Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Novel function discovery through sequence and structural data mining

Briallen Lobb and Andrew C Doxey

Large-scale sequence and structural data is a goldmine of novel proteins, but how can this data be effectively mined for new functions? Here, we review protein function prediction methods and recent studies that apply these methods to discover new functionality. Core approaches include sequence-based homology detection, phylogenetic analysis, structural bioinformatics, and inference of functional associations using genomic context and related methods. With such a wide range of approaches, sequences may reveal new functionality regardless of their similarity to a characterized reference. Homologs of known function may be identified in unexpected species or associations. Detection of functional shifts in sequences may reveal new activities and specificities. New protein functions may also be predicted in uncharacterized sequences and structures. Finally, methods and data may be integrated and applied at increasingly large scales due to improved protein domain knowledge and structural coverage, which amplifies the ability to predict and discover novel protein functions.

Address
Department of Biology, University of Waterloo, 200 University Ave. West, Waterloo, ON N2L 3G1, Canada

Corresponding author: Doxey, Andrew C (acdoxey@uwaterloo.ca)

Current Opinion in Structural Biology 2016, **38**:53–61

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

Edited by **L Aravind** and **Elizabeth Meiering**

http://dx.doi.org/10.1016/j.sbi.2016.05.017

Introduction

Modern sequencing technologies continue to accelerate the collection of new genes and genomes. This sequence information has become invaluable to protein researchers, fuelling new computational methods for structure and function prediction [1,2], analysis of protein family evolution [3,4•], and protein design [5,6]. Sequence databases are improving with regards to annotations [7,8] and coverage of protein domain space [9,10]. In addition, structural data is growing through structural genomics initiatives [11,12], further enabling large-scale homology modelling efforts [13,14].

The accuracy of protein function prediction has improved over recent years as a result of better methods as well as increased experimentally-based annotations [15,16]. Most proteins predicted from genomes can now be at least partially annotated [17] through detected homology to existing proteins (e.g., via BLAST search) or through matches to domain databases such as CDD [18], PFAM [9], CATH [10], and FIGFAMs [19]. These predictions form the initial landscape of functional annotations in newly sequenced genomes, upon which further questions may be investigated.

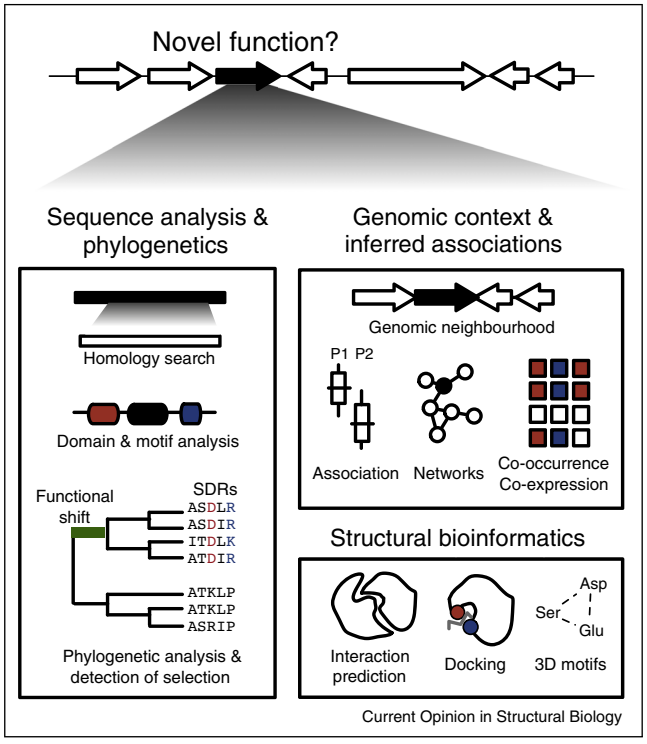
One important and common question following functional annotation is how to pinpoint the most functionally novel and biologically interesting predictions. This task is challenging due to the scale at which function predictions are often made and also because of the complexities surrounding the definition of 'function' [20]. As a result, expert biological knowledge is needed to interpret predictions and identify those providing particularly novel or unexpected biological functionality.

Here, we examine several key approaches to function prediction: sequence-based methods, structure-based methods, and inference of functional associations using genomic context and related methods. We also explore recent studies that apply these techniques to discover novel protein families or functions (Figure 1). The reviewed studies apply function prediction approaches to new genomes, metagenomes, and/or protein databases and typically confirm novel predictions using some form of experimental validation. The concept of function used here is broad and includes 'molecular function' as well as 'biological process', consistent with functional ontologies [21] and assessments of prediction methods [15].

