Chapter 1

55 Years of the Rossmann Fold

Woong-Hee Shin and Daisuke Kihara

Abstract

The Rossmann fold is one of the most commonly observed structural domains in proteins. The fold is composed of consecutive alternating β-strands and α-helices that form a layer of β-sheet with one (or two) layer(s) of α-helices. Here, we will discuss the Rossmann fold starting from its discovery 55 years ago, then overview entries of the fold in the major protein classification databases, SCOP and CATH, as well as the number of the occurrences of the fold in genomes. We also discuss the Rossmann fold as an interesting target of protein engineering as the site-directed mutagenesis of the fold can alter the ligand-binding specificity of the structure.

Key words Protein fold, Rossmann fold, Protein structure classification, Nucleotide-binding fold, CATH, SCOP, Protein engineering

1 Introduction

The Rossmann fold was originally called the nucleotide-binding fold by Michael Rossmann, who was first to discover it. It was later found to be present in many proteins of various functions and across many organisms. In this article, we review the first structures of this fold found 55 years ago and then examine how the population of this fold has grown in protein structure databases over the last 55 years. We also estimate the number of the Rossmann folds in various genomes through a database of homology models of proteins. At last, we discuss the Rossmann fold as an interesting protein design target.

2 The First Discovery of the Rossmann Fold

The Rossmann fold is the only domain structure that was named after its discoverer in the protein structure classification databases, CATH [1] and SCOP [2]. Dr. Michael Rossmann is one of the pioneers of structural biology. Besides the fold we will discuss here,

Alexander E. Kister (ed.), *Protein Supersecondary Structures: Methods and Protocols*, Methods in Molecular Biology, vol. 1958, https://doi.org/10.1007/978-1-4939-9161-7_1,

Dr. Rossmann is also well-known as the developer of the commonly used technique molecular replacement, which is used to solve the phase problem in X-ray crystallography. He has also solved many structures including hemoglobin and virus capsids. He started his research on crystallography of small molecules when he was a graduate student at the University of Glasgow, Scotland. Then he solved the structure of hemoglobin with Dr. Max Perutz at the University of Cambridge [3]. After he joined Purdue University in 1964 as a faculty member, he solved the structure of lactate dehydrogenase [4] in 1970, which would later be known to contain a Rossmann fold. This was the first protein structure that he solved at Purdue. A picture of the protein wire model he built at that time is shown in Fig. 1a. Figure 1b is the structure of lactate dehydrogenase in the PDB database (PDB ID: 3LDH). After the structure of lactate dehydrogenase was solved, a series of dehydrogenase structures were solved and published, including apo structures of lactate dehydrogenase [5] and D-glyceraldehyde-3-phosphate dehydrogenase [6]. From the series of structures, Dr. S.T. Rao and Rossmann identified a common set of the secondary structure elements in dinucleotide-binding proteins, which was the discovery of the Rossmann fold [7]. They compared crystal structures of a lactate dehydrogenase [5], two flavodoxin structures (one from Dr. Martha Ludwig of University of Michigan and the other from Dr. Lyle Jensen of University of Washington), subtilisin (PDB ID: 1SBT), and malate dehydrogenase (from Dr. Leonard J. Banaszak of Washington University of St. Louis) by minimizing the position difference of Cα-atoms after superimposition of the structures. They found a common structural motif that is composed of alternating three β-strands and two α-helices, i.e., a βαβαβ structure. In a lactate dehydrogenase structure, 60 residues are involved in this super-secondary structure. The three strands are paired in parallel, connected by helices. The loops that connect the C-terminus of β-strand and the N-terminus of α-helix construct a hydrophobic pocket, specifically formed by Val27, which is at the loop between the first α-helix and the first β-strand, and Val54, which is on the loop between the second α-helix and the second β-strand. These two valines interact with the adenine ring of nicotinamide adenine dinucleotide (NAD) or other substrates (Fig. 1c).

Currently in the SCOP database [2], the Rossmann fold is defined as a repeat of the βαβαβ structure, which has an approximate twofold symmetry. The six β-strands form a sheet in an order of 321456 counted from the N-terminus to the C-terminus (Fig. 2). The repeated fold binds to a dinucleotide, a coenzyme that contains two distinct nucleotides, such as NAD or flavin adenine dinucleotide (FAD). In 1982, Schulz, Schirmer, and Pai found that a tight loop that connects the first β-strand and α-helix has a conserved sequence GXGXXG in FAD-binding domains [8] by

