PROTEINS: Structure, Function, and Genetics 51:96–108 (2003)

Efficient Identification of Side-Chain Patterns Using a Multidimensional Index Tree

Thomas Hamelryck*
ULTR Department, Vrije Universiteit Brussel (VUB), Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Brussel, Belgium

ABSTRACT **Convergent evolution often produces similar functional sites in nonhomologous proteins. The identification of these sites can make it possible to infer function from structure, to pinpoint the location of a functional site, to identify enzymes with similar enzymatic mechanisms, or to discover putative functional sites. In this article, a novel method is presented that (a) queries a database of protein structures for the occurrence of a given side chain pattern and (b) identifies interesting side-chain patterns in a given structure. For efficiency and to make a robust statistical evaluation of the significance of a similarity possible, patterns of three residues (or triads) are considered. Each triad is encoded as a high-dimensional vector and stored in an SR (Sphere/Rectangle) tree, an efficient multidimensional index tree. Identifying similar triads can then be reformulated as identifying neighboring vectors. The method deals with many features that otherwise complicate the identification of meaningful patterns: shifted backbone positions, conservative substitutions, various atom label ambiguities and mirror imaged geometries. The combined treatment of these features leads to the identification of previously unidentified patterns. In particular, the identification of mirror imaged side-chain patterns is unique to the here-described method. Interesting triads in a given structure can be identified by extracting all triads and comparing them with a database of triads involved in ligand binding. The approach was tested by an all-against-all comparison of unique representatives of all SCOP superfamilies. New findings include mirror imaged metal binding and active sites, and a putative active site in bacterial luciferase. Proteins 2003;51:96–108.** © 2003 Wiley-Liss, Inc.

Key words: functional site; function from structure; mirror image; SR tree; structural bioinformatics; luciferase

INTRODUCTION

Novel, efficient methods are necessary to derive the maximum amount of knowledge from the enormous amount of macromolecular structure data that is currently being generated. With the advent of various structural genomics projects,[1] it will become more and more important to obtain as much information as possible from the 3D structure of a protein, as in many cases there might be little or no information available about biological function.

The 3D structure of a protein can provide clues about the protein's function through various comparisons with the set of previously solved structures present in the Brookhaven Protein Database (PDB).[2] This comparison can be done in many different ways and at different levels of structural detail.[3] One can use this approach from the level of the protein fold to the level of an active site in atomic detail. Fold recognition is a classic first step in examining a new structure, and numerous structural databases and classification methods have been described.[4–8] Often a common fold can provide clues about the protein's function, or the putative position of the functional site.[9]

As a logical next step, more detailed features can be considered. In contrast to the plethora of different fold recognition methods, only a few methods have been reported that can identify similarities at the atomic level (typically focusing on an active or ligand binding site).[10–13] More specifically, these methods identify recurring side-chain patterns in a database of structures. A typical use case of such a method, e.g., is querying a subset of the PDB database for the occurrence of (putative) catalytic Asp-His-Ser triads, as present in the trypsine superfamily. The scope of these methods is clearly more general but at the same time less powerful than those that deal with the identification of specific sites using carefully constructed templates.[14,15]

A related problem is the identification of putative functional sites in an entire protein structure.[11,13] Structural genomics projects will produce structures for which no putative active or binding sites are known. In this case, it

Abbreviations: 1,2-CTD, catechol 1,2-dioxygenase; DMSO, dimethylsulfoxide; FAH, fumarylacetoacetate hydrolase; FPAase, L-fuculose-1-phosphate aldolase; IPMDH, 3-isopropylmalate dehydrogenase; PDB, Brookhaven Protein Database; PheOH, phenylalanine hydroxylase; PPase, pyrophosphatase; rmsd, root mean square deviation; SCOP, structural classification of proteins; SR tree, sphere/rectangle tree.

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/2003/51/v51.96.html.

*Correspondence to: Thomas Hamelryck, ULTR Department, Vrije Universiteit Brussel, Vlaams Interuniversitair Instituut voor Biotechnologie, Pleinlaan 2, 1050 Brussel, Belgium. E-mail: thamelry@vub.ac.be

Received 29 August 2002; Accepted 24 October 2002

is desirable to identify putative functional sites by a large-scale comparison with the functional sites of previously solved structures. This is not a trivial problem, because of the poor annotation of active sites in PDB structures and possible over- and under-parameterization (i.e., using too few or too many residues to characterize a site).[13]

However, the above-mentioned methods that identify side-chain patterns make use of a number of simplifications for reasons of speed, simplicity, and memory use. This inevitably leaves interesting patterns undiscovered.

A first problem that is only partly handled in current state-of-the-art methods[10–13] is the ambiguity of the labeling of many side-chain atoms in crystal structures. This arises either because the side-chain atoms are chemically identical (for Glu, Asp, and Arg) or because the atoms cannot be distinguished from each other experimentally (for Asn, Gln, and His). Taking atom label ambiguity into account also played a critical role in determining the elusive rotamer preferences of Asn and Gln residues.[16] Current methods either assume that the atom labels are unambiguous[11,12] or represent the amino acid side-chains with dummy atoms.[10,13] For a typical pair of three His residues with a common geometry, e.g., the former methods would not correctly detect the pattern in all cases. The latter method will include many dissimilar patterns because of the lack of geometric detail in the dummy atom representation,[10,13] which also complicates calculating a suitable statistical significance of the rmsd. The above-mentioned methods use a subgraph isomorphism algorithm,[10] the geometric hashing method,[11] or a depth first search,[12,13] and dealing with atom label ambiguity would introduce a substantial speed or memory use penalty.

Current methods do not identify side-chain patterns that are mirror images of each other. Mirror-imaged active sites have recently been discovered by manual inspection in D-amino acid oxidase and flavocytochrome b2,[17] β-carbonic anhydrase and the α-carbonic anhydrases,[18] L- and D-amino acid oxidase[19] and AhpF and thioredoxin.[20] However, up to now, a systematic search for mirror-imaged side-chain patterns has never been performed. It is shown that such a systematic search indeed identifies new biologically relevant examples of side-chain patterns.

Here, a novel method is presented that can look for the occurrence of a given side-chain pattern consisting of three residues (here called *triads*) in a database of structures. By encoding the patterns as vectors, the method can deal with atom label ambiguities and mirror-imaged patterns without a large computational penalty. In the database, each triad is represented by a distance matrix vector, which allows efficient lookup of similar triads using a multidimensional index tree. This avoids a computationally intensive pairwise comparison of a pattern with each structure in the database,[10–13] at the expense of using a fixed pattern size. In addition, the method is also very modest in memory use.

The method simultaneously takes into account many factors that are crucial for the identification of many relevant patterns:

- Side-chain patterns that are mirror images of each other can be identified.
- The method deals with various atom label ambiguities in crystal structures.
- Only the chemically relevant atoms of a side chain are considered, which makes it possible to identify similarities that do not extend to the Cα positions of the involved residues and mirror-imaged similarities.
- Conservative substitutions that are often relevant (Asn/Gln, Asp/Glu, and Ser/Thr) are allowed.

The statistical significance of a similarity can be evaluated in different ways (using E, P, and Z values), taking into account the amino acid composition of the triad and the size of the database.

A second application of the method is the identification of putative functional sites in an entire structure. This is done by extracting all potentially interesting triads from the structure and comparing them with a database of triads involved in ligand binding. The latter database was constructed by extracting all triads near ligands from a set of unique representatives of all SCOP superfamilies. This approach is easy to implement and avoids problems with over- or underspecification of a binding site. To evaluate the usefulness of the approach, an all-against-all comparison of these representatives was done: each ligand-binding triad was used to query the set of all triads extracted from the SCOP superfamily representatives. The novel biologically relevant similarities described below prove the validity of this approach.

MATERIALS AND METHODS

Triad Representation

A triad of residues consists of a central amino acid of a certain type (the *core residue*) and two neighboring residues that have at least one side-chain atom within a radius of 6.0 Å of a prechosen atom (the *core atom*) of the side chain of the core residue. The latter atom roughly indicates the position of the chemically important group of the side chain of the core residue (e.g., Cγ for Asp). The core atoms are Cγ for Asp, Asn, and His, Sδ for Met, Nε1 for Trp, Oη for Tyr, Cδ for Glu and Gln, Cε for Lys, Cζ for Arg, Oγ for Ser and Thr, and Sγ for Cys. More formally, a triad is defined as an unordered set of three residues, of which at least one pair of residues interacts (judged by the 6.0-Å distance cutoff requirement as described above) with a third residue of the triad.

In principle, any of the 20 amino acid residues can be picked as core residue and/or neighboring residue. However, as in Russell et al.,[9] only amino acids whose side chains contain noncarbon atoms are considered here, because the exact orientation of these side chains is often critical in catalysis or ligand binding. Residues containing disordered atoms are left out.

