# Protein functional-group 3D motif and its applications

YE Yuzhen, XIE Tao & DING Dafu

Shanghai Institute of Biochemistry, Chinese Academy of Sciences, Shanghai 200031, China
Correspondence should be addressed to Ding Dafu (e-mail: dingdafu@server.shcnc.ac.cn)

**Abstract** Representing and recognizing protein active sites sequence motif (1D motif) and structural motif (3D motif) is an important topic for predicting and designing protein function. Prevalent methods for extracting and searching 3D motif always consider residue as the minimal unit, which have limited sensitivity. Here we present a new spatial representation of protein active sites, called "functional-group 3D motif", based on the fact that the functional groups inside a residue contribute mostly to its function. Relevant algorithm and computer program are developed, which could be widely used in the function prediction and the study of structural-function relationship of proteins. As a test, we defined a functional-group 3D motif of the catalytic triad and oxyanion hole with the structure of porcine trypsin (PDB code: 1mct) as the template. With our motif-searching program, we successfully found similar sub-structures in trypsins, subtilisins and $\alpha/\beta$ hydrolases, which show distinct folds but share similar catalytic mechanism. Moreover, this motif can be used to elucidate the structural basis of other proteins with variant catalytic triads by comparing it to those proteins. Finally, we scanned this motif against a non-redundant protein structure database to find its matches, and the results demonstrated the potential application of functional group 3D motif in function prediction. Above all, compared with the other 3D-motif representations on residues, the functional group 3D motif achieves better representation of protein active region, which is more sensitive for protein function prediction.

**Keywords: functional-group 3D motif, RMSD of distance matrix, random probability, expectation, functional prediction.**

Comparisons of protein sequences, structures and motifs (including 1D and 3D motifs) are of vital value in extracting information of protein function[1—3]. To gather functional information between remote homologues or even convergent proteins, quite a few programs have been developed for extracting spatial motifs and finding matches in protein structures[4—9], which facilitate not only functional prediction[4—9] but also functional protein design[10—12]. Residues are the minimal operational units in these methods, no matter whether their Cα atoms or side-chain atoms are used to represent active sites. However, sometimes these strategies fail to represent the active sites efficiently. For example, trypsins, subtilisins and some α/β hydrolases are good testing cases for different 3D motif methods because they possess similar catalytic triads but distinct folds[6, 7]. Recently, many proteases with variant catalytic triads have been discovered, in which acid-base-serine/threonine patterns of catalytic residues are conserved, while the individual components vary[13, 14]. Therefore, these catalytic triad variants cannot be recognized if we focus on the residual level. For example, residue lysine replaces histidine as the base in asparaginase. In an esterase from *streptomyces scabies*, backbone carbonyl oxygen of triptophan acts as the acid component instead of Glu/Asp to form carbonyl-oxygen-histidine hydrogen bond[13, 14]. On the other hand, "oxyanion hole" has been found to be the common active site for serine proteases, lipases and serine carboxypeptidases[15], which consists of two or more hydrogen-bond donors to stabilize the acyl-enzyme intermediate in catalysis. These hydrogen bonds are donated by main-chain amides or side-chain amides, on which no residual specificity needs to be imposed. All the above imply that a new spatial motif representation in smaller unit rather than the whole residue should be considered.

In this note, we propose a new 3D motif representation, called "functional-group 3D motif", considering functional groups as the units, to get more specific and sensitive tertiary description of protein active sites. Under such a definition, backbone groups may match side-chain ones; two groups within a residue may match two groups from two residues respectively; a residue matches another one owing to sharing similar groups, and so on. Meanwhile, we developed a depth-first search algorithm to find the occurrence of a particular functional motif in protein structure database, providing the statistical significance assessment of the hits. As a test, a functional group 3D motif of "catalytic-triad and oxyanion hole" is represented here and its applications in function-structure relationship elucidation and functional prediction are also shown.

## 1 Materials and methods

(ⅰ) Materials. (1) Porcine trypsin (PDB code: 1mctA), for extracting functional-group 3D motif. (2) Non-redundant protein set 1 (986 proteins), for the statistics of the random distribution of functional-group 3D motifs. We used the first item of each protein class derived from http://www-lmmb.ncifcrf.gov/~tsai/index.html to construct this set. (3) Test data sets: ⅰ) Proteins predicted possessing typical catalytic traids by Russell[7]. ⅱ) α/β hydrolases, taken from http://wwww.ensam.inra.fr/cholinesteras/. ⅲ) *Streptomyces scabies* esterases (PDB codes are 1esc, 1esd and 1ese)[14]. ⅵ) Non-redundant protein set 2 (2 098 proteins in total, the similarity between any two proteins is below 95%[8]).

