Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

The use of evolutionary patterns in protein annotation

Angela D Wilkins[1], Benjamin J Bachman[1,3], Serkan Erdin[1,2] and Olivier Lichtarge[1,2,3]

With genomic data skyrocketing, their biological interpretation remains a serious challenge. Diverse computational methods address this problem by pointing to the existence of recurrent patterns among sequence, structure, and function. These patterns emerge naturally from evolutionary variation, natural selection, and divergence — the defining features of biological systems — and they identify molecular events and shapes that underlie specificity of function and allosteric communication. Here we review these methods, and the patterns they identify in case studies and in proteome-wide applications, to infer and rationally redesign function.

Addresses
[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[2] CIBR Center for Computational and Integrative Biomedical Research, Baylor College of Medicine, Houston, TX 77030, USA
[3] Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, TX 77030, USA

Corresponding author: Lichtarge, Olivier (lichtarge@bcm.edu)

Current Opinion in Structural Biology 2012, **22**:316–325

This review comes from a themed issue on
Sequences and topology
Edited by Christine Orengo and James Whisstock

Available online 24th May 2012

0959-440X/$ – see front matter

http://dx.doi.org/10.1016/j.sbi.2012.05.001

Introduction

Proteins remain difficult to characterize functionally despite the exponential growth in experimental data on sequence, structure, and function. There are many reasons for this persistent challenge. Proteins have not a single molecular function but rather multiple features that cooperatively sustain their biological fitness. The details and parameters of these features, for example, folding, dynamics, cellular targeting, molecular interactions, catalytic activity, allosteric control, post-translational modifications, and degradation, to name a few, are often vague for a lack of laboratory assays to measure them accurately on a large scale and in their relevant cellular context. As a consequence, as of March 2012, fewer than 0.1% of the 21 million protein sequences from 3173 completely sequenced genomes [1] had experimentally tested functions, and only two-fifths had at least one automated computationally inferred annotation [2–4]. The number of genes without known function is 37% in eukaryotes, 24% in humans, 33% in the far simpler and much studied *Escherichia coli*, and 40% in other bacteria [2,5]. Although most of the 4225 *E. coli* genes were recently assigned putative annotations of functional associations, they were not assigned biochemical function [6]. Given concerns that some of these annotations may not be accurate [7], the problem of translating sequence into function, and more broadly of translating genotype into phenotype, remains daunting.

Computational methods have long sought to fill this role. A remarkable early success was to realize that sequence and structure diverge smoothly: the root mean square deviation of protein backbones increases exponentially with the sequence divergence of evolutionarily related proteins, or homologs [8]. This elegant observation is robust [9], and extends to other functional features besides folding [10] so that, in practice, it justifies homology-based predictions of structure and of function [11], arguably the two most widespread computational applications in biology. Other basic evolutionary principles are emerging from high throughput and systems biology [7]. Protein mutation rate and protein expression are inversely correlated [8], biological networks obey power-laws and are scale-free [12]; and the evolutionary rates of orthologs follow a Gaussian spread [13]. Despite their statistical power, because these principles involve ensemble averages over whole sequences, structures, families, genomes and networks, as well as very long time-scales, they carry limited information on the direct role of individual sequence positions to the function of a given protein.

Single residue variations may profoundly impact function and explain why homology-based function prediction can lead to incorrect annotations: although alike in sequence and structure, two homologs may harbor differences at one or just a few residues with disproportionate impact on function [14]. The identification of such key residues is therefore essential to distinguish meaningful variations of function. This review therefore focuses on methods to identify functionally relevant evolutionary patterns among sequence, structure, and function. Such patterns emerge naturally from random variations and natural selection; they identify molecular events and shapes that determine function and specificity; and they can be approached by focusing on sequences, on structures, and on evolutionary classification. In the second part of the review, the focus will shift to the combination of these techniques in a unifying Evolutionary Trace framework.

Throughout the review, we will refer to two popular functional classification systems. Gene Ontology (GO) [4] provides well-defined terms for the molecular function, cellular component, and biological process of a gene product, along with evidence codes that specify the basis for the annotation and therefore its reliability. Enzyme Commission classification designates enzymatic function into four (EC) numbers [15], indicating the mechanism of the enzyme, the type of bond, the catalyzed reaction, and the substrate, respectively.

Sequence-based patterns

The simplest and most widespread evolutionary pattern for defining function is homology between proteins or domains. The rationale is that homology implies that proteins share a common ancestry and hence the function of that common ancestor. Once it is recognized by similarity searches with BLAST or PSI-BLAST [16], function is transferred between close homologs. A concern is that these homologs may have already evolved distinct functions. Thus homology-based annotation errors are not uncommon: divergence of activity has been observed even between enzymes with as much as 70% sequence identity [17]. To compound this problem, these errors may in turn propagate across databases [7]. To reduce incorrect annotations, multiple techniques, including GOtcha [18], ESG [19], and GOPred [20], tally the GO terms of all of the most significant sequence similarity matches and identify those with the best statistics. For example, GOtcha weighs this tally by the significance of each PSI-BLAST match to a database of proteins with GO annotations, to generate a probability that the query protein performs a particular function.