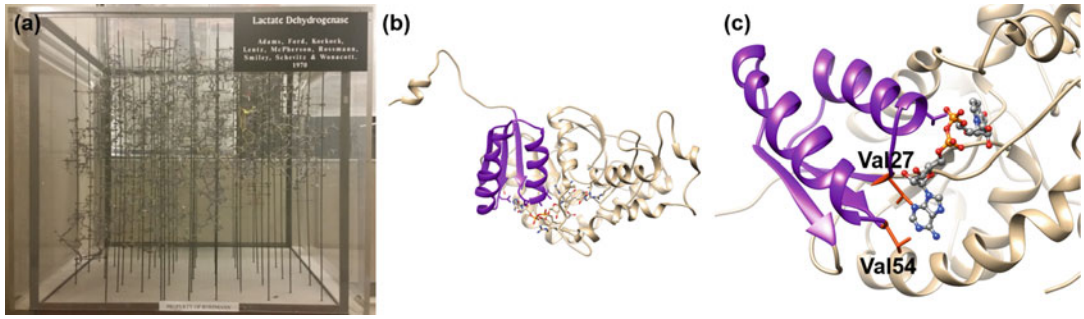

Fig. 1 The tertiary structure of lactate dehydrogenase. (**a**) A physical wire model, which is exhibited in the Hockmeyer Hall of Purdue University. (**b**) The structure in PDB with the oldest deposited date (PDB ID: 3LDH, deposited in 1977). The purple region is the βαβαβ structure that is identified by Rao and Rossmann. (**c**) The NAD-binding site (3LDH). Val27 and Val54, which form a hydrophobic pocket, are colored in red. NAD is shown in a ball-and-stick representation

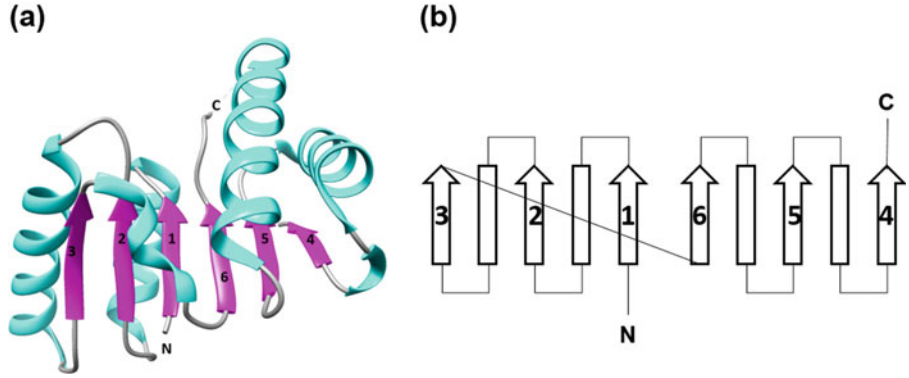

Fig. 2 The Rossmann fold definition in the SCOP database. (**a**) An example of the Rossmann fold (PDB ID: 2D4A). Helices are colored in cyan and strands are in magenta. (**b**) The two-dimensional model of the Rossmann fold. Arrows and squares are β-strands and α-helices, respectively. The numbering on the strands starts from the N-terminus of the protein

comparing three homologous protein sequences (pig heart lipoamide dehydrogenase, *p*-hydroxybenzoate hydroxylase, and D-amino acid oxidase). The loop with the consensus sequence makes a contact with negatively charged oxygens of two phosphate groups in both NAD-binding protein and FAD-binding proteins (Fig. 3) [9]. Israel Hanukoglu found that in nicotinamide adenine dinucleotide phosphate (NADP)-binding crystal structure, the last glycine of the loop is mutated to alanine [10]. It was revealed that the position determines binding coenzyme specificity by Purham and his colleagues [11].

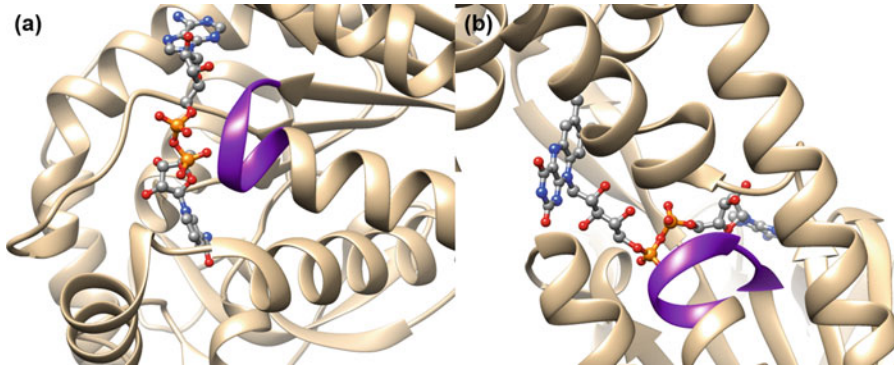

Fig. 3 Crystal structures of Rossmann fold proteins binding with cofactors. (**a**) NAD (PDB ID: 1I0Z); (**b**) FAD (PDB ID: 3GRS). The conserved GXGXXG motif is colored in purple, and the cofactors are represented in a ball-and-stick model

3 Database Entries of Structures of the Fold

Since the first lactate dehydrogenase solved in 1970, a number of Rossmann fold structures have been determined. We examined Rossmann fold entries in the SCOPe (a successor of SCOP) [12] and CATH [1] databases. The Rossmann fold is classified as one of the folds in SCOPe while as a topology in CATH.