All co-ordinate data were taken from RCSB (http://www.rcsb.org). All entries of those proteins are presented as four characters, while the fifth character labels the chain used in the computation.

(ⅱ) Methods. First we used the information of protein structure and function to construct functional-group 3D motif representing the active site. Then we simulated the random distribution of this motif in the space of protein structures, which could be further used to evaluate its matches in protein targets.

(1) Functional group representations of amino acids. Differed from the reduced representation of the amino acids using only their Cα atoms[8] or side-chain groups[7], we adopted a new representation of amino acids. In our method, one residue may be dissected into many functional groups including two main-chain functional groups (carbonyl, saying CO, and amide, saying NH) and one or more side-chain functional groups. Here, we consider that backbone groups of all amino acids are the same. Moreover, a functional group may have different components. Take the hydroxy group as an example. When it

functions as a nucleophile, the oxygen atom and its bonded hydrogen atom should be considered (see fig. 1(d) 1# rectangle), while if it acts as hydrogen bond donor, the oxygen atom and its bonded carbon atom should be used (see fig. 1(d) 2# rectangle). The co-ordinations of hydrogen atoms are calculated using standard bond lengths, bond angles and dihedral angle values from CHARMM force field[16].

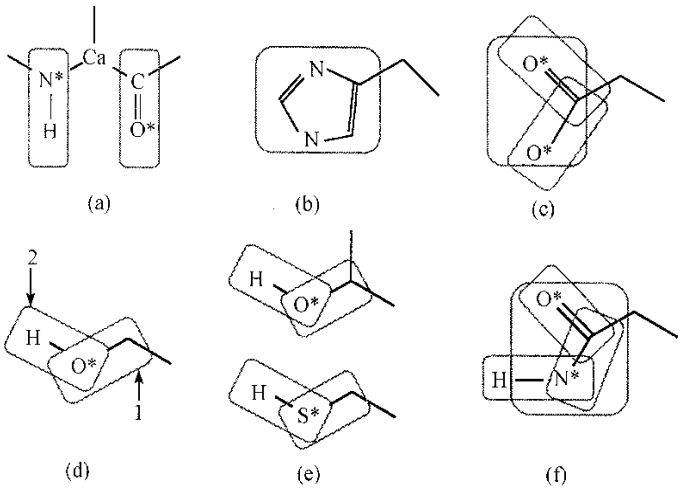


Fig. 1. Schematic diagrams showing the functional groups of some amino acids. The component atoms of a group are boxed in a rectangle, of which the atom marked with * is the center of each group. (a) Main-chain amide; (b) His side-chain; (c) Asp/Glu side-chain and carbonyl; (d) Ser side-chain; (e) Thr/Cys side-chain; (f) Asn/Gln side-chain.

We endow each group with a physical and chemical property profile to show its properties, such as hydrophobicity, hydrogen-bond donor or acceptor, aromaticity, charge, etc. And then a similar score between any two groups can be calculated based on their property profiles. This score is then normalized to a range between zero and one, where one means that the two groups are identical.

(2) Derivation of functional-group 3D motif. Assuming that for a protein, its key residues have been known. First we dissect these key residues into functional groups and then gather the vital ones to construct a functional-group 3D motif associated with a function. Such a motif typically contains the groups and their component atoms. Each group is endowed a similarity threshold which will be used in the later searching procedure. The value of 1 means that only an identical group matches this group, while a value smaller than 1 means that similar functional groups are also suited to this group.

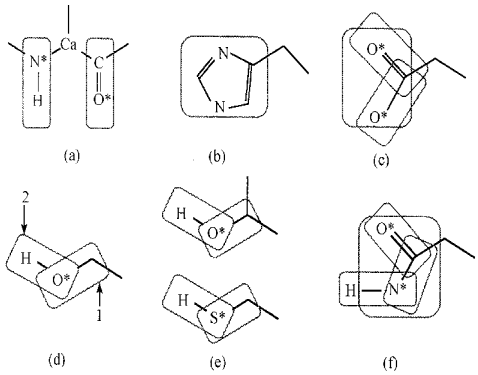


Fig. 2. Schematic diagrams showing functional-group 3D motifs. (a) Query 3D motif (composed of two goups). (b) A sub-structure matched with the query 3D motif. The dotted lines stand for the distances formed by component atoms from different groups.

