Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

The robustness and innovability of protein folds

Ágnes Tóth-Petróczy and Dan S Tawfik

Assignment of protein folds to functions indicates that >60% of folds carry out one or two enzymatic functions, while few folds, for example, the TIM-barrel and Rossmann folds, exhibit hundreds. Are there structural features that make a fold amenable to functional innovation (*innovability*)? Do these features relate to *robustness* — the ability to readily accumulate sequence changes? We discuss several hypotheses regarding the relationship between the architecture of a protein and its evolutionary potential. We describe how, in a seemingly paradoxical manner, opposite properties, such as high stability and rigidity versus conformational plasticity and structural order versus disorder, promote robustness and/or innovability. We hypothesize that polarity — differentiation and low connectivity between a protein's scaffold and its active-site — is a key prerequisite for innovability.

Addresses
Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

Corresponding author: Tawfik, Dan S (dan.tawfik@weizmann.ac.il)

Current Opinion in Structural Biology 2014, **26**:131–138

This review comes from a themed issue on **Sequences and topology**

Edited by **L Aravind** and **Christine Orengo**

http://dx.doi.org/10.1016/j.sbi.2014.06.007

0959-440X/Published by Elsevier Ltd.

Introduction

Most new proteins arise from tinkering with existing ones, mainly via sequence divergence and gene rearrangement processes [1–3]. The emergence of new topologies (specific orders of secondary-structure elements) may occur readily, via rearrangements, such as duplications, insertions, deletions or circular permutations [1,4–7]. Overall, nearly all known structures may relate via a relatively small number of structurally intermediates [3,8]. The persistently high fraction of orphan genes [9,10] suggests that the *de novo* emergence of proteins has occurred throughout evolution, for example, from noncoding genomic regions [11,12] or from overlapping frames with existing coding regions [13,14]. We have little information, let alone structural data regarding *de novo* created proteins [15,16]. However, as indicated by the rarity by which novel folds are identified relative to the exponentially growing number of newly solved structures, the birth of new folds (i.e., protein architectures distinctly different from known ones) is likely to be relatively rare [17]. Indeed, 'mix and match' of pre-existing domains with known folds underlines the birth of the vast majority of new proteins [18–22].

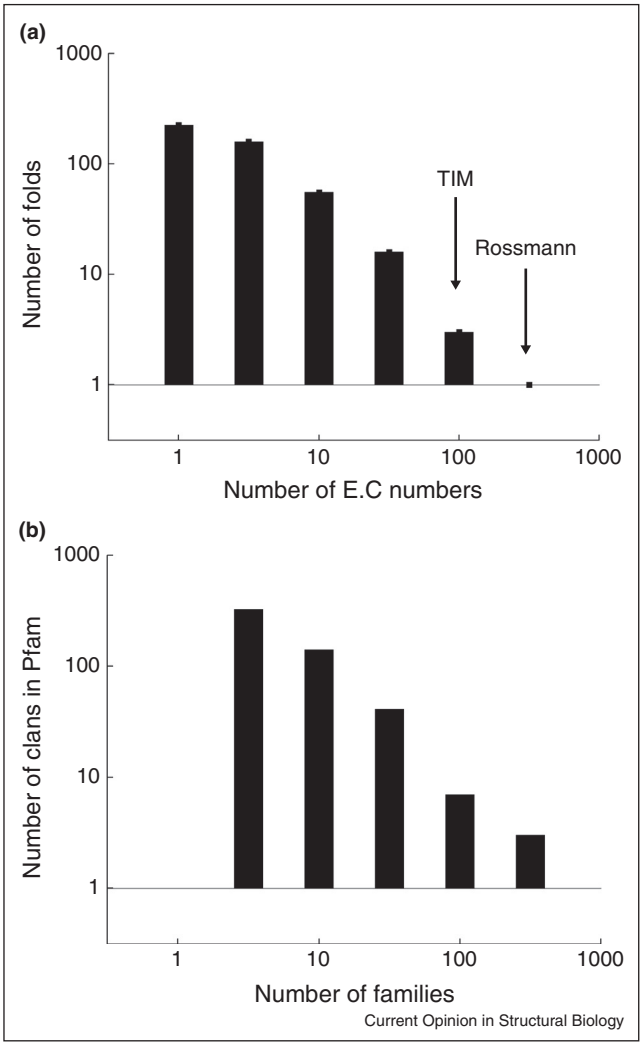
In the case of enzymes, the invention of new folds is likely to be extremely rare, as the degree of structural order and complexity of enzymatic folds is high. The vast majority of functional innovations appear to be driven by the ability of existing folds to accommodate new functions, via point mutations, short insertion and deletions [23,24] insertion of longer segments or domains [25] or by changes in oligomeric state [26]. Accordingly, this review focuses on the divergence of new functions within existing folds, and specifically on the acquisition of new catalytic functions within single domain enzymes.

