Journal of Invertebrate Pathology 142 (2017) 11–15

Contents lists available at ScienceDirect

Journal of Invertebrate Pathology

journal homepage: www.elsevier.com/locate/jip

Revisiting protein structure, function, and evolution in the genomic era

Joseph M. Jez

Department of Biology, Washington University in St. Louis, One Brookings Drive, CB1137 St. Louis, MO 63130, United States

ARTICLE INFO

Article history:
Received 13 February 2016
Revised 4 June 2016
Accepted 28 July 2016
Available online 30 July 2016

Keywords:
Protein structure
Protein evolution
Protein function
Domains
Sequence comparisons
Genomes

ABSTRACT

The expansion of genomic data, three-dimensional structures of proteins, and computing power continues to improve our understanding of the evolution of protein structure and function relationships. As of June 2016, publically available databases contain more than 60 million unique protein sequences that group into 16,295 protein families that adopt ~1400 different three-dimensional folds. This data supports the exploration of evolutionary relationships on protein structure and function to answer a basic question – how do changes in gene sequence lead to alterations in protein structure and to the tailoring of biological and chemical function? This mini-review aims to provide a primer on the basics of protein structure, how evolution of sequence leads to diversity in protein structure and function, how these changes occur, and the role of domains in protein evolution. Understanding how to use the vast amount of sequence and structural information may also aid in assessing if changes in protein sequence and/or structure are relevant for safety assessments of new commercial biotechnology products.

1. Introduction

The classic architectural phrase that "form ever follows function (Sullivan, 1956)" is also a cornerstone concept of both biology and chemistry. The interplay between three-dimensional structure and either biological and/or chemical function is evident at every level of life from the atomic to the macroscopic. For example, the three-dimensional structures of various proteins are central to their ability to perform diverse biochemical tasks, including enzymatic catalysis, recognition of other proteins, DNA, RNA, and small molecule ligands, and for formation of larger order assemblies both inside and outside of cells. The adaptability of protein structure (mediated through gene sequence changes that alter amino acid sequence) is essential for the evolution of function.

The explosion of genome information driven by advances in sequencing technology and computational power provides data from all kingdoms of life – bacteria, archaea, protozoa, chromista, plants, fungi, and animals (Ruggiero et al., 2015) – from diverse environments around the globe. In April 2016, GenBank (www.ncbi.nlm.nih.gov/genbank/statistics) contained 193,729,511 DNA sequences. Automated annotation of these sequences in the UniProt database (www.uniprot.org) identifies 62,148,086 unique protein sequences (The UniProt Consortium, 2015). These repositories provide foundational data for exploring evolutionary relationships that complement studies on protein structure and function to answer a basic question – how do changes in gene sequence lead to alterations in protein structure and to the tailoring of biological and chemical function?

Understanding the structure/function relationships and evolutionary history of proteins can provide new insight on the biological and molecular mechanisms of proteins with potential commercial value and for approaching safety assessments of those candidates. This mini-review aims to provide a primer on the basics of protein structure, how evolution of sequence leads to diversity in protein structure and function, and how these changes occur.

2. Basics of protein structure: Primary to quaternary structure, motifs, folds, and domains

The classic features of protein structure – primary (the amino acid sequence of a polypeptide), secondary (folding of a polypeptide into α-helices, β-strands, and random coil), tertiary (the overall organization of secondary structures of a polypeptide), and quaternary (the association of multiple polypeptides) – provide a context for additional descriptions of structural features, such as motifs, folds, and domains. It should be noted that the presence of motifs, folds, and domains in various protein structures may or may not correlate with amino acid sequence similarity, as primary structure can vary greatly between proteins sharing common tertiary structures.

Structural "motifs" refer to the arrangement of secondary structure and 'super-secondary' structures. Often defined by the connectivity of α-helices, β-structures, and unstructured regions, motifs can be simple, such as a helix-turn-helix motif, or more

E-mail address: jjez@biology2.wustl.edu

http://dx.doi.org/10.1016/j.jip.2016.07.013
0022-2011/© 2016 Elsevier Inc. All rights reserved.

complex, such as an $(\alpha/\beta)_8$-barrel consisting of 8 repeats of an α-helix followed by a β-strand in a barrel-like arrangement (Richardson, 1981; Branden and Tooze, 1999). A protein "fold" describes how various secondary structure features (or motifs) are arranged relative to each other in three-dimensions. For example, a "Rossmann-fold", which commonly occurs in proteins that bind nucleotides, consists of seven parallel β-strands with the first two strands connected by an α-helix (Rao and Rossmann, 1973). A "domain" is classically defined as a polypeptide that retains its structure and function independent of other three-dimensional features (Richardson, 1981). Protein domains vary widely in size, but average around 100 amino acids, and may have sequence and/or structural homology across different proteins (Wheelan et al., 2000; Orengo et al., 2002). Domains often have specific functions, such as protein-protein interaction and DNA binding. For example, the Src Homology 2 (SH2) domains found in proteins of tyrosine kinase signaling pathways bind target proteins containing phosphotyrosines (Liu et al., 2012).

3. What is the scope of diversity in protein sequence and structure?

Although more than 60 million unique protein sequences are publically available, this total includes homologs of proteins found across multiple species. Bioinformatic analysis of protein sequences aims to define relationships of proteins and to provide insights on their evolutionary history. For example, the Pfam database groups the available protein sequences into 16,295 distinct families (Finn et al., 2016). Each entry in Pfam is used to generate a set of matching sequences and a profile hidden Markov model (HMM). Comparison of the profile HMM against amino acid sequence databases identifies related sequences. Members of the resulting family are then aligned to generate a complete sequence alignment for a given family. Amino acid sequence-driven comparisons are powerful and reveal key features in a wide variety of protein families.