SCOP classifies protein domain structures in a hierarchical level, class, fold, superfamily, and family. The latter two are sequence-based classifications. In SCOPe, Rossmann fold, named as "NAD(P)-binding Rossmann-fold domains," is classified as one of a fold (c.2) under "α and β proteins (α/β)" class, class c. The fold (c.2.1) has only one superfamily, with the same name as the fold. The condition for classifying domains as Rossmann fold is "three layers, α/β/α; parallel beta-sheet of six strands, order 321456." The Rossmann fold is the fourth largest population in SCOPe in the fold level (5985 domains out of 246,157 domains, 2.4%) following immunoglobulin-like β-sandwich (b.1, 7.2%), TIM β/α-barrel (c.1, 4.5%), and N-terminal hydrolase-like (d.153, 2.6%). In the superfamily level, it is the third (2.4%) following immunoglobulin (b.1.1, 5.7%) and N-terminal nucleophile hydrolases (d.153.1, 2.6%) (classes of low-resolution protein structures, peptides, designed proteins, and artifacts were not considered in these statistics).

The Rossmann superfamily is composed of 13 families (proteins in the same family share sequence identity >30% or a lower sequence identity but perform the same function). Structures of 12 families are shown in Fig. 4a, except for c.2.1.0, classified as "not a true family" whose structures are classified as a Rossmann fold superfamily, but cannot be categorized to any of families by

