D260–D264 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 6 December 2006*
doi:10.1093/nar/gkl1043

TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes

Jeremy D. Selengut*, Daniel H. Haft, Tanja Davidsen, Anurhada Ganapathy, Michelle Gwinn-Giglio, William C. Nelson, Alexander R. Richter and Owen White

TIGR, Bioinformatics Department, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 18, 2006; Revised and Accepted November 3, 2006

ABSTRACT

TIGRFAMs is a collection of protein family definitions built to aid in high-throughput annotation of specific protein functions. Each family is based on a hidden Markov model (HMM), where both cutoff scores and membership in the seed alignment are chosen so that the HMMs can classify numerous proteins according to their specific molecular functions. Most TIGRFAMs models describe 'equivalog' families, where both orthology and lateral gene transfer may be part of the evolutionary history, but where a single molecular function has been conserved. The Genome Properties system contains a queriable set of metabolic reconstructions, genome metrics and extractions of information from the scientific literature. Its genome-by-genome assertions of whether or not specific structures, pathways or systems are present provide high-level conceptual descriptions of genomic content. These assertions enable comparative genomics, provide a meaningful biological context to aid in manual annotation, support assignments of Gene Ontology (GO) biological process terms and help validate HMM-based predictions of protein function. The Genome Properties system is particularly useful as a generator of phylogenetic profiles, through which new protein family functions may be discovered. The TIGRFAMs and Genome Properties systems can be accessed at http://www.tigr.org/TIGRFAMs and http://www.tigr.org/Genome_Properties.

TIGRFAMs AND GENOME PROPERTIES: OVERALL PHILOSOPHY

TIGRFAMs, a source of models that characterize proteins in terms of molecular function and Genome Properties, which makes assertions of biological processes, are both governed by a similar philosophy. These systems are informed by the need to provide the highest possible quality assertions to support the needs of annotation efforts. For this reason, each model and property is a manually curated computable object for use in automated processes and downstream comparative genomics analyses. A necessary drawback of this insistence on curation is that not all possible objects are created. What is gained is the combination of computational speed and curatorial quality. In those cases where these systems yield unambiguous results, a great deal less manual annotation effort need be expended.

TIGRFAMs

Assignments of molecular function to non-experimentally characterized proteins are made, ideally, based on sequence similarity directly to sequences that do have experimental characterizations. In actual practice, assignments frequently are made transitively, a significant source of error. However, in best practices for manual review of functional annotation, sequence similarity is not the only consideration. Other contributing factors include molecular phylogeny, conserved gene neighborhoods, paralogous families, bidirectional best hit matches across multiple pairs of genomes and metabolic context as implied by other annotations. This process may yield a group of trusted functional assignments for a homology family that may be mathematically represented as a profile hidden Markov model (HMM). Once created, such models can be added to the library of HMMs and used to evaluate novel sequences. Sequences which yield scores that are above curated 'trusted' cutoffs can be considered true matches to that family.

The models in the TIGRFAMs database (www.tigr.org/TIGRFAMs) have been built specifically to aid in automated annotation of prokaryotic genes, particularly by focusing on the creation of 'equivalog' family models. All members of 'equivalog' families are believed to have the same molecular function and to be related to a common ancestor having that

*To whom correspondence should be addressed. Tel: +1 301 795 7566; Fax: +1 301 838 0208; Email: selengut@tigr.org same function. In contrast to ortholog families, this may include examples of paralogs (duplicated genes) and laterally transferred genes, so long as they have the same function. Curators of TIGRFAMs models identify a number of criteria in order to determine whether a model qualifies as an equivalog. These fall into three areas: (i) the observation that two or more sequences within the family have experimentally characterized or highly trusted annotations as a particular function, (ii) that no sequences within the family are indicated by reasonable evidence to have a different function and (iii) that phylogenetic trees constructed from members of the family are consistent with the hypothesis that the most recent common ancestor of the characterized (or highly trusted) members of the family is an ancestor of all of the members of the family.

The TIGRFAMs database contains 3000 curated protein family models, of which 1700 are of equivalog type. These models enable accurate, thorough, automated assignment of molecular function. They are supplemented by over 100 'equivalog domain' models, which assign annotations to discrete domains of multi-functional proteins, and by over 350 'hypothetical equivalog' models, which describe uncharacterized proteins families in which the members are likely to be equivalogs. In addition, we have identified ~1000 models from the Pfam database (1) that we classify tentatively as equivalog models based on their performance, and that can supplement the library of TIGRFAMs equivalog-type HMMs.

Many homology families are inherently multi-functional or not practically modeled as equivalog HMMs. Even so, it may be practical to build 'subfamily' HMMs that perform meaningful classifications within larger protein families, such as the heavy metal sensor kinases (TIGR01386). This example is a subfamily among the sensor histidine kinases, which in turn, in the region recognized by Pfam model PF02518, represent a branch of a family of ATPases that also includes DNA gyrases and HSP90-like proteins. TIGR01386 does not behave as an equivalog model, in that it does not specify which metal or metals will bind, but still is quite useful for its fairly rich assertion of protein function. The TIGRFAMs database includes 600 models of the 'subfamily' type, many of which represent important divisions among broader protein families from Pfam.

