Genome-Wide Analysis of Enzyme Structure-Function Combination Across Three Domains of Life

Ziding Zhang* and Yu-Rong Tang

Bioinformatics Center, College of Biological Sciences, China Agricultural University, Beijing 100094, China

Abstract: To investigate diverse enzyme structure-function combination (SFC) types in different species, 34 different genome sequences were annotated using the protein catalytic domain database SCOPEC (http://www.enzome.com/enzome/), in which both the structure and function for each entry are known. Annotated enzymes with catalytic domains from the same SCOP superfamily are considered to have an identical structure. Annotated enzymes sharing the identical three-digit EC number are considered to have the same enzymatic function. Results reveal that the different SFC types for enzymes identified in archaea, bacteria and eukaryota are 137, 300 and 313, respectively. About 80% of the SFCs identified in archaea can be consistently found in bacteria and eukaryota species, whereas 28% and 35% combination types in bacteria and eukaryota respectively are unique to their corresponding groups. The number of functions per structure and the number of structures per function for the annotated sequences were measured in different species. Furthermore, a new concept was proposed to represent enzymatic structures as a functional similarity network. Thus, the current study will be helpful to enhance the global view on the evolution of enzymatic structure and function.

Keywords: Enzyme, structure, function, evolution, network.

INTRODUCTION

To discover new enzymes and design new drugs, *in silico* functional annotation of a large number of enzyme sequences obtained from genomic and proteomic studies is becoming increasingly important. The classical bioinformatics method of assigning functions to a query enzyme sequence requires the identification of homologous sequences with known functional annotation. Subsequently, the functional class (e.g. Enzyme Classification (EC) number) for the identified homolog can be transferred to the query sequence. However, in some cases such methods can be misleading due to the fact that enzyme functions are less conserved [1, 2]. It was observed that the conservation of function between a pair of enzymes becomes questionable when sequence identity drops below 40% [2, 3]. In addition to the sequence similarity based annotation (*i.e.* sequence alignment-based methods), prediction of an enzyme's family class based on its sequence property such as amino acid composition has been intensively investigated [4-10], which is also playing an important role in accelerating the annotation of functional unknown enzyme sequences.

Since protein 3D structure is much more conserved than sequence and closely related to protein function, structural comparisons were therefore able to identify functional relationships even when no clear sequence similarity was detectable provided that the 3D structure for the targeting sequence can be obtained by experimental or computational studies [11]. Such "structure-based functional annotation" can offer in-depth insight by often highlighting the 3D structural arrangements for the catalytic residues. Even so, the power of structure-based annotation is often weakened by the fact that a similar fold does not necessarily imply a similar function. For example, enzymes with the same fold, like TIM barrels, can have multiple functions [12]. On the other hand, proteins from different folds, such as subtilisin and trypsin, can share a similar function [13]. Therefore, further understanding of the relationship between enzyme structure and function remains an important topic in the field of structural biology.

The accumulated enzyme structures deposited in PDB database [14] have provided essential insight into the relationship between enzyme function and structure. Furthermore, the 3D structures for the catalytic residues have also been investigated and the corresponding database has been constructed [15]. Recently, a database of protein catalytic domains – SCOPEC was compiled (http://www.enzome.com/enzome/). By adding the verified functional information (*i.e.* EC number) into the SCOP structural domains, SCOPEC ensures that the domain-EC annotation is correct. Each EC number is defined as a four-digit code, which represents a hierarchy of functional classification of catalytic reaction. Similar to EC scheme, SCOP domain is also hierarchical with all domains classified by structural class, followed by fold, superfamily and family [16,17]. Furthermore, the elegant analysis of domain-EC relationships in SCOPEC highlights the evolution of protein structure and function [18]. Representing about 75% enzymes with known structures, SCOPEC can be a valuable resource in the analysis and prediction of protein structure and function.

