BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 449–462 doi:10.1093/bib/bbr042
Advance Access published on 27 August 2011

Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium

Pascale Gaudet, Michael S. Livstone, Suzanna E. Lewis and Paul D. Thomas

Submitted: 8th March 2011; Received (in revised form): 8th June 2011

Abstract

The goal of the Gene Ontology (GO) project is to provide a uniform way to describe the functions of gene products from organisms across all kingdoms of life and thereby enable analysis of genomic data. Protein annotations are either based on experiments or predicted from protein sequences. Since most sequences have not been experimentally characterized, most available annotations need to be based on predictions. To make as accurate inferences as possible, the GO Consortium's Reference Genome Project is using an explicit evolutionary framework to infer annotations of proteins from a broad set of genomes from experimental annotations in a semi-automated manner. Most components in the pipeline, such as selection of sequences, building multiple sequence alignments and phylogenetic trees, retrieving experimental annotations and depositing inferred annotations, are fully automated. However, the most crucial step in our pipeline relies on software-assisted curation by an expert biologist. This curation tool, Phylogenetic Annotation and INference Tool (PAINT) helps curators to infer annotations among members of a protein family. PAINT allows curators to make precise assertions as to when functions were gained and lost during evolution and record the evidence (e.g. experimentally supported GO annotations and phylogenetic information including orthology) for those assertions. In this article, we describe how we use PAINT to infer protein function in a phylogenetic context with emphasis on its strengths, limitations and guidelines. We also discuss specific examples showing how PAINT annotations compare with those generated by other highly used homology-based methods.

Keywords: *gene ontology; genome annotation; reference genome; gene function prediction; phylogenetics*

INTRODUCTION

The Gene Ontology (GO) project [1, 2] is a collaborative effort among multiple groups to develop a standardized and shared approach for describing biology in a species-independent manner. The ontology itself contains over 32 000 terms describing the sub-cellular localization [Cellular Component (CC): ~3000 terms], biochemical activity [Molecular Function (MF): ~9000 terms] and participation in larger processes [Biological Process (BP), ~20 000 terms] of proteins and other gene products. Each term is defined and placed in a directed acyclic graph with relations between terms: is a (for subclasses), part of and regulates. For example, superoxide dismutase (SOD) proteins are annotated with the term 'SOD

Corresponding author. Pascale Gaudet, CALIPHO group, Swiss Institute for Bioinformatics, CMU, 1 Rue Michel Servet, 1211 Geneva 4, Switzerland. E-mail: pascale.gaudet@isb-sib.ch

Pascale Gaudet is the manager of the Reference Genome Project of the Gene Ontology Consortium.

Mike Livstone is a curator in Princeton University's Genome Databases Group. In addition to the Reference Genome Project, Mike is a contributor to the BioGRID interaction database, the Princeton Protein Orthology Database (P-POD), and the *Saccharomyces* Genome Database (SGD).

Suzanna E. Lewis is a scientist at the Lawrence Berkeley National Laboratory and head of the Berkeley Bioinformatics Open-source Projects, involved in a number of projects, including the Gene Ontology, OBO Foundry, the Phenotypic Quality Ontology, modENCODE, and the Generic Model Organism Database Project.

Paul D. Thomas is associate professor and heads the Division of Bioinformatics in the Department of Preventive Medicine at the University of Southern California. His research interests lie in the evolution of genes and the biological systems they encode, and their involvement in disease.

activity' (MF, GO:0004784), which is a subclass of 'antioxidant activity' (GO:0016209); SOD proteins are also described by the term 'removal of superoxide radicals' (BP, GO:0019430) and—for different members of the family—the CC terms 'mitochondrion' (CC, GO:0005739) or 'extracellular space' (GO:0005615). For a recent review on GO (see du Plessis *et al.*, 2011 [3]). The GO database contains nearly 3 million annotations to over 466 000 proteins (In this article, we will generally refer to gene products simply as 'proteins', although the overwhelming majority of statements will apply to the various types of RNA gene products and protein complexes as well).

GO annotations are assigned using either of two general approaches: based on direct experimental results or by sequence analysis. In the experiment-based approach, biocurators make annotations that record the results of experimental work published in the biomedical literature. There are 375 000 experiment-based annotations in the GO database to more than 81 000 proteins. While these annotations describe proteins from over 900 different species, most of the data come from a small number of well-studied model organisms. As shown in Table 1, only 20 species have more than 1000 experiment-based GO annotations. The second annotation approach, sequence-based, uses bioinformatics techniques to infer a likely function for uncharacterized proteins from sequence characteristics. These can include short sequence motifs that can evolve by both convergent and divergent evolution (e.g. mitochondrial targeting sequences or helical transmembrane domains), or long regions of sequence similarity between two proteins that can only be reasonably explained by divergence from a common ancestor (homology).

The overwhelming majority of sequences in public databases remain experimentally uncharacterized, a trend which is increasing rapidly with the ease of modern sequencing technologies. To give a rough idea of the disparity between characterized and uncharacterized sequences, there are ~15 million protein sequences in the UniProt database that are candidates for annotation, while, as previously noted, only 81 000 (0.3%) have been annotated with a GO term based on experimental evidence. It is therefore indispensable to develop powerful and reliable methods for predicting protein function.

Table 1: Species with more than 1000 experimentally-based annotations (evidence codes: EXP, IDA, IEP, IMP, IGI and IPI[a])

Species name	Number of annotations based on experimental data
Mus musculus	54 131
Homo sapiens	53 428
Caenorhabditis elegans	50 291
Arabidopsis thaliana	37 367
Rattus norvegicus	32 320
Saccharomyces cerevisiae	29 169
Drosophila melanogaster	24 332
Mycobacterium tuberculosis	23 861
Schizosaccharomyces pombe	14 708
Danio reiro	9442
Escherichia coli str. K-12	6684
Candida albicans	5244
Dictyostelium discoideum	4350
Xenopus laevis	3720
Emericella nidulans	2307
Sus scrofa	1779
Magnaporthe grisea	1673
Oryctolagus cuniculus	1250
Thermoplasma acidophilum	1093
Pseudomonas aeruginosa PAO1	1081

[a] See http://geneontology.org/GO.evidence.shtml for evidence codes description.

The GO Consortium coordinates an effort to maximize the utility of a large and representative set of key genomes, which we refer to as reference genomes. The Reference Genome project has two aspects: (i) to encourage complete and precise annotations of the proteins for the species widely used as model organisms; and (ii) to provide inferred annotations for proteins for which no experimental data are available [4]. We describe here the homology-based method and software we have developed to achieve those goals.