Other methods go beyond whole sequence comparison to focus on alignment columns with significant conservation [21,22]. The results are generalized profiles to infer structural or functional similarities. Pfam [23•] is a widely used database of Hidden Markov Model profiles generated by HMMER [24] applied to the Uniprot database [2]. To enhance specificity, Pfam-A uses a smaller set of almost 12 000 sequences representative of individual families that were hand-curated with functional annotations from literature references; to achieve sensitivity, Pfam-B uses a larger set of nearly 140 000 families that were clustered automatically and without dedicated annotation or reference. While Pfam and methods such as Prosite [25] and Interpro [26] focus primarily on the entire protein domain, other sources, such as the ELM database [27•], focus instead on smaller motifs.

Even more refined searches focus on specific residues that together define a functional signature. Transfer of function based on these signatures can increase annotation specificity, that is lower false positives, by recognizing functionally inconsistent differences among key residues. Several sequence motif-based algorithms were designed specifically for this task, including Confunc [28], DME [29], and EFICAz2 [30]. All rely on discovering discriminatory sequence fragments shared by proteins with identical function and not others. ConFunc applies GO terms to partition homologs into multiple subsets. The sequences of each subset are then aligned to identify conserved residues. A GO term can then be transferred to a new homolog if it shares this residue signature. Controls suggest 24% greater accuracy of annotation compared to BLAST for homologs with less than 35% sequence identity. Likewise, DME and EFICAz2 use conservation to key in on functional residues specific to given enzyme functions.

Together these studies show that comparative sequence analyses identify evolutionary patterns at different levels of resolution, from whole sequence to profiles to motifs, that are all relevant to structure and function and useful to transfer annotations among proteins.

Structure-based patterns

Structural information adds another dimension to the search for functionally relevant similarities among proteins. First, global structure alignments will detect homologies that elude sequence searches [8]. Additionally, spatial correlation among key residues can reveal highly specific three-dimensional (3D) functional features [31]. Some structural comparisons treat the structure as a rigid body, as in DALI [32] and TM-align [33], while others tolerate flexibility, as in TOPS++FATCAT [34•]. A challenge for these structural alignments is the lack of a universally accepted definition of structural similarity [35]. In order to address this, CATH [36] and SCOP [37] created manually curated protein structure classification codes based on both domain and evolutionary similarities. These classifications enable functional inference of protein structure in many cases, but overall, and for the same reasons that a few amino acid prove determinant of function in sequence comparisons, the structure-to-function relationship over protein domains is not one-to-one [38].

This motivated searches for specific structural regions resembling previously characterized pockets for catalysis and ligand-binding or surface regions for macromolecular interactions [39]. In a control set of 332 ligand-binding proteins, ConCavity [40] correctly predicted the binding site in 80% of cases by searching jointly for the local conservation of sequence and structural topology. Similar methods [41,42] are listed in Table 1. FINDSITE [43•] and 3DLigandSite [44] extend these ideas to homology models and detect the functional determinants of a ligand binding site. FINDSITE specifically creates homology models of the query, structurally aligns these to determine a likely binding site, and then suggests ligands and other GO functional annotations. In controls with less than 35% sequence identity to the nearest target protein,

Table 1

Common methods to characterize proteins and the main evolutionary pattern they rely on (See text for citations).

Method	Website	Comments
Gene ontology	http://www.geneontology.org	Standard representation of gene and gene product attributes
Enzyme nomenclature	http://www.chem.qmul.ac.uk/iubmb/enzyme	Enzyme classification
BLAST/PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Sequence comparison
Gotcha	http://www.compbio.dundee.ac.uk /Software/GOtcha/gotcha.html	Assigns GO terms based on sequence comparison
ESG	http://kiharalab.org/web/esg.php	Assigns GO terms based on sequence comparison
GOPred	http://kinaz.fen.bilkent.edu.tr/gopred	Assigns GO terms based on sequence comparison
Pfam	http://pfam.sanger.ac.uk	Database of protein families and their MSA
HMMER	http://hmmer.janelia.org	Sequence comparison based on hidden Markov models
Prosite	http://prosite.expasy.org	Database of protein domains, families and functional sites
Interpro	http://www.ebi.ac.uk/interpro	Database of protein functional signatures
ELM	http://elm.eu.org/links.html	Resource to investigate functional sites in eukaryotic proteins
ConFunc	http://www.sbg.bio.ic.ac.uk/~confunc	Assigns GO terms based on sequence comparison
DME	http://adios.tau.ac.il/DME11.html	Assigns full EC number based on sequence comparison
Eficaz2	http://cssb.biology.gatech.edu/skolnick /webservice/EFICAz2/index.html	Assigns full EC number based on sequence comparison
Dali	http://ekhidna.biocenter.helsinki.fi/dali_server	3D protein structure comparison
TM-align	http://zhanglab.ccmb.med.umich.edu/TM-align	3D protein structure comparison
TOPS + +FATCAT	http://fatcat.burnham.org/TOPS	3D protein structure comparison
CATH	http://www.cathdb.info	Protein domain structure classification
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	Protein domain structure classification
ConCavity	http://compbio.cs.princeton.edu/concavity	Predicts ligand binding sites from protein structure
FTSite	http://ftsite.bu.edu	Predicts ligand binding sites from protein structure
LIGSITEcsc	http://projects.biotec.tu-dresden.de/pocket	Predicts ligand binding sites from protein structure
3DLigandSite	http://www.sbg.bio.ic.ac.uk/~3dligandsite/	A threading-based method to predict ligand binding site
FINDSITE	http://cssb.biology.gatech.edu/skolnick/files/FINDSITE	A threading-based method to predict binding site, ligand, and function
pevoSOAR	http://sts.bioengr.uic.edu/pevosoar	Assigns up to four digit EC numbers based on local structure similarities
Catalytic Site Atlas	http://www.ebi.ac.uk/thornton-srv/databases/CSA	Database of known and predicted catalytic residues in the protein structures
FunClust	http://pdbfun.uniroma2.it/funclust	Identifies local functional motifs in the protein structures
GASPSdb	http://gaspsdb.rbvi.ucsf.edu	Database of 3D motifs generated by GASPS algorithm
SuMo	http://sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome	3D structure comparison based on local structure similarity
Par-3D	http://sunserver.cdfd.org.in:8080/protease /PAR_3D/index.html	Detects active site residues using 3D templates
PINTS	http://www.russelllab.org/cgi-bin/tools/pints.pl	3D structure comparison based on non-sequential local motifs
Flora	http://www.mcsg.anl.gov/	Assigns three digit EC numbers based on local structural similarities
GeMMA	http://www.biochem.ucl.ac.uk/cgi-bin/dlee/GeMMA	Provides classification based on phylogenetic analysis
SCI-PHY	http://phylogenomics.berkeley.edu/	Provides classification based on phylogenetic analysis
PROTONET	http://www.protonet.cs.huji.ac.il	Classifies protein sequences based on phylogenetic analysis
SIFTER	http://sifter.berkeley.edu	Assigns GO terms based on phylogenetic analysis
PhylomeDB	http://phylomedb.org/	Database of phylogenetic trees with ortholog assignments
TreeFam	http://www.treefam.org/	Database of phylogenetic trees with ortholog assignments
ET	http://mammoth.bcm.tmc.edu/ETserver.html	Ranks amino acids based on phylgenetic analysis
ETA	http://mammoth.bcm.tmc.edu/eta	Assigns three digit EC numbers and GO terms based on local structural similarities

