Short Specialist Review

Functional prediction through phylogenetic inference and structural classification of proteins

Kimmen Sjölander and Chelsea Specht
University of California, Berkeley, CA, US

1. Functional classification of genes is a pressing need in the genome era

With tens of thousands of new genes being identified monthly, experimental determination of gene function for all new genes is not possible. Thus, computational prediction of gene function is an essential tool for modern biologists. The primary method of gene functional annotation employs transfer of annotation from the top hit in a database search. Because homology-based methods of function prediction have been shown to be prone to systematic errors (Galperin and Koonin, 1998; Koski and Golding, 2001), methods in recent years have focused on inclusion of information from evolutionary studies and structural analysis. In this paper, we present a perspective on the use of phylogenetic and structural analyses for improving the inference of protein function.

2. Structural and functional divergence in protein superfamilies

Evolutionary processes, particularly gene duplication and domain shuffling, produce protein superfamilies that often span a bewildering variety of functions and exhibit extreme structural and sequence divergence. Evolutionary pressures to conserve function (e.g., binding pocket positions that determine enzyme substrate specificity or receptor–ligand interaction) or structure (e.g., disulfide bridges in secreted proteins) can be relieved as gene duplication events permit the family as a whole to innovate novel functions and structures; one copy maintains the original function (with consistent constraints at key sites) while the other is enabled to achieve novel function (termed *neofunctionalization*) or to specialize for different tissue types or temporal needs (termed *subfunctionalization*). If not considered appropriately, the

resulting sequence divergence can cause complications in determining homology and interpreting phylogenetic signal. If the underlying assumptions of phylogenetic analyses are taken into consideration, however, they can provide powerful tools for inferring protein function.

3. Phylogenetics and protein superfamilies

The first assumption of a phylogenetic reconstruction is that the data being analyzed are *homologous*, that is, related by evolution. One of the major challenges in phylogenetic reconstruction of protein superfamilies is that not all positions are homologous across all family members; some regions of proteins may have been inserted or deleted over time (e.g., insertions at surface loops). These positions may not be easily distinguishable from distantly related regions that are homologous, albeit with low sequence similarity. It is critical to differentiate homologous and nonhomologous regions at the outset, perhaps by alignment masking protocols, else the phylogenetic analysis may be adversely affected. This remains a challenge in practice.

Other issues in phylogenetic analysis of protein superfamilies arise due to lineage-specific or site-specific rate variation. While most phylogenetic tree inference methods may allow site-specific rate variation, few attempt to model different types of conservation within different lineages at individual positions. Site-specific mutation rates can vary across sites within a lineage, and can be entirely different from rates observed in other lineages. Both lineage and site specificity are considered independently (Muse and Gaut, 1994; Yang, 1998; Yang and Nielsen, 1998; Nielsen and Yang 1998; Yang *et al*., 2000) and together (Yang and Nielsen, 2002) in a variety of models. These models may be appropriate when some positions are conserved within a particular lineage (e.g., substrate specificity determining residues) but vary across lineages. In other cases, residues observed to interact in the buried core of proteins often exhibit correlated or compensatory mutations. If appropriately incorporated in a phylogenetic analysis, these types of conservation can be used to help determine phylogenetic relatedness.

For example, G-protein coupled receptors (GPCRs) span dozens of different subtypes developed by multiple gene duplication events. Small modifications at key positions determine the ligands recognized by these receptors, for example, serotonin, dopamine, histamine, acetylcholine, chemical odorant signals, chemokines, and opiates. Residues at these positions determine the ligand-specificity of each protein, and thereby its molecular function. Phylogenetic methods that can identify these positions *a priori* could upweight these positions in tree topology estimation. The SATCHMO (Edgar and Sjolander, 2003) algorithm is an example of this type of phylogenetic inference approach.