Take a motif composed of two functional groups (A and B, each is composed of two atoms) as an example (see fig. 2(a)). The four distances ($d_1$, $d_2$, $d_3$ and $d_4$) between any two atoms from different groups form a distance matrix. Supposing that a matched sub-structure (composed of functional groups $A'$ and $B'$) is detected in a target protein, we use eq. (1) to calculate the root mean squared distance of distance matrix (*dmRMSD*) between them to define their structural similarity:

$$dmRMSD = \sqrt{\sum_{i=1}^{n} (d_i - d_i')^2 \,/\, n}\,, \tag{1}$$

where $n$ is the number of distances. In this example, $n = 4$. Lower *dmRMSD* means higher spatial similarity.

(3) Random distribution of a functional-group 3D motif in protein structure space. Wallace et al.[6] have shown that the structure of catalytic motif and non-catalytic motif are different, though they have similar compositions. This observation provides foundation for evaluating the hits detected in the searching procedure. Given a particular functional-group 3D motif, we find the occurrences of matched sub-structures in a non-redundant structure database to simulate the distribution of such component groups in the whole protein structure space. Later on, this distribution could be referred to evaluate the matches detected in targets.

First, we dissect each protein into the combinations of functional groups, which are classified by group type. Then, with depth-first searching algorithm, we detect the functional group combinations matching the query motif (here no spatial similarity is required). When each functional group combination is found, the *dmRMSD* between this sub-structure and the query motif is calculated with formula (1). After this procedure, *dmRMSD*s are ranked to derive the empirical distribution function $p(x)$, which represents the proportion of hits whose *dmRMSD*s are less than $x$. In the case of catalytic and oxyanion hole motif, $p(x)$ and $1/dmRMSD$ fit linear model (eq. (2)) well when $x < 3.0$ with the correlation coefficient >0.94.

$$\log(p) = a + \frac{b}{dmRMSD}. \tag{2}$$

We regard a sub-structure matching the query motif with *dmRMSD* >3.0 as an insignificant match. Therefore, we assign a set of $a$, $b$ values to each query motif. The probability of observing a specific *dmRMSD* value between a sub-structure and this query motif can be computed by the above formula.

(4) Detection of matches for query motif in protein targets. Depth-first algorithm is applied in this step, where the distance constraints are used to speed the searching procedure. Suppose that a query motif is composed of $n$ functional groups, $M_a$, $M_b$, ..., $M_n$, while a match is composed of $T_a$, $T_b$, ..., $T_n$. We denote their *dmRMSD*, calculated by formula (1), as $dmRMSD(M_i, T_i, 1 \leqslant i \leqslant n)$; the random probability of observing such *dmRMSD* value, calculated by formula (2), as $p(dmRMSD(M_i, T_i, 1 \leqslant i \leqslant n))$. Apparently, the probability of observing a specific sub-structure in a protein must be related to its group composition. The more functional groups there are in a protein, the more probably a sub-structure composed of these groups is observed. To avoid such a bias, we use expectation instead of random probability to describe the probability of randomly hitting a match in a protein. Suppose that there are totally $N(M_i)$ groups in a particular protein matching group $M_i$ of query motif. We get the number of functional group combinations in the target protein, denoted as $comb(N(M_i), 1 \leqslant i \leqslant n)$, where $n$ is the number of groups in the query motif. Thus, the expectation of randomly observing such a motif in this protein should be

$$E = comb(N(M_i), 1 \leqslant i \leqslant n) \times p(dmRMSD(M_i, T_i, 1 \leqslant i \leqslant n)). \tag{3}$$

Low probability and low expectation together mean that a hit is statistically significant (in the case of catalytic triad and oxyanion hole motif, expectation threshold 1.0 and the probability threshold 1.0e −10 are used). If a statistically significant match of the query motif is found in an unknown protein, the annotation of the protein from which the motif is derived could be transferred to it.

## 2 Results

With the catalytic triad and oxyanion hole motif as a test, we extract its functional group, and then estimate its $a$ and $b$ values. Detection of its matches in some protein structure sets shows the good effect of functional based motif strategy and its applications.

