the defined RMSD cut-off, there is no example of a hit revealing a new catalytic site. It should be noted that the functions of the vast majority of proteins in the PDB are understood and the PDB represents a small fraction of all proteins in the genome.

One of the major problems in genome sequence analysis is the lack of information available on functional sites. Even for proteins of known structure, the relevant information is often hidden in the literature and cannot be extracted easily. These catalytic templates represent one approach to encapsulating functional information, and relating it back from structure to sequence. As the number of templates grows, it will become a useful source of additional information on functionally important residues in a sequence, which may be useful in recognizing distant homologues and so assigning function.

Acknowledgments

We thank Roman Laskowski, Christine Orengo, and Laurence Pearl for their help and suggestions. A.C.W is supported by a BBSRC CASE studentship, sponsored by Roche Products Ltd.

References

Alber T, Dao-Pin S, Nye JA, Muchmore DC, Matthews BW. 1987. T4 temperature sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry 26*:3754–3758.

Alexandrov NN, Takahashi K, Go N. 1992. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J Mol Biol 225*:5–9.

Altschul SF, Gish W, Miller W, Eugene WM, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Arni RK, Pal GP, Ravichandran KG, Tulinsky A, Walz FG, Metcalf P Jr. 1992. Three-dimensional structure of Gln 25-ribonuclease T_1 at 184 Å resolution: Structural variations at the base recognition and catalytic sites. *Biochemistry 31*:3126–3135.

Artymiuk PJ, Poirette AR, Grindley HM, Rice DW, Willett P. 1994. Graph theoretic approach to the identification of three-dimensional patterns of amino-acid side-chains in protein structures. *J Mol Biol 243*:327–344.

Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ. 1994. PRINTS—A database of protein motif fingerprints. *Nucleic Acids Res 22*:3590–3696.

Bacon DJ, Anderson WF. 1988. A fast algorithm for rendering space-filling molecular pictures. *J Mol Graph 6*:219–220.

Bairoch A, Boeckmann B. 1994. The SWISS-PROT protein sequence databank: Current status. *Nucleic Acid Res 22*:3578–3589.

Bairoch A, Bucher P. 1994. PROSITE: Recent developments. *Nucleic Acids Res 22*:3583–3582.

Beintema JJ, Confalone E, Sasso MP, Furia A. 1990. Structure mechanism and function of ribonucleases. In: Cuchillo MC, Llorens R, Nogues MV, Pares X, eds. *Structure mechanism and function of ribonucleases*. Gerona, Italy. pp 275–281.

Beintema JJ, Schuller C, Masachika I, Carsana A. 1988. Molecular evolution of the ribonuclease superfamily. *Prog Biophys Mol Biol 51*:165–192.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shoimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bewley MC, Boustead CM, Walker JH, Waller DA, Huber R. 1993. Structure of chicken annexin V at 225 Å resolution. *Biochemistry 32*:3923–3929.

Blackburn P, Moore S. 1982. Pancreatic ribonuclease. In: Boyer PD, ed. *The enzymes*. New York: Academic Press. pp 317–433.

Blake CCF, Koenig DF, Mair GA, North ACT, Phillips DC, Sarma VR. 1965. Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. *Nature 206*:757–780.

Bleasby AJ, Akrigg D, Attwood TK. 1994. OWL—A non-redundant, composite protein sequence database. *Nucleic Acid Res 22*:3574–3577.

Blow DM, Birktoft JJ, Hartley BS. 1969. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature 221*:337–340.

Boel E, Brady L, Brzozowski AM, Derewenda Z, Dodson GG, Jensen VJ, Peterson SB, Swift H, Thim L, Woldike HF. 1990. Calcium binding in α-amylases: An X-ray diffraction study at 21 Å resolution of two enzymes from *Aspergillus*. *Biochemistry 29*:6244–6249.

Bone R, Fujishige A, Kettner CA, Agard DA. 1991. Structural basis for broad specificity in α-lytic protease mutants. *Biochemistry 30*:10388–10398.

Borkakoti N, Moss DS, Stanford MJ, Palmer RA. 1982. The refined structure of ribonuclease A at 145 Å resolution. *J Crystallogr Spectros Res 14*:467–471.

Brady L, Brzozowski AM, Derewenda ZS, Dodson E, Dodson G, Tolley S, Turkenburg JP, Christianson L, Huge Jensen B, Norskov L, Thim L, Menge U. 1990. A serine protease triad forms the catalytic centre of triacylglycerol lipase. *Nature 343*:767–770.