With a growing number of sequenced genomes, comparative studies have been carried out to identify the differences of protein sequence or structure among three domains of life [19-24]. By mapping genome sequences into functional known enzyme categories, a comprehensive analysis was

*Address correspondence to this author at the College of Biological Sciences, China Agricultural University, Beijing 100094, China; Tel: +86-10-62734376; Fax: +86-10-62731332; E-mail: zidingzhang@cau.edu.cn

0929-8665/07 $50.00+.00

recently performed to estimate the fraction of enzymes in genomes and to determine the extent of their functional redundancy in different genomes [23]. The availability of SCOPEC allows us to systematically investigate enzyme Structure-Function Combination (SFC) types across three domains of life.

By representing complex systems as networks of interactions between their components [25], the study of such networks is recently gaining importance in biological disciplines. Efforts have been made to apply network concepts to describe protein molecular world, such as protein-protein interactions [26-28], interactions within protein domain families [29, 30], residue contacts within protein structures [31, 32], conformational spaces of transition-states in protein folding [33], protein family/fold occurrence and distribution in genomes [34], protein structural similarity networks [35] and similarity networks of protein binding sites [36], etc. These investigations have provided systematic and deeper understanding of the evolution and diversity of proteins.

In this study, we first attempted to annotate different genome sequences by using BLAST searching [37] against SCOPEC sequence database. For each annotated protein sequence, the predicted function (EC 3-digit level) and predicted structure (SCOP superfamily level) is regarded as one enzymatic SFC type. Annotated enzymes with catalytic domains from the same superfamily are considered to have an identical structure. Annotated enzymes sharing the identical three-digit EC number are considered to have the same enzymatic function. Thus, the occurrence of different SFCs in different species can be analyzed, which allows us to have a global view on enzymatic SFC types in three different domains of life. Moreover, a novel network of enzyme structures is constructed by considering their functional similarity. In the last part of this paper, the potential applications of the obtained network are discussed.

MATERIALS AND METHODS

Data Sets

The SCOPEC database, downloaded from http://www.enzome.com/enzome/, was employed to derive enzyme SFC types in different genomes. Based on SCOP database (version 1.63) [16], the current SCOPEC database contains 15761 catalytic domains, covering 250 folds, 340 superfamilies and 593 families. Structural domains in the same superfamily share distinctive features that suggest a common evolutionary ancestor. Therefore, here enzymes from the same SCOP superfamily are considered to have the same structure. Concerning an EC number, the first digit indicates a general level of function. Subsequent digits indicate more specific features of the catalytic reaction through subclass, subsubclass and finally a serial number, often used to distinguish different substrate specificities. The current SCOPEC includes 141 and 771 different EC numbers at three-digit level and four-digit level, respectively. It has been well observed that substrate specificity, measured at the fourth–level EC number, is not conserved within homologues, whereas function, measured at the third EC level, is often conserved [18, 23]. Furthermore, the fourth EC number for some enzymes in SCOPEC database is unknown and it is denoted as "-". In this work, therefore enzymes sharing the identical three-digit EC number are considered to have the same function.

A set of 6 eukaryotic, 9 archaeal, and 17 bacterial genomes sampled over 100 finished genome sequences, as initially selected by Caetano-Anolles and Caetano-Anolles [20], was utilized for the genome-wide investigation of enzymatic SFC types. Additionally, two other eukaryotic species (*H. Sapiens* and *M. musculus*) were also included. Thus, the 34 different species were analyzed in this study. Their corresponding protein sequences were downloaded from the website of NCBI (http://www.ncbi.nlm.nih.gov/).

Mapping Genome Sequences into SCOPEC Database

For each sequence within the genomes under investigation, a BLAST searching [37] was performed against the SCOPEC sequence database. All the parameters for BLAST searching were set as the default values defined in the BLAST package. Since the BLAST searching only generates local alignments, the ClustalW algorithm [38] was further employed to obtain a global alignment between the query sequence and the top hit from BLAST searching. Finally, the sequence identity between the query sequence and the top hit was counted as follows. For instance, if the top hit (sequence length = N_h) shares N_{id} identical residues with the query sequence (sequence length = N_q), the sequence identity is defined as: SeqId%= $N_{id} / N_h \times 100\%$. In this study, if the sequence identity is $\geq$ 40%, the top hit is assigned as a confident annotation.