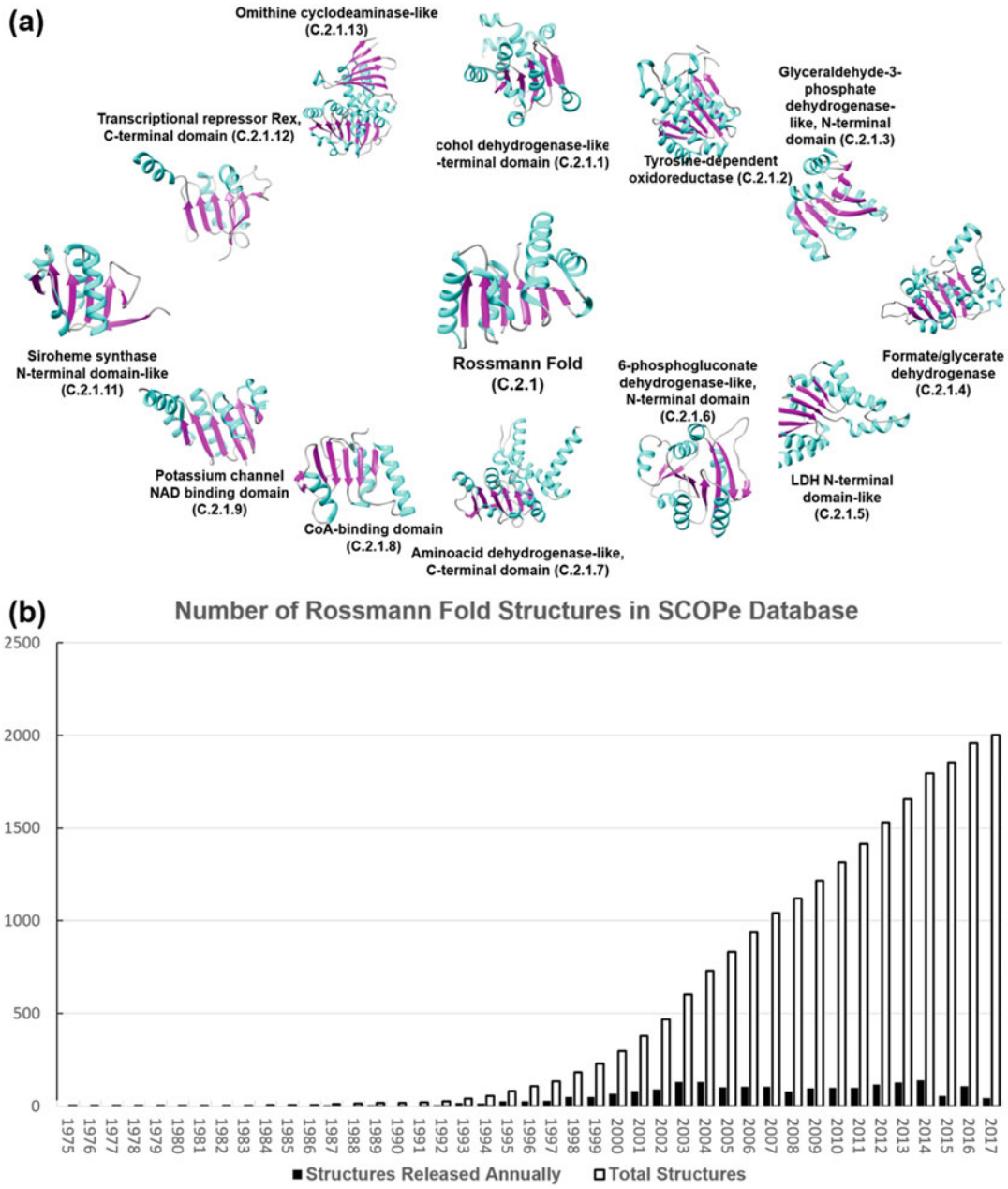

Fig. 4 The Rossmann fold in the SCOPe database. (**a**) Structures of 12 families of the Rossmann fold superfamily in SCOPe. Helices are colored in cyan, and strands are shown in magenta. (**b**) The growth of PDB structures of the Rossmann fold in the SCOPe database. Black bars are the number of PDB structures released in that year, and white bars are the total number of the structures in the database

sequence identity or protein function. Although all families share 321456 β-sheet as a core structure, there is a small difference in each family. For example, family c.2.1.2, tyrosine-dependent oxidoreductases, has an extra parallel seventh strand, forming a β-sheet order of 3214567. On the other hand, for family c.2.1.6, 6-phosphogluconate dehydrogenase-like, N-terminal domain, β-sheet is extended to eight strands with an order of 32145678, and the extra strands are antiparallel to the core six strands. The growth of PDB structures of the Rossmann fold in SCOPe released every year is plotted in Fig. 4b. As of the SCOPe version 2.07 (2018-09-06), the Rossmann fold superfamily has 2002 PDB structures.

In the SCOPe database, there are six folds that are classified as "Rossmann-like folds." They are nucleotide-binding domain (c.4), MurCD N-terminal domain (c.5), nucleoside phosphorylase/phosphoribosyltransferase catalytic domain (c.27), cytochrome/phosphorylase N-terminal domain (c.28), DHS-like NAD-/FAD-binding domain (c.31), and GckA/TtuD-like (c.118). The difference between the Rossmann fold and the Rossmann-like fold is the number of β-strands that form the β-sheet. The Rossmann-like folds have five parallel strands with an order of 32145, a combination of βαβαβ structure and βαβ structure. Just as in the Rossmann fold, the Rossmann-like fold contains the tight loop that connects the first strand and the first helix of βαβαβ motif interacts with negative oxygens of phosphate of a ligand molecule. Their structures are shown in Fig. 5a. As of the SCOPe version 2.07 (2018-09-06), 347 PDB structures belong to the six Rossmann-like fold superfamilies. The growth of the structures of the Rossmann-like fold is shown in Fig. 5b.

CATH is another protein structure classification database [1]. CATH classifies structures into a slightly different hierarchy than SCOP, which has four main levels: Class (C), which concerns the secondary structure content, Architecture (A), which concerns the spatial arrangement of the secondary structure elements, Topology (T), which corresponds to the fold level in SCOP, and Homology (H), a sequence identity-based classification where proteins need to have over 25% sequence identity to be in the same H classification. C, T, and H levels are automatically assigned by a protein structure comparison algorithm by Michie et al. [13] and the SSAP program [14]. Architecture is manually assigned after C, T, and H levels are determined. In CATH, the Rossmann fold is classified in the αβ class (labeled as 3), the architecture class of 40, which is three-layer (αβα) sandwich, and the topology class of 50; thus, the classification ID is 3.40.50. Comparing with CATH and SCOPe classification of Rossmann fold, the Rossmann fold topology in CATH includes not only the Rossmann fold in SCOPe, but also the Rossmann-like folds. CATH classifies the domain to 3.40.50 if the βαβαβ motif is included in the protein