In comparison to the depth and breadth of protein sequence data, the number of available three-dimensional structures is more limited, but also invaluable. As of June 2016, the Protein Data Bank (PDB; www.rcsb.org) contains 119,303 structures of 38,292 distinct proteins (Bernstein et al., 1977). The classification of those structures into different three-dimensional fold groups uses computational analyses that aim to identify structural similarities (Hadley and Jones, 1999). The exact number of distinct structural scaffolds varies based on the approach used, but the protein structures in the PDB represent ~1400 unique folds (Orengo et al., 2002; Murzin et al., 1995; Andreeva et al., 2014; Sillitoe et al., 2015).

The two major databases that classify protein structure are the Structural Classification of Proteins (SCOP; 13-14) and the Class, Architecture, Topology, and Homologous superfamily (CATH; 8, 15). SCOP (scop2.mrc-lmb.cam.ac.uk) initially organizes proteins into five classes by fold composition: (1) α proteins; (2) β proteins, (3) α/β proteins containing mixed α-helices and β-strands; (4) α + β proteins consisting of folds with α-helices and β-strands separated in the fold; and (5) multi-domain proteins containing distinct domains from one or more than one of the four other classes. In contrast to SCOP, CATH (http://www.cathdb.info) uses automated computer comparisons to classify and identify three-dimensional relationships. CATH uses a hierarchical grouping to determine the structural lineage of a protein based upon automated classification to assign similarities in topology and three-dimensional fold.

4. How is this diversity generated and what are the constraints?

The genomes of all organisms are in constant flux. Typically, genome sequence changes are gradual and occur over millions of years with the accumulation of mutations and recombination events driving evolution (Horowitz, 1945; Jensen, 1976; Jacob, 1977). A variety of genetic mechanisms can alter sequences encoding proteins and/or lead to the evolution of gene families. Gene duplication can lead to multiple isoforms; divergence of gene sequence through mutations can result in a protein with functions unlike the parent protein; and intronic recombination events that mix-and-match portions (i.e., domains) of different genes – all these processes can generate new functionality that may (or may not) be of value to an organism's fitness (Chothia and Gough, 2009; Weber et al., 2012). Genome-wide changes in sequence occur gradually and the evolution of new biochemical function is ultimately constrained by the physical properties of a protein.

Mutations in gene/protein sequences can lead to a loss of function (negative), no changes (neutral), or new function (positive). Comparison of protein homologs from different species provides a way of analyzing the accumulation of mutations and rates of mutation frequency (Wilson et al., 1977, 1987). Because mutations accumulate at comparable rates in different species, sequence comparisons indicate that most mutations are deleterious and eliminated by natural selection (Wilson et al., 1977, 1987; Creighton, 1992). Homologous proteins show different but characteristics rates of evolution. For example, cytochrome *c* undergoes 6.7 changes per 100 amino acids every 100,000,000 years and histone H4 shows 0.25 changes per 100 amino acids every 100,000,000 years (Wilson et al., 1977). Sequence comparisons reveal that neutral changes, which do not alter protein structure and biochemical function, occur less frequently than expected from average mutation rates.

Any of the various enzyme superfamilies provide examples of how gene duplication and sequence divergence leads to the evolution of new biochemical function (Chothia and Gough, 2009; Jörnvall et al., 1995; Jez et al., 1997; Gerlt and Babbitt, 2001; Penning and Jez, 2001; Eliot and Kirsch, 2004; Khersonsky et al., 2006; Redfern et al., 2008; Gulick, 2009). Within these superfamilies, member proteins sharing as low as 20% amino acid sequence identity can retain similar three-dimensional structures, highly conserved active site residues, and common chemical reaction mechanisms, yet recognize diverse substrates. Sequence changes can be tolerated, if the changes do not alter protein folding or compromise structural features required for protein function.

The aldo-keto reductase (AKR) superfamily provides an example (Jez et al., 1997). Members of this enzyme superfamily typically catalyze the NAD(P)(H)-dependent interconversion of ketones and alcohols on a variety of substrates (Fig. 1A). In addition, some AKRs can reduce carbon-carbon double bonds found in steroids. These enzymes commonly function as monomeric proteins of ~300 amino acids that fold into an $(\alpha/\beta)_8$-barrel (Fig. 1B). Of the 319 amino acids in the example structure (Fig. 1C), 20–30 (depending on the AKR) contact either NAD(P)(H) or the ketone/alcohol substrate (Fig. 1D). Only 4 residues are directly involved in the oxidation/reduction chemistry (Fig. 1A and E). The majority of amino acids in the protein are responsible for forming the overall three-dimensional structure that positions residues for recognition and binding of substrates and for catalysis. Regions beyond the active site (i.e., catalytic and ligand binding sites) are where nearly all substitutions occur and show the greatest variation across the AKR superfamily. With these changes, the requirement to maintain a stable, properly folded three-dimensional structure balances these basic functional requirements, as there are a limited number of energetically stable protein structure scaffolds available (Chothia and Gough, 2009; Todd et al., 1999).

Mutations in the catalytic residues usually will prevent the chemistry and are deleterious; however, it is possible that the mutations in these residues can subtly change the reaction mechanism. For example, mutation of the histidine in the AKR catalytic