Finding evolutionary relations beyond superfamilies: Fold-based superfamilies

KEIKO MATSUDA,¹ TAKAAKI NISHIOKA,² KENGO KINOSHITA,^{3,4}
TAKESHI KAWABATA,¹ AND NOBUHIRO GO^{1,5}

¹Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, 630-0101, Japan

²Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan

³Graduate School of Integrated Science, Yokohama City University, Yokohama 230-0045, Japan

⁴Structure and Function of Biomolecules, PRESTO, Japan Science and Technology Corporation, Kawaguchi, Saitama 332-0012, Japan

⁵Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, Kizu, Souraku, Kyoto, 619-0215, Japan

(RECEIVED May 7, 2003; FINAL REVISION July 8, 2003; ACCEPTED July 8, 2003)

Abstract

Superfamily classifications are based variably on similarity of sequences, global folds, local structures, or functions. We have examined the possibility of defining superfamilies purely from the viewpoint of the global fold/function relationship. For this purpose, we first classified protein domains according to the β -sheet topology. We then introduced the concept of kinship relations among the classified β -sheet topology by assuming that the major elementary event leading to creation of a new β -sheet topology is either an addition or deletion of one β -strand at the edge of an existing β -sheet during the molecular evolution. Based on this kinship relation, a network of protein domains was constructed so that the distance between a pair of domains represents the number of evolutionary events that lead one from the other domain. We then mapped on it all known domains with a specific core chemical function (here taken, as an example, that involving ATP or its analogs). Careful analyses revealed that the domains are found distributed on the network as >20 mutually disjointed clusters. The proteins in each cluster are defined to form a fold-based superfamily. The results indicate that >20 ATP-binding protein superfamilies have been invented independently in the process of molecular evolution, and the conservative evolutionary diffusion of global folds and functions is the origin of the relationship between them.

Keywords: ATP-binding domains; kinship relations of global folds; purine biosynthesis; structure/function relationship

Supplemental material: See www.proteinscience.org

Empirical relationships between the three-dimensional structures of proteins and their functions provide powerful tools for deducing functions from structures. Two types of such relationships appear to exist; the first is between local structures and functions (Wallace et al. 1996; Fetrow and

Skolnick 1998), and the other is between global folds and functions (Orengo et al. 1994; Lo Conte et al. 2000). Although the mechanisms of enzymatic reaction are likely to explain the former type of relationship, such direct mechanisms to explain the latter type are rather difficult to elucidate. The latter type of relationship may have its origin in molecular evolution; that is, both global folds and functions are conserved during the molecular evolution within certain ranges of variation (Todd et al. 2001), and hence, a relationship exists between global folds and functions both with certain ranges of variation. Mechanisms should exist only

Reprint requests to: Takaaki Nishioka, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan; e-mail: nishioka@scl.kyoto-u.ac.jp; fax: 81-75-753-6408.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.0383603>.

indirectly behind the conservation of global folds and their associated functions.

The validity of this picture can be attested if clear clusters of superfamilies of proteins with respective common core functions can be identified in a fold space of proteins with a certain measure of fold similarity. Each such cluster, if identified, should deserve to be called as a fold-based superfamily. The aim of this article is to show that this is actually possible with an objective procedure.

This aim was achieved by first constructing a network of kinship relations of global folds of protein domains. To limit the scope of the study, we focused on proteins with functions involving ATP or its analogs. This choice of proteins was made because of their functional importance and abundance of three-dimensional structures. We mapped all of the known ATP-binding domains on the constructed network. If the mapped domains formed mutually disjoint clear clusters, each with a similar core chemical function, then such clusters are examples of fold-based superfamilies. The fold-based superfamilies thus identified in this article will lead us to some interesting new findings.

Results and Discussion

Construction of a global fold kinship relations network

First, we must construct a fold space with a proper measure of fold similarity. Because domains rather than entire protein molecules are thought to be the units of long-range evolution and because we are interested in the extent of global fold diversity, we should construct a fold space in which protein domains with different sets of secondary structural elements can be mutually related. We introduce the concept of a network of kinship relations, which is intended to locate essentially all β -sheet-containing protein domains. We focus our attention on β -sheet topology (Richardson 1977; Ptitsyn and Finkelstein 1980), that is, the directions of individual β -strands and their connectivity (Fig. 1A). The kinship relations among various β -sheet topologies are defined based on the assumption that a predominant elementary event leading to the creation of a new β -sheet topology during the molecular evolution is either the addition or deletion of one β -strand at the edge of an existing β -sheet (Fig. 1B). β -Sheets are occasionally in contact with one or two α -helix layers. However, we disregarded α -helices from our consideration in this article.

Each domain from the 1011 nonhomologous β -sheet-containing domains within the CATH database (Orengo et al. 1997) was classified by focusing our attention on the topology of the largest, and sometimes the second largest, β -sheets. Domains with the same β -sheet topology but with different α -helix arrangements were collected into one group (Supplementary Table 1). Such a group is defined here as a β -topology group. We classified the global folds of

Figure 1. Definitions of β -sheet topology and kinship relation of β -topology groups. (A) Definition of β -sheet topology. β -Strands are numbered according to the order of the amino acid sequence. The underlined numbers represent β -strands that are antiparallel to strand 1. The first characters—P, A, and M—refer to parallel-, antiparallel-, and mixed-type β -sheets, respectively. The same β -sheet can be expressed in two ways, for example, M15423 or M32451, depending on the face of the β -sheet from which the sheet is viewed. However, we usually were not concerned with the face of the β -sheet. In such cases, we used the convention that the strand number at the left edge is lower than that at the right edge; for example, the β -topology group of the example is M15423, but not M32451. (B) Kinship 1 relation among β -topology groups. The pair of M15423 and M1432 is an example of a kinship 1 relation. Deleting strand 1 from M15423 and renumbering the rest of the strands gives M1432. (C) Examples of a network of kinship 1 relations. M15423 and A2134, M15423 and M1432, M15423 and M216534, and A21345 and A2134 are kinship 1 pairs. M15423 and A21345 are connected by two kinship 1 distances, because both of them form a kinship 1 pair with A2134. M216534 is connected to A21345 by three kinship 1 distances.

the nonhomologous domains into 428 β -topology groups (Supplementary Table 2).

Note that a β -topology group can be a mixture of evolutionarily related and unrelated domains. A pair of domains, which are closely related evolutionarily, should belong to the same β -topology group. However, an evolutionarily unrelated pair of domains with an accidentally identical β -sheet may also be members.