The residues in a triad are represented by a subset of their side-chain atoms (see Fig. 1), e.g., Arg is represented by the atoms belonging to its guanidinium group (Cζ, Nε, Nη1, and Nη2).

The atom subset representation makes it possible to detect side-chain patterns in which only the chemically

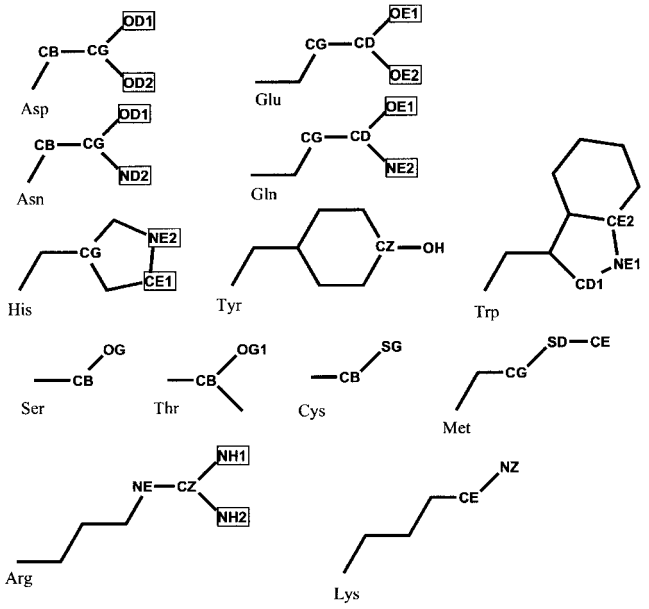

Fig. 1. The side chains of the 13 amino acids used in the triads. Atoms that are used to represent the side chain are labeled. The unboxed atoms can be superimposed unambiguously, based on the atom label. For boxed atoms, two possibilities need to be evaluated due to atom label ambiguities.

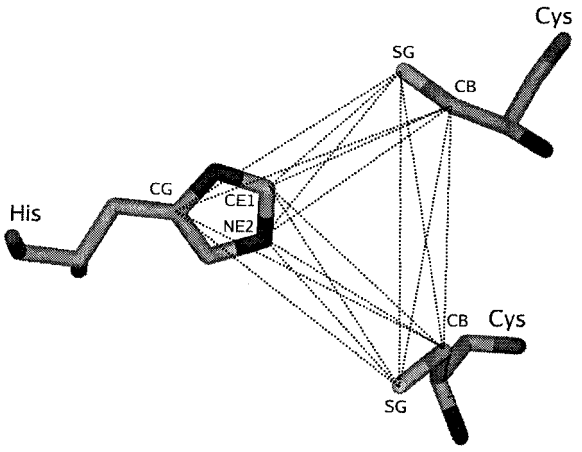

Fig. 2. Illustration of the vector representation of a triad consisting of two Cys and a His residue. The atom distances between the residues that are used to represent the triad are shown as dotted lines. The atoms used are labeled. The His, Cys, Cys triad is characterized by a $3 \times 2 + 3 \times 2 + 2 \times 2 = 16$-dimensional vector. Because His is subject to atom label ambiguity and the triad contains two Cys residues, this triad is represented by four vectors. Figures 2, 4, and 6–10 were made using MOLSCRIPT.[21]

important groups of the side chains coincide, whereas their Cα positions are shifted. This also plays an important role in the identification of mirror imaged triads, because the groups choosen here are reflection invariant. In addition, it allows to group amino acids together whose side chains have the same chemically relevant groups, namely Asp/Glu, Asn/Gln, and Ser/Thr. The 13 amino acids can thus be subdivided in 10 groups. Because the sequence order of the residues in a triad is irrelevant and the residues in the triads are sorted according to amino acid type, the triads can be grouped into 220 different triad types.

Triad Similarity Search

Each triad is represented by a set of vectors whose components consist of all inter-residue distances between the atoms in the chemically important groups (see Fig. 2). For each triad type, these distances are stored in a particular order in the vector. The dimension of the vector depends on the number of atoms used to represent the side chains in the triad. If the side chains are characterized by a, b, c atoms, the dimension of the vector will be $ab + ac + bc$.

Some triads (e.g., the Ser, Cys, Met triad) only need to be represented by a single vector. For other triad types, however, more than one vector needs to be used because of atom label equivalence. This can be due to experimental limitations (e.g., in the case of His) or chemical equivalence of the atoms (e.g., in the case of Asp). The equivalences of relevance here are: Oδ1, Oδ2 for Asp; Oε1, Oε2 for Glu; Nη1, Nη2 for Arg; Cε1, Nε2 for His; Oδ1, Nδ2 for Asn; and Oε1, Nε2 for Gln. An Asp, His, Arg triad is, e.g., represented by $2 \times 2 \times 2 = 8$ vectors, each corresponding to a particular permutation of the equivalent atoms and thus a particular order of the distances in the vector representation.

Another complication arises for triads that contain two or three residues that belong to the same amino acid type, because residues in a triad are sorted according to amino acid type. In those cases, the vectors are calculated for each permutation of the residues belonging to the same amino acid type. An X, Y, Y set of neighboring residues (where X and Y are amino acid types) is thus represented by two vectors sets, whereas an Y, Y, Y set is represented by six vector sets (see Fig. 2).

The vector representations of the triads can be used to quickly identify triads that are similar to a query triad. The vectors, whose dimensionality varies between 12 (e.g., a Ser, Ser, Ser triad) and 48 (e.g., an Arg, Glu, Asn triad), are stored in SR (Sphere/Rectangle) trees.[22] Multidimensional index trees like the SR tree are efficient data structures that can be used to perform nearest neighbor queries in high-dimensional vector spaces. They are used for example to implement content based retrieval of videos or images from multimedia libraries. The nodes of these index trees correspond to nested, possibly overlapping bounding regions in vector space. Leaf nodes point to a location on disk that stores the coordinates of all points within a bounding region. Thus, the point set does not have to fit in main memory. Typical examples of multidimensional index trees are R* and SS trees, where the bounding regions are rectangles and spheres, respectively. Rectangular and spherical bounding regions each have their advantages and disadvantages: the former have a low volume but a high diameter, and the latter have a low diameter but a high volume. The recently described SR tree[22] uses the intersection of a bounding sphere and a bounding rectangle, which leads to bounding regions with both low volume and low diameter. The SR tree outper-

forms both the R* and SS tree with respect to CPU usage and disk access, which increase sublinearly with increasing data set size. Another important point is that the set of SR tree vectors does not need to fit in memory. In view of the rapid growth of the PDB database, these are clearly important advantages.

Each of the 220 triad types has its own SR tree. For a vector representation of a given query triad, the N nearest neighbors are determined by lookup in the SR tree that corresponds to the query triad type. The parameter N should be greater than the expected number of similar triads for a given query, also taking into account that several nearest neighbor vectors can originate from one triad (when more than one relevant superposition is possible). N should thus increase with the degree of redundancy of the database.

For each of the nearest neighbor triads, the rmsd after superposition is calculated for all possible atom equivalences. The triads can then be ranked according to an rmsd based P-, E-, and Z-value (see next paragraph).

Two triads of the same type with geometries that are mirror images of each other will give rise to the same vector representation. The reflection invariance of the vector representation is an advantage here, because it allows the efficient identification of triads that are mirror images of each other. This is done by calculating the above rmsd's with the neighbor triads after reflection of the query triad.

Significance of the rmsd between a Triad Pair

The same rmsd value will have a different statistical significance for different triad types. Two factors will determine the significance of the rmsd between two triads: the triad type (which determines the number of atoms in the triad) and the number of triads of that type in the database.

The number of atoms used to characterize a side chain varies between 2 (e.g., Ser) and 4 (e.g., Arg), so a triad can thus contain from 6 (e.g., three Ser residues) to 12 atoms (e.g., three Arg residues). As can be expected, the number of triads in the database differs widely among the different triad types. The most and less frequently found triad types are Ser/Thr, Asp/Glu, Asn/Gln (2.5% of all triads), and Trp, Trp, Trp (0.01% of all triads).

The triads can be subdivided in 10 groups, based on the number of atoms used to represent each side chain in the triad. For example, a Ser, Ser, Ser triad is represented by (2, 2, 2) atoms, whereas an Asp/Glu, His, Ser/Thr triad is represented by (4, 3, 2) atoms. For each of these 10 groups, 1000 rmsd values between two random triads (not involved in ligand binding as defined below, and not belonging to domains with the same fold) of the same type were calculated. The distribution of the inverse rmsd values was well described by a gamma distribution in all 10 cases.