(ⅰ) Deriving functional-group 3D motif. The co-ordinations of five functional groups are extracted from porcine trypsin (PDB code: 1mct) to construct the motif of catalytic triad and oxyanion hole. The five groups are a nucleophile (hydroxyl group of Ser195, OH), a base (imidazole group of His57, GH), an acid (side-chain carboxyl of Asp102, GD) and two oxyanion hole forming groups (main-chain amides of Ser195 and of Gly193, NH) (see fig. 3(a)). Different similarity thresholds are assigned to those groups. Imidazole group can match the base; carboxyl can match the acid; hydroxyl of Ser and Thr, and sulfhydryl of Cys can match the nucleophile part; all hydrogen bond donor groups

are allowed to match the oxyanion hole groups. Because main-chain carbonyls may function as acid in some variants, we use the two side-chain carbonyl groups ($EO^1$ and $EO^2$) instead of carboxyl of Asp102 to define two other motifs, in which the remaining groups are kept. And their *a* and *b* values are assessed by distribution simulation, respectively (see table 1). Two "catalytic triad and oxyanion hole" motifs, Motif 1 and Motif 2 are defined ultimately, which only differ at the acid. Motif 1 possesses only carboxyl (GD) while Motif 2 processes either carboxyl or carbonyl ($EO^1$ or $EO^2$) at this position.

Table 1 Statistics of distribution of the motifs of catalytic triad and oxyanion hole

| Functional-group 3D motif | *a* | *b* | Coefficient |
|---|---|---|---|
| NH(Ser195) NH(Gly193) GH(His57) GD (Asp102) OH(Ser195) | –10.050 | –9.779 | 0.968 |
| NH(Ser195) NH(Gly193) GH(His57) $EO^1$(Asp102) OH(Ser195) | –10.986 | –8.530 | 0.941 |
| NH(Ser195) NH(Gly193) GH(His57) $EO^2$(Asp102) OH(Ser195) | –8.595 | –12.273 | 0.991 |

(ii) Results of scanning the motif against Russell's protein data set. The detailed searching results of Motif 1 against this protein set[7] can be found at http://dna.sibc.ac.cn/~ye/tesltable.html. In summary, matches for Motif 1 are found in trypsins (e.g. 1 ppfE, probability is 2.22e −17 and expectation is 6.61e −10), subtilisins (e.g. 1thm, probability is 3.26e −12 and expectation is 4.84e −4, see fig. 3(b)) and *a*/*b* hydrolases (e.g. 1tahB, probability is 1.71e −12 and expectation is 1.25e −3, see fig. 3(c)). Generally speaking, oxyanion holes are formed by two main-chain amides in trypsin and *a*/*b* hydrolase, while in subtilisin, one main-chain amide is replaced by side-chain amide of Asn, which accords with the previous conclusion very well[15]. The scanning results also demonstrate that no residue type is required on those residues donating hydrogen bond donor to oxyanion hole. Furthermore,