Crestfield AM, Stein WH, Moore S. 1963. Alkylation and identification of the histidine residues at the active site of ribonuclease. *J Biol Chem 238*:2413–2420.

deMel VS, Martin PD, Doscher MS, Edwards BF. 1992. Structural changes that accompany the reduced catalytic efficiency of two semisynthetic ribonuclease analogs. *J Biol Chem 267*:247–256.

Diamond R. 1975. Real-space refinement of the structure of hen egg-white lysozyme. *J Mol Biol 82*:371–396.

Drenth J, Jansonius JN, Koekoek R, Swen HM, Wolthers BG. 1968. Structure of papain. *Nature 218*:929–934.

Fischer D, Wolfson H, Shuo LL, Nussinov R. 1994. Three-dimensional, sequence order-independent comparison of a serine protease against the crys-

tallographic database reveals active site similarities: Potential implications to evolution and to protein folding. *Protein Sci 3*:769–778.

Fischmann TO, Bentley GA, Bhat TN, Boulot TN, Mariuzza RA Phillips SEV, Tello D, Poljak RJ. 1991. Crystallographic refinement of the 3D structure of the Fabdi13–lysozyme complex at 25 Å resolution. *J Biol Chem 266*:12915–12920.

Fujinaga M, James MNG. 1987. Rat submaxillary serine proteinase, tonin structure solution and refinement at 18 Å resolution. *J Mol Biol 195*:373–391.

Glusker JP. 1991. Structural aspects of metal liganding to functional groups in proteins. *Adv Protein Chem 42*:1–76.

Grochulski P, Li Y, Schrag JD, Bouthillier F, Smith P, Harrison D, Rubin B, Cygler M. 1993. Insights into interfacial activation from an "open" structure of *Candida rugosa* lipase. *J Biol Chem 268*:12843–12847.

Harata K, Muraki M, Jigami Y. 1993. Role of Arg 115 in the catalytic action of human lysozyme—X-ray structure of His 115 and Glu 115 mutants. *J Mol Biol 233*:524–535.

Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123–138

Hutchinson EG, Thornton JM. 1996. PROMTOIF—A program to identify and analyse structural motifs in proteins. *Protein Sci 5*:212–220.

Jernigan R, Raghunathan G, Bahar I. 1994. Characterization of interactions and metal-ion binding-sites in proteins. *Curr Opin Struct Biol 4*:256–263.

Johnson LN, Phillips DC. 1965. Structure of some crystalline-inhibitor complexes determined by X-ray analysis at 6 Å resolution. *Nature 206*:761–764.

Jongsun K, Rees DC. 1992. Crystallographic structure and functional implications of the nitrogenase molybdenum–iron protein from *Azotobacter vinelandii*. *Nature 360*:553–560.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Lamdan Y, Schwartz JT, Wolfson HJ. 1988. On recognising 3D objects from 2D images. *Proceedings of IEEE Int Conf on Robotics and Automation, Philadelphia, Pennsylvania*. pp 1407–1413.

Lawson CL, Van Montfort R, Strokkopytov B, Rozeboom HJ, Kalk KH, De Vries GE, Penninga D, Dijkhuizen L, Dijkstra BW. 1994. Nucleotide sequence and X-ray structure of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251 in a maltose-dependent crystal form. *J Mol Biol 236*:590–600.

Lemer CMR, Rooman MJ, Wodak SJ. 1995. Protein structure prediction by threading methods: Evaluation and current techniques. *Proteins Struct Funct Genet 23*:337–355.

Lenz A, Heinemann U, Maslowska M, Saenger W. 1991. X-ray analysis of cubic crystals of the complex formed between ribonuclease T_1 and guanosine-3′,5′-bisphosphate. *Acta Crystallogr B 47*:521–527.

Lundqvist T, Schneider G. 1989. Crystal structure of the complex of ribulose-1,5-bisphosphate carboxylase and a transition state analogue, 2-carboxy-D-arabinitol 1,5-bisphosphate. *J Biol Chem 264*:7078–7083.

Matthews BW, Remington SJ. 1974. The three-dimensional structure of the lysozyme from bacteriophage T4. *Proc Natl Acad Sci USA 71*:4178–4182.

Matthews BW, Rossmann MG 1985. Comparison of protein structures. *Methods Enzymol 115*:397–420.

Mitchell EM, Artymiuk PJ, Rice DW, Willet P. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol 212*:151–166.