For each annotated sequence, the predicted function (EC 3-digit level) and predicted structure (superfamily level) is regarded as one enzymatic SFC type. Thus, the total numbers of different enzymatic SFC types identified in archaea, bacteria and eukaryota were calculated. In each genome, two parameters were measured for all the annotated enzymes, 1) the number of functions per structure and 2) the number of structures per function.

Construction of a Functional Similarity Network for Enzyme Structures

Each SCOP superfamily is regarded as an enzyme structure provided that it adopts at least one enzymatic function. Then, all the enzyme structures (*i.e.* SCOP superfamilies) compose the nodes in the network, and an edge is assigned between two nodes if they can adopt at least one identical enzymatic function (*i.e.* EC 3-digit level). For instance, superfamily A adopts *M* different enzymatic functions, whereas superfamily B can have *N* different enzymatic functions. If at least one function can be overlapped in these *M* and *N* functions mentioned above, an edge is assigned between these two nodes (superfamily A and superfamily B). Based on the enzymatic structures and functions presented in the SCOPEC database, archaea species, bacteria species and eukaryota species, four functional similarity networks (SCOPEC-FSN, archaea-FSN, bacteria-FSN and eukaryota-FSN) were constructed.

RESULTS AND DISCUSSION

Occurrence of Enzymatic SFC Types in Different Genomes

To investigate enzymatic SFC types in different species, 34 different genomic sequences were annotated using SCOPEC database. The fraction of confidently annotated sequences within these 34 genomes is low, ranged from 2% to 13% (*cf.* Table **1**). The average annotation rate of the archaeal species is 0.030 ± 0.005, while the higher annotation rates are observed in the bacterial and eukaryotic species (0.062 ± 0.019 and 0.050 ± 0.011, respectively).

137 different SFC types are identified in archaea, 115 of which also consistently appeared in bacteria and eukaryota (Fig. **1**). A larger number of SFC types (300 types) occur in

Table 1. Size and Fraction of Structure and Function known Enzyme Sets in Different Species [a]

	Species	Proteome size	Number of annotated enzymes	Fraction of annotated enzymes
E	*H. sapiens*	27960	1561	0.056
	M. musculus	26650	1382	0.052
	D. melanogaster	18941	1034	0.055
	C. elegans	21124	607	0.029
	S. cerevisiae	5862	366	0.062
	S. pombe	5034	301	0.060
	N. crassa	11857	514	0.043
	A. thaliana	28860	1292	0.045
B	*B. burgdorferi*	851	36	0.051
	H. pylori	1576	90	0.057
	M. genitalium	484	28	0.058
	M. pneumoniae	689	34	0.049
	T. pallidum	1036	49	0.047
	C. pneumoniae	1112	61	0.055
	R. prowazekii	835	59	0.071
	D. radiodurans	2629	157	0.060
	E. coli	5379	373	0.069
	B. subtilis	4105	266	0.065
	M. tuberculosis	4189	203	0.048
	S. aureus	2615	177	0.068
	S.sp PCC 6803	3167	170	0.054
	A. aeolicus	1529	111	0.072
	H. influenzae	1657	214	0.129
	C. acetobutylicum	3672	171	0.047
	T. maritima	1858	112	0.060
A	*A. fulgidus*	2420	82	0.034
	M. thermautotrophicus	1873	60	0.032
	P. horikoshii	1955	59	0.030
	M. jannaschii	1729	58	0.034
	H.sp NRC-1	2075	72	0.035
	T. acidophilum	1482	53	0.035
	S. solfataricus	2977	67	0.023
	S. tokodaii	2825	59	0.021
	A. pernix	1841	52	0.028

[a] **E**, eukaryota; **B**, bacteria; **A**, archaea

bacterial species. Of them, about 28% combination types remain unique in bacteria. Interestingly, only a slightly larger number of SFC types (313 types) in eukaryotic species are observed, including about 35% unique types (*cf.* Fig. **1**). One possible explanation is that a much larger genome size in eukaryotic genomes can make use of multiple copies of some SFC types instead of inventing new ones.