Finding homologs in unexpected places

Homology search has been described as the single most powerful tool in bioinformatics and, for decades, has been the core strategy in protein annotation [22]. Beyond its utility in finding new members or relatives of existing families, homology search can reveal profound functional novelty when a homolog is found in a novel/unexpected biological setting (Figure 2). This setting may be a new species or environment [23,24••,25•,26], or an unexpected co-occurrence with other proteins/pathways [27••,28••]. The discovery of bacterial rhodopsins [23,29], archaeal ammonia monooxygenases [23,30], and, recently, complete nitrification by *Nitrospira* [27••,28••], are all examples of important biological phenomena predicted through sequence homology.

Figure 1



An overview of common approaches to prediction and analysis of protein function. The top panel depicts a genomic region containing predicted genes. Any given gene (example shown in black) can be further analyzed to predict its function through sequence analysis and phylogenetics, structural bioinformatics, or by inferring functional associations. Sequence-based methods include homology detection (query shown in black), domain and motif analysis (a hypothetical three domain architecture is shown), and analysis of phylogeny, evolutionary conservation and shifts in function. As an example, two hypothetical specificity-determining residues (SDRs) unique to a subfamily are depicted. Functional associations may also be inferred by analysis of genomic context (gene of interest shown in black and its function may be inferred from annotations of adjacent genes in white), by phenotype associations (e.g., differential protein abundance in two phenotypes, P1 and P2), and gene coexpression (e.g., using microarray data) or co-occurrence with other genes across genomes. Finally, structural bioinformatics methods include template-based prediction of protein–protein interactions, docking, and prediction of 3D motifs representing binding or catalytic sites. This is a sample of common function prediction methods and is not intended to be complete. These methods may be used individually or in combination to predict new functionality.

The discovery of complete nitrification [27••,28••] illustrates the power of detecting unexpected enzyme combinations. By identifying genes encoding ammonia monooxygenase and hydroxylamine dehydrogenase together in a single genome, two recent studies [27••,28••] were able to identify the microbial basis for the long-sought-after process of complete nitrification (oxidation of ammonia to nitrate, 'comammox'). Undersampled phyla from the tree of life are a likely hotspot for functional novelty of this kind as their genomes have been less explored. Indeed, recent analyses of hundreds of new microbial 'dark matter' genomes obtained by single-cell genome sequencing have revealed novel and unexpected metabolic features such as archaeal sigma factors previously considered exclusive to bacteria [24••]. Ultimately, even if molecular function is completely conserved in newly detected homologs, finding homologs in unexpected biological settings can reveal profound novelty at the pathway to organismal to ecological level [23,24••,27••,28••].

Detecting functional shifts in sequences

On the other hand, newly identified homologs may have diverged in function with respect to their reference (Figure 2). Finding functional shifts in sequences or families, for example through detection of site-specific changes in evolutionary rate or amino acid preference [31], is another way to uncover new functionality.

Several recent studies have applied the evolutionary trace (ET) method [32] to identify conserved and likely functional sites that differ between protein subfamilies [33–35]. Applications of these methods to families of G protein-coupled receptors have uncovered specificity-determining residues (SDRs, see Figure 1) that differentiate substrate affinity and specificity [33–35]. These studies also highlight the important role of changes to allosteric pathways in shaping the evolution of specificity.

Analyses of functional diversification have also been expanded to entire protein superfamilies [3,36–38]. An effective approach has been to map structural and functional properties onto large-scale sequence similarity networks of enzyme superfamilies, thus revealing broad-scale differentiation of substrate specificity and how it correlates with sequence and structural features [3]. Such approaches have revealed functional differentiation in ligases [36], cytosolic glutathione transferases [37], dipeptide epimerases [39], and diverse trans-polyprenyl transferases [40]. In a recent study, Furnham *et al.* [38] examined changes in enzymatic function within 379 protein domain superfamilies, revealing how both subtle and large-scale changes in enzymatic machinery can lead to functional changes in chemistry and substrate specificity.

Building on past approaches [41], recent databases have attempted to subdivide known protein families into functionally distinct subfamilies [10,42]. The FunFHMMer method has subdivided 2735 CATH superfamilies into 110 439 subfamilies (FunFams) with increased functional coherency [42]. Similarly, the Selectome database has predicted positive selection across thousands of vertebrate protein phylogenies, facilitating large-scale exploration of adaptive evolution [43].

While the above approaches tend to examine functional shifts over macroevolutionary time scales, others are