FINDSITE reached 67% accuracy. A related method, pevoSOAR [45], annotates structures for enzymatic function with 80% accuracy in limited controls. Together these studies show that patterns of local structural similarities add important information for functional inference.

Further following the logic of sequence comparisons, structural searches can also focus on just the few residues that mediate the most essential aspects of catalysis or binding. The example of the Ser-His-Asp catalytic triad of serine proteases illustrates that only a few amino acids in a well-defined structural conformation are sufficient to annotate function in structures [46]. This suggests a general strategy in which a small but functionally essential structural motif, called a 3D template, is matched geometrically in other protein structures. A matched protein may then potentially perform the function associated with the template [47]. Several methods, including FunClust [48], GASPS [49], SuMo [50], PAR-3D [51], and PINTS [52] follow this strategy. They typically rely on a source of structural motifs that are functionally relevant, such as The Catalytic Site Atlas [53] database, which compiles templates for enzyme activity taken from the experimental literature. To identify enzymatic templates more generally, FLORA defines

them in terms of recurrent structural patterns in the superimposed structures of enzyme homologs [54].

Phylogenomic patterns

Molecular function may also be inferred from phylogenomic classifications. Starting with an alignment of homologs and an associated phylogenetic tree, annotations are transferred within branches following the topology of the tree [55]. Typically, uncharacterized proteins can inherit the annotation of the ortholog subfamily to which they belong. GeMMA [56••], SCI-PHY [57], PROTONET [58••], and SIFTER [59,60] reflect these ideas. The phylogenetic tree of PROTONET [58••] has nearly 10 million sequences, and a user can retrieve the evolutionary tree relevant to a query protein of their choice, and navigate its branches to search for functional information. In a more automated approach, SIFTER models protein evolution to propagate GO annotations within the tree [59,60]. This is a slow process, but limiting the number of possible combinations of molecular functions for individual proteins significantly raises efficiency without loss of prediction accuracy [60].

Because paralogs arise from gene duplication and usually evolve different functions, it is important to distinguish them from orthologs. Algorithms that detect orthology often rely on tree reconciliation approaches. Typically, a phylogenetic tree of homologs is compared to a speciation tree, allowing paralogs and orthologs to be identified by inferring the order of events for gene loss and duplication. TreeFam [61] provides ortholog and paralog assignments based on this approach, as well as phylogenetic trees for individual proteins for mammal families. PhylomeDB [62] uses a different species-overlap algorithm, which compares the species identity of closely related branches to decide whether their parental node is a duplication or a speciation. It provides orthology predictions, alignments, and phylogenetic trees for human, *Saccharomyces cerevisiae*, and *E. coli*.

Synthesis through evolutionary trace patterns

It is possible to integrate the diverse evolutionary patterns seen in sequences, motifs, templates, and phylogenies through Evolutionary Trace (ET) analysis [63]. This approach applies proteome-wide and has been extensively validated in experimental case studies. It yields tools to map functional sites in proteins, identify their key determinants, guide protein redesign studies, and extract 3D functional motifs with which to annotate protein function in novel structures. In view of this variety of applications, ET patterns arise from a surprisingly basic classification procedure.