Function inference by homology: theory and implementation in PAINT

Our method starts by treating each gene function (in this case, a GO term, or group of related terms) as a 'character', in the standard sense used for evolutionary inference [5]. These functional characters are not used to reconstruct the phylogeny of each gene family (amino acid or nucleotide sequence characters are used for that purpose as described above). Rather, given the phylogeny, and the known functions of some subset of the extant genes (leaves of the tree), the goal is to reconstruct the functional evolution events (e.g. gain, loss and inheritance) that most likely led to the functions observed in extant sequences. We have developed a software application, called Phylogenetic Annotation and Inference Tool (PAINT), which allows a biocurator to implement this explicit phylogenetic paradigm. In PAINT, gain and loss events are represented as annotations of ancestral nodes in the phylogenetic tree. Inheritance of an annotation from each ancestor to its descendants is then automatically inferred to occur unless stopped by an explicit annotation of a loss event. This inheritance enables the inference of GO annotations for extant sequences that have not been characterized experimentally. In short, our process represents homology inference in terms of a gene family-specific model of the evolution of function within that family.

Our general approach is similar to the 'phylogenomic' method proposed by Eisen [6] and further developed into a probabilistic form by Engelhardt *et al.* [7], but with important differences. Eisen proposed a conceptual approach for predicting protein function using a phylogenetic tree together with available experimental knowledge of proteins. The original approach relied on manual curation to identify gene duplication events and to find and assimilate the literature for characterized members of the family. Engelhardt *et al.* used automated reconciliation with the species tree [8] to identify gene duplication events, and experimental GO terms (MF only) to capture the experimental literature. Using this information, they defined a probabilistic model of evolution of MF involving transitions between different molecular functions.

From these previous studies, we adopt the basic approach of function evolution through a phylogenetic tree and the use of GO annotations to represent function. However, unlike these other phylogenomic methods, we represent the evolution in terms of discrete gain and loss events. In Eisen's original model, an annotation does not necessarily represent a gain of function (it could have been inherited from an earlier ancestor), and losses are not explicitly annotated. The transition-based model of Engelhardt *et al.* assumes replacement of one function by another (gain of one function coupled to the loss of another), and does not capture uncoupled events, which is particularly important for BP annotations and cases where a protein has multiple molecular functions (see examples below). In addition, we make no a priori assumptions about conservation of function within versus between orthologous groups, or about the relationship between evolutionary distance and functional conservation (as the distance may not necessarily reflect every given function). While, as described below, gene duplication events and relatively long tree branches are important clues for curators to locate functional divergence (gain and/or loss), in our paradigm an ancestral function can be inherited by both descendants following a duplication (resulting in paralogs with the same function) or gained/lost by one descendant following a speciation event (resulting in orthologs with different functions). Evolution of each function is evaluated on a case-by-case basis, using many different sources of information about a given protein family.

METHODS AND RESULTS

Phylogenetic trees

The first element necessary for PAINT curation is the generation of phylogenetic trees to be annotated with functional evolution events. Currently we annotate the reference trees from the PANTHER database [9], which include protein-coding genes from all of the 12 GO Reference Genomes, plus an additional 36 fully sequenced genomes. The phylogenetic trees were constructed using the GIGA algorithm [10], which explicitly identifies gene duplication and speciation events. GIGA estimates relative branch lengths immediately following gene duplication events, as functional gain and loss events may be associated with an increased evolutionary rate due to adaptation or relaxation of selective constraints.

The PAINT curator interface

PAINT presents the biocurator with a phylogenetic tree and a multiple sequence alignment dynamically retrieved from the PANTHER database, and auxiliary information such as gene and protein names and identifiers. In addition it displays all the experimentally based annotations dynamically retrieved from the live GO database. PAINT allows querying and retrieval of protein family trees, multiple sequence alignments and sequence annotation data from the PANTHER database [9]. PAINT also provides linkouts to major databases displaying annotations of protein domains and sequence features such as active sites in UniProt records. These sequence features play an important role in the functional inference process, helping the curator to decide which nodes to annotate with functional gain and loss events. PAINT portrays duplications as square internal nodes and speciation events as circles, and estimates of evolutionary distances as different branch lengths. Curators use both duplication events and accelerated evolutionary rate as important pieces of evidence when attempting to identify and locate functional evolution events. GO annotations are represented in a matrix view to help the curator integrate experimentally based annotations from a wide range of organisms and to group annotations that are related in the ontology structure.

PAINT inference process

In PAINT, annotation transfer is an explicit two-step process (Figure 1). In the first step, a biocurator infers a GO annotation for an ancestral gene based on the GO annotations for the descendants of that gene. Each experiment-based GO term is treated as a different 'character', and the curator attempts to infer when each function most likely first evolved, capturing the inference as an annotation of the appropriate ancestral gene. Note that only experimental, experiment-based annotations can be used to support ancestral inferences. Thus, an ancestral gene can be annotated only with those functions that have been experimentally determined in at least one of its descendants. The power of this paradigm is that it enables experimental evidence from many sequences, and even across different aspects of the ontology, to be integrated into ancestral inferences. GO annotations are supported by evidence codes, as described on the GO consortium website (http://geneontology.org/GO.evidence.shtml).
PAINT records the annotation using an evidence code indicating that the annotation is inferred from biological descendant(s) (IBD), which is a subclass of inferred from sequence similarity, and captures the database identifiers of all the extant descendants with experimental data for the function as evidence for the ancestral annotation. Since GO is a directed acyclic graph, a protein annotated with a child term is implicitly annotated to its parent terms. Moreover, if an annotation is too specific for propagation, the annotator can choose to propagate a parent (less granular) term instead.

In the second step, PAINT automatically takes each curated annotation of an ancestral gene (from the first step), and propagates it by inheritance to all of the gene's descendants in the phylogenetic tree. For this step, PAINT uses an evidence code indicating that the annotation is inferred from biological