Phylogenies, or evolutionary trees, are the fundamental structures that enable us to visualize evolution in a hierarchical and comparative framework. Molecular phylogenetic methods have been primarily developed for and validated with respect to inferring the evolutionary history of organisms and gene families based on DNA sequence data. When protein evolution has been examined, it has been in the context of a single group of orthologous genes. In recent years, the challenges of inferring

molecular function for hundreds of thousands of unknown genes have driven the application of phylogenetic approaches to elucidating the evolutionary history of protein superfamilies. Within a family of proteins, sequences with known functions can be grouped with proteins of unknown function in a phylogenetic context, and proteins that group together phylogenetically may be predicted to share a common function. This approach is called *phylogenomic inference of protein function* (Brown and Sjölander, 2006).

4. The objective of protein superfamily phylogenetic analysis

Unlike the primary objective of organismal phylogenetic analysis – elucidating the evolutionary history of an organism or lineage of organisms – the primary objective for most phylogenetic analyses of protein families is protein function prediction. Phylogenomic inference of protein function – the analysis of a protein sequence in the context of a larger protein family with potentially numerous functional subtypes – is a prime example of this type of analysis.

Phylogenomic inference of protein function has several distinct challenges. First, accurate phylogenomic inference depends on homologs with the same domain structure; this is not normally known for sequences found in database search. Most database searches return sequences that may have only local homologs (i.e., have a different domain architecture from the seed). (Brown and Sjölander, 2006). Second, many protein families span large evolutionary distances, resulting in significant structural divergence among homologs and reduced alignment accuracy (Baker and Sali, 2001). While closely related sequences (e.g., with pairwise identities >50%) are normally clustered into subtrees consistently by most phylogenetic tree methods, the branching order between these conserved clades can be very different from one phylogenetic tree method to the next. This is particularly problematic if function inference is based on the branching order observed in a single phylogenetic tree, and other phylogenetic trees for the same dataset are not examined.

In contrast to the issues involved in reconstructing phylogenies for groups of single orthologous genes, phylogenetic reconstruction of protein superfamilies has a greater degree of uncertainty and variability, due to choices made in gathering homologs, constructing a multiple sequence alignment (MSA), alignment masking, choice of phylogenetic method, and phylogenetic tree interpretation. We discuss each of these points in turn.

4.1. Selecting homologs

Accurate phylogenetic reconstruction requires adequate sampling of all the major evolutionary subgroups. Engineered sequences should be excluded, and sequences should be filtered to remove those not sharing the same overall domain structure. The Berkeley Phylogenomics Group FlowerPower server enables selection of

sequences that can be predicted to share the same domain structure as an input seed sequence.

4.2. Constructing a multiple sequence alignment

Phylogenetic analysis and alignment construction are integrally related: most phylogenetic analyses are based on input MSAs, which encode the primary source of evolutionary signal upon which the phylogenetic tree construction method operates. The MSA represents the primary statement of homology, such that each column of DNA sequence or amino acid data is considered to be derived from a common ancestor as a single evolutionary event. All methods for alignment therefore inherently make inferences about evolution, and errors in the input alignment can be positively misleading to the phylogenetic analysis.

Progressive alignment methods, such as ClustalW (Thompson *et al*., 1994), construct a "guide tree" developed by distance-based methods to determine the order in which sequences are aligned; alignments are fixed within subgroups as sequences are progressively aligned. Progressive methods are among the most popular, but are primarily appropriate for closely related sequences with few required *indel* characters. Assessment of sequence alignment accuracy relative to structural superposition of 3D solved protein structures shows alignment accuracy drops sharply for pairs of sequences with $<30\%$ identity. Datasets including such highly divergent sequences will require more advanced and computationally intensive alignment methods. More recently, iterative methods have been developed that allow sequences to realign and thereby improve the overall accuracy (e.g., MUSCLE (Edgar, 2004), ProbCons (Do *et al*., 2005), and MAFFT (Katoh *et al*., 2005).