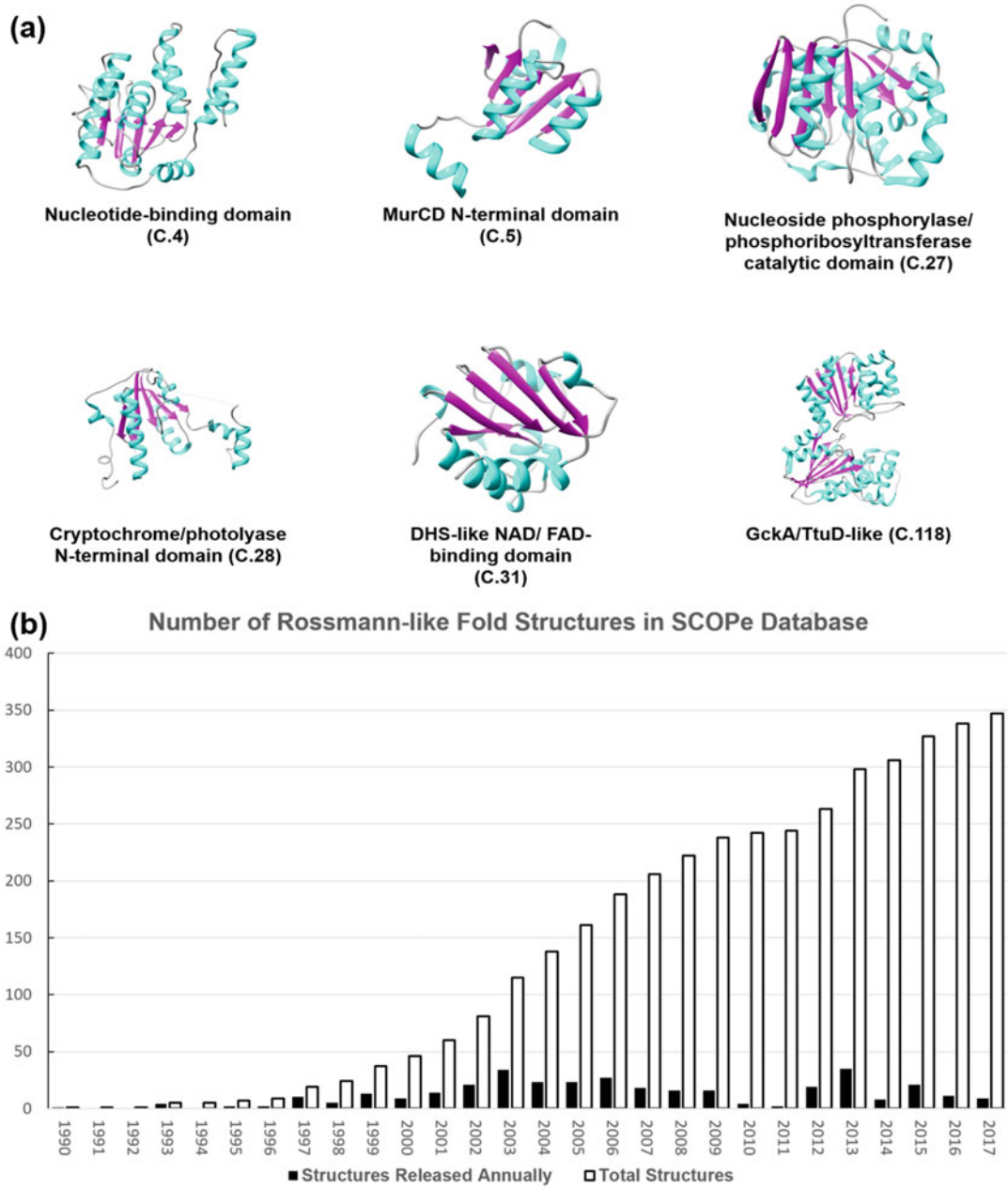

Fig. 5 Rossmann-like superfamilies in SCOPe. (**a**) The tertiary structures of six Rossmann-like superfamilies in SCOPe. (**b**) The growth of Rossmann-like fold structures in SCOPe

domain structure not restricting it to structures with a twofold symmetry, as shown in the center of Fig. 6a.

In the topology level, the Rossmann fold has the largest number of domains (52,880 domains out of 434,857 domains, 12.2%). The Rossmann fold topology is composed of 232 homologous