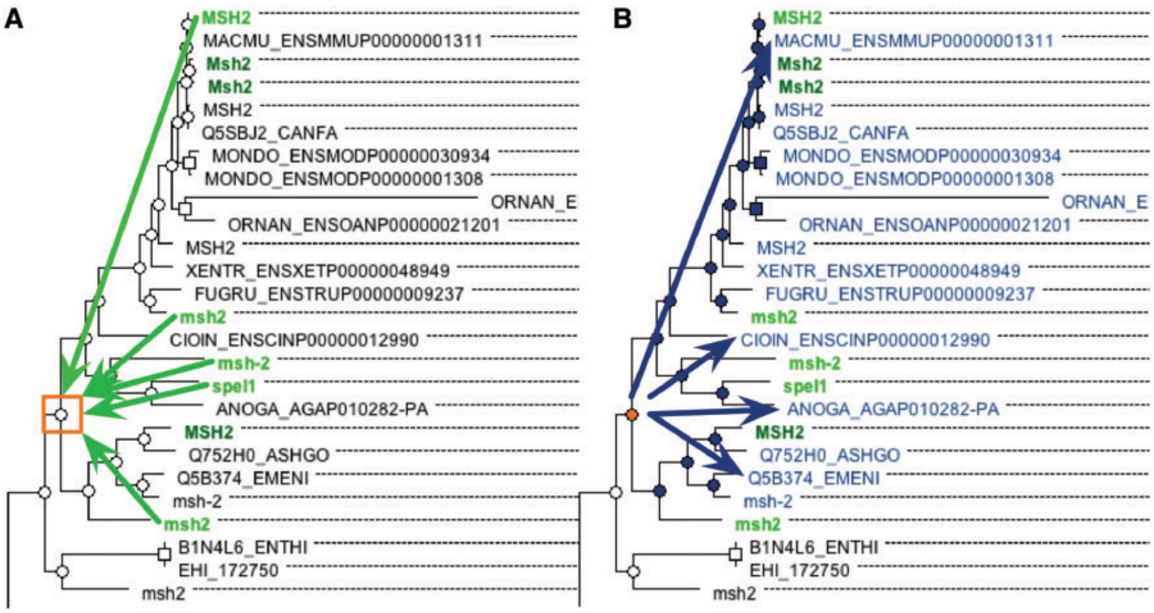


Figure 1: The concept of PAINT. This example presents a MutS homolog family showing experimental evidence for 'GO term'. (**A**) Primary experimentally based annotations to one term or any of its ancestors (light green labels) are used to infer that the most recent common ancestor (CA) of the all those proteins also had that function. The curator notes this by dragging the term onto the node of the MCRA (orange box). (**B**) Subsequently, PAINT propagated this annotation forward to other descendant leaves (blue labels).

ancestry (IBA), and captures the identifier of the ancestral gene from which the annotation was inferred as the basis of evidence. Curators can manually block propagation to descendants either by annotating a loss of the function at some point in the tree (loss of function is discussed below), or by removing a clade of sequences from the tree ('pruning'). Pruning is used when the curator believes the sequence(s) may be misplaced in the tree, or may not belong to the family at all.

Taken together, these two steps generate a complete evidence trail for each inferred annotation of an extant protein.

Functional evolution events captured by PAINT

The two 'elemental' functional evolution events we wish to capture in PAINT are gain of function and loss of function relative to an ancestor. PAINT annotates ancestral genes with these events, but the actual semantics are that the functional evolution occurred on the branch of the tree leading to the annotated node, rather than at the node itself, and may have occurred earlier.

More complex events are construed as the combined effects of gain and loss of function, often at gene duplications. Gene duplication provides an opportunity for functional divergence [11], so orthologs (genes that diverged via a speciation event) are often considered to be more likely than paralogs (genes that diverged via a gene duplication event) to inherit a function in common. However, this assumption continues to be debated [12]. Curators are particularly sensitive to the possibility for functional gains or losses in one or both duplicates when there is a gene duplication event in a protein family. However, they do not assume that orthologs have the same functions, nor do they assume that a particular ancestral function must be lost after a gene duplication event. Rather, to infer the most likely phylogenetic locations for functional evolution events, they integrate evidence from multiple sources, including GO and UniProtKB annotations, tree topology, sequence features (including active sites and protein domains), organismal biology, and evolutionary rates.

Gain of function

A gain of function is the addition of a function to a protein, while retaining its other existing functions. In PAINT, a biocurator is presented with all of the experiment-based GO annotations for the genes in a given family. For each annotation, the curator infers when in the evolutionary history of the family a given function was most likely to have first evolved, i.e. which ancestor 'gained' the function. This is recorded as an annotation of a gene at an internal node in the phylogenetic tree and means that the function is inferred to have evolved along the branch leading to that gene. The location of the inferred annotation determines the possible 'phylogenetic span' of the inferred annotations, since only direct descendants of the annotated ancestral gene can inherit that annotation. Gain of function may occur after a speciation event, meaning that orthologous genes will not share all functions in common. One example occurs in the MSH2 subfamily of PTHR11361, where a gene originally involved in recognizing DNA mismatches and recruiting the DNA repair machinery was co-opted in animals to regulate apoptosis and in vertebrates to mediate somatic hypermutation of immunoglobulin genes (Figure 2).

Loss of function

When a biological characteristic was lost during evolution, we annotate an ancestral (or extant) gene with the 'NOT' qualifier prefixed to the relevant annotation. 'NOT' annotations are inherited by descendants just like other GO annotations, in addition to preventing the inheritance of the corresponding positive annotation. 'NOT' annotations of ancestral genes must be supported by evidence, either: (i) an experiment-based annotation of a descendant sequence indicating it lacks this function; or (ii) absence of specific residues in the sequence, e.g. a missing active site residue; long branch lengths indicating rapid sequence evolution. Loss of function can be observed in the phosphoglucomutase (PGM) family, PTHR22573 in the PANTHER database (Figure 3). Based on the phylogeny and experimental annotations, phosphoglucomutase activity most likely evolved prior to the last universal common ancestor and is found in most eubacteria and eukaryotes. A gene duplication event in the vertebrate ancestor in this family resulted in two genes that would become PGM1 and PGM5 in humans. Both mouse and human PGM5 have been demonstrated experimentally to have lost phosphoglucomutase activity. These experimental annotations strongly suggest that the loss occurred before the mouse–human common ancestor, but how long