Fitting the inverse rmsd data to a statistical distribution was done with the program ExpertFit (Averill M. Law & Associates, Inc.). Using this program, a total of 32 standard statistical distributions were each in turn fitted to the data and ranked according to how well they described the distribution of the data. The gamma distribution (described by three parameters: shape α, scale β, and location γ) described the distribution sufficiently well (evaluation either "good" and "borderline") in all cases. Figure 3 shows the histogram of the inverse of the rmsd values and the fitted gamma distribution function for triads containing (4, 3, 2) atoms, as is the case for Asp/Glu, His, Ser/Thr triads. The gamma distributions were used to calculate the cumulative distribution function $P(x)$. $P(x)$ is the probability that the rmsd between two randomly picked triads is below or equal to x. The same $P(x)$ is used for the rmsd's between a triad and a mirrored triad, because the corresponding rmsd distributions were similar.

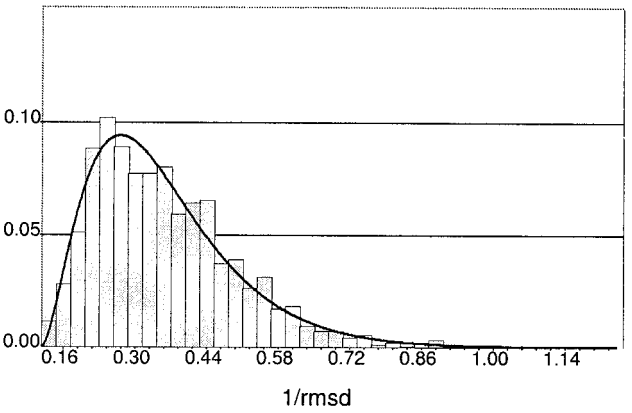

Fig. 3. Histogram of the inverse rmsd values and fitted gamma distribution function for triads containing (2,3,4) atoms. For the histogram, the interval size is 0.028 and the number of intervals is 40. The fitted gamma distribution function ($\alpha = 2.82$, $\beta = 0.084$, $\gamma = 0.15$) is shown as a black curve.

In addition to the P-value, the size of the database for each triad type was taken into account by calculating an E-value, which equals $P(x)$ times the number of triads in the database of a given type. $E(x)$ represents the expected number of triads with an rmsd below x resulting from a database search that are due to random similarities. The E-value is especially suited to compare the results of an all-against-all comparison, because it takes into account the number of atoms used to calculate the rmsd and the size of the database. In other words, the E-value can be used to roughly rank triad pairs of different types.

In addition, a third way to asses the significance of an rmsd was used. An indication of the significance of the rmsd between two triads of a certain type A,B,C can be obtained by calculating the following Z-value:

$$Z_{ABC}(rmsd) = \frac{rmsd - \mu_{ABC}}{\sigma_{ABC}}$$

where μ_{ABC}, σ_{ABC} are the mean and the standard deviation of a random set of nearest neighbor rmsd's for a given triad type ABC. The latter set of distances is generated as follows. For each of the 220 triad types, 300 random triads are selected from the database. If there are less than 300 triads present in the database for a certain triad type, the maximum amount of triads is selected. For each of these triads, the lowest rmsd to a triad in the database is

calculated as described above. In order to obtain a more unbiased set of distances, triads that are in the neighborhood of a ligand (see below) and triad pairs that belong to domains with the same fold are left out. For each of the 220 rmsd sets generated, the mean μ_{ABC} and the standard deviation σ_{ABC} are calculated. The *Z*-values for triads that are mirror images of each other are calculated in the same way. The μ_{ABC}, σ_{ABC} values are in this case calculated starting from the mirror images of (at most) 300 random triads.

The *Z*-value is useful to evaluate the results of a query with an interesting template triad because it has a clear geometric interpretation. A *Z*-value below zero indicates a similarity that is better than the average similarity between a random template triad of the same type and its nearest neighbor in the database.

Database Construction

The triad database was built using a representative domain from each super-family in the ASTRAL/SCOP classification[23] (version 1.55, http://astral.stanford.edu/scopseq-gd-1.55.html). In total, 941 domains were used. The average number of triads per domain was 381, and in total 358,552 triads were extracted from the domains. Only the core residue in the triad needs to belong to the domain as defined in ASTRAL/SCOP; the two other residues in the triad are allowed to belong to other polypeptide chains present in the PDB file that contains the specified domain. This was the case for 10,822 triads. In 124 cases, the latter triads were also near ligands (as defined in the next paragraph). For NMR structures consisting of an ensemble of models, the first model present in the file was used.

All-against-all Comparison of Superfamily Representatives

All triads were extracted from the above database that have at least one heteroatom (not belonging to a water molecule) within a radius of 4.5 Å of the core atoms of the three residues in the triad, resulting in a set of 7971 triads. For each of these triads, a triad search was performed. The value of parameter *N* (which should be greater than the number of expected hits) was 30. The whole procedure takes about 3 h on a 1-GHz PC.

In order to minimize the number of trivial hits, only triad pairs with the triads originating from domains with a different fold were retained. The remaining results were sorted according to their *E*-value, and all matches with an *E*-value below 0.01 were visually inspected. Both the *E*- and *Z*-value proved to be useful to assess the significance of the similarities: most interesting triad pairs occurred for *E*-values below 0.01 and *Z*-values below −1.5. A complete list of the found triads is available online as supplementary information.

RESULTS AND DISCUSSION

Test Case: The Classic Asp-His-Ser Catalytic Triad

The classic Asp-His -Ser triad (for a review, see Dodson and Wlodawer[24]) was first discovered in the active site of the protease chymotrypsin and is present in members of the trypsine superfamily. The same triad with a similar geometry is found in two other superfamilies: the subtilase and the α/β-hydrolase superfamilies (including the cutinase family, which is classified as belonging to the flavodoxin superfamily in SCOP).

A search was performed using the catalytic triad of bovine trypsine (pdb code 1mct, residues Asp 102, His 57, and Ser 195), to test the method and compare the results with those obtained with the method of Russel.[12] The parameter *N* was set to 30, as in the all-against-all comparison. The used set of SCOP superfamily representatives contains four domains in which a catalytic Asp-His-Ser triad is present: hydroxynitrile lyase (PDB code 1qj4), acetylxylan esterase (PDB code 1g66), subtilisin from *B. lentus* (PDB code 1gci) and trypsinogen from *F. oxysporum* (PDB code 1gdn). The results of the query are shown in Table 1. The hit with the highest rank is trypsinogen, closely followed by substilisin in the fourth position (see Fig. 4). The two triads present in acetylxylan esterase and hydroxynitrile lyase were, however, not discovered. Inspection of the triads of these two structures showed that in both cases one or more of the catalytic residues contain disordered atoms. Because disordered residues are not included in the triad database, these triads are not detected.

In contrast to the method of Russell et al.,[9] the current method identifies new putative triads with a *E*-, *P*-, and *Z*-values that are comparable to those for true catalytic triads (see Fig. 4). In both cases, the Asp residue is replaced by a Glu residue, whose carboxyl group exactly coincides with the carboxyl group of the Asp residue in the query triad. These triads are present in the NAD-binding domain of human HMG-COA reductase and in dimethylsulfoxide (DMSO) reductase from *R. sphaeroides.* Replacement of Asp by Glu in a true catalytic triad has also been observed for an acetylcholinesterase and a lipase.[25] The significance of these two previously undescribed putative catalytic triads is unclear, but they might well be biologically relevant. This illustrates how the identification of side-chain patterns in a structure can provide interesting starting points for further experimental research.

The four hits discussed above are the only ones that occur with an *E*-value below 1.0, a *Z*-value below 0.0, or a *P*-value below 1.0e-4. Note that there is a sharp increase of *E*- and *P*-values and that the *Z*-value jumps from negative to positive between the last true catalytic triad (subtilisin) and the next putative catalytic triad (Ni-Fe hydrogenase), suggesting that the hits below subtilisin are not biologically significant.