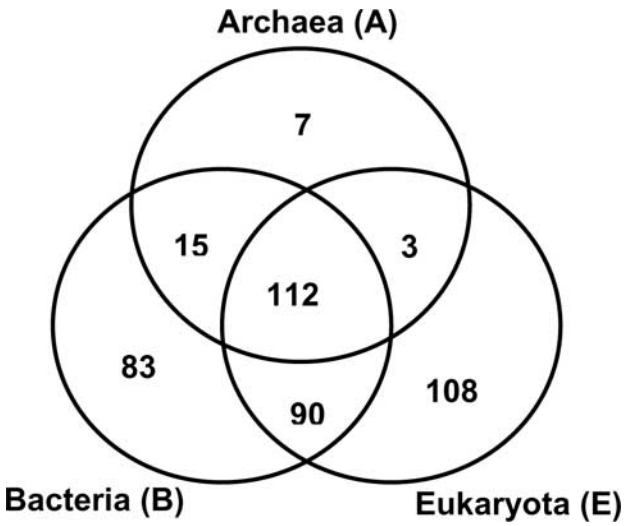

Figure 1. The Venn diagram shows the distribution of different enzymatic structure and function combination types in three different domains of life.

Inferring Structure and Function from Sequence

As the structures and functions have been experimentally characterized for only a very limited number of enzymes, investigation on the enzyme SFC types per species requires the inference of structural and functional information from sequence. Previous studies have shown that enzymes with ≥ 40% sequence identity share in most cases the same function (*e.g.* identical 3-digit EC number) [2, 3]. Therefore, using a cut-off of 40% sequence identity allows us to transfer the functional annotation from the top BLAST hit to the query sequence with reasonable confidence. Considering only local alignments obtained from BLAST program, the sequence identity based on a global alignment between the query sequence and the top BLAST hit was calculated using ClustalW algorithm. Furthermore, the sequence length of the top BLAST hit was used as the reference to calculate the identity. Since each entry in the SCOPEC database represents one catalytic domain, the query sequence should contain a domain with the same function as the top BLAST hit if they share ≥ 40% sequence identity. Generally protein structure is much more conserved than sequence, therefore a 40% cut-off sequence identity can also be reasonable to claim that the query sequence should contain a similar structural domain as the top BLAST hit. In some cases, the query sequence may contain multiple catalytic domains. Thus, attention on the second or other lower rank hits may be helpful for annotating other catalytic domains in the query sequence. It should be pointed out, however, only the top BLAST hit was analyzed in the current analysis.

Similar strategy was previously used by Freilich *et al.* [23] to annotate enzymatic functions for protein sequences with the purpose of investigating the complement of enzymatic sets in different species. To study enzyme SFC types in different species, it needs to be emphasized that in this work only SCOPEC was used to annotate genome sequences. It has been well accepted that other sensitive profile-based sequence searching algorithms (*e.g.* PSI-BLAST [37]) may identify more distantly related homologues and increase the rate of annotation. Such distant relatives may have often evolved into new functions [3], therefore the profile-based searching was not used in this work.

Evolution of Enzyme Structure and Function

For the annotated proteins in each genome, the structural diversity in each enzymatic function and the functional diversity in each superfamily are observed. The larger number of superfamilies per function versus functions per superfamily observed in each genome (*cf.* Fig. **2**) suggests that nature re-invents function (convergent evolution). Following a re-invention, it is likely that modification leads to new specificities of function (divergent evolution) [39]. The overall evolutionary relationship of enzymatic structure and function in each genome is in line with that reported by George *et al.* [18], which in turn is based on the whole SCOPEC database.

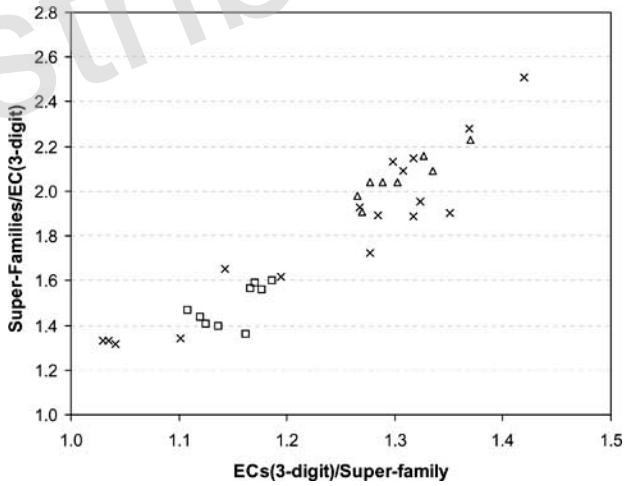

Figure 2. Measures of the ECs (3-digit level) per SCOP superfamily and the SCOP superfamilies per EC (3-digit level) in eukaryotic (Δ), archaeal (□), and bacterial (×) genomes.