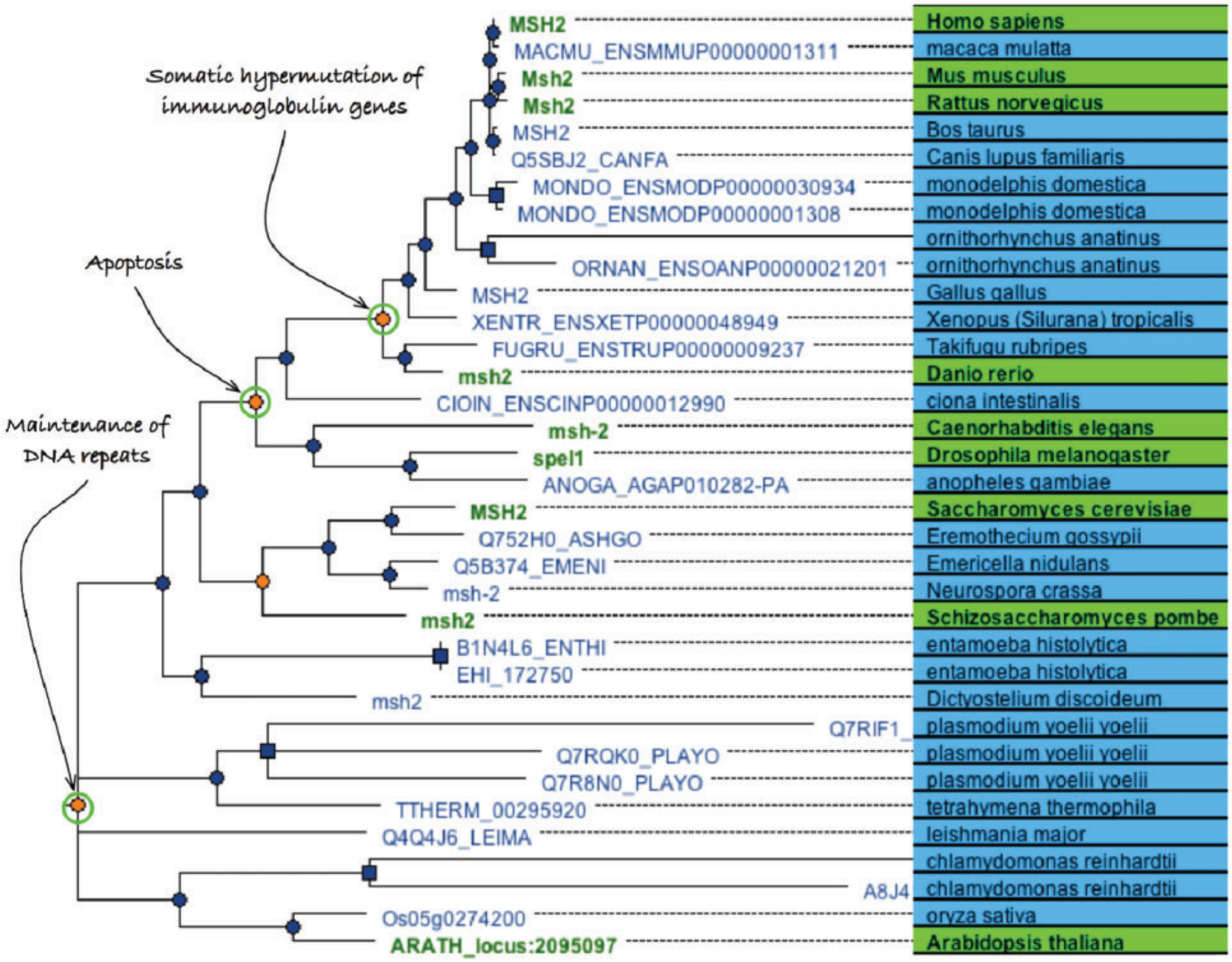


Figure 2: Gain of function. The MRCA of all eukaryotic MSH2 orthologs (leftmost orange circle) already likely functioned in DNA repair (inherited from LUCA, data not shown) and maintenance of DNA repeats. The gene was then coopted in the animal MRCA for a role in apoptosis, and later, in the vertebrate MRCA for a role in somatic hypermutation of immunoglobulin genes. Inferences for ancestral genes (orange circles) are based on experimental GO annotations for the genes shown in green, which are inferred by inheritance for descendants including uncharacterized genes in extant organisms shown in blue. Thus, the ortholog in *Bos taurus*, for example, will be annotated by PAINT with different functions than the ortholog in *Saccharomyces cerevisiae*.

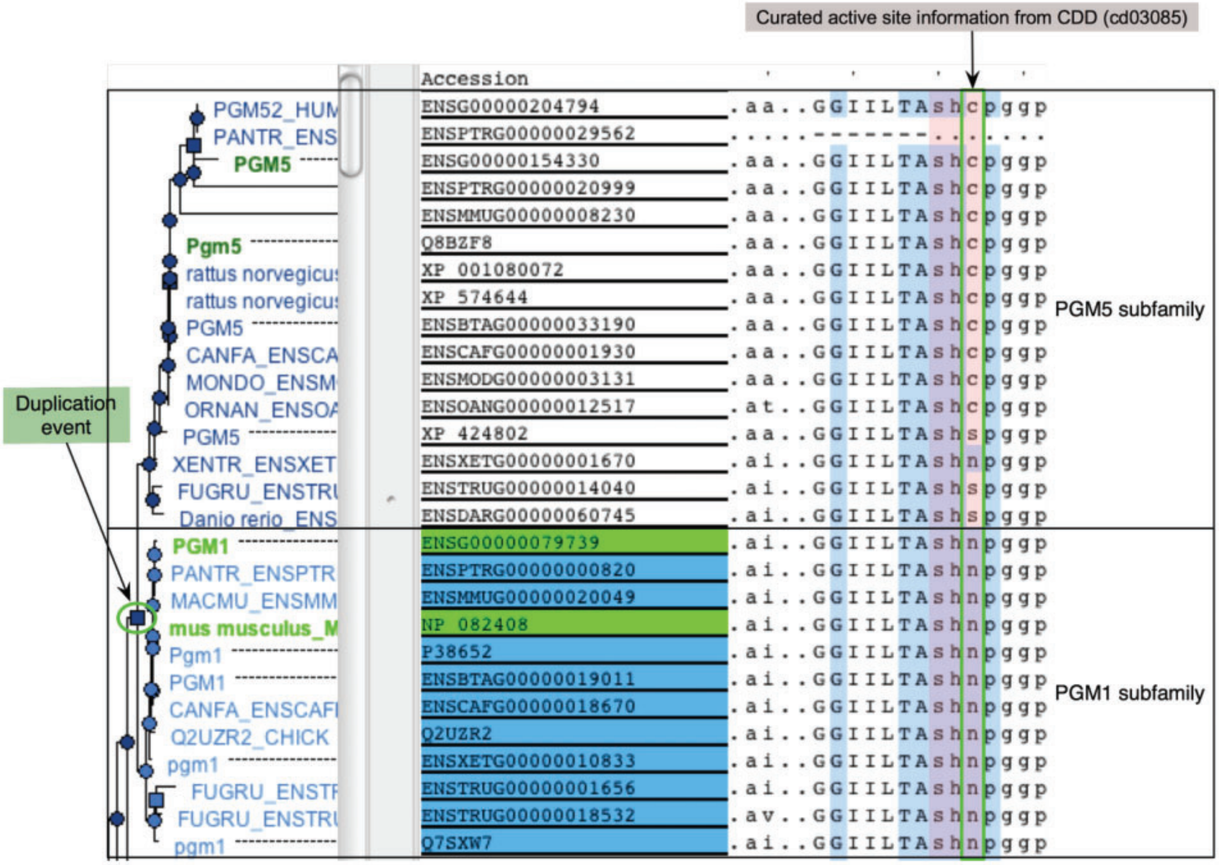
before? Based on active site mutations present in almost all of the vertebrate PGM5 proteins, the biocurator determined that the loss of function occurred in the vertebrate common ancestor.