To compare the results with the method of Kleywegt,[13] the freely available program SPASM was run with the same triad (PDB code 1mct, residues Asp 102, His 57, Ser 195) as input. Two searches were performed: one using a dummy atom to represent the side-chain atoms and one using the Cα position in addition to the dummy atom. In both cases the recommended 1.0 rmsd threshold for three-residue patterns was used. Asp/Glu substitutions were allowed. Neither of the two putative triads were identified,

TABLE I. Results of the Search with the Asp-His-Ser Catalytic Triad of Bovine Trypsine

Protein	PDB code	rmsd	*E*	*P*	*Z*	Triad
Trypsinogen	**1gdn**	**0.16**	**< 1E-14**	**<1E-16**	**−4.02**	**D102A, H57A, S195A**
DMSO reductase	**1eu1**	**0.56**	**6.76E-03**	**1.2E-6**	**−1.52**	**E193A, H40A, S646A**
HMG-CoA reductase	**1dqa**	**0.64**	**7.39E-02**	**1.3E-5**	**−0.98**	**E700A, H635A, S637A**
Subtilisin	**1gci**	**0.69**	**1.95E-01**	**3.5E-5**	**−0.71**	**D32, H64, S221**
Ni-Fe hydrogenase*	1h2r	0.86	3.31	6.0E-4	0.41	E16S, H13S, T47S
Substilisin*	1gci	0.86	3.32	6.0E-4	0.42	D32, H64, S221
Carbamate kinase	1e19	0.88	4.52	8.2E-4	0.52	E225A, H173A, S192A
Nitrite reductase*	1qks	0.89	5.15	9.3E-4	0.67	D411A, H408A, T382A
Transducin*	1got	0.95	9.32	1.7E-3	1.05	D247B, H225B, S201B
p41*	1icf	0.98	1.25	2.3E-3	1.26	E141A, H2531, S2491
Transducin*	1got	0.98	1.35	2.4E-3	1.31	D205B, H183B, S201B
Ni-Fe hydrogenase*	1h2r	1.00	1.52	2.7E-3	1.40	E75S, H13S, T 47S
Trypsinogen*	1gdn	1.01	1.75	3.2E-3	1.51	D 102A, H57A, S195A
Trp synthase	1qop	1.04	2.28	4.1E-3	1.51	E109B, H115B, S119B
DMSO reductase*	1eu1	1.01	1.81	3.3E-3	1.54	E193A, H40A, S646A
Ni-Fe hydrogenase*	1h2r	1.04	2.23	4.0E-3	1.71	E16S, H143S, S21S
Ni-Fe hydrogenase	1h2r	1.08	3.22	5.8E-3	1.77	E75S, H13S, T47S
Carbamate kinase*	1e19	1.06	2.56	4.6E-3	1.82	E225A, H173A, S192A
Benzoylformate decarboxylase	1bfd	1.11	3.83	6.9E-3	1.92	E28, H70, S26

All triad pairs with an *E*-value below 4.0 are listed. The results are ranked according to their *E*-value. An asterisk next to the protein name indicates a match with the mirrored triad. The four triad pairs with a negative *Z*-value are shown in bold.

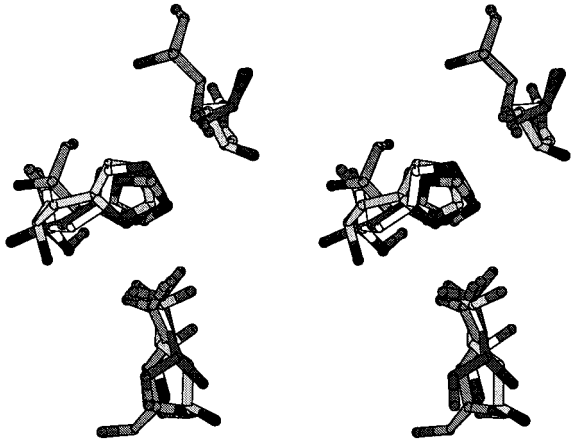

Fig. 4. Stereo figure of the three highest scoring Asp-His-Ser triads not belonging to a trypsine homologue. Each triads was superimposed on the query triad. The carbon atoms of the different triads are colored with increasing shades of grey: the query triad (white), DMSO reductase, HMG-CoA reductase, and subtilisin (darkest grey).

although both structures in which these triads are found (PDB codes leu1 and 1dqa) were present in the SPASM library.

On the same desktop PC, SPASM is considerably slower (26 s for SPASM and under 2 s for the here-described method) for the same query. This comparison, admittedly simplistic because the methods have a different scope and because the database size should be taken into account, indicates that the here-described method is speed efficient, especially because the current implementation was not optimized for speed. Because the performance of the here-used SR tree data structure scales well with data set size,[22] a larger database of structures will only have a limited effect on its performance.

To evaluate specificity and sensitivity issues in more detail, a database consisting of 44 structures containing (putative or catalytic) Asp-His-Ser triads was used.[9] The database consists of 30 structures that contain a true, functional catalytic triad, and 14 structures that contain a putative, most probably nonfunctional catalytic triad. In two cases, a PDB file used in Russell (4ptp, 1amg)[9] was replaced by an up-to-date version of the same structure (5ptp, 2amg). The question addressed here is wether there the *P*- and *Z*-values can be used to distinguish the putative, probably nonfunctional catalytic triads from the functional ones and whether all expected Asp-His-Ser triads are correctly identified.

Again, a query was performed with the catalytic triad of bovine trypsine (pdb code 1mct). For this query, the parameter *N* was set to 200, because at least 44 matches are expected. All matching triads with a *Z*-value below 6.0 were selected. To facilitate the discussion, secondary hits found in structures containing a true catalytic triad were omitted. These secondary hits mostly involve mirror imaged catalytic triads or are due to catalytic tetrads[26] (which involve a second Ser residue). All matches with structures containing a putative triad were included.

The great majority (28 of 29) of the functional triads in the database are found in the 29 lowest *Z*- or *P*-ranking triad pairs (see Table II). Plots of the rank of the triad matches against their *Z*- and *P*-values are shown in Fig. 5. The *Z*- and *P*-values for these triads range between −3.99 and $<10^{-16}$ (for human leukocyte elastase, pdb code 1ppf) and 1.07 and 2.2×10^{-3} (for *Vibrio harveyi* thioesterase, pdb code 1tht). The only putative triad found among these is present in 1rib with *Z*- and *P*-values of 0.63 and $1.0 \times$

TABLE II. *Z*-Values, *P*-Values, and rmsd of the 50 Asp-His-Ser Triads with Lowest *Z*-Values

PDB code	Triad	*Z*	*P*	Rmsd (Å)	PDB code	Triad	*Z*	*P*	Rmsd (Å)
1ppf	**S195E,D102E,H57E**	**−3.99**	**<1E-16**	**0.2**	**2sga**	**S195,D102,H57**	**−3.81**	**<1E-16**	**0.2**
2alp	**S195,D102,H57**	**−3.78**	**<1E-16**	**0.2**	**3rp2**	**S195A,D102A,H57A**	**−3.73**	**<1E-16**	**0.2**
3sgb	**S195E,D102E,H57E**	**−3.68**	**<1E-16**	**0.2**	**1rtf**	**S195B,D102B,H57B**	**−3.43**	**1.1E-16**	**0.2**
1hyl	**S195A,D102A,H57A**	**−3.36**	**7.8E-16**	**0.3**	**1sgt**	**S195,D102,H57**	**−3.33**	**1.8E-15**	**0.3**
1kxf	**S215,D163,H141**	**−3.17**	**6.0E-14**	**0.3**	**2sfa**	**S147,D65,H35**	**−3.14**	**1.0E-13**	**0.3**
3gct	**S195A,D102A,H57A**	**−2.97**	**2.1E-12**	**0.3**	**1pfx**	**S195C,D102C,H57C**	**−2.90**	**5.9E-12**	**0.3**
1hpg	**S195A,D102A,H57A**	**−2.09**	**3.4E-08**	**0.5**	**1hcg**	**S195A,D102A,H57A**	**−1.70**	**4.3E-07**	**0.5**
1lmw	**S195B,D102B,H57B**	**−1.64**	**6.3E-07**	**0.5**	**1fuj**	**S195A,D102A,H57A**	**−1.44**	**1.8E-06**	**0.6**
1thm	**S225,D38,H71**	**−1.42**	**2.0E-06**	**0.6**	**2prk**	**S224,D39,H69**	**−1.03**	**1.1E-05**	**0.6**
1svn	**S221,D32,H64**	**−0.99**	**1.3E-05**	**0.6**	**1cse**	**S221E,D32E,H64E**	**−0.96**	**1.4E-05**	**0.7**
1mee	**S221A,D32A,H64A**	**−0.83**	**2.4E-05**	**0.7**	**1st3**	**S215,D32,H62**	**−0.50**	**6.8E-05**	**0.7**
1fon	**S187A,D93A,H43A**	**−0.25**	**1.4E-04**	**0.8**	**1cus**	**S120,D175,H188**	**0.00**	**2.7E-04**	**0.8**
1mpt	**S221,D32,H64**	**0.11**	**3.5E-04**	**0.8**	**1bro**	**S98A,D228A,H257A**	**0.29**	**5.2E-04**	**0.8**
1rib	S114A,E204A,H241A	0.63	1.0E-03	0.9	**1tah**	**S87B,D263B,H285B**	**0.67**	**1.1E-03**	**0.9**
1tht	**S114A,D211A,H241A**	**1.07**	**2.2E-03**	**1.0**	1rib*	S114A,D84A,H118A	2.12	6.5E-03	1.1
1rib	S114A,E238A,H241A	2.33	1.1E-02	1.2	2rmc*	S133A,D157A,H126A	2.48	9.2E-03	1.1
1mio	S183A,D374A,H371A	2.53	1.4E-02	1.2	2cpl*	S99,D123,H92	2.55	9.9E-03	1.2
1amp	S77,D160,H70	2.59	1.4E-02	1.2	1pta	S231,D253,H254	2.71	1.6E-02	1.2
1amp*	S77,D160,H70	3.00	1.5E-02	1.2	1pta	T202,D235,H230	3.03	2.1E-02	1.2
1pta*	S231,D253,H254	3.03	1.5E-02	1.2	1rib*	S114A,E204A,H241A	3.04	1.5E-02	1.2
1vhh	S185,D148,H183	3.04	2.1E-02	1.3	1mio	T22B,D320B,H146B	3.11	2.3E-02	1.3
1rib	S114A,D84A,H118A	3.22	2.5E-02	1.3	1rib*	S114A,E238A,H241A	3.25	1.8E-02	1.3
1cel	S174A,D214A,H228A	3.34	2.7E-02	1.3	1cel	T226A,D257A,H228A	3.39	2.8E-02	1.3
1cel*	T226A,E212A,H228A	3.43	2.1E-02	1.3	1mio*	T22B,D320B,H146B	3.52	2.2E-02	1.3
1mio	S69D,D113D,H43D	3.53	3.E-02	1.4	2cpl	S99,D123,H92	3.58	3.3E-02	1.4