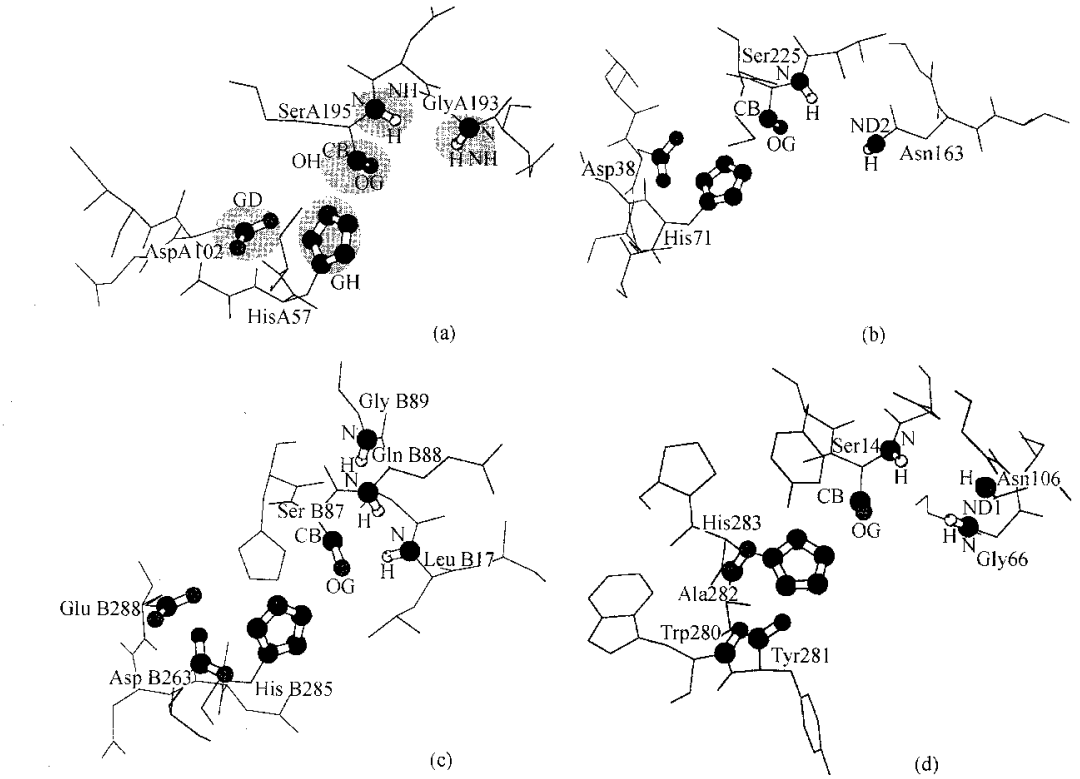


Fig. 3. Functional group 3D motifs of catalytic triad and oxyanion hole. Only residues providing functional groups and their neighbors are shown in bonds, with functional groups are highlighted with ball-and-sticks. All figures are prepared with program MOLSCRIPT[17]. (a) Porcine trypsin (PDB code: 1mctA), for query motif extraction. Five component functional groups are highlighted with filled circles; (b) subtilisin (PDB code: 1thm); (c) *a*/*b* hydrolase (PDB code: 1tahB), which has an alternative catalytic triad; (d) streptomyces scabies esterase (PDB code: 1ese), in which main-chain carbonyls function as acid.

we distinguish 13 proteins (PDB codes are 1vnc, 1tlcA, 1ribA, 1fatA, 1mioA, 1amp, 1celA, 1jbc, 1vhh, 2cpl, 1pta, 2rmc and 1cynA) with "false" catalytic triad detected by Russell[7] from the real ones, showing high specificity of recognition. The improvements should benefit from more detailed motif, which contains not only the catalytic triad but also the oxyanion hole. It is worth pointing out that we find no hit in trypsin (PDB code: 1ton) because of its slightly distorted active site in its crystal structure[18].

In *pseudomonas glumae* triacylglycerol hydrolase (PDB code: 1tahB), not only the original catalytic triad Asp263-His141-Ser215 but also the "alternative" catalytic triad Glu288-His285-Ser87 is recognized, which fits well with the experimental conclusion[19]. And we find alternative catalytic triad Glu263-His141-Ser215 besides the common catalytic triad Asp163-His141-Ser215 in *sindibis virus* capsid protein (PDB code: 1kxf).

There are some conflicts in the characterization of the active site of *vibrio harveyi* thioesterase (PDB code: 1tht). The experiments by Ferri et al.[20] showed that Ser71 is the nucleophile while the results of Lawson et al.[21] showed that Ser114-His241-Asp211 form its catalytic triad and the backbone amide of Leu115 participates in forming its oxyanion hole. Our motif searching result support Lawson's result, and we even found another possible oxyanion hole component, saying backbone amide of Ser116.

(iii) Results of scanning the motif against $\alpha/\beta$ hydrolases. $\alpha/\beta$ hydrolase fold proteins belong to an important, diverse, widespread group of enzymes, which have eight-stranded mostly parallel $\alpha/\beta$ structure. Heikinheimo et al.[22] made a classification of $\alpha/\beta$ hydrolase. Database ESTHER lists more than 100 $\alpha/\beta$ hydrolase structures. We gathered these proteins from ESTHER, excluding those proteins with only co-ordinations of Cα atoms available (PDB codes: 1tia, 1tic, 1tg and 5tgl), theoretically modeled structures (PDB codes: 3ace and 4ace) and one protein with Ser120Ala mutation (PDB code: 1cui), and then scanned those proteins for occurrences of Motif 1. No matches are found in total 13 haloalkane dehalogenases (with aspartate at nucleophile position) and 29 zinc-dependent exopeptidases, which have no common catalytic-triads[22]. And matches are detected in most of the remaining 115 $\alpha/\beta$ hydrolases (99 in total). The detailed results are shown in table 2.