Then, we introduce the kinship relations among the β -topology groups, based on the assumption described above. A pair of β -topology groups, related by the deletion or addition of one β -strand at the edge of a β -sheet, is defined as having the first degree of kinship, kinship 1 (Fig. 1B). We note here that a kinship relation between a pair of β -topology groups can, in some cases, stand for two or more different evolutionary lineages.

When the kinship 1 relationship was introduced into the 428 β -topology groups, the majority of them were con-

nected into one large network. The remaining β -topology groups were found either as an isolated orphans or as members of small networks with up to three members. However, a large fraction of these orphans and small networks can be connected to the one large network via a β -topology group that is not a member of the 428 β -topology groups. There are 101 such β -topology groups that can act as a bridge, called bridging β -topology groups. We added these 101 β -topology groups to the 428 β -topology groups in the following analyses. The resulting 529 β -topology groups now belonged to one major network, except for 36 orphans and seven groups in three isolated small networks (Table 1; Supplementary Table 2). Even though the majority of the β -topology groups were connected into one main network, this does not necessarily mean that they are all mutually evolutionarily related. This is because a kinship relation can stand for two or more different evolutionary lineages. In fact, later in this article, we will describe how the major network can be decomposed into ≥ 20 disjointed clusters, by an analysis that dissects different evolutionary lineages.

The 36 orphans discussed above are those lacking any intermediates connecting them to the network. This lack of an intermediate group might be due to the limited number of known structures. The occurrence of orphans is rare in β -topology groups of small β -sheets; they comprise only 3.5% (13 of 376) among the β -topology groups of two- to eight-stranded β -sheets (Table 1). This rarity supports the validity of the assumption that the addition or deletion of one β -strand at the existing β -sheet is the major elementary event that took place during evolution. The lack of intermediates is either due to a paucity of known structures or due to a consequence of minor events during evolution, such as domain swapping (Fukami-Kobayashi et al. 1999). In contrast, orphans are found among almost half (44% = 23

of 52) of the β -topology groups of 9- to 18-stranded β -sheets. Some of these orphans may have evolved by the fusion of two smaller β -sheets rather than by the stepwise addition of a β -strand (Table 2).

Basic features of the network of global fold kinship relations

On the network, parallel and antiparallel β -topology groups are not separated from each other but are connected via mixed type β -topology groups (Fig. 2); for example, the P12 β -topology group is connected to the A12 β -topology group by six mixed-type, three-stranded β -topology groups such as M123.

Most of the β -topology groups, ~79% (337 of 428), have only one or two domains in each group. Some β -topology groups are more populated. The most populous β -topology groups are A12 (e.g., mainly β -hairpin; 91), A123 (e.g., mainly meander; 88), A2134 (55), A1234 (51), A1243 (50), P12345678 (TIM-barrel; 48), A132 (34), A1423 (32), and A213 (27) in which the number of domains classified in each β -topology group is given in the parentheses. The total number of domains that can be classified in these nine groups is about one third (476 of 1444) of all of the domains analyzed in the present study. These populated β -topology groups apparently tolerate various kinds of amino acid sequences and, therefore, should be highly flexible for creating new functions. The well-populated β -topology groups also frequently appear as substructures of larger sheets; for example, A123 also appears also in A2134, A1234, and A1243. The TIM-barrel P12345678, which is one of the nine most populated groups, is an exception and does not occur as a substructure of other larger sheets.

Table 1. Distribution of domains, β -topology groups, and kinship relation

Number of strands in a β -topology group	Number of domains	Number of β -topology groups	Number of β -topology groups having at least one kinship 1 pair	Number of β -topology groups connected to a major network via bridging β -topology group	Orphans
2	104	2	2	0	0
3	190	12	12	0	0
4	350	43	43	0	0
5	309	94	92	2	0
6	206	105	90	13	2
7	113	70	45	23	2
8	114	50	26	15	9
9	20	20	9	4	7
10	19	15	3	5	7
11	8	5	2	0	3
12–18	15	12	6	0	6
Total ^a	448	428	330	62	36

^a Bridging β -topology groups are listed in Supplementary Table 2.

Table 2. List of β -topology groups that might be created by a fusion of two β -sheets and examples of proteins with their PDB entry codes

β -Sheet topology	Component topologies	PDB entry codes (protein name, chain, and domain)	
M125346987	A12534 + A1432	1gof	(galactose oxidase, domain 1)
M876514329	A12345 + M3214	1dpg	(glucose 6-phosphate dehydrogenase, chain A, domain 2)
M876514329	A12345 + M3214	1ofg	(glucose-fructose oxidoreductase, chain A, domain 2)
M21310945867	M21354 + M12534	1xaa	(3-isopropylmalate dehydrogenase)
M65437189210	M1234 + M2134	1fsu	(<i>N</i> -acetylgalactosamine-4-sulfatase, domain 1)
M3216549871011	M21543 + M32145	1ema	(green fluorescent protein)
M13212134561171098	A14523 + M23415	1epm	(aspartic proteinase, chain E, domain 1)
M56748321109121311	M23415 + A32154	3psg	(porcen pepsinogen, domain 1)
		1eag	(aspartic proteinase, chain A, domain 1)

About 44% (188 of 428) of the β -topology groups are connected to the network by a single kinship relation, whereas the rest are multiply connected by more than one kinship relation.

Mapping ATP-binding domains on the network

The next step is to analyze the distribution of the β -topology groups of the ATP-binding domains on this network. For

this purpose, we systematically collected as many three-dimensional structures of ATP-binding proteins as possible, by including those within the CATH nonhomologous domain set. According to the LIGAND chemical database for enzymatic reactions in KEGG (Goto et al. 2000), 406 enzymes use ATP as a substrate or an effector. For 75 among these enzymes, crystal structures have been registered in the Protein Data Bank (PDB, April 2000; Berman et al. 2000).

Figure 2. Network of global fold kinship relations. This figure is the network of β -sheets with two to five strands. Circles are two-stranded β -sheets; triangles, three-stranded β -sheets; squares, four-stranded β -sheets; and pentagons, five-stranded β -sheets. The symbols are colored red, blue, or green, indicating a parallel, antiparallel, or mixed β -sheet, respectively. The connecting lines are kinship 1 relations. This graph was drawn by using Graphviz (<http://www.research.att.com/sw/tools/graphviz/>).