True catalytic triads are shown in bold. Mirror imaged triads are indicated with an asterisk next to the PDB code.

10^{-3}. Interestingly, the *Z*-and *P*-values of the two putative triads identified in HMG-COA reductase and DMSO reductase (with $Z = -1.52$ and -0.98, and $P = 1.2 \times 10^{-6}$ and 1.3×10^{-5}, respectively) lie well within the range of the *Z*- and *P*-values of the true catalytic triads.

The triad in tonin (pdb code 1ton, Asp 102, His 57, Ser 195) is the only functional triad that occurs with much higher *Z*- and *P*-values ($Z = 4.48$, $P = 5.9 \times 10^{-2}$). However, the catalytic site of the tonin structure is distorted due to the presence of a Zn-ion,[37] which explains why this triad is identified with higher *Z*- and *P*-values. The catalytic triad in 5ptp (trypsine) was not found because its catalytic Ser residue is covalently modified to mono-isopropyl-phosphoryl-serine. Hence, the Ser residue is flagged as a hetero residue in the PDB file and thus not recognized as an amino acid during database construction.

Below a *Z*-value of 0.0 or a *P*-value of about 10^{-4}, 24 of 29 (83%) functional triads are found, whereas no false positives occur. A *Z*-value of zero or a *P*-value of 10^{-4} thus seems to be a conservative cutoff to distinguish genuine from random similarities for a functional triad.

Overview of the All-against-all Comparison Results

In total, 2733 triad pairs with an *E*-value below 0.01 were found. Of these, 1199 matches involved mirror symmetry.

The 10 top scoring triad pairs are shown in Table III, together with a short description of the triads. The top 10 hits include five metal binding sites, several similarities due to structural reasons (one salt bridge, one helix N-cap, one glycosylated loop and one similar beta sheet region)

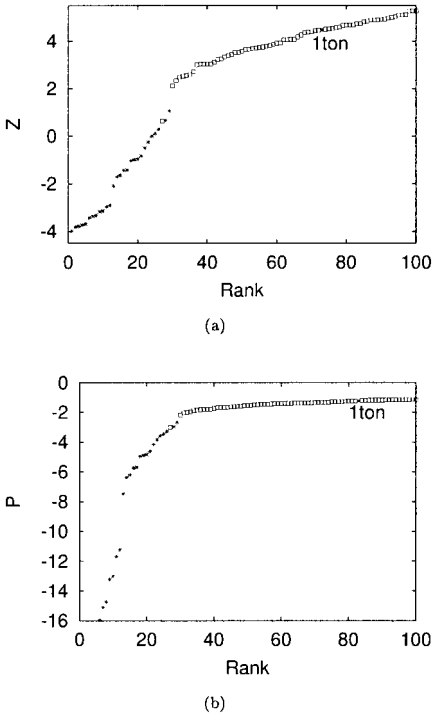

Fig. 5. *Z*- and *P*-values of the Asp-His-Ser triad matches. Putative triads are plotted as boxes, and catalytic triads are plotted as asterisks. The catalytic triad in tonin (PDB code 1ton), which has high *Z*- and *P*-values, is labeled. The 100 triads with lowest *Z*- and *P*-values are plotted. (a) *Z*-value versus rank. (b) *P*-value versus rank.

TABLE III. The 10 Highest Scoring Triads

Rmsd	*Z*	PDB code	Protein	Triads	Description
0.20*	−4.35	2mhr	Myohemerythin	H106,H77,H73	Metal binding site
		le4c	L-Fuc-1-phosphate aldolase	H92P,H94P,H155P	(Fe in 2mhr, Zn in le4c)
0.14	−4.29	1zme	PUT3	S36C,C37C,C34C	Zn binding site
		1vfy	vps27p	S224A,C225A,C222A	
0.18	−3.54	1hxr	RabGEF	C23A,C97A,C94A	Zn binding site
		1ee8	MutM	C258A,C241A,C238A	
0.22	−3.38	1qnf	DNA photolyase	S244,D380,R352	ST-DE-R salt bridge
		1svb	Envelope glycoprotein	T325,E373,R9	
0.23	−3.23	1d0q	DNA primase	C61A,C64A,C40A	Zn binding site
		1adn	Ada DNA repair protein	C42,C38,C72	
0.18	−3.05	2uag	MurD	S160A,S159A,Q162A	Helix N-cap
		1aop	Sulfite reductase	T93,T92,Q95	
0.22	−3.05	1ali	ZIF268	C107A,C112A,H125A	Zn finger
		1esk	HIV nucleocapsid	C36A,C39A,H44A	
0.19	−3.02	1d7b	cellobiose dehydrogenase	S112A,T113A,N111A	Similar loops. *N*-glycosylated
		1hvb	D-ala carboxypeptidase/ transpeptidase	S276A,T277A,N275A	at N111A in 1d7b.
0.22	−2.87	1fnd	Ferredoxin reductase	S96,S75,Y95	FAD binding site in 1fnd.
		1bow	BmrR	S62A, T61A,Y60A	FAD position is occupied by E65A in BmrR
0.21	−1.90	1qqq	Thymidylate synthase	T46A,T47A,S254A	Similar beta sheet regions.
		1lla	Hemocyanin	S182,T183,S311	

The 10 triad pairs (occurring between different structures) with lowest *E*-values. The asterisk in the first row indicates that this match is a mirror image match. Because all *E*-values are very small (below 1e-14), the triad pairs are sorted according to their *Z*-values, and only the latter value is shown.

TABLE IV. Number of triad pairs containing His and/or Cys residues.

Triad	Number of pairs	Number of mirrored pairs
HHH	28	6
CHH	5	3
CCH	121	40
CCC	1433	583
Total	1587	642

For each triad type, the total number of triad pairs and the number of triad pairs involving mirror symmetry is shown.

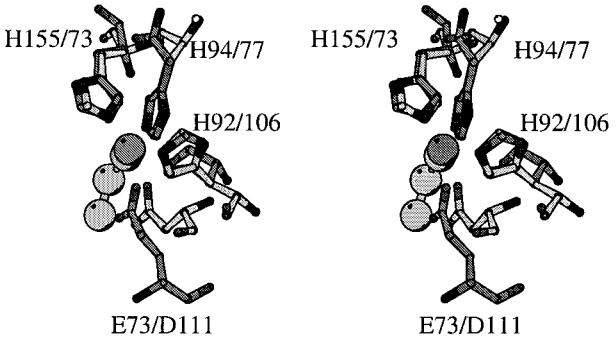

Fig. 6. The superimposed triad pairs in FPAase (dark grey, first label) and myohemerythrin (light grey, second label). The zinc and Fe_2O atoms are shown as spheres. The FPAase residues are mirrored.

and a similarity with an FAD binding site whose significance is unclear.

In the next subsections, various biologically relevant examples are given that illustrate the capabilities of the method. These include mirror-imaged metal binding sites, active site similarities, active site similarities involving mirror symmetry, and the identification of a putative active site in luciferase.

Mirror-imaged Metal-binding Sites

As can be expected, many similarities involve metal binding sites. Of the 2733 triad pairs, about 58% were metal-binding sites containing only His and/or Cys residues (see Table IV). It is interesting that metal binding sites that are mirror images of each other are quite common: about 40% of the metal binding triad pairs involve mirror symmetry.