In order to discover which residues are important to structure and function, ET systematically ranks amino acid positions by their phylogenetic patterns of variation. Starting with a protein family alignment and the corresponding evolutionary divergence tree, ET ranks residue positions better, or worse, depending on whether the substitutions in their alignment column correlate with larger, or smaller, tree divergences (Figure 1). Thus, by definition, variations of top-ranked ET residues entail big evolutionary steps, suggesting that they contribute importantly to structure and function. Variations of poorly ranked residues, by contrast, entail small evolutionary steps and suggest at best a limited influence on structure and function. Thus, by systematizing these comparisons between alignment and tree, ET ranks residue positions relative to each other by the size of their phylogenetic variations. This procedure mimics the laboratory strategy of measuring with assays which substitutions disrupt function, replacing assays and mutations in the wet lab with divergences and variations, respectively, *in silico* [63].

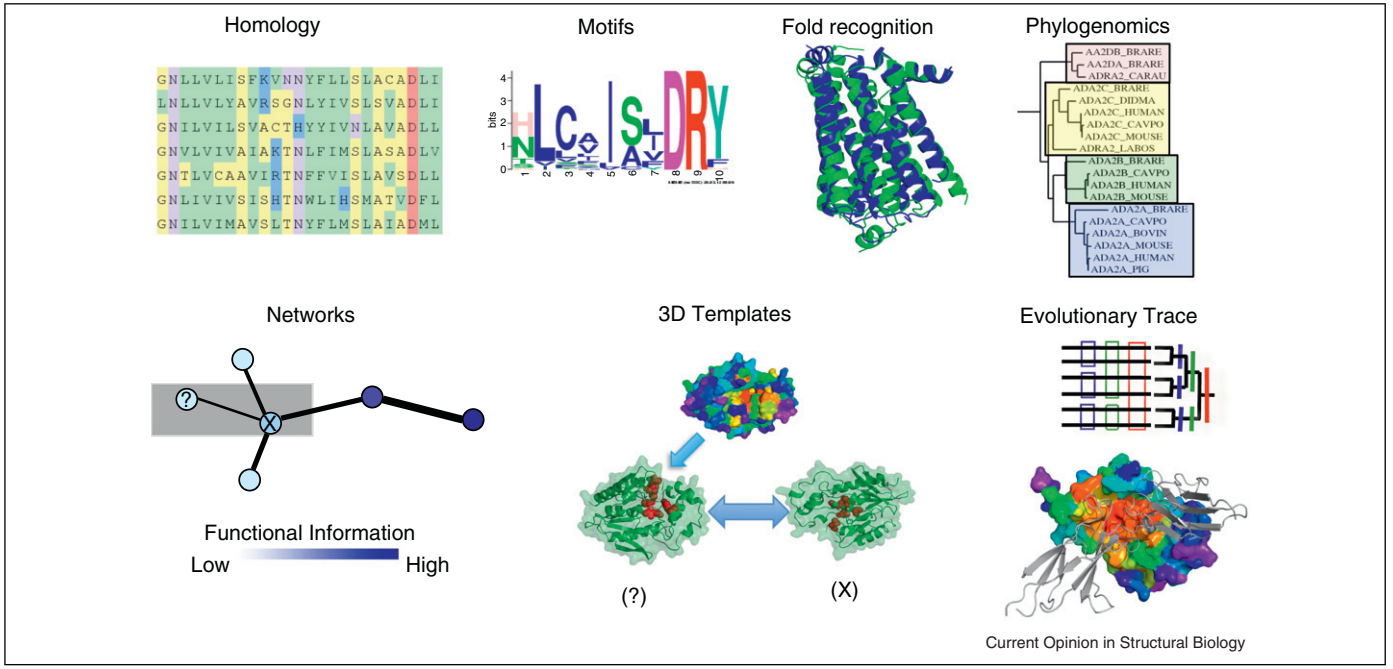
A series of technical studies show that the ET rank of evolutionary importance reveals structurally and functionally relevant patterns (Table 2). First, top-ranked ET residues cluster spatially in protein structure [63–65]. Second, this clustering is widespread in the structural genome and greater than expected by chance as measured with a z-score to yield an overall measure of structural clustering of important residues (Figure 2). When no structure is available, sequence-based quality measures can also assess the significance of ET patterns [66]. Third, these clusters overlap with functional sites as shown in 37 of 38 proteins with known ligand binding sites, and so can yield insights into the regions of a protein that mediate function most directly [64,67]. Fourth, the ET link between sequence and structure is such that better clustering z-score strongly correlates with more accurate functional sites discovery [67], as shown in 50 diverse proteins by varying the input parameters of ET and observing correlations mostly above 0.7 [68]. Mapping evolutionarily important residues to the structure has also been useful in other studies. Spatial clustering of important residues formed presumed functional sites useful for protein–protein docking [69] and the prediction of catalytic residues [70]. Thus phylogenetic patterns of residue variations in sequences are linked to a clustering bias in structures that reveals functional sites. As discussed next,

Table 2

Recurrent observations regarding ET residues suggest general rules of proteome evolution which link sequence, structure and function.