Table 3. Summary of ATP-binding domains: their SCOP superfamilies (S) and families (F), and β -topology groups

S1. P-loop containing nucleotide triphosphate hydrolases (13 families, 47 proteins, 69 structures, 72 domains, 23 β -topology groups)
S1F1. Nucleotide and nucleoside kinases (8 proteins, 16 structures, 16 domains, 1 β -topology group) P23145
S1F2. Shikimate kinase (1 protein, 1 structure, 1 domain, 1 β -topology group) P23145
S1F3. Adenosine-5'-phosphosulfate kinase (APS kinase) (1 protein, 1 structure, 1 domain, 1 β -topology group) P23145
S1F4. PAPS sulfotransferase (3 proteins, 2 structures, 3 domains, 2 β -topology groups) P2134, P23145
S1F5. Phosphoribulokinase (1 protein, 1 structure, 1 domain, 1 β -topology group) M23145687
S1F6. 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase, kinase domain (1 protein, 1 structure, 1 domain, 1 β -topology group) M231456
S1F7. G proteins (14 proteins, 20 structures, 20 domains, 4 β -topology groups) M231456, M2314576, M2314567, M21453678
S1F8. Motor proteins (3 proteins, 7 structures, 7 domains, 3 β -topology groups) M1273645, M1726534, M21837645
S1F9. Nitrogenase iron protein-like (5 proteins, 6 structures, 6 domains, 4 β -topology groups) P231456, M32415678, M34251678, M432516789
S1F10. RecA protein-like (ATPase-domain) (4 proteins, 6 structures, 8 domains, 5 β -topology groups) P13245, P1435627, M3245167, M14356278, M324516789
S1F11. ABC transporter ATPase domain (1 protein, 1 structure, 1 domain, 1 β -topology group) M234156
S1F12. Extended AAA-ATPase domain (4 proteins, 5 structures, 5 domains, 1 β -topology group) P23415, P324516
S1F13. RNA helicase (1 protein, 1 structure, 2 domains, 2 β -topology groups) P165243, M1762534
S2. Nucleotidyl transferase (2 families, 7 proteins, 8 structures, 8 domains, 3 β -topology groups)
S2F1. Class I aminoacyl-tRNA synthetases (RS), catalytic domain (6 proteins, 7 structures, 7 domains, 3 β -topology groups) P32145, P321564, M165234
S2F2. CTP:glycerol-3-phosphate cytidyltransferase (1 protein, 1 structure, 1 domain, 1 β -topology group) P32145
S3. Adenine nucleotide α -hydrolases (2 families, 4 proteins, 4 structures, 4 domains, 3 β -topology groups)
S3F1. N-type ATP pyrophosphatases (3 proteins, 3 structures, 3 domains, 2 β -topology groups) P3214, P32145
S3F2. Phosphoadenylyl sulphate (PAPS) reductase (1 protein, 1 structure, 1 domains, 1 β -topology group) M321465
S4. β -Glucosyltransferase and glycogen phosphorylase (2 families, 3 proteins, 4 structures, 7 domains, 2 β -topology groups)
S4F1. β -Glucosyltransferase (DNA-modifying) (1 protein, 1 structure, 1 domain, 1 β -topology group) P321456
S4F2. Oligosaccharide phosphorylase (2 proteins, 3 structures, 6 domains, 2 β -topology groups) P321456, M345216789
S5. Phosphoglycerate kinase (1 family, 1 protein, 4 structures, 4 domains, 1 β -topology group)
S5F1. Phosphoglycerate kinase (1 protein, 4 structures, 4 domains, 1 β -topology group) P321456
S6. Phosphofructokinase (1 family, 1 protein, 2 structures, 4 domains, 3 β -topology groups)
S6F1. Phosphofructokinase (1 protein, 2 structures, 4 domains, 3 β -topology groups)
Domain 1: P2314
Domain 2: M32145, M3214567
S7. S-adenosyl-L-methionine-dependent methyltransferases (6 families, 10 protein, 11 structures, 11 domains, 5 β -topology groups)
S7F1. Catechol O-methyltransferase, COMT (1 protein, 1 structure, 1 domain, 1 β -topology group) M3214576
S7F2. Glycine N-methyltransferase (1 protein, 1 structure, 1 domain, 1 β -topology group) M3214576
S7F3. Chemotaxis receptor methyltransferase CheR, C-terminal domain (1 protein, 1 structure, 1 domain, 1 β -topology group) M3214576
S7F4. RNA methylases (2 proteins, 3 structures, 3 domains, 2 β -topology groups) M213465, M3214576
S7F5. DNA methylases (4 proteins, 4 structures, 4 domains, 3 β -topology groups) M213465, M3214576, M321457689
S7F6. Type II DNA methylase (1 protein, 1 structure, 1 domain, 1 β -topology group) M75643892110
S8. Ribokinase-like (2 families, 3 proteins, 3 structures, 3 domains, 2 β -topology groups)
S8F1. Ribokinase-like (2 proteins, 2 structures, 2 domains, 1 β -topology group) M321456789
S8F2. Thiazole kinase (1 protein, 1 structure, 1 domain, 1 β -topology group) M213456789
S9. Protein kinase-like (PK-like) (4 families, 20 proteins, 27 structures, 27 domains, 4 β -topology groups)
S9F1. Serine/threonine kinases (12 proteins, 18 structures, 18 domains, 3 β -topology groups) A12354, A123465, A156432
S9F2. Tyrosine kinase (6 proteins, 7 structures, 7 domains, 2 β -topology groups) A12345, A12354
S9F3. Atypical protein kinases (1 protein, 1 structure, 1 domain, 1 β -topology group) A12354
S9F4. Phosphoinositide 3-kinase (PI3K), catalytic domain (1 protein, 1 structure, 1 domain, 1 β -topology group) A123465
S10. Class II aaRS and biotin synthetases (2 families, 10 proteins, 13 structures, 13 domains, 5 β -topology groups)
S10F1. Class II aminoacyl-tRNA synthetase (aaRS)-like, catalytic domain (9 proteins, 12 structures, 12 domains, 5 β -topology groups) M1237654, M12387645, M12387654, M156784329, M546783219
S10F2. Biotin holoenzyme synthetase (1 protein, 1 structure, 1 domain, 1 β -topology group) M1237654
S11. Aspartate carbamoyltransferase, Regulatory-chain, C-terminal domain (1 family, 1 protein, 1 structure, 1 domain, 1 β -topology group)
S11F1. Aspartate carbamoyltransferase, Regulatory-chain, C-terminal domain (1 protein, 1 structure, 1 domains, 1 β -topology group) A23145
S12. Glutathione synthetase ATP-binding domain-like (6 families, 10 proteins, 11 structures, 22 domains, 6 β -topology groups)
S12F1. Peptide synthetases, C-terminal domain (2 proteins, 2 structures, 4 domains, 2 β -topology groups)
Domain 1: A1423
Domain 2: A32145
S12F2. Biotin carboxylase/Carbamoyl phosphate synthetase (4 proteins, 4 structures, 8 domains, 4 β -topology groups)
Domain 1: A1423
Domain 2: A32145, A5412367, A54123687
S12F3. Synapsin Ia, C-terminal domain (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: A1423
Domain 2: A32145