4.3. Masking protocols

The general consensus among computational biologists is that masking protocols ought to be employed, so that ambiguously aligned columns, or regions of questionable homology, are excluded from the phylogenetic analysis. The relative pros and cons of different masking protocols are not well understood.

4.4. Selection of phylogenetic tree method

There are two main classes of molecular phylogenetic analysis: character-based (e.g., maximum parsimony (MP), maximum likelihood (ML) and Bayesian approaches) and distance methods (e.g., neighbor-joining). Distance methods work by estimating a matrix of pairwise distances between input sequence data. These distances are then used to construct a tree topology that is most consistent with all observed pairwise distances. Character-based approaches, by contrast, retain the character state information in estimating a phylogenetic tree that is based on some optimality criterion (i.e., MP searches for the shortest tree consistent with the data, and ML methods attempt to find the most likely tree given the data and some model

of evolution). Distance methods have the primary advantage of being computationally efficient, and are therefore often used by biologists estimating phylogenies for large datasets. However, they are not believed to be as accurate as character-based approaches in estimating tree topology and branch lengths (Yang, 1994), both of which are essential to accurate functional inference by phylogenetic analysis.

Despite inherent controversy, recent years have seen Bayesian methods making a large contribution to phylogenetic inference (Yang and Rannala, 1997; Li *et al*., 2000; Huelsenbeck *et al*., 2001). Controversy surrounds the selection of a prior and its overall contribution to the resulting tree topology; however, the use of Markov-chain Monte Carlo algorithms to rapidly explore tree space makes model-based phylogenetic reconstruction and determination of statistical topological support attainable, even for large and variable datasets. The ability of Bayesian methods to elucidate the evolutionary history of protein superfamilies has not been fully explored, but we believe this class of phylogenetic method has great potential in this area due to the probabilistic nature of the analyses combined with computational speed.

4.5. Assessing phylogenetic method accuracy

The accuracy of phylogenetic methods has traditionally been assessed by simulation studies. In a simulation study, the parameters of a phylogenetic model are predetermined (forming the "true" tree topology), and data is generated from those parameters. These data are then used as input to a phyologenetic tree estimation, and the estimated tree is compared to the "true" tree. The sensitivity of a phylogenetic method to violations of assumptions can then be assessed. Simulation studies have demonstrated the sensitivity of both distance and character-based methods to variations in evolutionary rates across taxa, to site-specific mutation rates and to shifts in evolutionary rates over time (see Swofford *et al*., 1996; Felsenstein, 2004).

4.6. How informative are these simulation studies to phylogenetic reconstruction of protein superfamilies, especially in cases of significant structural divergence due to gene duplication?

Structural studies have shown that as sequence similarity decreases, the degree of structural superposability (of solved 3D structures) also drops. Extremes of structural and sequence divergence is common in protein superfamilies. However, to our knowledge no simulation studies have assessed the impact of the types of evolutionary innovations and dramatic changes observed in multigene families on the accuracy of phylogenetic reconstruction. In fact, most simulation studies and comparisons of methods for phylogeny reconstruction are performed based on two assumptions; that the data being analyzed come from a single orthologous gene family and that the alignment is correct. The first assumption is automatically nullified for protein superfamilies, and studies of alignment accuracy under significant sequence divergence show that the second assumption must also be

assumed to be false. The degree to which errors in the input MSA affect the phylogenetic tree topology accuracy for large protein superfamilies is not known.

4.7. Using multiple approaches

Because different tree methods can produce dramatically different tree topologies for the same input MSA, careful attention must be paid to the assumptions made during phylogenetic inference. We recommend applying more than one discrete tree-building method to the data in order to better understand the data and their response to phylogenetic analysis. The wide range of available models provides tools for exploring data while developing a robust phylogenetic hypothesis. If well-supported nodes are in conflict when different tree-building algorithms are used, manipulations of the data (removing characters) and the taxa/proteins (removing taxa) may be necessary to test for underlying conflict in the data set. Davis *et al*. (1998) and Gatesy *et al*. (1999) give reviews of various means of testing for phylogenetic incongruence, data decisiveness, and hidden support in phylogenetic data in a parsimony framework, while Kishino and Hasegawa (1989) review support and conflict in a likelihood framework.