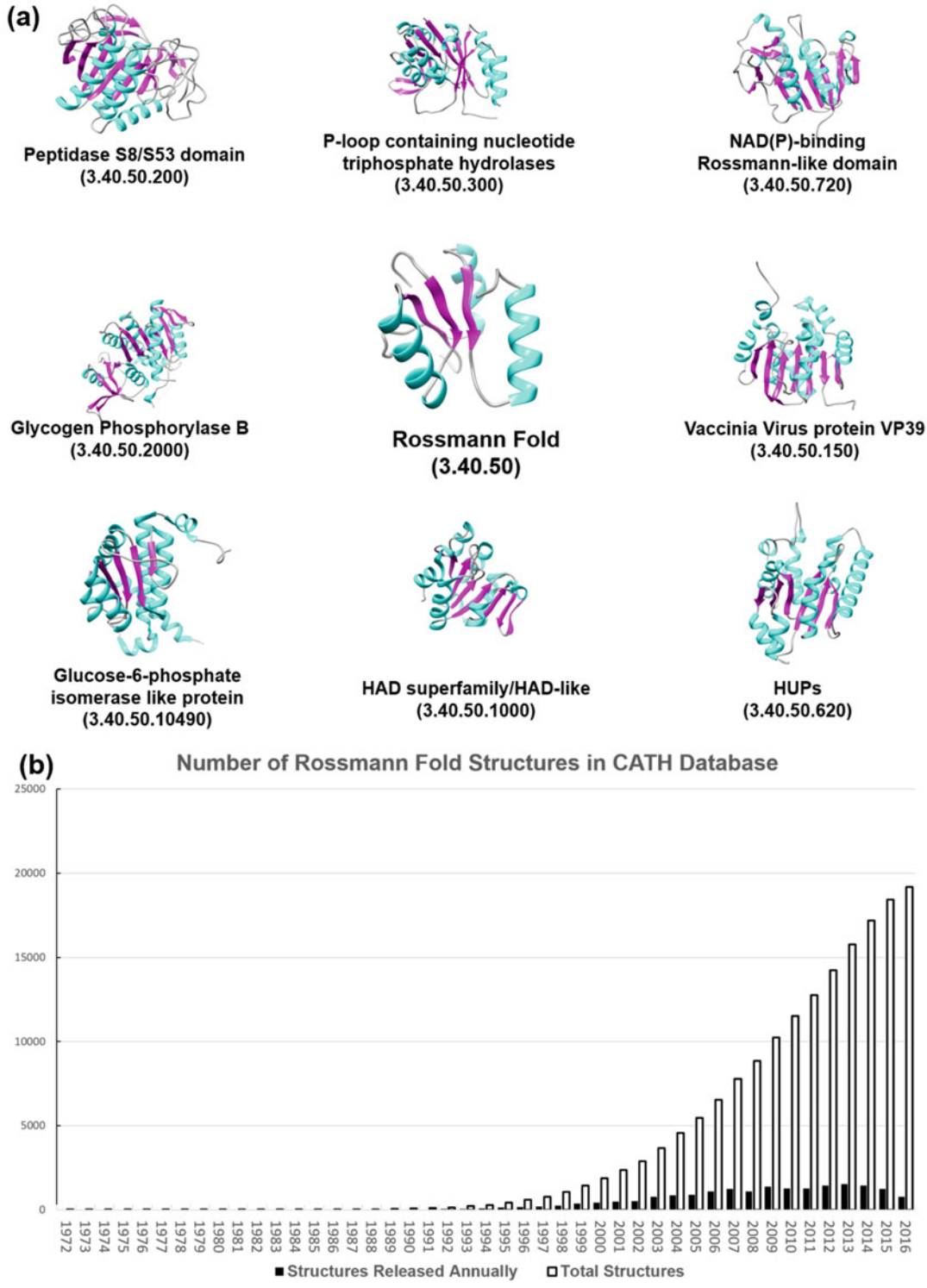

Fig. 6 The Rossmann fold in CATH. (**a**) Structures that belong to the Rossmann fold topology. (**b**) The growth of Rossmann fold structures in the CATH database

superfamilies (H level). The eight largest superfamilies of the Rossmann fold are illustrated in Fig. 6a. The main difference across the homologous superfamilies is the number of β-strands forming the β-sheet and their directions. For example, Vaccinia virus protein (3.40.50.150) VP39 has seven β-strands with one strand antiparallel, whereas glucose-6-phosphate isomerase-like protein (3.40.50.10490) has five parallel β-strands. As of the latest update of the CATH database (2018-09-06), the number of PDB entries that belong to the Rossmann fold (3.40.50) has reached 19,184. Figure 6b shows the growth curve.

4 Rossmann Folds in Genomes

The number of Rossmann fold proteins in genomes has been estimated a couple of times by using bioinformatics predictions. Wolf et al. estimated the protein fold distribution on a genomic scale in 1999 [15]. They assigned protein structures to 13 complete genome sequences (*H. influenzae*, *M. genitalium*, *M. pneumoniae*, *Synechocystis* sp., *H. pylori*, *E. coli*, *B. subtilis*, *B. burgdorferi*, *A. aeolicus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. fulgidus*, and *S. cerevisiae* and *C. elegans*), which were available at that time (April 1998) using a sequence database search method, PSI-BLAST [16] ran against SCOP ver. 1.35. A fold was assigned to a protein if the E-value was less than or equal to 10^{-2}. They found that P-loop NTPases are the most abundant fold in all the genomes, and distribution of top 30 folds follows an exponential distribution. The Rossmann fold had a proportion of 3.6% (6th), 3.5% (8th), and 3.2% (9th) in bacterial, archaeal, and eukaryotic genomes, respectively. Although their aim was to assign protein folds to the each protein in the genomes, the percentages of the fold assignment in a genome ranged from 19.2 to 39.0% leaving 60% or more of the proteins unassigned.