(continued)

Table 3. *Continued*

S12F4. Succinyl-CoA synthetase, β -chain, N-terminal domain (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: A1423
Domain 2: A32145
S12F5. Pyruvate phosphate dikinase, N-terminal domain (1 protein, 2 structures, 4 domains, 2 β -topology groups)
Domain 1: A12534
Domain 2: A32145
S12F6. Eukaryotic glutathione synthetase (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: A1423
Domain 2: A2134
S13. DNA ligase/m RNA capping enzyme, catalytic domain (3 families, 3 proteins, 3 structures, 6 domains, 4 β -topology groups)
S13F1. ATP-dependent DNA ligase, N-terminal domain (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: M15423
Domain 2: M321456
S13F2. Adenylation domain of NAD ⁺ -dependent DNA ligase (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: M15423
Domain 2: A2134
S13F3. RNA guanylyltransferase (mRNA capping enzyme), N-terminal domain (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: M15423
Domain 2: A321456
S14. Glutamine synthetase, N-terminal domain (2 families, 3 proteins, 6 structures, 6 domains, 2 β -topology groups)
S14F1. Glutamine synthetase, N-terminal domain (1 protein, 1 structure, 1 domain, 1 β -topology group) M2314765
S14F2. Guanido kinase (2 proteins, 5 structures, 5 domains, 1 β -topology group) A23451687
S15. Nucleoside diphosphate kinases (1 family, 1 protein, 5 structures, 5 domains, 1 β -topology group)
S15F1. Nucleoside diphosphate kinases (1 protein, 5 structures, 5 domains, 1 β -topology group) A2314
S16. Ribonuclease H-like (6 families, 8 proteins, 14 structures, 14 domains, 3 β -topology groups)
S16F1. Ribonuclease H (2 proteins, 4 structures, 4 domains, 1 β -topology group) M32145
S16F2. Retroviral integrase (1 protein, 1 structure, 1 domain, 1 β -topology group) M32145
S16F3. μ transposase, core domain (1 protein, 1 structure, 1 domain, 1 β -topology group) M32145
S16F4. Exonuclease domain of DNA polymerase (1 protein, 2 structures, 2 domains, 1 β -topology group) M143256
S16F5. 3'-5' exonuclease domain of archaeal and phage DNA polymerases (2 proteins, 5 structures, 5 domains, 2 β -topology groups) M32145, M432156
S16F6. RuvC resolvase (1 protein, 1 structure, 1 domain, 1 β -topology group) M32145
S17. Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-lyase) (1 family, 1 protein, 1 structure, 2 domains, 2 β -topology groups)
S17F1. Phosphoenolpyruvate carboxykinase (ATP-oxaloacetate carboxy-lyase) (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: M312465
Domain 2: M14568723
S18. Actin-like ATPase domain (3 families, 5 proteins, 11 structures, 14 domains, 2 β -topology groups)
S18F1. Actin/HSP70 (2 proteins, 7 structures, 9 domains, 2 β -topology groups) M32145, M321456
S18F2. Hexokinase (2 proteins, 3 structures, 4 domains, 2 β -topology groups) M32145, M143256
S18F3. Glycerol kinase (1 protein, 1 structure, 1 domain, 1 β -topology group) M32145
S19. Nucleotidyltransferases (2 families, 2 proteins, 3 structures, 3 domains, 2 β -topology groups)
S19F1. DNA polymerase β , catalytic (31kD) fragment (1 protein, 2 structures, 2 domains, 1 β -topology group) M12543
S19F2. Kanamycin nucleotidyltransferase (KNTase) (1 protein, 1 structure, 1 domain, 1 β -topology group) M34215
S20. DNA/RNA polymerases (2 families, 5 proteins, 9 structures, 9 domains, 3 β -topology groups)
S20F1. DNA polymerase I (3 proteins, 7 structures, 7 domains, 3 β -topology groups) A2314, A231456, A1342567
S20F2. Reverse transcriptase (2 proteins, 2 structures, 2 domains, 1 β -topology group) A2314
S21. Adenylyl and guanylyl cyclase catalytic domain (1 family, 3 proteins, 3 structures, 3 domains, 2 β -topology groups)
S21F1. Adenylyl and guanylyl cyclase catalytic domain (3 proteins, 3 structures, 3 domains, 2 β -topology groups) A23145, A2314576
S22. S-adenosylmethionine synthetase. MAT (1 family, 1 protein, 1 structure, 1 domain, 1 β -topology group)
S22F1. S-adenosylmethionine synthetase. MAT (1 protein, 1 structure, 1 domain, 1 β -topology group) A2134
S23. Firefly luciferase-like (1 family, 2 proteins, 2 structures, 4 domains, 4 β -topology groups)
S23F1. Firefly luciferase-like (2 proteins, 2 structures, 4 domains, 4 β -topology groups)
Domain 1: A1234, A12345
Domain 2: M12345, M214356789
S24. Pyruvate kinase β -barrel domain (1 family, 1 protein, 5 structures, 5 domains, 2 β -topology groups)
S24F1. Pyruvate kinase β -barrel domain (1 protein, 5 structures, 5 domains, 2 β -topology groups) P123456, P12345678
S25. SAICAR synthase-like (2 families, 2 proteins, 2 structures, 4 domains, 4 β -topology groups)
S25F1. SAICAR synthase (1 protein, 1 structure, 2 domains, 21 β -topology groups)
Domain 1: A1234
Domain 2: A2134
S25F2. Phosphatidylinositol phosphate kinase II β , PIPK II β (1 protein, 1 structure, 2 domains, 2 β -topology groups)
Domain 1: A1267543
Domain 2: M21354

(continued)