5. Assessing the accuracy of phylogenetic methods and phylogenetic trees

Both parametric and nonparametric means are used for assessing the amount of uncertainty in a phylogenetic tree. The likelihood ratio test (Goldman, 1993) provides a means of the testing assertions about the parameters of the chosen model of evolution, using the ML estimate and comparing it with alternative topologies that cannot be rejected based on their likelihood score. Interior branch tests (Nei *et al*., 1985; Li, 1989; Tajima, 1992; Rzhetsky and Nei 1992) use the variance on the estimate of the length of a branch in the interior of a tree to determine the reliability of the branch. Both of these are limited by the fact that they are parametric in nature and may underestimate the uncertainty for a given topology. Resampling techniques such as the bootstrap (resampling with replacement: Felsenstein, 1985) and jackknife (resampling without replacement: Mueller and Ayala, 1982; Farris *et al*., 1996), decay indices (Bremer, 1992; Baker *et al*., 1998) and data decisiveness tests (Davis *et al*., 1998) use empirical information about character variation to infer confidence in tree topology. In Bayesian analyses, support for relationships is determined by posterior probabilities of the identified nodes. These values tend to be inflated relative to bootstrap/jackknife values and may not be directly correlated with respect to indication of nodal support (Douady *et al*., 2003).

6. Computational efficiency

The complexity of protein superfamily evolution begs for more complex models of molecular evolution. Unfortunately, most protein families have large numbers

of taxa (in the hundreds or thousands) as well as significant sequence divergence (many pairs with <15% identity) while the sequence length is relatively short (at most a few hundred residues in length). This leaves very little actual information to estimate complex phylogenetic models. In addition, model complexity increases the risk of overparameterization, leading to decreased support for any phylogenetic hypothesis.

Many of the phylogenetic questions typically addressed by functional and structural biologists require supercomputers or clusters to run the algorithms necessary to adequately align the data and construct a reliable phylogenetic tree in a reasonable amount of time. Bayesian methods promise to reduce the time necessary for a model-based approach using Markov-chain Monte Carlo simulations of character space.

While these evolutionary reconstructions of protein superfamilies are undoubtly computationally demanding, we believe that the need for function prediction accuracy warrants the time and resources spent on phylogenetic reconstruction. The field as a whole will benefit from simulation studies specifically designed to assess the expected accuracy of existing methods for protein superfamily reconstruction.

7. The role of structure prediction and analysis in protein function prediction

Structural analysis efforts and structural classification databases often have a similar intent: enabling biologists to infer molecular function on the basis of structural similarities. For example, SCOP and CATH enable biologists to make inferences of protein function based on structural similarity, and the DALI and VAST servers provide on-the-fly identification of structural neighbors for newly solved structures. A variety of webservers enable prediction of structural domains through the use of profile or hidden Markov model approaches (e.g., 3d-pssm, the NCBI Conserved Domain Database, and Superfamily). Several methods that integrate structural and phylogenetic analyses have been successfully used to predict functional epitopes in proteins (Lichtarge *et al*., 1996; Glaser *et al*., 2003). A number of protein sequence databases serve as repositories of vast amounts of data and associated bioinformatics predictions; these can be extremely valuable in inferring function and structure for proteins (e.g., UniProt, InterPro, SMART, and PhyloFacts).