Three years later, Gerstein and his colleagues also performed a fold assignment for 20 complete genomes [17]. Six more genomes, *C. pneumoniae*, *C. trachomatis*, *M. tuberculosis*, *P. horikoshii*, *R. prowazekii*, and *T. pallidum*, were added to the 14 genomes which Wolf et al. have analyzed. PSI-BLAST was used for the assignment, which was ran against SCOP ver. 1.39 with an E-value cutoff of 10^{-4}. The assignment coverages ranged from 17.6 to 34.6%. In their assignment, the Rossmann fold did not appear in three genomes, *C. elegans*, *B. burgdorferi*, and *T. pallidum*. On the other hand, in *M. tuberculosis* and *B. subtilis*, the Rossmann fold was the most abundant fold, 16% in *M. tuberculosis* and 14% for *B. subtilis*. In the remaining 15 genomes, the proportion of the Rossmann fold was 4–12%.

In 2004, Kihara and Skolnick predicted protein folds of five complete genomes, *M. genitalium*, *E. coli*., *B. subtilis*, *A. aeolicus*,

Table 1
The number of proteins with a structure model that have the Rossmann fold in five genomes

Genomes	The number of proteins	The number of proteins with a model	The proteins of the Rossmann fold
Homo Sapiens	170,418	140,104	21,809 (15.6%)
Mus musculus	57,825	52,444	7117 (13.6%)
Canis familiaris	37,262	34,447	4915 (14.3%)
Caenorhabditis elegans	27,954	24,011	3587 (14.9%)
Drosophila melanogaster	21,986	19,550	3023 (15.5%)

We counted the number of proteins with a structure model built from a template structure of the Rossmann fold. The percentage is computed relative to the number of proteins with a model

and *S. cerevisiae*, using a template-based structure prediction method (threading), PROSPECTOR_Q [18]. PROSPECTOR_Q searches structures that fit to a query protein sequence in a similar way as sequence database search methods, e.g., PSI-BLAST, but uses structure information, which is predicted secondary structure information and predicted residue-residue contacts, in addition to sequence similarity information from FASTA [19]. Structures in CATH at the topology levels were used as a reference dataset to be searched. The coverage of the fold assignment was higher than the two previous methods, 73–85%. In their results, the Rossmann fold was the most abundant topology in all five genomes, around 18% except for *S. cerevisiae* (12.8%).

In Table 1, we have newly analyzed the number of the Rossmann folds in genomes in ModBase [20], which is a database of annotated homology models of proteins in genomes. In ModBase, structure models were generated by a modeling pipeline called ModPipe, which uses PSI-BLAST to find template structures for modeling and uses the Modeller [21] program to build models. We searched the number of the Rossmann folds in five genomes, *Homo Sapiens*, *M. musculus*, *C. familiaris*, *C. elegans*, and *D. melanogaster*, which have over 10,000 proteins that have structure models. The number of Rossmann fold proteins is summarized in Table 1. The percentage of the Rossmann folds among proteins with models in the five genomes was around 14%, slightly lower than the estimate by Kihara and Skolnick [18].

5 The Rossmann Fold and Protein Engineering

Since the Rossmann fold binds nucleoside cofactors and catalyzes various reactions, it is considered an interesting target for protein

engineering. The focus of this engineering has been to change the specificity of cofactors that bind to enzymes [22].

The first protein engineering performed on the Rossmann fold was by Scrutton et al. [11] in 1990, who mutated the last glycine of the GXGXXG motif of glutathione reductase to alanine so that the engineered protein binds to NADP, but not NAD, the protein's original cofactor of the protein. In 2015 Gerth and her colleagues attempted to modify the behavior of primary-secondary alcohol dehydrogenase to use NADPH as a cofactor in contrast to the original NADH cofactor. Based on observations of the conservation and the difference between structures of NADPH-dependent enzymes and NADH-dependent enzymes, they introduced a quadruple mutant, G198D/S199 V/P201E/Y218A, which significantly reduced the enzymatic activity upon NADPH binding while showing the catalytic activity upon binding of NADH. The observed change of the enzymatic activity was explained by the interactions between side chains and NADPH [23]. Brinkmann-Chen et al. conducted a thorough study on cofactor specificity of ketol-acid reductoisomerase [24]. By comparing a few hundred sequences and seven structures of the enzyme, particularly around the GXGXXG motif, they found "DDV" switch residues that change the cofactor specificity. The native enzyme is activated upon binding NADPH. They found that two mutations to D from S make up for missing phosphates in NAD and substitution to V from I makes a deeper pocket for an adenine fragment, which resulted in a higher catalytic efficiency with NADH than the wild-type enzyme with NADPH.

6 Summary

The Rossmann fold is probably one of the best known folds because of its abundance and the functional divergence of proteins that adopt this fold. Here we briefly reviewed the history, the database entries, and the population of this fold. We also introduced works that used the Rossmann fold as a target of engineering, which lead us to a better understanding of the ways to control compound binding specificity of proteins.