Table 2 Results of scanning Motif 1 against $\alpha/\beta$ hydrolases

| Protein | Target with hit | Total | Targets without hit | Total |
|---|---|---|---|---|
| AChE (Acetylcholinesterase) | 1acl, 1cfjA, 1eea, 1eve, 1maaA, 1oce, 1somA, 1mah, 1vot, 2ace, 2ack, 2dfpA, | 12 | 1acj, 1amn, 1ax9, 2clj, 1fss | 5 |
| Bacterial lipase | 1cvl, 1oilA, 1tahB, 2lip, 3lip, 4lip, 5lip | 7 | | 0 |
| Carboxypeptidase | 1ac5, 1bcr, 1bcs, 1ivyA, 1whs, 1wht, 3sc2AB | 7 | 1cpy, 1ysc | 2 |
| Cholesterol esterase | 1akn, 1aqlA, 2bce | 3 | | 0 |
| Cutinase | 1agy, 1cex, 1cua, 1cub, 1cuc, 1cudA, 1cue, 1cuf, 1cug, 1cuh, 1cuj, 1cus, 1cuu, 1cuv, 1cuwA, 1cux, 1cuy, 1cuz, 1ffa, 1ffc, 1ffd, 1ffe, 1xomA, 1xza, 1xzb, 1xzc, 1xze, 1xzf, 1xzg, 1xzh, 1xzi, 1xzj, 1xzkA, 1xzl, 1xzm, 2cut | 36 | 1ffb | 1 |
| Dienelactone hydrolase | | 0 | 1din | 1 |
| Fungal carboxylesterase lipase | 1cleA, 1crl, 1lpm, 1lpo, 1lps, 1thg, 1trh | 7 | 1lpn, 1lpp | 2 |
| Fungal triacylglycerol lipase | 1lbs, 1lbt, 1lgyA, 1tca, 1tcbA, 1tccA, 1tib, 3tgl, 4tgl | 9 | | 0 |
| Haloperoxidase | 1a7uA, 1a88A, 1a8q, 1a8s, 1a8uA, 1broA, 1brt | 7 | | 0 |
| Hormone-sensitive lipase like | 1jkmA | 1 | | 0 |
| Hydroxynitrile lyase | 1yas | 1 | | 0 |
| PAF-Acetylhydrolase | 1jfr | 1 | | 0 |
| Pancreatic lipase | 1bu8, 1ethAB, 1lpbAB, | 3 | 1gpl, 1hplA, 1lpaAB, 1rpl | 4 |
| Proline iminopeptidase | | 0 | 1azm | 1 |
| Proline endopeptidase | 1qfmA, 1qfsA | 2 | | 0 |
| Psudomonas carboxylesterase | 1auo, 1aur | 2 | | 0 |
| Thioesterase | 1thtA | 1 | | 0 |
| Total | | 99 | | 16 |

The details of the matches of different $a/b$ hydrolase families with Motif 1 are not completely the same. In most cutinases, good hits with low probability and expectation are found, which show that the active sites of cutinases are quite similar to those of trypsins. In one cutinase mutation (Ser120Cys), the active site is formed by Asp175-His188-Cys120, with the backbone amide of Gln121, backbone amide of Ser42 and/or side-chain amide of Asn84 forming an oxyanion hole. Furthermore, Motif-1-like sub-structures are found in all 7 bacterial lipases, all 7 haloperoxidases, and some of acetycholine- sterase, carboxypeptidase, cholesterol esterase, fungal carboxylesterase lipase, etc. No hit is found in proline iminopeptidase and dienelactone hydrolase. Hits are also found in acetycholinesterases and fungal carboxylesterases, which use Glu instead of Asp at acid position, since the query motif is defined by functional groups here. For example, in acetycholinesterase (PDB code: 1eve), sub-structure with Glu327-His441-Ser200 forming catalytic triad and backbone amide of Ala201 and Gly119 forming oxyanion hole is recognized. The searching results confirm the conclusion that though the folds of most $a/b$ hydrolases and trypsins are distinct, their active sites are quite similar[22].

In triacylglycerol lipase (PDB code: 1cvl), we found the previously mentioned "alternative" catalytic triad. Both Glu288-His285-Ser87 and Asp263-His285-Ser87 are found spatially similar to the catalytic triad of Motif 1.

(iv) Structure-function elucidation of some proteins with variant catalytic triads. Here we just consider those proteins with variant at acid, i.e. main-chain carbonyl instead of Asp/Glu at this position (protein codes are 1esc, 1esd and 1ese) [14]. The matches found in the three proteins for Motif 2 are shown in table 3. From the results, the backbone amide of Ser14 and Gly66, and the side-chain amide of Asn106 participate in forming oxyanion hole, which coincides with the conclusion of Wei[14] well. Wei[14] and Dodson[13] hypothesized that the main-chain carbonyl of Trp280 functions as "acid" in catalytic triad. The searching results here show that not only this carbonyl, but also the main-chain carbonyls of Tyr281 and Ala282 may function as "acid" for their suitable geometry. We put forward such a hypothesis that the multiple main-chain carbonyls can compensate their weak acidic property.