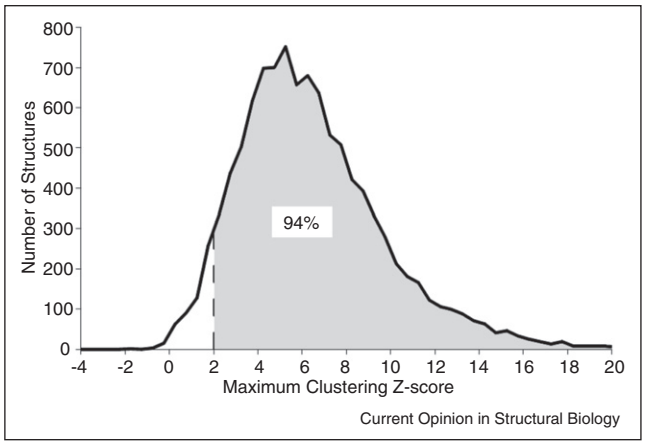
Proteomic rules
1. Amino acids may be ranked by evolutionary importance
2. Top-ranked residues cluster in the protein structure
3. Clusters predict functional sites
4. Clustering quality correlates with prediction quality
5. Maximizing clusters improves the quality of predictions

Figure 1



Evolutionary approaches to characterize protein function rely on both global or local patterns. These include global sequence similarity (Homology) and local residue conservation (Motifs), or global structural similarity (Fold Recognition) and local structural similarity (3D Templates). Another pattern is evolutionary classification (Phylogenomics). The Evolutionary Trace (ET) combines these approaches by defining key structural or functional positions based on whether their evolutionary variations couple to small (blue, where the breaks between rectangles indicate residue variations), or large evolutionary divergences (red). Top-ranked positions typically map out functional sites to guide targeted mutations or extract functional motifs, such as for 3D templates. Proteome-wide 3D template matches between structures give rise to a proteomic network that can be analyzed for global function prediction.

Figure 2



Distribution of the statistical clustering *z*-score of ET residues in 10 417 proteins from the PDB90. This *z*-score is the difference between the observed and the expected random clustering pattern in units of standard deviation. A *z*-score can be obtained at any ET coverage of a protein. This histogram shows the maximum clustering *z*-score between 0% and 50% coverage, which is representative of *z*-scores over most of this interval. The high values (94% with *z*-score > 2) show that evolutionarily important residues cluster together in the protein, as a general rule.

one may then interrogate a novel structure with ET to identify its functional sites and its residue determinants. In a variety of prospective experimental case studies, this guided the design of separation-of-function mutations; the rewiring of functional specificity, such as the discovery and reprogramming of an allosteric pathway; and the design of peptide inhibitors. On a structural proteomic scale, top-ranked ET residues enable large-scale function prediction.

Case studies: evolutionary patterns and functional redesign

Selective separation of function mutations helped clarify in the eukaryotic Ku70/80 heterodimer how different and antagonistic functions co-exist in the same complex, and suggested a long-sought interaction site with the gene repressor LexA in the prokaryotic protein RecA. The former study identified two structurally distant clusters of top-ranked ET residues that suggested distinct functional sites in Ku70/80. Targeted mutations to one of the clusters disrupted end-joining but not telomere-maintenance, and mutations of the other cluster did the reverse. Thus double-strand break DNA repair and telomere maintenance segregate to opposite ends of the Ku structure which explains how both functions may be performed without risking end to end chromosome fusion [71]. Likewise, in RecA, ET revealed a number of new functional sites that were then mutated. These mutations disrupted either DNA repair by recombination, or LexA interaction, but not both. Thus, even though RecA is a heavily mutagenized, classic example for homologous DNA repair, ET patterns of evolutionary importance revealed previously unrecognized functional regions including the potential trigger of LexA-mediated error-prone DNA repair — one of the root causes of antibiotic resistance [72••].

ET patterns typically identify functional sites on protein surfaces, but they can also suggest internal mechanisms. An ET study mapped key functional residues in the seven-helical transmembrane core of G-protein-coupled receptors (GPCRs) and suggested that distinct internal functional modules couple allosterically the binding of extracellular ligands to intracellular signaling through G proteins or β-arrestin-mediated internalization. Consistent with predictions, mutations of top-ranked ET residues in each module variously inhibited ligand binding, caused constitutive activity [73], and could even block G protein signaling while leaving β-arrestin signaling intact [74]. More recently, a difference analysis of ET applied solely to bioamine receptors and applied to all rhodopsin-related receptors suggested a set of residues uniquely important to bioamine function. Single point mutations then transferred these putative bioamine specificity determinants from the 5HT-2A serotonin receptor into the D2R dopamine receptor and, as a result, increased serotonin signaling and decreased dopamine signaling independent of changes in binding affinity [75••]. These mutations, located deep in the GPCR transmembrane core, show that the GPCR allosteric pathway can encode signaling response specificity independently of binding, demonstrating the concept of allosteric specificity, and that this specificity code can be traced back and rekeyed, at least in part, by swapping top-ranked ET residues between paralogs.

Besides point mutations, ET patterns have been moved whole into a new scaffold to create functional mimetics. A cluster of ET residues suggested a novel binding site on surface exposed helices of G-protein-coupled receptor kinases (GRK), proteins that phosphorylate the intracellular loops of GPCRs to regulate their activity [67]. This site was then mimicked with peptides designed to keep the evolutionarily important residues intact, while less important amino acids were substituted in order to stabilize a helical structure. Some of these peptides inhibited GPCR phosphorylation by 80% [67]. Together these studies show that in diverse proteins and in diverse types of experimental manipulation, top-ranked ET residues consistently identify the key determinants of functional sites. They should therefore be useful for 3D functional motifs to annotate function in novel protein structures.