Table 3. Continued

S26.	PH domain-like (5 families, 15 proteins, 16 structures, 16 domains, 5 β -topology groups)
S26F1.	Pleckstrin-homology domain (PH domain) (7 proteins, 8 structures, 8 domains, 3 β -topology groups) A1234, A4321765, M43217658
S26F2.	Phosphotyrosine-binding domain (PTB, 3 proteins, 3 structures, 3 domains, 2 β -topology groups) A4321765, A43217658
S26F3.	Ran-binding domain of nuclear pore complex protein Nup358 (1 protein, 1 structure, 1 domain, 1 β -topology group) A432165
S26F4.	Enabled/VASP homology 1 domain (EVH1 domain) (3 proteins, 3 structures, 3 domains, 2 β -topology groups) A432165, A4321765
S26F5.	Third domain of FERM (1 protein, 1 structure, 1 domain, 1 β -topology group) A1234
S27.	Carbamate kinase (1 family, 1 protein, 1 structure, 1 domain, 1 β -topology group)
S27F1.	Carbamate kinase (1 protein, 1 structure, 1 domain, 1 β -topology group) M34215786
S28.	6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase, HPPK (1 family, 1 protein, 2 structures, 2 domains, 1 β -topology group)
S28F1.	6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase, HPPK (1 protein, 2 structures, 2 domains, 1 β -topology group) A41325
S29.	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase (3 families, 5 proteins, 6 structures, 6 domains, 3 β -topology groups)
S29F1.	Heat shock protein 90, HSP90, N-terminal domain (1 protein, 2 structures, 2 domains, 1 β -topology group) M16547328
S29F2.	DNA gyrase/MutL, N-terminal domain (2 proteins, 2 structures, 2 domains, 1 β -topology group) M54362178
S29F3.	Histidine kinase (2 proteins, 2 structures, 2 domains, 1 β -topology group) M12354
S30.	MurD-like peptide ligases, catalytic domain (2 families, 2 proteins, 2 structures, 2 domains, 2 β -topology groups)
S30F1.	UDP-N-acetylmuramoyl-L-alanine: D-glutamate ligase, gene MurD (1 protein, 1 structure, 1 domain, 1 β -topology group) P231456
S30F2.	Folylpolyglutamate synthetase (1 protein, 1 structure, 1 domain, 1 β -topology group) M43251678
S31.	Aminoimidazole ribonucleotide synthetase (PurM) N-terminal domain (1 family, 1 protein, 1 structure, 2 domains, 2 β -topology groups)
S31F1.	Aminoimidazole ribonucleotide synthetase (PurM) N-terminal domain (1 protein, 1 structure, 2 domains, 2 β -topology groups)
	Domain 1: M1423
	Domain 2: M241536
S32.	EPT/RTPC-like (2 families, 3 proteins, 4 structures, 4 domains, 1 β -topology group)
S32F1.	RNA 3'-terminal phosphate cyclase, RTPC (1 protein, 1 structure, 1 domain, 1 β -topology group) M1243
S32F2.	Enolpyruvate transferase, EPT (2 proteins, 3 structures, 3 domains, 1 β -topology group) M1243
Total	32 superfamilies, 83 families, 182 proteins, 254 structures, 285 domains, 93 β -topology groups ^a

SCOP superfamilies and families are arranged and numbered arbitrarily in the present study. For example, S1F1 refer to the first family F1 (SCOP family name is Nucleotide and nucleoside kinases) in the superfamily S1 (SCOP superfamily name is P-loop-containing nucleotide triphosphate hydrolases). The number of proteins, PDB structures, domains, and β -topology groups in each family and superfamily are given in the parentheses. Domain 1 and Domain 2 are a pair of two domains that constitute an ATP-binding site.

^a See Supplementary Table 3 for the PDB entry code and the β -topology group of each domain.

Their ATP-binding domains belong to 32 different Structural Classification of Proteins (SCOP) superfamilies (Hubbard et al. 1997). To enrich the analysis of the distribution, we extended the data from the 75 enzymes to 182 proteins belonging to the 32 superfamilies (see Materials and Methods). In this process of enrichment, some GTP-binding domains were also included. Finally, we obtained 93 β -topology groups from the PDB structures of 285 ATP-binding (and exceptionally some GTP-binding) domains (Table 3). Some of the SCOP domains consist of two substructures, which are each defined as a domain in our treatment, for example, the superfamily “Phosphofructokinase”. Some of the structures were determined after the collection of the CATH data set. For this reason, their β -topology groups might not be found in the network.

Upon the network of kinship relations, we mapped the 93 β -topology groups of the ATP-binding domains. Of them, 59 β -topology groups were found on the network. Another 27 groups were not found on the network but were connected to the network either by a kinship 1 relation or through a bridging β -topology group. A total of 86 (59 + 27) β -topology groups were widely distributed on the network (Fig. 3A,B). The remaining seven groups were orphans to the network and were not included in the following analysis.

With the exception of just two cases, the “motor proteins” family in the “P-loop-containing nucleotide triphosphate hydrolases” SCOP superfamily and the “SAICAR synthase-like” SCOP superfamily (S1F8 and S25, respectively, in Table 3), all of the β -topology groups in the same SCOP superfamily (and domain, if applicable) are connected within three kinship 1 distances to form a cluster. This verifies the assumption that each global fold evolves by adding or deleting one β -strand at the edge of an existing β -sheet. The above two exceptions are probably due to the lack of known structures that would connect them to the other members within three kinship 1 distances, or due to incorrect classification of evolutionarily unrelated proteins into a SCOP superfamily.

Chemical reactions catalyzed by ATP-binding domains

To study the relationship between the global folds and the molecular functions, we introduce five coarse-grained chemical categories of ATP-binding proteins (given in Table 4). The reactions are first classified into two types: phosphoryl-transfer reactions on the γ -phosphorus atom, and nucleotidyl-transfer reactions on the α -phosphorus atom. Both of the reactions are further classified into chemi-

Figure 3. (Continued on next page)

cal energy-retaining and -releasing reactions. Chemical energy-retaining reactions produce high-energy compounds, such as acyl phosphates, as a product or reaction intermediate, whereas chemical energy-releasing ones proceed by ATP hydrolysis or produce low-energy phosphate compounds (Walsh 1979; Voet et al. 1999; Berg et al. 2001). We included a fifth category, "other," for the various reactions that occur on atoms other than the α - and γ -phosphorus atoms and for ATP binding as an effector.

Identifying fold-based superfamilies

The distribution of ATP-binding domains was examined by assigning the category and the topological position of binding site to each domain (shown in Fig. 3 by a color-coded distribution). If a clear cluster with common chemical categories was identified on this network, then we define it as a fold-based superfamily. We operationally define each cluster connected within three kinship 1 distances (cutoff kinship distance = 4) as a fold-based superfamily. Furthermore, the mode of ATP binding to the fold is assumed to be conserved in each fold-based superfamily, at least at the level of the topological position of the binding site with

respect to the plane of the β -sheet, for example, whether it is on one particular face or edge of the sheet out of the two possible faces or edges.

This definition is based on the assumption that no pair of proteins of the same evolutionary origin catalyzes reactions belonging to different categories. Although we think that this is a reasonable assumption, it remains as an assumption in this article and needs to be critically examined in the future. Todd et al. (2001) found that a few number of proteins changed their reaction chemistry during evolution. However, no ATP-binding domains collected in the present study changed the chemical category during evolution. Chemical category is definable at different level of granularity. Reaction chemistry in the study by Todd et al. (2001) was defined at the level of EC-number classification, but core chemical function in our present study is more roughly defined at the level of superfamily classification.