Few folds foster functional innovation

There seem to be a huge variability in the number of different enzymatic functions associated with different folds [27]. For example, dihydrofolate reductase (DHFR), an enzyme found throughout the three kingdoms of life, has only one enzymatic function, namely, no other known enzyme family, or another enzymatic activity relates to this fold. In contrast, the TIM-barrel fold is observed in 57 functionally diverse superfamilies in Pfam (clan CL0036). To illustrate the level of variability, we extracted the number of different catalytic functions (different reactions and/or different substrates) associated with different folds in CATH database [28], and plotted the distribution in Figure 1a. Similarly, the distribution of clans in the Pfam database [29], namely of superfamilies that group all families related by sequence and structure, with respect to the number of different families (and hence of different functions) per clan, is wide (Figure 1b). Both these distributions are skewed — essentially, ~a quarter of the total number of known folds carry out ~80% of all known enzymatic functions.

The skewed fold-functions distributions could be simply due to different folds having had different evolutionary times for innovation [30]. Additionally, as any other evolutionary phenomenon, this distribution is the outcome of both chance (the rich get richer trend, namely folds that initially diverged by chance are more likely to continue and diverge) and necessity, namely, of certain folds having higher potential for functional innovation than others. Assuming that the latter underlines a significant part of the observed, skewed distribution, we ask what structural features might affect the potential of folds to acquire new functions.

Figure 1



The innovability of folds differs dramatically. **(a)** Log scale distribution of the number of different enzymatic functions associated with each fold as indicated by the number of different E.C. numbers. Data were derived from CATH [28] for single domain enzymes (N^{folds} = 457). **(b)** Log scale distribution of sizes of clans (superfamilies), that is of the number of different families within each clan. Data were extracted from Pfam v27.0 [29] (N^{clans} = 515).

Box 1 Definitions.

- **Evolvability** is the ability to change along evolutionary time. Evolvability relates to the capacity to accommodate sequence changes over time, as well as to adopt new functions[31]. The latter is also driven by sequence changes. Evolvability has therefore two components that are interlinked: robustness and innovability [32,33].
- **Robustness** relates to the ability to preserve a phenotype in the face of genotype changes. Robustness of proteins is defined as the ability to tolerate mutations whilst maintaining the original structure and function, and thus have the sequence change over evolutionary time (drift) at a relatively fast rate. Alternative terms found in the literature are **genetic robustness**, **designability** [64] or **neutrality**. The latter defines robustness of a genotype by the fraction of its neighboring genotypes (sequences that deviate by a given number of amino acid exchanges) that exhibit the same phenotype [43]. Of the related above terms [56], we chose to use robustness.
- **Evolutionary rates**. Robustness is manifested in evolutionary rates. The latter are obtained by examining alignments of protein families within a given phylogeny, say all vertebrate orthologs of a given protein, and calculating the average rate of amino acid exchanges per position. Since all alignments reflect sequence divergence within the same evolutionary time, the average rates are comparable [47]. The variability in evolutionary rates is high, and proteins of one given species may show up to 100-fold different rates, and thus, very different degrees of robustness (e.g., see [47]).
- **Innovability** is defined as the ability to acquire new functions [33]. Mutations are rare, and their combinations are extremely rare [34]. Innovability therefore relates to the ability of relatively few sequences changes to induce large changes in function and/or structure of a protein.
- **Superfamilies** comprise functionally and structurally related proteins, typically having the same fold and the same key catalytic residues, and that are likely to have all diverged from a common ancestor. Superfamilies are comprised of different families representing different **paralogs** (though the sequence identity between families can be nondetectable). Each family groups many different **orthologs** — proteins belonging to different species yet sharing the same structure and function. The sequence variability between **orthologs** represents robustness, while **paralogs**, that is evolutionary related proteins with different functions, represent innovability.

Evolvability: robustness and innovability

Evolvability, namely the ability to change along evolutionary time, has two crucial components [31,32]: robustness and innovability. At the protein level, robustness relates to mutations having a minimal effect on function and structure (Box 1). On the other hand, innovability relates to the ability to adapt and acquire new functions, that is, new substrate and/or reaction specificities in enzymes [33]. Both these features are critical for long-term survival. Mutations occur constantly, regardless of and more frequently than adaptation. Hence, low robustness means a high fraction of nonfunctional proteins, and hence a smaller number of fit offspring. However, adaptation depends on one, or a few mutations at most providing a selective advantage via a new or improved function [34]. Thus, by default, the more robust a protein is (i.e., most mutations have no, or mild effects) the less innovable it is. How do proteins reconcile this dichotomy? Or perhaps, many do not — that is, they might be robust but not innovable, or neither? And, are there structural and biophysical features of proteins that promote robustness and/or innovability?

The structural order–disorder paradox

Experimental and computational analyses indicate that configurational stability, driven typically by a higher degree of structural order and compactness, confers tolerance to mutations and thereby promotes protein evolvability [35–41]. A highly ordered, well-packed protein affords a higher stability threshold, and enables more