L-fuculose-1-phosphate aldolase (FPAase, PDB code 1e4c), an enzyme that plays a role in carbohydrate metabolism and myohemerythrin (PDB code 2mhr), an oxygen carrier present in worms, contain a representative mirror symmetric metal-binding triad pair. The triads consist of the residues His 73, His 77, and His 106 for myohemerythrin and His 155P, His 94P, and His 92P for FPAase (rmsd = 0.20 Å, *P*-value $< 10^{-16}$, *E*-value $< 10^{-14}$, *Z*-value = −4.35, see Fig. 6). Visual inspection revealed an additional matching residue pair: Asp 111 (myohemerythrin) and Glu 73P (FPAase). In the case of myohemerythrin, these residues are involved in the coordination of an Fe_2O ligand. In the case of FPAase, they coordinate a Zn^{2+} ion, whose position coincides exactly with one of the Fe^{2+} ions in myohemerythrin. The Fe_2O ligand in myohemerythrin is further coordinated by three residues (His 25, His 54, Glu 58) that have no structural counterpart in FPAase.

TABLE V. Mirror Imaged Metal-binding Sites

Rmsd	*E*	*Z*	*P*	PDB code	Name	Triad	Metals
0.20	<1E-14	−4.35	<1E-16	2mhr	Myohemerythin	H106,H77,H73	Fe
				1e4c	L-Fuc-1-phosphate aldolase	H92P,H94P,H155P	Zn
0.19	<1E-14	−3.54	<1E-16	1rb9	Rubredoxin	C6,C39,C9	Fe
				4mt2	Metallothionein	C24,C29,C19	Zn
0.21	<1E-14	−3.45	<1E-16	1rb9	Rubredoxin	C6,C42,C39	Fe
				1rmd	RAG1	C29,C46,C26	Zn
0.16	<1E-14	−3.77	<1E-16	1rmd	RAG1	C29,C46,C26	Zn
				1d09	Asp carbamoyltransferase	C109B,C141B,C138B	Zn
0.22	1.06E-13	−3.36	2.2E-16	1ee8	MutM	C258A,C241A,C238A	Zn
				4mt2	Metallothionein	C24,C15,C29	Zn
0.22	1.60E-13	−3.34	3.3E-16	4mt2	Metallothionein	C29,C15,C24	Zn
				1d09	Asp carbamoyltransferase	C138B,C141B,C109B	Zn
0.26	2.35E-11	−3.15	4.9E-14	1qf8	Casein kinase-II	C140A,C137A,C114A	Zn
				4mt2	Metallothionein	C26,C7,C15	Zn
0.26	4.48E-11	−3.12	9.4E-14	1hxr	RabGEF Mss4	C23A,C26A,C94A	Zn
				4mt2	Metallothionein	C24,C19,C29	Zn
0.27	8.53E-11	−3.09	1.8E-13	1hxr	RabGEF Mss4	C23A,C97A,C94A	Zn
				1d09	Asp carbamoyltransferase	C114B,C138B,C109B	Zn
0.28	2.75E-10	−3.03	5.8E-13	1hxr	RabGEF Mss4	C23A,C97A,C94A	Zn
				1rmd	RAG1	C29,C46,C26	Zn

The 10 mirror-imaged metal-binding triad pairs (occurring between different pairs of structures) with the lowest *E*-values are listed. The triad pairs are sorted according to their *E*-values.

The triad His 73/His 155P, His 106/His 92P, Asp 111/Glu 73P is also detected with a very low *E*-value (*E*-value $<10^{-14}$, *Z*-value = −4.08). This illustrates that patterns consisting of more than three residues could be identified by looking for overlapping and/or neighboring triads (see Conclusions).

The 10 mirror-imaged metal binding with the highest *E*-value, although not occurring between the same two structures are shown in Table V.

Elastase and Phe Hydroxylase

A His-His-Glu/Asp triad that ligates a Fe^{3+} ion occurs in the active sites of a number of unrelated enzymes. The Fe^{3+} ion is only ligated by amino acids on one side, whereas the remaining coordination sites are taken in by water and substrate molecules. This type of triad has been called a 2-His-1-carboxylate facial triad.[28] A potentially functional 2-His-1-carboxylate facial triad was found in bacterial luciferase (see next paragraph), and a Zn^{2+}-binding variant of the triad is present in elastase.

Elastase[29] from *Pseudomonas aeruginosa* is a member of the thermolysin family of zinc-dependent metalloproteases. Phenylalanine hydroxylase (PheOH)[30] catalyzes the hydroxylation of Phe to Tyr using molecular oxygen and tetrahydrobiopterin as co-substrates.

Elastase (PDB code 1ezm) and PheOH (PDB code 3pah) contain a triad pair (rmsd = 0.52 Å, *P*-value = 1.1×10^{-7}, *E*-value = 8.4×10^{-5}, *Z*-value = −2.46, see Fig. 7) that consists of the residues His 140, His 144, and Glu 164 (elastase) and His 285, His 290, and Glu 330 (PheOH). In both cases the three residues are involved in metal coordination: Zn^{2+} in the case of elastase and Fe^{3+} in the case of PheOH. Visual inspection revealed two additional matches: Glu 141 and Tyr 155 in elastase occupy similar positions as Glu 286 and Tyr 325 in PheOH.

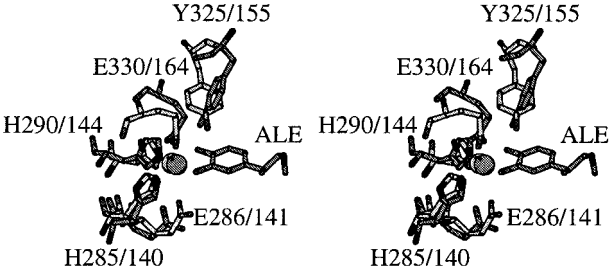

Fig. 7. The superimposed triad pair found in PheOH (dark grey, first label) and elastase (light grey, second label). The metal ions (Fe^{3+} in the case of PheOH and Zn^{2+} in the case of elastase) are shown as spheres. An adrenaline molecule bound in PheOH is labeled ALE.

In elastase, the Zn^{2+} ion (coordinated by His 140, His 144, and Glu 164) polarizes the carbonyl group of the protein backbone and facilitates the deprotonation of the nucleophilic water molecule by Glu 141. Tyr 155 hydrogen bonds to the hydrated peptide, together with His 223 that has no structural counterpart in PheOH (not shown in Fig. 7). The reaction mechanism of PheOH is unknown, but probably involves the creation of "activated" O_2 by ferrous iron. In elastase, the Zn^{2+} ion and Tyr 155 are part of an oxyanion hole.

A Putative Active Site in Bacterial Luciferase

Bacterial luciferase[31] is used by bacteria to produce blue-green light. Luciferase catalyzes a monooxygenase reaction, in which molecular oxygen reacts with reduced flavin and an aldehyde. Dioxygenases such as catechol 1,2-dioxygenase (1,2-CTD)[32] play a role in the aerobic catabolism of aromatic compounds by bacteria and catalyze ring cleavage by incorporation of molecular oxygen.

Luciferase (PDB code 1luc) and 1,2-CTD (PDB code 1dmh) contain a mirror symmetric triad pair (rmsd = 0.73

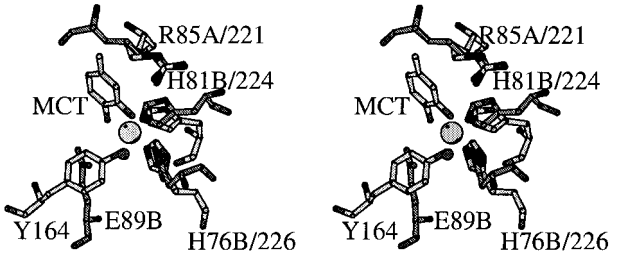

Fig. 8. The superimposed triad pair found in luciferase (dark grey, first label) and 1,2-CTD (light grey, second label). The bound Fe^{3+} ion in 1,2-CTD is shown as a light grey sphere. Wat 3135 in luciferase is shown as a dark grey sphere. The luciferase residues are mirrored. The 4-methylcatechol molecule bound in 1,2-CTD is labelled MCT.

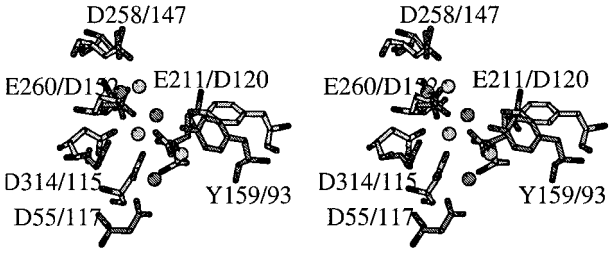

Fig. 9. The superimposed triad pair of phytase (dark grey, first label) and PPase (light grey, second label). The bound metal ions are shown as spheres. The PPase residues are mirrored.