Table 3 Results of scanning Motif 2 against *streptomyces scabies* esterases

| PDB code | dmp | comb | E | Matched functional groups | | | |
|---|---|---|---|---|---|---|---|
| 1esc | 5.13e−12 | 1.88e+09 | 9.66e−03 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(281TYR) OH(14SER) |
| | 1.11e−11 | 1.88e+09 | 2.09e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(280TRP) OH(14SER) |
| | 1.13e−11 | 1.88e+09 | 2.13e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(282ALA) OH(14SER) |
| 1esd | 6.66e−12 | 1.88e+09 | 1.26e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(281TYR) OH(14SER) |
| | 1.08e−11 | 1.88e+09 | 2.03e−02 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(281TYR) OH(14SER) |
| | 1.43e−11 | 1.88e+09 | 2.70e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(282ALA) OH(14SER) |
| | 1.45e−11 | 1.88e+09 | 2.73e−02 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(282ALA) OH(14SER) |
| | 2.48e−11 | 1.88e+09 | 4.67e−02 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(280TRP) OH(14SER) |
| 1ese | 5.04e−12 | 1.88e+09 | 9.49e−03 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(282ALA) OH(14SER) |
| | 7.86e−12 | 1.88e+09 | 1.48e−02 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(281TYR) OH(14SER) |
| | 7.86e−12 | 1.88e+09 | 1.48e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(282ALA) OH(14SER) |
| | 8.99e−12 | 1.88e+09 | 1.69e−02 | NH(14SER) | NQ(106ASN) | GH(283HIS) | CO(281TYR) OH(14SER) |
| | 2.45e−11 | 1.88e+09 | 4.62e−02 | NH(14SER) | NH(66GLY) | GH(283HIS) | CO(280TRP) OH(14SER) |

dmp, Random probability of dmRMSD; comb, the combination number; E, expectation. For details see text.

(v) Function prediction by motif detection. By comparing a motif to protein targets, we may assign its associated function to those targets with matches. Here we detect matches for Motif 1 and Motif 2 respectively in the non-redundant protein database 2.

Totally 77 proteins were detected, which contain Motif-1-like sub-structures. Thus these proteins could be predicted to have one of the corresponding functions of protease, lipase, esterase, etc. Among them, at least 58 proteins are supposed to have such functions according to the description in the PDB file and the literature information, including a mutation of trypsin (Ser120Cys, PDB code: 1dpo) with catalytic triad Asp102-His57-Cys195. In triacylglycerol lipase (PDB code: 1cvl), an alternative triad Asp263-His285-Ser87 was detected besides the catalytic triad Glu288-His285-Ser87.

Those hits without any experimental identification to date are listed at http://dna.sibc.ac.cn/~ye/test4table.html. They have similar sub-structures with the 3D functional "catalytic triad and oxyanion

hole" motif, but additional criteria are required for further verification. Certainly experiments offer the most convincing and ultimate proofs.

Motif 2 is less specific than Motif 1. Thus more candidates with hits were detected by scanning Motif 2 against this data set, including more false hits. However, some variants (like 1esc, with main-chain carbonyl at the acid) may only be recognized by Motif 2. In this sense, Motif 1 is superior to Motif 2 for its high sensitivity. The searching results of Motif 2 are not shown due to space limit.

## 3 Discussion

In this note, the functional-group motif strategy, together with the functional group 3D motif definition, the searching algorithm and the statistical evaluation of the hits, are represented. They are tested by many cases, including the $a/b$ hydrolases, the "catalytic triad" variants and so on. Those proteins having distinct folds but similar sub-structures are supposed to be caused by convergence evolution. With the functional group motif strategy, not only those proteins with common catalytic triads but also many variants can be detected though they have distinct folds. These variants include proteins with main-chain carbonyl at the acid (like PDB 1esc), with Glu instead of Asp at the acid position and with Cys as nucleophile instead of Ser (like PDB code 1cuj). In this sense, the functional-group-based 3D motif representation is better than those methods based on whole residues[7]. Meanwhile, we use the evolutionary trace method developed by our laboratory[3] to extract the sequence trace of catalytic triad and then use it as the template to search in the PDB database for the purpose of comparing their efficiency in functional inference. The results show that sequence motifs are efficient in detecting proteins with common catalytic triads, while failed in recognizing remote homologies and proteins with catalytic triad variants. Structural motif searching may solve this problem. Therefore, functional-group motif strategy is a sort of improvement of evolutionary method from the sequence level to the structural level.

How to balance the specificity and sensitivity in structural motif strategy deserves further discussion. It is often difficult to provide a perfect structural motif definition with deficient knowledge of structure-function relationship and even faulty crystal structures. For example, in some variants, Lys instead of His acts as base. The side-chain amide of lysine must reside in the de-protonated state for lysine to act as a general base[13]. In this case, the structural environment of lysine is vital for its function, which is not easy to be described as a part of the structural motif currently. Our note did not consider this case for the above difficulties. Further properties, like the interface character, the evolutionary information and so on, are required to detect these variants.