The fold-based superfamilies thus obtained depend on the used cutoff kinship distance. When we tried to form clusters with different cutoff kinship distances, from three to six, a few different clusters appeared, depending on the cutoff kinship distances. Almost the same picture for the evolution of ATP-binding domains emerged. In the following, we

Table 4. Chemical categories and fold-based superfamilies of ATP-binding domains

Chemical Categories	Names of fold-based superfamilies (SCOP superfamily, family, domain, and their β -topology groups) ^a
Category 1: phosphoryl transfer energy-retaining	<ol style="list-style-type: none"> Adenylate kinase fold-based superfamily S1F1, S1F6, S1F9, S30 (P23145, P231456, M231456, M32415678, M34251678, M43251678, M432516789) Glutathione synthetase fold-based superfamily S12Domain2, S14, S15, S25F1Domain2, S31Domain2 (A2134, A2314, A32145, A32145, M241536, A5412367, M2314765, A23451687, A54123687) Phosphoglycerate kinase fold-based superfamily S5 (P321456) Phosphoenolpyruvate carboxykinase fold-based superfamily S17 (M312465, M14568723) Pyruvate kinase β-barrel domain fold-based superfamily S24 (P123456, P12345678) Carbamate kinase fold-based superfamily S27 (M34215786) Histidine kinase fold-based superfamily S12Domain1, S29F3 (A1423, A12534, M12354) <p>Other isolated domains S25F1Domain1, S31Domain1 (A1234, M1423)</p>
Category 2: phosphoryl transfer energy-releasing	<ol style="list-style-type: none"> Phosphosugar hydroxykinase I fold-based superfamily S1F2, S1F3, S1F5, S1F10, S1F11, S1F12, S1F13, S6Domain2 (P2314, P13245, P23145, P23415, M32145, P165243, P324516, M234156, P1435627, M1762534, M3214567, M3245167, M14356278, M23145687, M324516789) Phosphosugar hydroxykinase II fold-based superfamily S4F1, S6Domain1, S8F1 (P2314, P321456, M321456789) GTPase I fold-based superfamily S1F7 (M231456, M2314576, M2341567, M21453678) Myosin fold-based superfamily S1F8 (M1273645, M1726534, M21837645) GTPase II fold-based superfamily S1F9:GTPase domain of Ffh, FtsY (M32415678) Thiazole kinase fold-based superfamily S8F2 (M213456789) Protein kinase fold-based superfamily S9, S25F2 (A12345, A12354, A123465, A156432) Exonuclease fold-based superfamily S16, S18, (M32145, M143256, M432156) HSP90 chaperone fold-based superfamily S29F1, S29F2 (M16547328, M54362178)
Category 3: nucleotidyl transfer energy-retaining	<ol style="list-style-type: none"> Class I aminoacyl-tRNA synthetase fold-based superfamily S2, S3F1, (P3214, P32145, P321564, M165234) Class II aminoacyl-tRNA synthetase fold-based superfamily S10, S23Domain1 (A1234, A12345, M1237654, M12387645, M12387654, M156784329, M546783219) <p>Other isolated domain S23Domain2 (M12345, M214356789)</p>
Category 4: nucleotidyl transfer energy-releasing	<ol style="list-style-type: none"> Nucleotidyl polymerase fold-based superfamily S13Domain2, S20, S21 (A2134, M321456) Mononucleotidyl transferase fold-based superfamily S19 (M12543, M34215) RNA 3'-terminal phosphate cyclase fold-based superfamily S32F1 (M1243) <p>Other isolated domain S13Domain1 (M15423)</p>
Category 5: others	<ol style="list-style-type: none"> PAPS sulfotransferase fold-based superfamily S1F4 (P2134, P23145) PAPS reductase fold-based superfamily S3F2 (M321465) Oligosaccharide phosphorylase fold-based superfamily S4F2 (P321456, M345216789) SAM Methyltransferase fold-based superfamily S7 (M213465, M3214576, M321457689, M75643892110) Synapsin fold-based superfamily S12F3 (A1423, A32145)

(continued)

Table 4. Continued

Chemical Categories	Names of fold-based superfamilies (SCOP superfamily, family, domain, and their β -topology groups) ^a
	27. SAM synthetase fold-based superfamily S22 (A2134)
	28. PH domain-like fold-based superfamily S26 (A1234, A432165, A4321765, A43217658, M43217658)
	29. Hydroxymethyl dihydropterin pyrophosphokinase fold-based superfamily S28 (A41325)
	Other isolated domains S11 (A23145)

^a Fold-based families are arranged in the order of chemical category, then SCOP superfamily (S) and family (F) names that are referred to Table 3. Domain1 and Domain2 are a pair of two domains that constitute an ATP-binding site. When one SCOP family contains enzymes that catalyze reactions of different chemical categories, the respective protein names are given in each chemical category.

involve information contained in conventional fold representations based on atom positions), then we would be able to identify the fold-based superfamilies as mutually disjointed clear clusters, without resorting to a classification of functions. If separation into two clusters is really impossible in any fold space, then it means that the two should in fact be regarded as defining one cluster and, therefore, exist in one fold-based superfamily with reactions of both categories. Our research has not yet attained this horizon. At the level of using the coarse-grained fold space of the network of kinship relations, the categorization of reactions is a very powerful method to complement its coarse-grained nature.

New findings in the fold-based superfamilies

The glutathione synthetase fold-based superfamily 2 in Table 4, contains two SCOP single-domain superfamilies (S14 and S15) and three domains that are one CATH domain of the SCOP double-domain superfamilies S12Domain2, S25F1Domain2, and S31Domain2. Each of these SCOP superfamilies is characterized by a reaction that proceeds via a carboxyphosphate or phosphohistidine intermediate, by an energy-retaining phosphoryl-transfer mechanism. One isolated domain, S25F1Domain1, and the histidine kinase fold-based superfamily 7 are in the same chemical reaction category and within a cutoff distance of four from each other. However, these two are treated as independent fold-based superfamilies, because ATP binds to the β -sheet differently.