Å, *P*-value = 4.4×10^{-5}, *E*-value = 1.0×10^{-2}, *Z*-value = -2.44, see Fig. 8). The triad consists of the residues Arg 221A, His 224A, and His 226A for 1,2-CTD, and Arg 85A, His 81B, and His 76B for luciferase. In 1,2-CTD, the two His residues are part of a modified 2-His-1-carboxylate facial triad, in which the carboxylic residue is replaced by a Tyr residue (Tyr 164A), whose Oη atom ligates the Fe^{3+} ion. The Oη atom of the iron-coordinating Tyr 164A residue in 1,2-CTD coincides exactly with a water molecule (Wat 3135) that is coordinated by Glu 89B in luciferase. The His 76B, His 81B, Glu 89B triad in luciferase is thus remarkably similar to the modified 2-His-1-carboxylate facial triad in 1,2-CTD.

No metal ions are coordinated by this triad in luciferase, but possibly metal binding depends on substrate binding. In uncomplexed 1,2-CTD, a fourth residue (Tyr 200) ligates the Fe^{3+} ion. This residue dissociates from the Fe^{3+} ion upon ligand binding. The fact that this residue has no structural counterpart in luciferase might explain why ligand free luciferase does not contain a bound metal ion, whereas 1,2-CTD contains a constitutive iron ion.

Bacterial luciferase is a heterodimer that consists of two closely related subunits termed α and β. The single active site per dimer is thought to reside in the α-subunit. However, the β_2 homodimer still shows considerable activity. The active site in the β subunit cannot be located in the same position as the putative active site of the α-subunit, because of structural differences.[33] Tanner et al.[33] proposed an alternative active site for the β_2 homodimer, near the dimer interface. This site is a large, solvent-accessible cavity lined with an unusual high density of basic residues. Because a comparable site is also present in the native αβ heterodimer,[31] it was proposed that this site could be the actual active site in the native αβ heterodimer as well. The low activity of the β_2 heterodimer would then be due to the absence of a 29-residue flexible loop that would shield the active site from the solvent upon substrate binding and that is only present in the α subunit.

The above described triad (and the Tyr 164A residue) is found at exactly this site and thus provides additional evidence for the location of the active site at the dimer interface. Alternatively, the site might be an evolutionary remnant of an ancient active site. In an alignment of seven α and β subunit amino acid sequences,[34] Arg 85A and His 81B are completely conserved, His 76B is replaced in 5 cases by Gln, and Glu 89B is replaced in a single case by Gln. All these mutations could potentially preserve the site's structure and function. Similar to 1,2-CTD, the active site of luciferase could consist of a modified iron-binding 2-His-1-carboxylate facial triad (consisting of His 76B, His 81B, and Glu 89B). In 1,2-CTD, Arg 221A is thought to position the substrate in the active site and to stabilize the increased electron density on a carbon atom of the aromatic ring, which might be the case for Arg 85A in luciferase as well.

Phytase and Pyrophosphatase

Pyrophosphatase (PPase)[35] hydrolyzes inorganic pyrophosphate. Phytases[36] hydrolyze phytate, which is the main storage form of phosphorus in many plants.

A mirror symmetric triad pair (rmsd = 0.6 Å, *P*-value = 6.6×10^{-6}, *E*-value = 9.5×10^{-3}, *Z*-value = -1.7) was found in the structures of a thermostable phytase (PDB code 1poo) and a PPase from yeast (PDB code 8prk). The triads consist of the residues Glu 211, Glu 260, and Asp 258 of phytase, and Asp 120, Asp 152, and Asp 147 of PPase (see Fig. 9). In addition, visual inspection revealed an additional match between Tyr 159 of phytase and Tyr 93 of PPase. The residues in the triads and both Tyr residues are part of the active sites of the two enzymes. In the active site of PPase three Mn^{2+} ions are bound, whereas in the 1poo phytase structure no metal ions were present. Phytase, however, has a number of low-affinity Ca^{2+}-binding sites in its active site, three of which are occupied in the phytase structure with PDB code 2poo. The latter structure was thus used in the comparison.

The catalytic mechanisms of phytase[36] and PPase[35] have been studied in detail and turn out to be quite similar. In both cases, a water nucleophile that is coordinated to two metal ions attacks the phosphorus atom of the substrate. The phosphate group on which the attack occurs is charge shielded by the metal ions in the active site to facilitate the attack of the water nucleophile. A general acid donates a proton to the leaving group: an amino acid (Lys 76) or a water molecule in the case of phytase, and a water molecule for PPase.

The two metals that activate the water nucleophile interact with Glu 211 in phytase and with the corresponding residue Asp 120 in PPase. The two other residues in the triads ligate the various metal ions around the bound phosphate groups in both enzymes. Mutation of Tyr 159 to

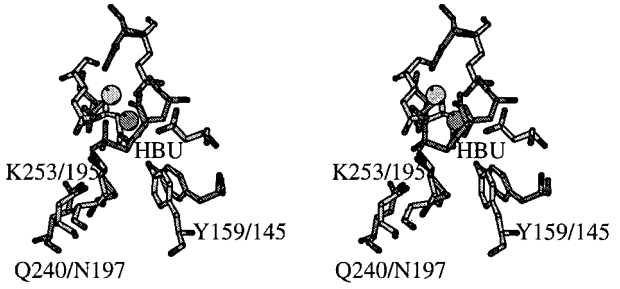

Fig. 10. The superimposed triad pair found in FAH (dark grey, first label) and IPMDH (light grey, second label). Metal ions are shown as spheres. The acidic residues that form the metal binding sites are unlabeled in order to simplify the figure. The IPMDH residues are mirrored. An inhibitor (4-hydroxymethyl-phosphinoyl-3-oxo-butanoic acid) in the active site of FAH is labeled HBU.

Phe in phytase results in complete loss of activity.[37] The Tyr residue plays a critical role in coordinating the water binding network around the Ca^{2+} ions for phytase and thus quite likely in PPase as well.

Both PPase and phytase contain a "cleavage site," where the nucleophilic attack by the water molecule occurs, and an "affinity site," which binds a second phosphate group (either a second phosphate group on the phytase ring or the second phosphate group in pyrophosphate). The residues in the two triads and the Tyr residues belong largely to the "cleavage site" for both enzymes. Visual inspection reveals some additional residues in roughly similar positions in this region of the active site: Asp 314 and Asp 55 in phytase, and Asp 115 and Asp 117 in PPase.

Fumarylacetoacetate Hydrolase and 3-Isopropylmalate Dehydrogenase

Fumarylacetoacetate hydrolase (FAH)[38] plays a role in the degradation of Tyr and Phe by catalyzing the hydrolytic cleavage of a carbon-carbon bond in fumarylacetoacetate. 3-Isopropylmalate dehydrogenase (IPMDH)[39] is involved in the leucine biosynthesis pathway and catalyzes the dehydrogenation and decarboxylation of 2-isopropylmalate.

A mirror symmetric triad pair (rmsd = 0.59 Å, P-value = 1.2×10^{-5}, E-value = 7.1×10^{-3}, Z-value = -1.46) was found between FAH (PDB code 1hyo) and IPMDH (PDB code 1cnz). The residues in the triad were Tyr 145B, Lys 195A, and Asn 197A for IPMDH, and Tyr 159A, Lys 253A, and Gln 240A for FAH (see Fig. 10). In addition, visual inspection revealed two similar clusters of acidic residues consisting of Glu 199A, Asp 233A, Glu 201A, and Asp 126A (FAH) and Asp 227A, Asp 255B, and Asp 251B (IPMDH). Both clusters coordinate metal ions: Ca^{2+} in the case of FAH and Mn^{2+} in the case of IPMDH.

The similarities between the roles of the active site residues of the two enzymes are striking. For both enzymes, the three residues in the triad are part of an oxyanion hole.[40,41] The Lys residues are thought to be involved in transferring a proton to a carbanion.[41,42] Finally, the metal ions aid in binding the negatively charged substrates.

CONCLUSION

A novel method to detect side-chain patterns was described, which is both computationally efficient, memory efficient, and capable of finding patterns that are not detected by other methods with a similar scope. The success of the method depends on simultaneously taking into account atom label ambiguities, shifted Cα positions, the chemically important groups of the side chains and mirror-imaged side-chain arrangements. By making use of patterns consisting of three residues, the pattern finding problem can be recast as a spatial lookup problem, for which efficient solutions (like the SR tree data structure used here) exist. The significance of the rmsd between a triad pair can be evaluated using Z-, P-, and E-values. In particular, the E-value depends on both the rmsd and the number of triads of a particular type in the database and is especially suited for identifying interesting similarities.