Complex evolutionary events

More complicated phenomena can be represented as the combined or coordinated effects of gain and loss of function. Subfunctionalization, the partitioning of ancestral functions, is the loss of different ancestral functions in different descendants. Neofunctionalization is the loss of one function concomitant with the gain of another. Co-option, the use of an existing protein for a new purpose, can be viewed as a gain of function without losing ancestral functions. In our model, these events are represented in terms of more elemental gain and loss events. Importantly, the model allows us to capture the effects of these more complex events on gene function and homology inference, which is our main goal.

PAINT annotation guidelines

The PAINT curation process is a manual process based on manual annotations. To some extent, those manual procedures are subjective and subject to variability due to various factors such as the completeness of the annotations and differences in curators' expertise. Moreover, the manual annotations are extracted from the literature, which lacks



Accession	
ENSG00000204794	.aa..GGIILTAshcpggp
ENSPTRG00000029562	.....--------......
ENSG00000154330	.aa..GGIILTAshcpggp
ENSPTRG00000020999	.aa..GGIILTAshcpggp
ENSMMUG00000008230	.aa..GGIILTAshcpggp
Q8BZF8	.aa..GGIILTAshcpggp
XP 001080072	.aa..GGIILTAshcpggp
XP 574644	.aa..GGIILTAshcpggp
ENSBTAG00000033190	.aa..GGIILTAshcpggp
ENSCAFG00000001930	.aa..GGIILTAshcpggp
ENSMODG00000003131	.aa..GGIILTAshcpggp
ENSOANG00000012517	.at..GGIILTAshcpggp
XP 424802	.aa..GGIILTAshspggp
ENSXETG00000001670	.ai..GGIILTAshnpggp
ENSTRUG00000014040	.ai..GGIILTAshspggp
ENSDARG00000060745	.ai..GGIILTAshspggp
ENSG00000079739	.ai..GGIILTAshnpggp
ENSPTRG00000000820	.ai..GGIILTAshnpggp
ENSMMUG00000020049	.ai..GGIILTAshnpggp
NP 082408	.ai..GGIILTAshnpggp
P38652	.ai..GGIILTAshnpggp
ENSBTAG00000019011	.ai..GGIILTAshnpggp
ENSCAFG00000018670	.ai..GGIILTAshnpggp
Q2UZR2	.ai..GGIILTAshnpggp
ENSXETG00000010833	.ai..GGIILTAshnpggp
ENSTRUG00000001656	.ai..GGIILTAshnpggp
ENSTRUG00000018532	.av..GGIILTAshnpggp
Q7SXW7	.ai..GGIILTAshnpggp

Figure 3: Loss of Function. The active site residues of PGMI relatives have been annotated in the CDD database based on the 3D protein structure for PGM from *Paramecium tetraurelia*. In PAINT, the biocurator used the integrated multiple sequence alignment viewer to determine that key active site residues are mutated in all of the vertebrate PGM5 orthologs, suggesting that phosphoglucomutase activity was lost shortly after duplication. The biocurator correspondingly annotated the vertebrate ancestor of PGM5 with 'NOT phosphoglucomutase activity', which PAINT then propagated to all vertebrate orthologs of PGM5.

standardization in terms of experimental descriptions and data interpretation. This results in some inconsistencies even in the experiment-based annotations from which PAINT annotations are produced.

To increase the consistency and reproducibility of annotations, we have elaborated detailed annotation guidelines, available at http://wiki.geneontology.org/index.php/PAINT_SOP.

Overview of literature on protein family function and phylogeny

The first step in PAINT curation is to identify any published literature on the family as a whole (recent reviews are particularly helpful when available) and its phylogeny. These papers are reviewed and PubMed identifiers are recorded by the curator in the Notes box in PAINT.

Verification of the tree topology and composition

Next, the curator assesses the quality of the tree. PAINT displays orthologous clusters determined by OrthoMCL and imported from the PPOD database [13]. The curator verifies that the PANTHER tree topology is consistent with those orthologous clusters, and with any published phylogenetic analyses. Also, the curator verifies that no proteins that should obviously be in the family are missing; for example if all mammals have two paralogs of a gene, except for humans, the curator investigates whether an ortholog of this protein can be found in the public databases. In the rare cases where there are inconsistencies that may affect PAINT annotations, the phylogeny is reviewed and reconstructed again to resolve the issues. On the other hand, if the errors are small and do not affect the

PAINT annotations, proteins that are mistakenly groups in the family can be pruned (see above) either before or during curation.

Ensuring sufficient annotation coverage

One limitation of the PAINT curation process is the fact that for almost all model organisms, due to limited resources, not all proteins that have been experimentally characterized are completely annotated. Moreover, in several cases the most recent literature is annotated first, while the most basic functions of certain proteins might be known for decades. To address this, before beginning to annotate a protein family the curator reviews the relevant literature and skims the existing annotations. Based on this background knowledge, the PAINT curator may request curators from one or more of the GO Reference Genomes to assign additional experimental annotations before starting the annotation of the family.

Annotating ancestral genes

The decision process involved in making annotations using PAINT is shown in Figure 4. Step 1 is to determine which ancestor would be annotated based on the experiment-based annotations to a given term, or its related terms in the ontology. The initial hypothesis is that the term was inherited from a common ancestor, so PAINT assists in this process by automatically highlighting the node in the tree corresponding to the most recent common ancestor (MRCA) of all sequences annotated by experiment with a particular term or its children. The curator may adjust this ancestor by considering all additional annotations, either ones that are directly related by GO relations (such as class-subclass relations), or those that may be biologically related but in a different part or even aspect of the ontology.

Given this initial hypothesis, the curator needs to decide between three possibilities:

Option A: The initial hypothesis is likely to be correct, i.e. the MRCA of the experimentally annotated sequences is where it likely first evolved.

Option B. The actual annotation should be more ancient; in other words, the MRCA most likely inherited this function from a more ancient ancestor. In making this decision, the curator takes into account information such as duplication events/orthology, sequence conservation, the presence of essential/active site residues, branch length, and genes having inconsistent experimental annotations