Acknowledgments

The authors are thankful for Lyman Monroe for proofreading the manuscript. This work was partly supported by the National Institute of General Medical Sciences of the NIH (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941).

References

1. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I (2017) CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45(Database issue): D282–D295. https://doi.org/10.1093/nar/gkw1098
2. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247 (7):536–540. https://doi.org/10.1016/S0022-2836(05)80134-2
3. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North AC (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. Nature 185(4711):416–422. https://doi.org/10.1038/185416a0
4. Adams MJ, Ford GC, Koekeok R, Lentz PJ, McPherson A, Rossmann MG, Smiley IE, Schevitz RW, Wonacott AJ (1970) Structure of lactate dehydrogenase at 2.8 Å resolution. Nature 227:1098–1103. https://doi.org/10.1038/2271098a0
5. Rossmann MG, Adams MJ, Buehner M, Ford GC, Hackert ML, Lentz PJ, McPherson A, Schevitz RW, Smiley IE (1972) Structural constraints on possible mechanisms of lactate dehydrogenase as shown by high resolution studies of the apoenzyme and a variety of enzyme complexes. Cold Spring Harb Symp Quant Biol 36:176–191
6. Buehner M, Ford GC, Moras D, Olsen KW, Rossmann MG (1973) D-glyceraldehyde-3-phosphate dehydrogenase: three-dimensional structure and evolutionary significance. Proc Natl Acad Sci U S A 70(11):3052–3054. https://doi.org/10.1073/pnas.70.11.3052
7. Rao ST, Rossmann MG (1973) Comparison of super-secondary structures in proteins. J Mol Biol 76(2):241–256. https://doi.org/10.1016/0022-2836(73)90388-4
8. Schulz GE, Schirmer RH, Pai EF (1982) FAD-binding site of glutathione reductase. J Mol Biol 160(2):287–308. https://doi.org/10.1016/0022-2836(82)90177-2
9. Hanukoglu I (2015) Proteopedia: Rossmann fold: a beta-alpha-beta fold at dinucleotide binding sites. Biochem Mol Biol Educ 43 (3):206–209. https://doi.org/10.1002/bmb.20849
10. Hanukoglu I, Gutfinger T (1989) cDNA sequence of adrenodoxin reductase. Identification of NADP-binding sites in oxidoreductases. Eur J Biochem 180 (2):479–484. https://doi.org/10.1111/j.1432-1033.1989.tb14671.x
11. Scrutton NS, Berry A, Perham RN (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. Nature 343 (6253):38–43. https://doi.org/10.1038/343038a0
12. Fox NK, Brenner SE, Chandonia JM (2014) SCOPe: structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42(Database issue): D304–D309. https://doi.org/10.1093/nar/gkt1240
13. Michie AD, Orengo CA, Thornton JM (1996) Analysis of domain structural class using an automated class assignment protocol. J Mol Biol 262(2):168–185. https://doi.org/10.1006/j,bi.1996.0506
14. Orengo CA, Brown NP, Taylor WR (1992) Fast structure assignment for protein databank searching. Proteins 14(2):139–167. https://doi.org/10.1002/prot.340140203
15. Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res 9 (1):17–26. https://doi.org/10.1101/gr.9.1.17
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389
17. Hegyi H, Lin J, Greenbaum D, Gerstein M (2002) Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. Proteins 47(2):126–141. https://doi.org/10.1002/prot.10078
18. Kihara D, Skolnick J (2004) Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. Proteins 55(2):464–473. https://doi.org/10.1002/prot.20044
19. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988 (85):2444–2448. https://doi.org/10.1073/pnas.85.8.2444
20. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A (2006) MODBASE: a database of annotated comparative proteins

structure models and associated resources. Nucleic Acids Res 34(Database issue): D291–D295. https://doi.org/10.1093/nar/gkj059

21. Webb B, Sali A (2016) Comparative protein structure modeling using modeller. Curr Protoc Bioinformatics 54:5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3
22. Li Y, Cirino PC (2014) Recent advances in engineering proteins for biocatalysis. Biotechnol Bioeng 111(7):1273–1287. https://doi.org/10.1002/bit.25240
23. Maddock DJ, Patrick WM, Gerth ML (2015) Substitutions at the cofactor phosphate-binding site of a clostridial alcohol dehydrogenase lead to unexpected changes in substrate specificity. Protein Eng Des Sel 28 (8):251–258. https://doi.org/10.1093/protein/gzv028
24. Brinkmann-Chen S, Flock T, Cahn JKB, Snow CD, Brustad EM, McIntosh JA, Meinhold P, Zhang L, Arnold FH (2013) General approach to reversing ketol-acid reductoisomerase cofactor dependence from NADPH to NAD. Proc Natl Acad Sci U S A 110(27):10946–10951. https://doi.org/10.1073/pnas.1306073110