A current limitation of the method is that only side chain patterns of three residues are detected. However, this limitation is not as severe as it might seem. First, in a previous all-against-all comparison of a subset of the structures in the PDB,[12] the great majority of all found patterns consisted of three residues (R. Russell, personal communication). Second, similarities between active sites of enzymes are most often confined to a few residues. This is especially well illustrated by the luciferase/1,2-CTD and FAH/IPMDH cases. Third, clustering of overlapping or neighboring triad pairs could be added as a next step to reveal larger patterns. Clustering of triads is an attractive alternative to using tetrads or pentads, because the latter approach would lead to a combinatorial explosions of the number of vectors. Clustering could also lead to the identification of similarities that are more extended in space, such as protein binding sites, or functional sites that are only locally similar. Obviously a statistical scoring scheme should be developed to assess the significance of similarities between clusters of triads, e.g., by combining P-values.[43] The above-described examples of functional site similarities in any case already corroborate the usefulness of the triad approach in itself.

For the identification of putative active or ligand binding sites, the use of side chain triads provides a convenient framework. In the above-described approach, triads whose residues are near a bound ligand are used as a ligand-binding site database. These triads are used to look for similar triads in other structures. The use of three residues to describe an active site avoids "over-specification" and makes the automated generation of a set of functional sites possible. This eliminates many problems associated with manual construction or automated use of unreliable PDB SITE identifiers.[13] In the current implementation, one representative from each SCOP superfamily was used to facilitate the construction of the statistical scoring scheme and to minimize duplicate hits. To make the database of ligand binding triads more complete, the complete PDB database can be used in a future version. In addition, sites that are potential binding sites (based on geometrical criteria,[44] energy calculations,[45] clustering of conserved residues,[46,47] or computational mapping using

molecular probes[48]) could be added as well. Finally, disordered residues could be dealt with by representing all the possible combinations of alternative residue positions in seperate triads.

Many of the found triad pairs are due to seemingly random similarities or structurally favorable interactions like salt bridges between the residues in the triad. Many other cases where the significance of the similarity is unclear can serve as useful starting points for further research. In conclusion, the method described here is a powerful tool to discover active site similarities for well-characterized protein structures, to determine the putative location of active sites or to identify putative active sites associated with an unknown function.

ACKNOWLEDGEMENTS

I thank Megan Thomas and Dr. Joe Hellerstein for help with using the libgist implementation of the SR tree; Dr. Robert Russell for sharing additional information; Dr. Bente Vestergaard, Dr. Remy Loris, Dr. Wim Versees, and Lieven Buts for reading the manuscript; and three anonymous referees for helpful suggestions. T. Hamelryck is a postdoctoral fellow of the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO).

REFERENCES

1. Teichmann SA, Murzin AG, Chothia C. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 2001;11:354–363.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001;11:348–353.
4. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
5. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
6. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
7. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
8. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 2001;29:228–229.
9. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
10. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graphtheoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
11. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
12. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
13. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;285:1887–1897.
14. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
15. Zhao S, Morris GM, Olson AJ, Goodsell DS. Recognition templates for predicting adenylate-binding sites in proteins. J Mol Biol 2001;314:1245–1255.
16. Lovell SC, Word JM, Richardson JS, Richardson DC. Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering. Proc Natl Acad Sci USA 1999;96:400–405.
17. Mattevi A, Vanoni MA, Todone F, Rizzi M, Teplyakov A, Coda A, Bolognesi M, Curti B. Crystal structure of D-amino acid oxidase: a case of active site mirror-image convergent evolution with flavocytochrome b2. Proc Natl Acad Sci USA 1996;93:7496–7501.
18. Kimber MS, Pai EF. The active site architecture of *Pisum sativum* beta-carbonic anhydrase is a mirror image of that of alpha-carbonic anhydrases. EMBO J 2000;19:1407–1418.
19. Pawelek PD, Cheah J, Coulombe R, Macheroux P, Ghisla S, Vrielink A. The structure of L-amino acid oxidase reveals the substrate trajectory into an enantiomerically conserved active site. EMBO J 2000;19:4204–4215.
20. Wood ZA, Poole LB, Karplus PA. Structure of intact AhpF reveals a mirrored thioredoxin-like active site and implies large domain rotations during catalysis. Biochemistry 2001;40:3900–3911.
21. Kraulis P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.
22. Katayama N, Satoh S. The SR-tree: an index structure for high-dimensional nearest neighbor queries. Proceedings of ACM SIGMOD, Tucson, Arizona, 1997;May 13–15:369–380.
23. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000; 28:254–256.
24. Dodson G, Wlodawer A. Catalytic triads and their relatives. Trends Biochem Sci 1998;23:347–352.
25. Schrag JD, Li YG, Wu S, Cygler M. Ser-His-Glu triad forms the catalytic site of the lipase from Geotrichum candidum. Nature 1991;351:761–764.
26. Barth A, Frost K, Wahab M, Brandt W, Schadler HD, Franke R. Classification of serine proteases derived from steric comparisons of their active sites, part II: Ser, His, Asp arrangements in proteolytic and nonproteolytic proteins. Drug Des Discov 1994;12: 89–111.
27. Fujinaga M, James MN. Rat submaxillary gland serine protease, tonin. Structure solution and refinement at 1.8 A resolution. J Mol Biol 1987;195:373–396.
28. Hegg EL, Que L Jr. The 2-His-1-carboxylate facial triad-an emerging structural motif in mononuclear non-heme iron(II) enzymes. Eur J Biochem 1997;250:625–629.
29. Thayer MM, Flaherty KM, McKay DB. Three-dimensional structure of the elastase of *Pseudomonas aeruginosa* at 1.5-Å resolution. J Biol Chem 1991;266:2864–2871.
30. Goodwill KE, Sabatier C, Marks C, Raag R, Fitzpatrick PF, Stevens RC. Crystal structure of tyrosine hydroxylase at 2.3 Å and its implications for inherited neurodegenerative diseases. Nat Struct Biol 1997;4:578–585.
31. Fisher AJ, Thompson TB, Thoden JB, Baldwin TO, Rayment I. The 1.5-Å resolution crystal structure of bacterial luciferase in low salt conditions. J Biol Chem 1996;271:21956–21968.
32. Vetting MW, Ohlendorf DH. The 1.8 Å crystal structure of catechol 1,2-dioxygenase reveals a novel hydrophobic helical zipper as a subunit linker. Structure Fold Des 2000;8:429–440.
33. Tanner JJ, Miller MD, Wilson KS, Tu SC, Krause KL. Structure of bacterial luciferase beta 2 homodimer: implications for flavin binding. Biochemistry 1997;36:665–672.
34. Meighen EA. Molecular biology of bacterial bioluminescence. Microbiol Rev 1991;55:123–142.
35. Heikinheimo P, Lehtonen J, Baykov A, Lahti R, Cooperman BS, Goldman A. The structural basis for pyrophosphatase catalysis. Structure 1996;4:1491–1508.
36. Shin S, Ha NC, Oh BC, Oh TK, Oh BH. Enzyme mechanism and catalytic property of beta propeller phytase. Structure 2001;9:851–858.
37. Oh BC, Chang BS, Park KH, Ha NC, Kim HK, Oh BH, Oh TK. Calcium-dependent catalytic activity of a novel phytase from *Bacillus amyloliquefaciens* DS11. Biochemistry 2001;40:9669–9676.
38. Bateman RL, Bhanumoorthy P, Witte JF, McClard RW, Grompe M, Timm DE. Mechanistic inferences from the crystal structure of fumarylacetoacetate hydrolase with a bound phosphorus-based inhibitor. J Biol Chem 2001;276:15284–15291.

39. Wallon G, Kryger G, Lovett ST, Oshima T, Ringe D, Petsko GA. Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus.* J Mol Biol 1997;266:1016–1031.
40. Hurley JH, Thorsness PE, Ramalingam V, Helmers NH, Koshland DE Jr, Stroud RM. Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. Proc Natl Acad Sci USA 1989;86:8635–8639.
41. Timm DE, Mueller HA, Bhanumoorthy P, Harp JM, Bunick GJ. Crystal structure and mechanism of a carbon-carbon bond hydrolase. Structure Fold Des 1999;7:1023–1033.
42. Lee ME, Dyer DH, Klein OD, Bolduc JM, Stoddard BL, Koshland DE Jr. Mutational analysis of the catalytic residues lysine 230 and tyrosine 160 in the NADP(+)-dependent isocitrate dehydrogenase from *Escherichia coli*. Biochemistry 1995;34:378–384.
43. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics 1998; 14:48–54.
44. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
45. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
46. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
47. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
48. Dennis S, Kortvelyesi T, Vajda S. Computational mapping identifies the binding sites of organic solvents on proteins. Proc Natl Acad Sci USA 2002;99:4290–4295.