8. Website references

http://phylogenomics.berkeley.edu/UniversalProteome/	PhyloFacts Universal Proteome Explorer
http://phylogenomics.berkeley.edu/resources/	Other Berkeley Phylogenomics Group resources, including FlowerPower, SATCHMO, and SCI-PHY
http://www.pantherdb.org	Celera Panther tools

http://smart.embl-heidelberg.de/	SMART
http://www.ncbi.nlm.nih.gov/	NCBI
http://www.uniprot.org/	UniProt
http://www.ebi.ac.uk/interpro/	InterPro
http://www.sanger.ac.uk/Software/Pfam/	PFAM HMM library at the Sanger Institute
http://www.rcsb.org/pdb/	Protein Data Bank (PDB)
http://scop.berkeley.edu/	Structural Classification of Proteins (SCOP)
http://cathwww.biochem.ucl.ac.uk/	CATH
http://ekhidna.biocenter.helsinki.fi/dali/	DALI/FSP
http://www.sbg.bio.ic.ac.uk/3dpssm/	3d-pssm
http://supfam.org/SUPERFAMILY/	Superfamily

Acknowledgments

This work was supported in part by Grant #0238311 from the National Science Foundation, and by Grant #R01 HG002769-01 from the National Institutes of Health to KS.

References

Baker D and Sali A (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Baker R, Yu XB and DeSalle R (1998) Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. *Molecular Phylogenetics and Evolution*, **9**, 427–436.

Bremer K (1992) Branch support and tree stability. *Cladistics*, **10**, 295–304.

Brown D and Sjölander K (2006) Functional classification using phylogenomics inference. *PLOS Computational Biology*, **2**, e77.

Davis JI, Simmons MP, Stevenson DW and Wendel JF (1998) Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from monocotyledons using mitochondrial atpA sequences. *Systematic Biology*, **47**, 282–310.

Do CB, Mahabhashyam MS, Brudno M and Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF and Douzery EJP (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, **20**, 248–254.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edgar RC and Sjölander K (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.

Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, **12**, 99–124.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Felsenstein J (2004) *Inferring Phylogenies*, Sinauer Associates: Sunderland, MA.

Galperin MY and Koonin EV (1998) Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biology*, **1**, 55–67.

Gatesy J, O'Grady P and Baker R (1999) Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*, **15**, 271–313.

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E and Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.

Goldman M (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Huelsenbeck JP, Ronquist F, Nielsen R and Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Katoh K, Kuma K, Miyata T and Toh H (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform Ser Workshop Genome Inform*, **16**, 22–33.

Kishino H and Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, **29**, 170–179.

Koski LB and Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, **52**, 540–542.

Li S, Pearl DK and Doss H (2000) Phylogenetic tree construction using Markov Chain Monte Carlo. *Journal of the American Statistical Association*, **95**, 493–508.

Li WH (1989) A statistical test of phylogenies estimated from sequence data. *Molecular Biology and Evolution*, **6**, 424–435.

Lichtarge O, Bourne HR and Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, **257**, 342–358.

Mueller LD and Ayala FJ (1982) Estimation and interpretation of genetic distance in empirical studies. *Genetical Research*, **40**, 127–137.

Muse S and Gaut BS (1994) A likelihood method for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, **11**, 715–724.

Nei M, Stephens JC and Saitou N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution*, **2**, 66–85.

Nielsen R and Yang Z (1998) Likelihood models for detecting positively selected amino acids sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Rzhetsky A and Nei M (1992) Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, **35**, 367–375.

Swofford DLGJ, Olsen PJ, Waddell DM Hillis (1996) *Phylogenetic Inference in Molecular Systematics*, Second Edition, Hillis DM, Moritz C and Mabel BK (Eds.), Sinauer Associates: Sunderland, MA.

Tajima F (1992) Statistical method for estimating the standard errors of branch lengths in a phylogenetic tree reconstructed without assuming equal rates of nucleotide substitution among different lineages. *Molecular Biology and Evolution*, **9**, 168–181.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Yang Z (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, **43**, 329–342.

Yang Z (1998) On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, **47**, 125–133.

Yang Z and Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, **46**, 409–418.

Yang Z and Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**, 908–917.

Yang Z, Nielsen R, Goldman N and Pedersen A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang Z and Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.