The finding that domains related to the five SCOP superfamilies are in fact members of one fold-based superfamily leads to new insight into the evolution of enzymes in the de novo purine biosynthesis. Purine is synthesized in microorganisms by a pathway involving 10 successive enzymes (Buchanan 1973): PurF, PurD, PurT, PurL, PurM, PurK, PurE, PurC, PurB, and PurH, in which the six underlined enzymes use ATP as a substrate. The three-dimensional structures have been determined for all of these ATP-using enzymes, except PurL. A PSI-BLAST sequence analysis of

PurL revealed PurM as its sequence homolog. The structures of PurD (Wang et al. 1998), PurK (Thoden et al. 1999), PurT (Thoden et al. 2000), PurC (Levdikov et al. 1998), and PurM (and by inference, also PurL; Li et al. 1999) are distributed in three different SCOP double-domain superfamilies (S12, S25F1, and S31 in Table 3). The ATP-binding cleft in each enzyme is formed by the two domains. One domain in each of these SCOP double-domain superfamilies was found in our analysis to belong to one common glutathione synthetase fold-based superfamily. The β -strands in these common domains are shown in red in Figure 4. The fact that these domains have a common evolutionary origin implies that an ancestral form of this domain alone was capable of catalyzing the chemical reaction of the same category. In fact, a single-domain enzyme, glutamine synthetase, which belongs to the SCOP single-domain superfamily S14 in Table 3 and catalyzes the chemical reaction of the same category (Gill and Eisenberg 2001), is a member of our glutathione synthetase fold-based superfamily. Our present analysis indicates that all the six of the ATP-using enzymes in de novo purine biosynthesis have evolved from an ancestral glutamine synthetase as the common evolutionary origin, by combining with another domain to improve their catalytic specificities.

This is the first evidence that as many as six enzymes catalyzing the reactions in one metabolic pathway are evolutionarily related from each other. Previous structure and function analyses of other enzymes indicated that they could evolve from an ancestral enzyme by retaining its core chemical reaction and adjusting its substrate specificity flexibly (Petsko et al. 1993; Babbitt and Gerlt 1997; Gerlt and Babbitt 1998; Hasson et al. 1998). However, these cases have mostly applied to only two or three enzymes in one metabolic pathway (Farber and Petsko 1990; Fani et al. 1995).

The green-colored fold-based superfamily 19 in Table 4, nucleotidyl polymerase fold-based superfamily, includes two SCOP single-domain superfamilies, S20 and S21 in Table 3, and one domain, S13Domain2 in Table 3, from one

Figure 4. Five de novo purine biosynthesis enzymes and glutamine synthetase, all in the glutathione synthetase fold-based superfamily. The ATP-binding cleft in each de novo purine biosynthesis enzyme is formed by two domains: domain 1 on the *right* and domain 2 on the *left*. Only domain 2, which is shown by β -strands in red, is a member of the glutathione synthetase fold-based superfamily. The ATP-binding cleft in glutamine synthetase is formed by a single domain. Their PDB entry codes and SCOP classification codes (Table 3) are as follows: PurT (1ez1, S12F2), PurK (1b6s, S12F2), PurD (1gso, S12F2), PurM (1cli, S31), PurC (1a48, S25F1), and glutamine synthetase (1lgr, S14). Structures were drawn by MOLSCRIPT (Kraulis 1991).

SCOP double-domain superfamily. This fold-based superfamily includes all of the enzymes with known structures involved in nucleotidyl polymerization, transcription and capping of polynucleotides, and cyclization of mononucleotides. DNA polymerase β (S19 in Table 3), which is incorrectly named as a polymerase but is actually a repair enzyme (Sawaya et al. 1997; Arndt et al 2001), is clustered into a separate fold-based superfamily.

Phosphorylation of sugars is catalyzed by kinases that have been classified into four SCOP superfamilies (S1F3, F5, F12, F13; S4F1; S6Domain1; S8F1 in Table 3). The domains in these superfamilies are now found to belong to two different fold-based superfamilies, superfamilies 8 and 9 in Table 4. These superfamilies are treated here as separate families, because of the different topological positions of the binding sites.

SCOP superfamily P-loop-containing nucleotide triphosphate hydrolases (S1 in Table 3) now appears to be a mixture of six evolutionarily unrelated proteins: adenylate kinase 1, phosphosugar hydroxykinase I 8, GTPase I 10, GTPase II 12, myosin 11, and PAPS sulfotransferase 22 fold-based superfamilies (Table 4). On the other hand, the adenylate kinase fold-based superfamily 1 in Table 4 includes S1F1, S1F6, S1F9, and S30 in Table 3, among which S30 was somehow not assigned to be a member of S1 in SCOP, even though they share the P-loop motif (Saraste et al. 1990; Bertrand et al. 1997).

Some of the interesting new findings are described above. Descriptions of the 29-fold-based superfamilies are provided given in Table 4, which contains numerous new findings not described above.

Implications for the evolution of proteins

An important aspect of understanding biological systems at the molecular level is not only to know the structure and function of enzymes but also to be able to relate enzymes to each other, from an evolutionary viewpoint. In this work, we found that there are $\sim \geq 20$ independently invented ATP-binding protein families, each with its specific core chemical reaction. During the course of evolution, each fold-based superfamily diverged somewhat in both the fold-space and the function-space, as summarized in Figure 3 and Table 4. Even during the course of divergence, the specific core chemical reaction was maintained. Because the use of ATP should always be necessary for the core chemical reaction, it leads us to realize that the mechanisms of ATP-binding and ATP-usage in the core chemical reaction are both created at the conception of a fold-based superfamily.

By closely examining the process of divergence, we should be able to establish a phylogeny of proteins in each fold-based superfamily. Through such a process, we could reexamine the currently popular, but often very confusing, concepts of family and superfamily.

Materials and methods

The CATH S-reps data set (June 1999), which is the set of non-homologous representatives with $<35\%$ sequence identities, consists of 1173 domains that were members of the mainly- β and the α - β classes. We selected 1011 domains by removing those that were low in resolution or determined by NMR. The topology of each β -sheet assigned by TOPS (Flores et al. 1994) was examined by inspecting the structures on a graphic display. Topologies of

barrels and sandwich layers were assigned by assuming that the corresponding flat β -sheet is first formed and then bent. When one β -sheet is composed of strands from two or more different chains, we separated it into β -sheets that are composed of the strands from the same peptide and assigned the topology for each sheet. Among the 1011 domains, 77% consisted of two or more β -sheets. When the largest β -sheet in each domain differed from the remaining ones by only one β -strand, the same domain was analyzed multiple times from the viewpoints of the individual β -sheets. Because of this treatment, our set of 1011 domains now appears to consist of 1444 domains (Supplementary Table 1).