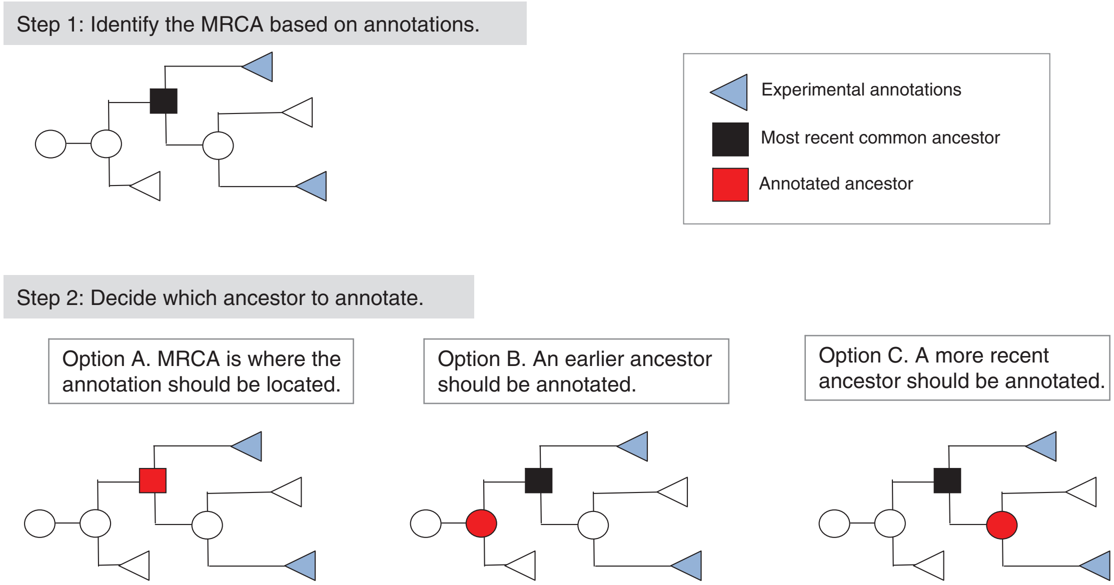

Step 1: Identify the MRCA based on annotations.
Experimental annotations
Most recent common ancestor
Annotated ancestor
Step 2: Decide which ancestor to annotate.
Option A. MRCA is where the annotation should be located.
Option B. An earlier ancestor should be annotated.
Option C. A more recent ancestor should be annotated.

Figure 4: General workflow for annotation of functional evolution events using PAINT. StepI: The curator uses experimental-based annotations to give an initial hypothesis that the function first appeared in the MRCA of all genes with a related experiment-based annotation. Step 2: The curator decides which ancestor is most appropriate for annotation: either the initially hypothesized MRCA (Option A); an earlier ancestor (Option B), meaning that the MRCA from Step I likely inherited its annotation from an earlier ancestor; or more recent ancestor(s) (Option C), meaning that there was homoplasy and the MRCA from Step I is not where the function first appeared.

(i.e. descendants with annotations, or missing annotations in well-characterized genes, that are most likely not compatible with the annotation). Determining compatibility or mutual-exclusivity of annotations requires careful curator judgment. Finally, the actual term propagated is also important: annotators are more conservative for BP annotations than for MF. Curators actively look for whether the data are consistent with functional divergence occurring after duplication events or long branches.

Option C. The annotation should be more recent, and probably arose more than once (homoplasy or convergent evolution). The curator considers this possibility to be more likely for functions that are mechanistically more likely to evolve convergently, such as targeting to the mitochondrion in eukaryotes (gain or loss of a relatively short N-terminal targeting peptide) or loss of an enzymatic function by substitutions in the active site. Again, conflicting annotations among descendants is helpful, and this, as well as assessing the likelihood of independent evolutionary events, requires curator judgment.

Achieving high specificity in annotation

Curators attempt to propagate the most specific term possible. For example, if a human protein is annotated to 'DNA binding' and its mouse ortholog is annotated to 'double-stranded DNA binding', the curator may infer, based on the evidence, that the human annotation refers to double-stranded DNA and may propagate the more specific term. Those types of annotation transfers may result in increasing levels of specificity of annotations, even for proteins already having experimentally supported annotations.

Avoiding over-propagation and uncertain statements

Molecular functions are usually more conserved than biological processes: for example, members of the MAP kinase family have 'protein kinase activity', but regulate a large number of varied processes. Therefore, the PAINT guidelines advise curators to be particularly conservative when annotating biological processes. This often means that cellular processes can be confidently transferred, and only very limited organismal processes may be transferred. Also, curators try not to propagate terms to ancestral organisms in which they are clearly inappropriate, such as 'nucleus' for a gene present in the last universal common ancestor (LUCA). GO has begun to perform taxonomic checks on annotations [14]. It is a high priority in the development of PAINT to integrate the taxonomic checks within the software.

Comparison with existing high-throughput methods of functional inference: case studies

The PAINT approach of constructing an explicit model of functional evolution, guided by a human curator, and using it to infer the functions of uncharacterized genes has some advantages over existing, fully automated sequence-based algorithms. Two highly used algorithms exemplify the two general approaches to automated function prediction by homology: family-based and close ortholog-based. In one protein family/motif-based approach, InterPro curators manually annotate groups of related sequences (either by family or domains) represented as a Hidden Markov model (HMM), with the functions they likely have in common, including GO terms [15]. The manually assigned GO terms for a family is automatically transferred to each protein belonging to the family. Since the GO assignments are automated, the evidence assigned for this is inferred from electronic annotation (IEA; GO_REF:0000002). This method is very accurate and rapid. The main limitation is that since families can contain very divergent sequences with divergent functions, the GO assignments tend to be to high level terms to avoid incorrect annotations.

In contrast, Compara [16] produces pairwise ortholog relationships among proteins from all sequenced vertebrate species, as well as a few important non-vertebrate species. GO annotations supported by experimental data from human and mouse are transferred automatically to other vertebrate species. To minimize false assignments, GO annotation transfers are limited to groups containing one-to-one orthologs (i.e. with no duplication events following speciation). As for InterPro2GO, since the step of assigning GO terms to proteins is automated the evidence code assigned is IEA (GO_REF:0000019).

We present two case studies to illustrate how PAINT compares those two high-throughput methods for annotation inference, summarized in Table 2. These examples were chosen because they are multigene families composed of several paralogous and orthologous groups. Annotations from Compara and InterPro were obtained from QuickGO in April 2011 GOA gene association file.

Table 2: GO annotations inferred for different human genes by InterPro2GO, Compara and PAINT

Human Gene	Aspect	InterPro2GO	Compara	PAINT
SODI	MF	Metal ion binding	SOD activity, chaperone binding	SOD activity, zinc ion binding, copper ion binding
	CC		Nucleus, cytoplasm, mitochondrion, neuronal cell body	Nucleus, cytosol, mitochondrion, extracellular region
	BP	Superoxide metabolic process, oxidation-reduction process,	Activation of MAPK activity, response to reactive oxygen species, ovarian follicle development, myeloid cell homeostasis, retina homeostasis, anti-apoptosis, spermatogenesis, aging, locomotory behavior, response to drug, 3l others	Removal of superoxide radicals
CCS	MF	Metal ion binding		SOD copper chaperone activity, zinc ion binding, copper ion binding, NOT SOD activity
	CC			Cytosol, mitochondrion, nucleus
	BP	Superoxide metabolic process, oxidation-reduction process, metal ion transport		Removal of superoxide radicals, intracellular copper ion transport
PGMI	MF	Magnesium ion binding, intramolecular transferase activity, phosphotransferases		Phosphoglucomutase activity
	CC			Cytosol
	BP	Carbohydrate metabolic process		Glycogen biosynthetic process, glucose-1-phosphate metabolic process
PGM5	MF	Magnesium ion binding, intramolecular transferase activity, phosphotransferases		NOT phosphoglucomutase activity
	CC		Spot adherens junction, Z disc, focal adhesion	Cytosol, spot adherens junction, Z disc, stress fiber, focal adhesion, intercalated disc
	BP	Carbohydrate metabolic process		NOT glycogen biosynthetic process, NOT glucose-1-phosphate metabolic process

These are arranged by aspect in the GO: MF, CC and BP.