ETA functional annotation

In order to annotate function of novel protein structures solved by structural genomics, ET annotation (ETA)

follows the 3D motifs strategies reviewed above. Uniquely, this approach repeatedly exploits ET patterns to select motifs and to filter acceptable matches. ETA applies ET ranks to the structure of an unknown protein, the query, to identify six best clustering, top-ranked ET residues at or near a protein structure's surface: the 3D template. Simple geometric matches of such templates to protein structures of known function, the targets, often prove too non-specific to suggest identical functions accurately. However, false positives can be reduced dramatically by requiring that the matched sites in the target be composed of top-ranked residues [76]; that a 3D template from the target reciprocally match the query [77]; and that a plurality of targets concur in suggesting the same function [76]. If so, this functional annotation may be reliably transferred to the query in high throughput fashion, with 92% accuracy for enzymes at three-digit EC numbers; and 94% accuracy for non-enzymes at the third GO depth level in over a thousand Structural Genomics protein controls [78]. These studies confirm, on a large scale, that phylogenetic residue variation patterns convey highly specific structure–function information.

A recent extension of ETA exploits graph-based semi-supervised learning to improve function annotation specificity and coverage. The approach ties all-against-all ETA matches among all known protein structures into a network, in which nodes represent protein structures and links indicate ETA 3D structural template matches between proteins [79•]. Labels that indicate function are then diffused globally following the topology of this network. Although all labels reach nearly all nodes, only a fraction does so with any statistical significance. This global analysis improves accuracy by 6% (to 96% accuracy) at 65% coverage over all four EC numbers compared to ETA, and it also performs favorably against other methods [54]. As further validation, a novel and nontrivial ETA network annotation was experimentally confirmed as a carboxylesterase (EC 3.1.1.1) in a vancomycin resistant strain of *Staphylococcus aureus* [79•]. This annotation was based on matches to three structures with sequence identities ranging between 11 and 13%. These data show that global comparison of phylogenetic variations patterns of 6 residues, in a well-defined structural arrangement, uncovers accurate and specific functional information, including the resolution of substrate specificity, far into the twilight zone of protein sequence similarity.

Conclusions

The relationship between sequence, structure and function is part of the broad effort to understand how genotype is linked to phenotype. Some approaches rely on biophysical modeling and others are purely experimental. However, because genotype information is constantly in flux and a gene's survival depends on the fitness that it encodes, evolutionary analysis is another central approach to understand how genotype relates to phenotype. The exponential dependence of deviations in structure and function as a result of deviations in sequence among homologs suggests that evolution proceeds smoothly following regular processes over long time periods. A challenge is to complement these statistical observations of evolutionary regularity with equally precise molecular level patterns that help to recover biological meaning from high throughput sequence, structure, and function data. This review shows that different approaches that compare sequences and structures, motifs and templates, correlations and phylogenetic classification are able to identify general patterns that contain precise information on molecular function.

Many of the benefits of each of these approaches are naturally contained in Evolutionary Trace analysis. This approach scores sequence positions by their relative evolutionary impact, as judged from the size of the evolutionary steps associated with their variations. Thus, residues are ranked by how well their own evolution correlated with the evolution of all other sequence positions, represented by the phylogenetic tree. Critically, residues with variations that correlate with root divergences are more important and have remarkable structural and functional properties: they cluster structurally; these clusters map functional sites; clustering quality correlates with functional site prediction; experimental mutations at top-ranked residues control function and specificity; and their mimicry enable the transfer of function to a peptide, or to other protein structures on a proteomic scale *in silico*. Thus, top-ranked ET residues embody features in the sequence, in the structure, in the protein function, and in the phylogeny that are reproducible as general across the proteome. This suggests that they capture basic patterns linking genotype to phenotype during evolution. To fully support this view, however, it remains to reframe evolutionary trace analysis in a formal and extensible framework to make explicit the genotype to phenotype relationship. Such a relationship might then, in turn, help clarify the impact of missense mutations on protein function.

Acknowledgements

We wish to thank Rhonald Lua and Eric Venner for helpful discussions, and gratefully acknowledge grant support from the National Institute of Health through R01GM079656 and R01GM066099, and from the National Science Foundation, through CCF 0905536, NSF DBI-0851393, CCF 0905536, as well as from the Cancer Prevention Research Institute of Texas, through CPRIT RP120258.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Res* 2010, **38**:D346-D354.

2. The Universal Protein Resource (UniProt): **Nucleic Acids Res**2009, **37**:D169-D174.

3. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata**. *Nucleic Acids Res* 2008, **36**:D475-D479.

4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25-29.

5. Kuznetsova E, Proudfoot M, Sanders SA, Reinking J, Savchenko A, Arrowsmith CH, Edwards AM, Yakunin AF: **Enzyme genomics: application of general enzymatic screens to discover new enzymes**. *FEMS Microbiol Rev* 2005, **29**:263-279.

6. Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P *et al.*: **Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins**. *PLoS Biol* 2009, **7**:e96.

7. Devos D, Valencia A: **Intrinsic errors in genome annotation**. *Trends Genet* 2001, **17**:429-431.

8. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins**. *EMBO J* 1986, **5**:823-826.

9. Hegyi H, Gerstein M: **The relationship between protein structure and function: a comprehensive survey with application to the yeast genome**. *J Mol Biol* 1999, **288**:147-164.

10. Devos D, Valencia A: **Practical limits of function prediction**. *Proteins* 2000, **41**:98-107.

11. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction**. *Nat Protoc* 2010, **5**:725-738.

12. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**:651-654.

13. Lemoine F, Lespinet O, Labedan B: **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data**. *BMC Evol Biol* 2007, **7**:237.

14. Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, Babbitt PC, Minshull J, Gerlt JA: **Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily**. *Biochemistry* 2003, **42**:8387-8393.