A similar relationship between functional diversity and structural diversity exists in each archaea genome (Fig. **2**). Compared with archaea species, eukaryota species demonstrate a significant larger number of superfamilies per function as well as a larger number of functions per superfamily. Compared with archaea and eukaryota species, the relationships between functional diversity and structural diversity are quite different within different bacterial genomes. The ratio ($R_{c/d}$) between the number of superfamilies per function and the number of functions per superfamily in each genome can reflect the relative effect of convergent and di-

vergent evolutions. For all the eukaryotic species under scrutiny, the average $R_{c/d}$ is 1.58 ± 0.03, which is larger than that of the archaeal genomes (1.29 ± 0.05). Compared with archaea, the convergent evolution is more dominant in eukaryota. It has been observed that the $R_{c/d}$ value of bacterial genomes is ranging from 1.22 to 1.77 (the average value is 1.46 ± 0.13). The highly diverse distribution of $R_{c/d}$ in bacterial genomes suggests the dual behaviours of convergent and divergent evolutions exist in bacteria. The $R_{c/d}$ values of some bacterial genomes are close to those of archaeal genomes, while the $R_{c/d}$ values of some other bacterial genomes are close to those of eukaryotic genomes (*cf.* Fig. **2**). Additionally, it is also interesting to mention that the statistical analyses of functions per structure and structures per function in *E. coli* were previously measured by Tsoka and Ouzounis [40], which are in good agreement with our current results.

Connectivity within Functional Similarity Network of Enzyme Structure

A novel way to represent the enzyme structures as a functional similarity network has been proposed. Based on the enzyme structures and functions presented in SCOPEC database, archaea species, bacteria species and eukaryota species, four networks (SCOPEC-FSN, archaea-FSN, bacteria-FSN and eukaryota-FSN) are constructed, which are further characterized in Table **2**. The average connectivity of a network is defined as the average value of the number of links (edges) for each node within the network. Generally, the average connectivity values in bacteria-FSN and eukaryota-FSN are close, which are much larger than that of archaea-FSN. To a certain extent, the average connectivity for archaea-FSN, bacteria-FSN and eukaryota-FSN reflects the convergent evolution in different kingdoms of life. Therefore, a larger number of the average connectivity in bacteria-FSN and eukaryota-FSN is in line with more significant convergent evolution observed in bacteria and eukaryota. Considering the rule for deriving functional similarity network, it is not surprising that the number of edges for each node (*i.e.* connectivity) in the network is linearly correlated with the number of ECs for the corresponding superfamily (the correlation coefficient is 0.765). The characteristic path length is defined as the number of links in the shortest path between two nodes averaged over all pairs of nodes, which is also characterized in Table **2**. Interestingly, the characteristic path lengths for bacteria-FSN and eukaryota-FSN are similar, which are significantly larger than that of archaea-FSN.

Subsequently, we attempted to determine if this functional similarity network is a scale-free network. Scale-free networks typically have many nodes with few links and have only few highly connected ones [25]. In contrast to a random network in which the connectivity distribution obeys a Poisson distribution, the probability $P(k)$ of nodes having k edges, decays as a power law $P(k) = k^{-\gamma}$ in scale-free networks. As shown in Fig. **3**, SCOPEC-FSN is significantly deviated from a Poisson distribution as well as poorly fitted with a power-law distribution. Therefore, SCOPEC-FSN can not be assigned as a scale-free network. The similar distributions were also observed in the archaea-FSN, bacteria-FSN and eukaryota-FSN networks. It can be envisaged that such functional similarity network could be constructed for each genome, thus it will provide a deeper investigation on the differences of enzyme structure and function relationship in different species. However, the low fraction of sequences annotated by SCOPEC makes the network based on each genome far to be complete, thus a higher fraction of annotated sequences is required in the future to facilitate such analysis.