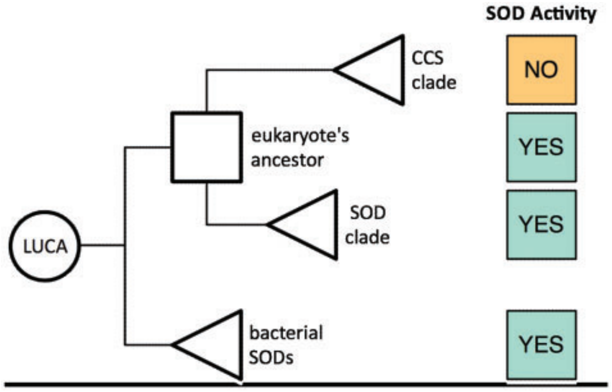


Figure 5: A simplified phylogeny of the SOD family (PTHR10003). The last universal common ancestor, LUCA, was duplicated in the ancestors to eukaryotes (square node). The descendents of the duplication that shows the least divergence from its ancestor also retained the SOD activity. That was lost in the CCS clade.

SOD1/CCS

We first consider two paralogous human genes from the SOD-related family, SOD1 and CCS. SOD1 encodes a SOD, and CCS is a copper 'chaperone' that delivers copper to SODs (Figure 5). InterPro2GO has annotated them both with the following GO terms: 'superoxide metabolic process, oxidation-reduction process, metal ion binding'.

These are the functions in common to all family members. InterPro2GO does not associate SOD1 with its mainMF, 'SOD' activity, because this function is not shared by all family members, in particular with the CCS clade.

Compara annotates SOD1 and CCS very differently

For SOD1, Compara makes 41 BP annotations, two MF annotations ('SOD activity' and 'chaperone binding') and five CC annotations. On the other hand, Compara does not make any annotations for CCS because CCS orthologs have not been characterized in the mouse or rat.

With the PAINT process, SOD1 was annotated with three molecular functions: 'SOD activity, zinc ion binding, copper ion binding', four CCs and one process: 'removal of superoxide radicals'. PAINT can capture the fact that SOD activity is present in only some family members, in contrast to InterPro2GO. PAINT curators chose to propagate fewer annotations that those transferred by Compara, especially those thought to be several steps downstream from the known molecular function, such as 'negative regulation of neuron apoptosis and spermatogenesis'.

For CCS, PAINT curation assigned three MF annotations: 'SOD copper chaperone activity, zinc ion binding, copper ion binding', three CC annotations and two BP annotations: 'removal of superoxide radicals and intracellular copper ion transport'. These annotations are more specific and complete than those for InterPro2GO, because PAINT is able to assign annotations to only a subset of the proteins in the family. In addition, PAINT explicitly records a negative annotation, 'NOT SOD activity', ensuring that the sequence similarity of CCS to SOD1 will not lead to erroneous functional inference.

PGM1/PGM5 family

We take our other examples from the phosphoglucomutase-related family. Human PGM1 encodes a functional phosphoglucomutase. Proto-PGM1 was duplicated prior to the vertebrate radiation and one copy evolved into PGM5, which as discussed above lost its phosphoglucomutase activity. Nevertheless, InterPro2GO annotates both PGM1 and PGM5 as 'magnesium ion binding' (MF), 'intramolecular transferase activity, phosphotransferases' (MF) and 'carbohydrate metabolic process' (BP). Compara does not annotate PGM1, but annotates PGM5 with three CC terms. PAINT annotation associates PGM1 with a number of additional CC terms, as well as 'phosphoglucomutase activity' (MF), and two biological processes ('glycogen biosynthetic process, glucose-1-phosphate metabolic process'), all of which provide greater specificity than InterPro2GO. PGM5, on the other hand, was annotated in PAINT with the same additional CC terms as Compara. In addition, PAINT curation provides several negative annotations arising from the loss of 'phosphoglucomutase activity'. In this way, PAINT avoids making the false positive assertions for PGM5 that are found in this case for InterPro2GO.

Each method has different advantages and limitations. Both PAINT and InterPro2GO benefit from (i) manual review by expert biocurators, allowing for selection of the experiment-based annotations used as the basis for homology inferences; and (ii) consideration of information about distantly related genes, allowing for additional annotations. However, when different family members have different functions, InterPro2GO can have incorrect, missing or less specific function predictions than PAINT, because PAINT is designed to capture functional divergence events. On the other hand, both PAINT and Compara benefit from the specificity of information about closely related genes, which has the advantage of providing very precise annotations when the function of a close ortholog is known. However, unlike PAINT, Compara will fail to annotate genes completely if additional functions have been characterized in more distantly related family members. The fact that PAINT makes inferences through ancestral sequences rather than in a pairwise manner, allows it to make precise assertions in a more flexible manner than either interPro2GO or Compara. To assess more precisely how PAINT compares with other methods, we plan to undertake a quantitative analysis once a sufficient number of families have been annotated.

Extending annotations to additional species

New trees are built periodically to include improved sequences or sequences from additional organisms. Currently the PANTHER trees contain genes from 48 completely sequenced genomes, with plans to increase this number to the emerging standard being developed by the UniProt team in collaboration with the wider ortholog prediction community (http://www.ebi.ac.uk/reference_proteomes/). PAINT-derived GO annotations are already available for genes in these 48 organisms. The PANTHER tree building process assigns stable identifiers to the nodes so that when new releases of the PANTHER database are produced, PANTHER will report which tree nodes have undergone topological changes. When this occurs, trees will be flagged for verification that the annotations are still valid, and re-annotated is appropriate. We have shown that the algorithm used for tree building is very robust when adding more sequences, with over 85% of the trees being completely unchanged and only 2% with major changes [10]. Therefore, there should be very few revisions in the annotations due to changed tree topology. We expect those re-annotations should easily be integrated in the regular annotation updates that need to be done regularly as new data are published.