Murzin AG, Brenner SE, 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.

Ollis DL, Cheah E, Miroslaw C, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A. 1992. The α/β hydrolase fold. *Protein Eng 5*:197–211.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.

Ohkawa H, Hogue C, Bryant S, Kans J, Epstein J, Schuler G, Ostler J. 1997. *MMDB: Entrez's Structure Database—FAQ*. http://www.ncbi.nlm.nih.gov/structure/struchelp.html.

Parry N, Fox G, Rowlands D, Brown F, Fry E, Acharaya R, Logan D, Stuart D. 1990. Structural and serological evidence for a novel mechanism of antigenic variation in foot-and-mouth disease virus. *Nature 347*:569–572.

Phillips DC. 1966. The three-dimensional structure of an enzyme molecule. *Sci Am 215*:78–90.

Rao ST, Hogle J, Sundaralingam M. 1983. Studies of monoclinic hen-egg white lysozyme. The refinement at 25 Å resolution—Conformational variability between the two independent molecules. *Acta Crystallogr C 39*:237–242.

Rao Z, Esnouf R, Isaacs N, Stuart D. 1995. A strategy for rapid and effective refinement applied to black swan lysozyme. *Acta Crystallogr D 51*:331–335.

Remington SJ, Matthews BW. 1978. Method to assess the similarity of protein structures, with applications to T4 bacteriophage lysozyme. *Proc Natl Acad Sci USA 75*:2180–2184.

Sheriff S, Silverton EW, Padlan EA, Cohen GH, Smith-Gill SJ, Finzel BC, Davies DR. 1987. Three-dimensional structure of an antibody–antigen complex. *Proc Natl Acad Sci USA 84*:8075–8079.

Schneider G, Lindqvist Y, Lundqvist T. 1990. Crystallographic refinement and structure of ribulose-1,5-bisphosphate carboxylase from *Rhodospirillum rubrum* at 17 Å resolution. *J Mol Biol 211*:989–1008.

Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I. 1991. Atomic structure of acetylcholinesterase from *Torpedo californica*: A prototypic acetylcholine-binding protein. *Science 253*:872–879.

Timkovich R, Dickerson RE. 1976. The structure of *Paracoccus denitrificans* cytochrome C500. *J Biol Chem 251*:4033–4055.

Thayer MM, Flaherty KM, McKay DB. 1991. Three-dimensional structure of the elastase of *Pseudomonas aeroginosa* at 15 Å resolution. *J Biol Chem 266*:2864–2871.

Tsuge H, Ago H, Noma M, Nitta K, Sugai S, Miyano M. 1992. Lysozyme from equine milk at 25 Å resolution. *J Biochem 111*:141–143.

Varadarajan R, Richards FM. 1992. Crystallographic structures of ribonuclease S variants with nonpolar substitutions at position 12: Packing and cavities. *Biochemistry 31*:12315–12327.

Verschueren HG, Seljee F, Rozeboom HJ, Kalk KH, Dijkstra BW. 1993. Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature 363*:693–698.

Wallace AC, Laskowski RA, Thornton JM. 1996. Derivation of 3D coordinate templates for searching structural databases: Application to the Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci 5*:1001–1013.

Weaver LH, Grutter MG, Matthews BW. 1995. The refined structures of goose lysozyme and its complex with a bound trisaccharide show that the "goose-type" lysozymes lack a catalytic aspartate residue. *J Mol Biol 245*:54–68.

Weaver LH, Matthews BW. 1987. Structure of bacteriophage T4 lysozyme refined at 17 Å resolution. *J Mol Biol 193*:189–199.

Wei Y, Schottel JL, Derewenda U, Swenson L, Patkar S, Derewenda ZS. 1995. A novel variant of the catalytic triad in the *Streptomyces scabies* esterase. *Nature Struct Biol 2*:218–223.

Wright CS, Alden RA, Kraut J. 1969. Structure of subtilisin *bpn* at 25 Å resolution. *Nature 221*:235–242.

Young ACM, Dewan JC, Nave C, Tilton RF. 1993. Comparison of radiation-induced and structure refinement from X-ray data collected from lysozyme crystals at low and ambient temperatures. *J Appl Crystallogr 26*:309–319.

Zegers I, Verhelst P, Choe HW, Steyaert J, Heinemann U, Saenger W, Wyns L. 1992. The role of histidine-40 in ribonuclease t1 catalysis: 3-Dimensional structures of the partially active His40Lys mutant. *Biochemistry 31*:11317–11325.