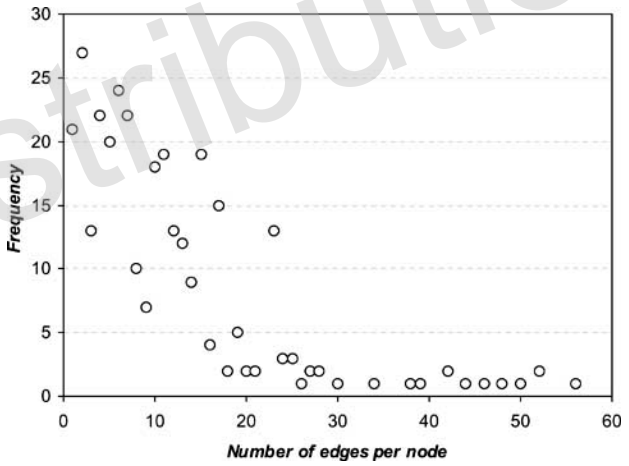

Figure 3. The distribution of node connectivity in the SCOPEC-FSN network.

Highly Connected Hubs in Functional Similarity Network

Several highly connected hubs have been identified in these newly derived functional similarity networks. The top ten hubs in each of these four networks are overlapped to a

Table 2. The Average Connectivity and Characteristic Path Length of Different Functional Similarity Networks of Enzyme Structures

	Number of nodes	*Average connectivity*	*Characteristic path length*
SCOPEC-FSN	324	11.0	3.2
Archaea-FSN	85	4.1	2.3
Bacteria-FSN	183	7.8	3.5
Eukaryota-FSN	195	7.6	3.6

Table 3. **The Ten Most Highly Connected Superfamilies within the Functional Similarity Network of SCOPEC-FSN.**[a]

	Superfamily	Description of superfamily	Description of fold	Number of connectivity
1	c.1.2	Ribulose-phoshate binding barrel	TIM beta/alpha-barrel	56
2	c.67.1	PLP-dependent transferases	PLP-dependent transferases	52
3	c.2.1	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains	52
4	c.69.1	Alpha/beta-Hydrolases	Alpha/beta-Hydrolases	50
5	c.37.1	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases	48
6	c.1.10	Aldolase	TIM beta/alpha-barrel	46
7	c.14.1	ClpP/crotonase	ClpP/crotonase	44
8	c.79.1	Tryptophan synthase beta subunit-like PLP-dependent enzymes	Tryptophan synthase beta subunit-like PLP-dependent enzymes	42
9	b.81.1	Trimeric LpxA-like enzymes	Single-stranded left-handed beta-helix	42
10	c.1.12	Phosphoenolpyruvate/pyruvate domain	TIM beta/alpha-barrel	39

[a] The descriptions of fold and superfamily are extracted from SCOP database.

certain extent. For the top ten hubs in SCOPEC-FSN (*cf.* Table **3**), six, eight and seven hubs appear as the top hubs in archaea-FSN, bacteria-FSN and eukaryota-FSN, respectively.

Of these top ten hubs in SCOPEC-FSN, nine hubs belong to α/β structural class and the remained one is from β structural class. The reason for the hubs highly occurred in α/β structural class could be due to that the combination of rigid surface formed by β-sheets with the conformational flexibility provided by α-helices makes these scaffolds more suitable for enzymatic function [41]. As reported by George *et al.* [18], the most popular enzymatic structures in SCOPEC database are the PLP-dependent transferases (c.67.1), α/β hydrolases (c.69.1), P-loop containing nucleotide triphosphate hydrolases (c.37.1) and NAD(P)-binding Rossmann-fold domains (c.2.1). It is not surprising that the above four superfamilies are ranked as the top ten hubs in the SCOPEC-FSN network.

Three of these top ten hubs are grouped into the fold of TIM β/α-barrel, but none of them can adopt more than 6 different ECs (three-digit level) [18]. This means that the top hubs are not necessary to be the superfamilies with multiple functions. To be one of the top hubs in SCOPEC-FSN, the superfamily should satisfy with at least one of the following two criteria: 1) The superfamily should host multiple functions; 2) The superfamily should have functions which are widely adopted by other superfamilies.