Limitations of functional predictions with PAINT

A major limitation of PAINT is the manual curator time required. To estimate the time required, we performed a pilot study in which families covering approximately 1% of the genes in the GO Reference Genomes were annotated. This covered 70 protein families and approximately 9100 proteins in the 48 species from PantherDB version 7.0. This required ~40 days of biocurator time, making annotation of all genes (in families with at least one experimentally characterized gene) a feasible ambition for the GO Consortium. Also, although we have developed numerous guidelines for PAINT curation, as in any manual curation process we expect variability in annotations due to differences in training and expertise of individual curators, as well as dependencies on the amount of time available for curation of a given family. Finally, as with any function prediction method, the primary limitation is the comprehensiveness of experimental annotations. For instance, for human PGM5, if we did not have any information about the residues necessary for phosphoglucomutase activity, nor any experimental results for PGM5 orthologs in vertebrates, our process would have incorrectly annotated human PGM5 as having phosphoglucomutase activity, as InterPro2GO does. The complete evidence trail for PAINT inferences is very important in this regard, as it allows us to know precisely which inferences were made by a curator. This will simplify updating and correcting annotations as additional experimental evidence accumulates in the future. We already have a software pipeline in place to detect annotation changes in families that have undergone PAINT annotation and update them accordingly.

Data availability

PAINTannotation tool

PAINT can be downloaded at Source Forge (http://sourceforge.net/projects/pantherdb/). GO annotations are available from the GO database (http://geneontology.org) and ancestral annotations are available from PanTree (http://pantree.org). PANTHER families, phylogenetic trees and multiple sequence alignments are available at http://pantherdb.org.

CONCLUSIONS

We report the development of a process for curated homology inference of gene function on a large scale. The process begins with evolutionary relationships among genes represented as phylogenetic trees, and the annotated functions of those genes represented as GO terms with experimental evidence. We have developed a software application, PAINT that integrates this information together with additional data such as sequence features, and allow a curator to create a reconstruction of the evolution of gene function within the family. This reconstruction explicitly captures inferred functional gain and loss events in specific branches of the tree, which are then used to predict functions for genes that have not yet been characterized. While orthology is one piece of evidence used to reconstruct functional evolution, no assumptions are made a priori about the relationship between orthology and functional conservation.

In essence, PAINT enables a biocurator to construct and record a (generally) parsimonious model of the evolution of function in the family that can be tested against, and modified by, new experimental data as it emerges. The aim is to provide as much data as possible for the biocurator to construct this evolutionary model. These data comprise not only the tree topology and branch lengths used in existing phylogenomic methods, but also general biological knowledge, knowledge of the protein family (our standard operating procedure includes reading published reviews of the family), specific sequence features and knowledge of other experimental annotations (both more or less specific within the GO or even apparently distantly related within the

GO). Importantly for users of GO annotations, our method allows the prediction of not only molecular function, but also BP and CC annotations, as these characters are also gained, lost and inherited over evolutionary time. These aspects are treated with care by a curator, as reflected by our standard operating procedures of (i) annotating evolution of MF first, and (ii) generally considering only the evolution of cellular processes and processes in which the molecular mechanisms are characterized to some degree.

While PAINT curation requires substantial manual input from a trained biocurator, it is both accurate and tractable on a large scale. We have performed a pilot project annotating approximately 1% of the genes across a broad set of genomes, and shown that the curation process is relatively efficient and feasible for entire genomes. We have compared the annotations from our approach in two case studies to those generated by the most widely used methods. In our examples, the differences in predictions are due largely to assumptions about the relationship between sequence and function. InterPro2GO assumes that some functions are conserved among all family members recognized by a given HMM. Thus, functional divergence within a family results in either false positive (e.g. phosphoglucomutase activity for PGM5) or false negative (e.g. SOD activity for SOD1) predictions or in some cases less specific predictions. Compara assumes that there is essentially no functional divergence between orthologs that are separated only by recent speciation events, but that functional divergence is common enough otherwise to render predictions unreliable. Thus, lack of experimental knowledge from close orthologs results in false negative predictions (e.g. CCS and PGM1). In PAINT, these issues are addressed in two ways. The first is in the model for explicitly representing functional gain and loss at any point in the evolutionary tree, which allows handling of conservation and divergence for each function on a case-by-case basis. The second is in the use of an expert curator to make the inferences, allowing multiple types of information to be integrated into the evolutionary model.

Finally, we have made the PAINT software and annotations available online, along with extensive documentation and standard operating procedures for GO annotation of functional evolution events in gene families, to encourage use by the wider community.

Key Points

- With the constant acceleration in the number of genome sequences available, it is indispensable to have powerful methods for predicting protein function.
- The GO offers a method to uniformly describe the functions of gene products in a species-independent manner; GO is being used extensively to 'annotate' genes from many different organisms, based on experimental evidence.
- We describe a method for inference of gene function by homology, based annotating gain/loss of function events directly onto a phylogenetic tree.
- We have developed a software tool, PAINT, that assists curators in annotating nodes (ancestral genes) in the tree with GO terms describing these gain and loss events, and then automatically propagates GO annotations to descendants of the annotated ancestral genes.

FUNDING

This work is funded by the National Institute of General Medical Sciences (R01-GM081084, to P.D.T., M.L., P.G. and S.L.) receive additional support from a National Institute of Human Genome Research grant (P41-HG002273) and supplements (M.L.) in support of the Gene Ontology Consortium.

References

1. Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
2. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* 2010;**38**: D331–5.
3. du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology–a primer for bioinformaticians. *Brief Bioinform* 2011. In press doi: 10.1093/bib/bbr002.
4. Gaudet P and the Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* 2009;**5**(7): e1000431.
5. Felsenstein J. *Inferring Phylogenies*. Massachusetts: Sinauer Associates Inc., 2004. ISBN 0-87893-177-5.
6. Eisen JA. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* 1998;**26**(18):4291–300.
7. Engelhardt BE, Jordan MI, Muratore KE, *et al*. Protein molecular function prediction by Bayesian phylogenomics. PLoS Computat Biol*,* 2005;**1**:432–45.
8. Page RD. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 1998; **14**(9):819–20.
9. Mi H, Dong Q, Muruganujan A, *et al*. PANTHER version 7: improved phylogenetic trees, orthologs, and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* 2010;**38**:D204–10.

10. Thomas PD. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics* 2010; **11**:312.
11. Fitch WM. Distinguishing homologous from analogous proteins. *Syst. Zool* 1970;**19**:99–113.
12. Studer RA, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* 2009;**25**(5):210–6.
13. Heinicke S, Livstone MS, Lu C, *et al*. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One* 2007;**2**(1):e766.
14. Deegan JI, Dimmer EC, Mungall CJ. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics* 2010; **11**:530.
15. McDowall J, Hunter S. InterPro protein classification. *Methods Mol Biol* 2011;**694**:37–47.
16. Vilella AJ, Severin J, Ureta-Vidal A, *et al*. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 2009;**19**(2):327–35.