The 32 SCOP superfamilies contain 83 SCOP families and 182 proteins (Supplementary Table 3). In the cases in which a protein had two or more PDB structures derived from different biological species, we selected one PDB structure from each different species. We selected 254 PDB structures from 182 proteins. When the ATP-binding site was composed of two domains, two ATP-binding domains were selected from one PDB structure. Thus, we obtained a total of 263 ATP- and 22 GTP-binding domains from 254 PDB structures. The identification of the ATP-binding domains is confirmed by the references cited in the PDB database and by the distribution of the residues interacting with the bound ATP or its analog. By LIGPLOT (Wallace et al. 1995), we further confirmed that such interacting residues are mainly located on the β -strands of the largest β -sheet in the ATP-binding domain or on the loops connecting the β -strands.

Acknowledgments

We thank Dr. K. Yura for helpful discussions. We also thank Dr. G. Basu for reading the manuscript carefully. This work was partly supported by the Special Coordination Funds Promoting Science and Technology from the MEXT (Ministry of Education, Culture, Sports, Science, and Technology, Japan).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Arndt, J.W., Gong, W., Zhong, X., Showalter, A.K., Liu, J., Dunlap, C.A., Lin, Z., Paxson, C., Tsai, M.D., and Chan, M.K. 2001. Insight into the catalytic mechanism of DNA polymerase beta: structures of intermediate complex. *Biochemistry* **40**: 5368–5375.
- Babbitt, P.C. and Gerlt, J.A. 1997. Understanding enzyme superfamilies: Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**: 30591–30594.
- Berg, J.M., Tymoczko, J.L., and Stryer, L. 2001. *Biochemistry*. W.H. Freeman and Co., New York.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- Bertrand, J.A., Auger, G., Fanchon, E., Martin, L., Blanot, D., van Heijenoort, J., and Dideberg, O. 1997. Crystal structure of UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli*. *EMBO J.* **16**: 3416–3425.
- Buchanan, J.M. 1973. The amidotransferases. *Adv. Enzymol. Relat. Areas Mol. Biol.* **39**: 91–183.
- Fani, R., Lio, P., and Lazcano, A. 1995. Molecular evolution of the histidine biosynthetic pathway. *J. Mol. Evol.* **41**: 760–774.
- Farber, G.K. and Petsko, G.A. 1990. The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* **15**: 228–234.
- Fetrow, J.S. and Skolnick, J. 1998. 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.* **281**: 949–968.
- Flores, T.P., Moss, D.S., and Thornton, J.M. 1994. An algorithm for automatically generating protein topology cartoons. *Protein Eng.* **7**: 31–37.
- Fukami-Kobayashi, K., Tateno, Y., and Nishikawa, K. 1999. Domain dislocation: A change of core structure in periplasmic binding proteins in their evolutionary history. *J. Mol. Biol.* **286**: 279–290.
- Gerlt, J.A. and Babbitt, P.C. 1998. Mechanistically diverse enzyme superfamilies: The importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **2**: 607–612.
- Gill, H.S. and Eisenberg, D. 2001. The crystal structure of phosphothricin in the active site of glutamine synthetase illuminates the mechanism of enzymatic inhibition. *Biochemistry* **40**: 1903–1912.
- Goto, S., Nishioka, T., and Kanehisa, M. 2000. LIGAND: Chemical database of enzyme reactions. *Nucleic Acids Res.* **28**: 380–382.
- Hasson, M.S., Schlichting, I., Moulai, L., Taylor, K., Barrett, W., Kenyon, G.L., Babbitt, P.C., Gerlt, J.A., Petsko, G.A., and Ringe, D. 1998. Evolution of an enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proc. Natl. Acad. Sci.* **95**: 10396–10401.
- Hubbard, T.P.J., Murzin, A.G., Brenner, S.E., and Chothia, C. 1997. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **25**: 236–239.
- Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**: 946–950.
- Levdikov, V.M., Barynin, V.V., Grebenko, A.I., Melik-Adamyan, W.R., Lamin, V.S., and Wilson, K.S. 1998. The structure of SAICAR synthase: An enzyme in the de novo pathway of purine nucleotide biosynthesis. *Structure* **6**: 363–376.
- Li, C., Kappock, T.J., Stubbe, J.A., Weaver, T.M., and Ealick, S.E. 1999. X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the *Escherichia coli* purine biosynthetic pathway at 2.5 Å resolution. *Structure* **7**: 1155–1166.
- Lo Conte, L., Ailey, B., Hubbard, T.P.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **28**: 257–259.
- Orengo, C.A., Jones, D.T., and Thornton, J.M. 1994. Protein superfamilies and domain superfolds. *Nature* **372**: 631–634.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
- Petsko, G.A., Kenyon, G.L., Gerlt, J.A., Ringe, D., and Kozarich, J. W. 1993. On the origin of enzymatic species. *Trends Biochem. Sci.* **18**: 372–376.
- Pitts, O.B. and Finkelstein, A.V. 1980. Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophysics* **13**: 339–386.
- Richardson, J.S. 1977. β -Sheet topology and the relatedness of proteins. *Nature* **268**: 495–500.
- Saraste, M., Sibbald, P.R., and Wittinghofer, A. 1990. The P-loop: A common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**: 430–434.
- Sawaya, M.R., Prasad, R., Wilson, S.H., Kraut, J., and Pelletier, H. 1997. Crystal structures of human DNA polymerase β complexed with gapped and nicked DNA: Evidence for an induced fit mechanism. *Biochemistry* **36**: 11205–11215.
- Thoden, J.B., Kappock, T.J., Stubbe, J., and Holden, H.M. 1999. Three-dimensional structure of N5-carboxyaminoimidazole ribonucleotide synthetase: A member of the ATP grasp protein superfamily. *Biochemistry* **38**: 15480–15492.
- Thoden, J.B., Firestone, S., Nixon, A., Benkovic, S.J., and Holden, H.M. 2000. Molecular structure of *Escherichia coli* purt-encoded glycinamide ribonucleotide transformylase. *Biochemistry* **39**: 8791–8802.
- Todd, A.E., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**: 1113–1143.
- Voet, D., Voet, J.G., and Pratt, C.W. 1999. *Fundamentals of biochemistry*. John Wiley & Sons, New York.
- Wallace, A.C., Laskowski, R.A., and Thornton, J.M. 1995. LIGPLOT: A program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.* **8**: 127–134.
- . 1996. Derivation of three-dimensional coordinate templates for searching structural databases. *Protein Sci.* **5**: 1001–1013.
- Walsh, C. 1979. *Enzymatic reaction mechanisms*. W.H. Freeman and Co., San Francisco.
- Wang, W., Kappock, T.J., Stubbe, J., and Ealick, S.E. 1998. X-ray crystal structure of glycinamide ribonucleotide synthetase from *Escherichia coli*. *Biochemistry* **37**: 15647–15652.