15. Bairoch A: **The ENZYME database in 2000**. *Nucleic Acids Res* 2000, **28**:304-305.

16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

17. Rost B: **Enzyme function less conserved than anticipated**. *J Mol Biol* 2002, **318**:595-608.

18. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes**. *BMC Bioinformatics* 2004, **5**:178.

19. Chitale M, Hawkins T, Park C, Kihara D: **ESG: extended similarity group method for automated protein function prediction**. *Bioinformatics* 2009, **25**:1739-1745.

20. Sarac OS, Atalay V, Cetin-Atalay R: **GOPred: GO molecular function prediction by combined classifiers**. *PLoS ONE* 2010, **5**:e12382.

21. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation**. *Bioinformatics* 2007, **23**:1875-1882.

22. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment**. *Bioinformatics* 2001, **17**:700-712.

23. • Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2012, **40**:D290-D301.

With over 8000 citations, the Pfam database is used extensively to characterize protein domains based on sequence motif matches.

24. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Res* 2011, **39**:W29-W37.

25. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors**. *Brief Bioinform* 2002, **3**:265-274.

26. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S *et al.*: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic Acids Res* 2012, **40**:D306-D312.

27. • Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A *et al.*: **ELM — the database of eukaryotic linear motifs**. *Nucleic Acids Res* 2011, **40**:D242-D251.

ELM differs from other motifs database by focusing on functional regions without domain specific considerations.

28. Wass MN, Sternberg MJ: **ConFunc — functional annotation in the twilight zone**. *Bioinformatics* 2008, **24**:798-806.

29. Weingart U, Lavi Y, Horn D: **Data mining of enzymes using specific peptides**. *BMC Bioinformatics* 2009, **10**:446.

30. Arakaki AK, Huang Y, Skolnick J: **EFICAz2: enzyme function inference by a combined approach enhanced by machine learning**. *BMC Bioinformatics* 2009, **10**:107.

31. Ma B, Elkayam T, Wolfson H, Nussinov R: **Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces**. *Proc Natl Acad Sci U S A* 2003, **100**:5772-5777.

32. Holm L, Rosenstrom P: **Dali server: conservation mapping in 3D**. *Nucleic Acids Res* 2010, **38(Suppl)**:W545-W549.

33. Zhang Y, Skolnick J: **TM-align: a protein structure alignment algorithm based on the TM-score**. *Nucleic Acids Res* 2005, **33**:2302-2309.

34. • Veeramalai M, Ye Y, Godzik A: **TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model**. *BMC Bioinformatics* 2008, **9**:358.

TOPS++FATCAT is a speedy search for structural neighbors against either a filtered PBD or CASP database.

35. Hasegawa H, Holm L: **Advances and pitfalls of protein structural alignment**. *Curr Opin Struct Biol* 2009, **19**:341-348.

36. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A *et al.*: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution**. *Nucleic Acids Res* 2007, **35**:D291-D297.

37. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **Data growth and its impact on the SCOP database: new developments**. *Nucleic Acids Res* 2008, **36**:D419-D425.

38. Friedberg I, Godzik A: **Functional differentiation of proteins: implications for structural genomics**. *Structure* 2007, **15**:405-415.

39. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function**. *Protein Sci* 1996, **5**:2438-2452.

40. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA: **Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure**. *PLoS Comput Biol* 2009, **5**:e1000585.

41. Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S: **FTSite: high accuracy detection of ligand binding sites on unbound protein structures**. *Bioinformatics* 2012, **28**:286-287.

42. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation**. *BMC Struct Biol* 2006, **6**:19.

43. • Brylinski M, Skolnick J: **FINDSITE: a threading-based approach to ligand homology modeling**. *PLoS Comput Biol* 2009, **5**:e1000405.
FINDSITE predicts binding sites, ligands, and functional annotations using homology models of structures in complex with ligands.

44. Wass MN, Sternberg MJ: **Prediction of ligand binding sites using homologous structures and conservation at CASP8**. *Proteins* 2009, **77(Suppl 9)**:147-151.

45. Tseng YY, Dundas J, Liang J: **Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns**. *J Mol Biol* 2009, **387**:451-464.

46. Wallace AC, Laskowski RA, Thornton JM: **Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases**. *Protein Sci* 1996, **5**:1001-1013.

47. Meng EC, Polacco BJ, Babbitt PC, Rigden DJ: **3D Motifs**. In *From Protein Structure to Function with Bioinformatics.* Edited by Rigden DJ. Netherlands: Springer; 2009:187-216.

48. Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M: **FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures**. *BMC Bioinformatics* 2008, **9(Suppl 2)**:S2.

49. Polacco BJ, Babbitt PC: **Automated discovery of 3D motifs for protein function annotation**. *Bioinformatics* 2006, **22**:723-730.

50. Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C: **The SuMo server: 3D search for protein functional sites**. *Bioinformatics* 2005, **21**:3929-3930.

51. Goyal K, Mohanty D, Mande SC: **PAR-3D: a server to predict protein active site residues**. *Nucleic Acids Res* 2007, **35**:W503-W505.

52. Stark A, Russell RB: **Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures**. *Nucleic Acids Res* 2003, **31**:3341-3344.

53. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data**. *Nucleic Acids Res* 2004, **32**:D129-D133.