It has also been observed that in some cases the evolution of protein structure, function and domain-domain interaction is interconnected. As reported in [20], the phylogenetic tree of protein architectures identified three α/β folds as the most ancestral. They are c.37, c.1 and c.2, ordered from more to less ancestral. Interestingly, 5 out of the top ten hubs in SCOPEC-FSN can be grouped into these three most ancestral folds (*cf.* Table **3**). Recently, the large-scale protein structural interactome suggested 19 SCOP superfamilies as the most interactive superfamilies [42]. Also interestingly enough, two out of these 19 superfamilies (*i.e.* c.37.1 and c.2.1) appear as top-ranked hubs in SCOPEC-FSN network. All together these data suggest that some ancestral folds (*e.g.* c.37.1 and c.2.1) are favored to host different enzymatic functions. Meanwhile, they are able to easily "combine" with many other protein domains to support diverse biological functions.

Applications of Functional Similarity Networks

The established functional similarity network of enzymatic structures is opening avenues for several potential applications. Historically, both the function and structure databases of enzymes are organized in hierarchical ways. In the future, the enzyme structure and function database can be presented in a network graph, which may provide a more intuitive understanding in the relationship of enzymatic structure and function. A second application could be identified in the area of *de novo* enzyme molecular design [43]. Undoubtedly, the availability of such an extensive similarity network of enzymatic structures will provide a *priori* knowledge of *de novo* designability for a specific enzymatic structure under scrutiny. A third immediate application concerns enzyme function prediction by searching enzymatic structures for 3D residue patterns resembling known catalytic sites [44]. The structural genomics projects are determining the structures of many proteins with unknown functions. Therefore, searching for 3D residue patterns is a useful complement to the classical methods based on sequence or overall structural similarities. Mapping a structure under scrutiny into the above functional similarity network may provide an optimal searching path. For example, the searching priority should be given to those catalytic sites represented in the neighboring nodes of the query structure.

ACKNOWLEDGEMENTS

We thank Dr. Marco VENTURA (University of Parma, Italy) and Dr. Carlos A. CANCHAYA (National University of Ireland) for their critical reading of this manuscript. ZZ is also grateful for many stimulating discussions with Dr. Martin G. GRIGOROV (Nestle Research Centre, Switzerland) on the general topic of Protein Bioinformatics.