With the development of structural genomics, more and more protein structures are resolved, which provides a wide prospect for functional prediction and functional protein design with structural motif. The structural motif strategy provides a good background for functional protein design by transferring active sites[10—12], which in reverse is of great benefit to functional reference verification.

**Acknowledgements** Thanks are due to Dr. Tang Haixu and Master Sheng Quanhu for their helpful advice. This work was supported by the National High Technology "863" Programs of China (Grant No. 863-103-03-03), the National Natural Science Foundation of China (Grant No. 39990600-03) and a grant from the Shanghai Key Program of Basic Research.

## References

1. Murzin, A. G., Patthy, L., Sequences and topology from sequence to structure to function, Curr. Opin. Struc. Biol., 1999, 9: 359.
2. Orengo, C. A., Todd, A. E., Thronton, J. M., From protein structure to function, Curr. Opin. Struc. Biol., 1999, 9: 374.
3. Xie Tao, Chen Jie, Ding Dafu, An evolutionary trace method for functional prediction of genomes, Acta Biochimica et Biophysica Sinica (in Chinese), 1999, 31(4): 433.
4. Artymiuk, P. J., Poirrette, A. R., Grindley, H. M. et al., A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures, J. Mol. Biol., 1994, 243: 327.
5. Chen Jie, Tang Haixu, Ding Dafu, Search for 3-D motif in proteins, Acta Biophysica Sinica (in Chinese), 1997, 13(4): 639.
6. Wallace, A. C., Laskowski, R. A., Thronton, J. M., Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases, Protein Science, 1996, 5: 1001.
7. Russell, R. B., Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution, J. Mol. Biol., 1998, 279: 1211.
8. Kleywegt, G. J., Recognition of spatial motifs in protein structures, J. Mol. Biol., 1999, 285: 1887.
9. Li Zhang, Godzik, A., Skolnick, J. et al., Functional analysis of the *Escherichia coli* genome for members of the $a/b$ hydrolase family, Folding & Design, 1998, 3: 535.

# NOTES


10. Iengar, P., Ramakrishnan, C., Knowledge-based modeling of the serine protease triad into non-protease, Protein Eng., 1999, 12(8): 649.
11. Quemeneur, E., Moutiez, M., Engineering cyclophilin into a proline-specific endopeptidase, Nature, 1998, 391(6664): 301.
12. Ye Yuzhen, Tang Haixu, Ding Dafu, Engineering novel functional proteins: Grafting active sites into natural scaffolds, Acta Biochimica et Biophysica Sinica (in Chinese), 1999, 31(3): 303.
13. Dodson, G., Wlodawer, A., Catalytic triads and their relatives, Trends Bioche. Sci., 1998, 23(9): 347.
14. Wei, Y., Derewemda, Z. S., A novel variant of the catalytic triad in the streptomyces scabies esterase, Nat. Struc. Biol., 1995, 2: 218.
15. Whiting, A. K., Peticolas, W. L., Details of the Acyl-enzyme intermediate and the oxyanion hole in serine protease catalysis, Biochemistry, 1994, 33(2): 552.
16. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D. et al., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, J. Comp. Chem., 1983, 4: 187.
17. Kraulis, P. J., MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, J. Appl. Crystallogr., 1991, 24: 946.
18. Fujinaga, M., James, M. N., Rat submaxillary gland serine protease, tonin: Structure solution and refinement at 1.8 Å resulution, J. Mol. Biol., 1997, 195(2): 373.
19. Noble, M. E., Cleasby, A., Johnson, L. N. et al., The crystal structure of triacylglycerol lipase from pseudomonas glumae reveals a partially redundant catalytic aspartate, FEBS Lett., 1993, 331(1-2): 123.
20. Ferri, S. R., Meighen, E. A., A lux-specific myristoyl transferase in luminescent bacteria related to eukaryotic serine esterases, J. Biol. Chem., 1991, 266(20): 12852.
21. Lawson, D. M., Derewenda, Z. S., Structure of a myristoyl-ACP-specific thioesterase from *Vibrio harveyi*, Biochemistry, 1994, 33(32): 9382.
22. Heikinheimo, P., Goldman, A., Jeffries, C. et al., Of barn owls and bankers: a lush variety of $\alpha/\beta$ hydrolases, Structure, 1999, 7: R141.


(Received April 24, 2000)