54. Redfern OC, Dessailly BH, Dallman TJ, Sillitoe I, Orengo CA: **FLORA: a novel method to predict protein function from structure in diverse superfamilies**. *PLoS Comput Biol* 2009, **5**:e1000485.

55. Eisen JA, Sweder KS, Hanawalt PC: **Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions**. *Nucleic Acids Res* 1995, **23**:2715-2723.

56. •• Lee DA, Rentzsch R, Orengo C: **GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains**. *Nucleic Acids Res* 2010, **38**:720-737.
GeMMA provides high throughput classification of functional subfamilies through pattern recognition and clustering of local sequence conservation.

57. Brown DP, Krishnamurthy N, Sjolander K: **Automated protein subfamily identification and classification**. *PLoS Comput Biol* 2007, **3**:e160.

58. •• Rappoport N, Karsenty S, Stern A, Linial N, Linial M: **ProtoNet 6. 0: organizing 10 million protein sequences in a compact hierarchical family tree**. *Nucleic Acids Res* 2012, **40**:D313-D320.
ProtoNet lets users traverse hierarchically classified and annotated sequences in search of functional information.

59. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics**. *PLoS Comput Biol* 2005, **1**:e45.

60. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE: **Genome-scale phylogenetic function annotation of large and diverse protein families**. *Genome Res* 2011, **21**:1969-1980.

61. Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L *et al.*: **TreeFam: a curated database of phylogenetic trees of animal gene families**. *Nucleic Acids Res* 2006, **34**:D572-D580.

62. Huerta-Cepas J, Bueno A, Dopazo J, Gabaldon T, Phylome DB: **a database for genome-wide collections of gene phylogenies**. *Nucleic Acids Res* 2008, **36**:D491-D496.

63. Lichtarge O, Bourne HR, Cohen FE: **Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex**. *Proc Natl Acad Sci U S A* 1996, **93**:7507-7511.

64. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O: **Structural clusters of evolutionary trace residues are statistically significant and common in proteins**. *J Mol Biol* 2002, **316**:139-154.

65. Baranski TJ, Herzmark P, Lichtarge O, Gerber BO, Trueheart J, Meng EC, Iiri T, Sheikh SP, Bourne HR: **C5a receptor activation. Genetic identification of critical residues in four transmembrane helices**. *J Biol Chem* 1999, **274**:15757-15765.

66. Wilkins AD, Lua R, Erdin S, Ward RM, Lichtarge O: **Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation**. *Protein Sci* 2010, **19**:1296-1311.

67. Baameur F, Morgan DH, Yao H, Tran TM, Hammitt RA, Sabui S, McMurray JS, Lichtarge O, Clark RB: **Role for the regulator of G-protein signaling homology domain of G protein-coupled receptor kinases 5 and 6 in beta 2-adrenergic receptor and rhodopsin phosphorylation**. *Mol Pharmacol* 2010, **77**:405-415.

68. Mihalek I, Res I, Lichtarge O: **A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins**. *Bioinformatics* 2006, **22**:149-156.

69. Aloy P, Querol E, Aviles FX, Sternberg MJ: **Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking**. *J Mol Biol* 2001, **311**:395-408.

70. Gutteridge A, Bartlett GJ, Thornton JM: **Using a neural network and spatial clustering to predict the location of active sites in enzymes**. *J Mol Biol* 2003, **330**:719-734.

71. Ribes-Zamora A, Mihalek I, Lichtarge O, Bertuch AA: **Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions**. *Nat Struct Mol Biol* 2007, **14**:301-307.

72. •• Adikesavan AK, Katsonis P, Marciano DC, Lua R, Herman C, Lichtarge O: **Separation of recombination and SOS response in *Escherichia coli* RecA suggests LexA interaction sites**. *PLoS Genet* 2011, **7**:e1002244.
A long-sought RecA site that mediates LexA proteolysis, and thus triggers error-prone DNA repair, was found with ET.

73. Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O: **Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions**. *J Biol Chem* 2004, **279**:8126-8132.

74. Shenoy SK, Drake MT, Nelson CD, Houtz DA, Xiao K, Madabushi S, Reiter E, Premont RT, Lichtarge O, Lefkowitz RJ: **beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor**. *J Biol Chem* 2006, **281**:1261-1273.

75. •• Rodriguez GJ, Yao R, Lichtarge O, Wensel TG: **Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors**. *Proc Natl Acad Sci U S A* 2010, **107**:7787-7792.
This work demonstrates allosteric pathway specificity: single point mutations of cognate ET residues create serotonin responsive mutant receptor mutants with wild type binding affinity to either dopamine or serotonin.

76. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O: **Prediction of enzyme function based on 3D templates of evolutionarily important amino acids**. *BMC Bioinformatics* 2008, **9**:17.

77. Ward RM, Erdin S, Tran TA, Kristensen DM, Lisewski AM, Lichtarge O: **De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features**. *PLoS ONE* 2008, **3**:e2136.

78. Erdin S, Ward RM, Venner E, Lichtarge O: **Evolutionary trace annotation of protein function in the structural proteome**. *J Mol Biol* 2010, **396**:1451-1473.

79. Venner E, Lisewski AM, Erdin S, Ward RM, Amin SR, Lichtarge O:
• **Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities**. *PLoS ONE* 2010, **5**:e14286.
A diffusion model was applied to a protein networks of local structural and evolutionary similarities in order to predict enzymatic function and substrate. Experiments documented accurate matches down to negligeable sequence identity, in the low teens.