REFERENCES

[1] Rost, B. (**2002**) *J. Mol. Biol., 318*: 595-608.
[2] Tian, W. and Skolnick, J. (**2003**) *J. Mol. Biol., 333*: 863-882.
[3] Todd, A.E., Orengo, C.A. and Thornton, J.M. (**2001**) *J. Mol. Biol., 307*: 1113-1143.
[4] Chou, K.C. and Elrod, D.W. (**2003**) *J. Proteome Res., 2*: 183-190.
[5] Chou, K.C. and Cai, Y.D. (**2004**) *Protein Sci., 13:* 2857-2863.
[6] Chou, K.C. (**2005**) *Bioinformatics, 21:* 10-19.
[7] Cai, Y.D. and Chou, K.C. (**2005**) *J. Proteome Res., 4:* 109-111.
[8] Chou, K.C. and Cai, Y.D. (**2004**) *Biochem. Biophys. Res. Commun., 325:* 506-509.
[9] Cai, Y.D., Zhou, G.P. and Chou, K.C. (**2005**) *J. Theor. Biol., 234:* 145-149.
[10] Cai, Y.D. and Chou, K.C. (**2005**) *J. Proteome Res., 4:* 967-971.
[11] Orengo, C.A., Todd, A.E. and Thornton, J.M. (**1999**) *Curr. Opin. Struct. Biol., 9*: 374-382.
[12] Nagano, N., Orengo, C.A. and Thornton, J.M. (**2002**) *J. Mol. Biol., 321*: 741-765.
[13] Shulman-Peleg, A., Nussinov, R. and Wolfson, H.J. (**2004**) *J. Mol. Biol., 339*: 607-633.
[14] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (**2000**) *Nucl. Acids Res., 28*: 235-242.
[15] Porter, C., Bartlett, G. and Thornton, J.M. (**2004**) *Nucl. Acids Res., 32*: D129-D133.
[16] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (**1995**) *J. Mol. Biol., 247*: 536-540.
[17] Chou, K.C. and Cai, Y.D. (**2004**) *Biochem. Biophys. Res. Commun., 321:* 1007-1009.
[18] George, R.A., Spriggs, R.V., Thornton, J.M., AI-Lazikani, B. and Swindells, M.B. (**2004**) *Bioinformatics, 20 (Suppl.)*: i130-i136.
[19] Koonin, E.V., Wolf, Y.I. and Karev, G.P. (**2002**) *Nature, 420*: 218-222.
[20] Caetano-Anolles, G. and Caetano-Anolles, D. (**2003**) *Genome Res., 13*: 1563-1571.
[21] Peregrin-Alvarez, J.M., Toska, S. and Ouzounis, C.A. (**2003**) *Genome Res., 13*: 422-427.
[22] Aguilar, D., Aviles, F.X., Querol, E. and Sternberg, M.J. (**2004**) *J. Mol. Biol., 340*: 491-512.
[23] Freilich, S., Spriggs, R.V., George, R.A., AI-Lazikani, B., Swindell, M. and Thornton, J.M. (**2005**) *J. Mol. Biol., 349*: 745-763.
[24] Yang, S., Doolittle, R.F. and Bourne, P.E. (**2005**) *Proc. Natl. Acad. Sci. USA, 102*: 373-378.
[25] Strogatz, S.H. (**2001**) *Nature, 410*: 268-276.
[26] Bork, P. Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (**2004**) *Curr. Opin. Struct. Biol., 14*: 292-299.
[27] Wuchty, S. (**2004**) *Genome Res., 14*: 1310-1314.
[28] Chou, K.C. and Cai, Y.D. (**2006**) *J. Proteome Res., 5:* 316-322.
[29] Park, J., Lappe, M. and Teichmann, S.A. (**2001**) *J. Mol. Biol., 307*: 929-938.
[30] Wuchty, S. (**2001**) *Mol. Biol. Evol., 18*: 1694-1702.
[31] Greene, L.H. and Higman, V.A. (**2003**) *J. Mol. Biol., 334*: 781-791.
[32] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. (2004) *J. Mol. Biol., 344*: 1135-1146.
[33] Rao, F. and Caflisch, A. (**2004**) *J. Mol. Biol., 342*: 299-306.
[34] Qian, J., Luscombe, N.M. and Gerstein, M. (**2001**) *J. Mol. Biol., 313*: 673-681.
[35] Dokholyan, N.V., Shakhnovich, B. and Shakhnovich, E.I. (**2002**) *Proc. Natl. Acad. Sci. USA, 99*: 14132-14136.
[36] Zhang, Z. and Grigorov, M.G. (**2006**) *Proteins, 62*: 470-478.
[37] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (**1997**) *Nucl. Acids Res., 25*: 3389-3402.
[38] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (**1994**) *Nucl. Acids Res., 22*: 4673-4680.
[39] Galperin, M.Y., Walker, D.R. and Koonin, E.V. (**1998**) *Genome Res., 8*: 779-790.
[40] Tsoka, S. and Ouzounis, C.A. (**2001**) *Genome Res., 11*: 1503-1510.
[41] Anantharaman, V., Aravind, L. and Koonin, E.V. (**2003**) *Curr. Opin. Chem. Biol., 7:* 12-20.
[42] Bolser, D., Dafas, P., Harrington, R., Park, J. and Shroeder, M. (**2003**) *BMC Bioinformatics, 4*: 45.
[43] Dwyer, M.A., Looger, L.L. and Hellinga, H.W. (**2004**) *Science, 304*: 1967-1971.
[44] Torrance, J.W., Bartlett, G.J., Porter, C.T. and Thornton, J.M. (**2005**) *J. Mol. Biol., 347*: 565-581.

Received: December 12, 2006 Revised: January 31, 2007 Accepted: February 01, 2007