TIGRFAMs was last described in *Nucleic Acids Research* in 2003 (2). Since then, the size of the library has nearly doubled. Aside from this growth, many models have been extensively revised as new sequence data have been deposited and new experimental information has accumulated in the literature. Over the whole of the Comprehensive Microbial Resource (CMR, cmr.tigr.org) (3), containing ~1.2 million protein sequences, ~20% have above-trusted cutoff matches to TIGRFAMs (and Pfam) equivalogs. This translates to ~590 proteins identifications per prokaryotic genome. Note that this 20% figure refers only to equivalog models; ~75% of all proteins have hits to some kind Pfam or TIGRFAMs model.

Equivalog models are linked to numerous informational fields such as standard names, numerical classifiers (Enzyme Commission or Transporter Commission numbers), controlled vocabulary terms from Gene Ontology (GO), the GO system (4), literature references and links to experimentally characterized homologs. A trusted match to an equivalog model justifies the automated transfer of these data to the sequence in question, generally reducing the need for manual review to a bare minimum. The annotation that results has a high degree of consistency from one genome to another, and in principle can improve consistency for annotation projects between different annotation centers. When, on occasion, errors or changes in scientific formalisms require modifications to the terms associated with a TIGRFAMs model, the changes can be propagated uniformly over all affected genes in a database (such as the CMR).

TIGRFAMs models, including equivalog models, have been used extensively in genome annotation at The Institute for Genomic Research (TIGR) for over 7 years. Manual review of annotations suggested by TIGRFAMs models have led to considerable feedback and to the improvement of the thresholds and annotations of many models. Currently, the AutoAnnotate program in TIGR's prokaryotic annotation pipeline uses HMM evidence extensively. AutoAnnotate weighs the evidence from HMMs, pair-wise homology searches and other analyses, makes tentative assertions as to molecular function and present these data to human curators. AutoAnnotate gives highest priority to trusted hits to TIGRFAMs equivalog models.

Once the molecular function of a protein is known, it may become necessary to understand the genomic context in order to assign the correct GO process term to the protein. For example, the equivalog model TIGR00658 identifies ornithine carbamoyltransferase. However, this enzyme may act in an arginine biosynthesis pathway from ornithine, along with members of families TIGR00032 and TIGR00838, as in *Yersinia pestis*. Alternatively (or additionally), it may act in an arginine degradation via citrulline, along with members of families TIGR00746 and TIGR01078 (usually in a three-gene operon), as in *Streptococcus pneumoniae*. This need to understand genomic context in order to complete annotation was an important motivation for the creation of the Genome Properties system (5).

Genome Properties

The Genome Properties system is a comparative genomics system that incorporates both machine-calculated and human-curated assertions about features of completely sequenced prokaryotic genomes. These features include taxonomic class, metrics such as genome size and GC-content and, if available, phenotypic information such as optimal growth temperature. More importantly, they include the results of metabolic and non-metabolic system reconstructions, produced according to rules that can be applied with full automation and without a requirement for prior annotation of gene functions. These reconstructions are expressed in a controlled vocabulary, and thus become computable objects, with uses that include creating high-level descriptions of metabolic capability or for comparing of one genome to another. As Genome Properties assertions are available for ~200 properties and 400 genomes, they are a rich substrate for the development and testing of hypotheses about the relationships of one property to another, or between properties and protein families, an exercise collectively referred to as data mining.

The assertions loaded into the Genome Properties database are produced by a flexible and powerful rules engine. In principle, these rules can use nearly any data type, including manually populated annotations of specific EC numbers. In practice, the rules that perform metabolic reconstructions for Genome Properties rely primarily on hits to TIGRFAMs HMMs and some Pfam HMMs, over 900 of which are currently incorporated into the system. These rules can be applied in the absence of any annotation, automated or manual, as long as HMM search results are available. The finding that all necessary proteins for some system are present becomes a 'YES' assertion, meaning that the system itself is present, at least in principle. It need not be demonstrated *in vivo*. The controlled vocabulary term 'some evidence' means that not every required component of a system has been established, but enough have been detected to suggest the system is present. The terms 'not supported' and 'none found' are self-explanatory, while the term 'NO' represents an even stronger negative assertion that usually reflects additional manual review. The rules engine does not require that every HMM be an equivalog model. It may instead require the presence in a genome of at least one member of some larger protein family, and may add a secondary requirement that genes for the various components of the property be near one another.

An interface into the Genome Properties data is provided through the Genome Properties home page (http://www.tigr.org/Genome_Properties) as well as through the Comprehensive Microbial Resource (3) which has, since 2004, integrated these data with the other analyses it provides. The web interface provides the ability to execute simple and complex queries and features multiple layers of data views, many of which have been recently upgraded and enhanced. The Genome Properties may be traversed through a descriptive hierarchy or by text searching. The underlying evidence for each assertion may be accessed through links to the individual gene pages in the CMR.

Synergy between TIGRFAMs and Genome Properties

Not all protein families lend themselves to the assignment of specific molecular functions based solely on homology to multiple experimentally characterized proteins. Common cases include families with only a single characterized member, families with many, but heterogeneous, characterized members and those with no characterized members but are subsets of larger families with established generic function. Additional information is required in these cases to define the proper boundaries of homogeneous function and create reliable equivalog models.

An evaluation of the genomic context of candidate family members can aid in this process. The Genome Properties system is an excellent tool for examining a protein within its genomic context. For example, a genome may have most components of a particular metabolic pathway identified unambiguously by equivalog HMMs. One step in the pathway may have several candidate genes identified by a less specific 'family' or 'domain' model, but there may be only one of these candidates within an apparent operon with other members of the pathway. One would infer that the embedded gene completes the pathway. In *Mycobacterium smegmatis* MC2, for instance, 58 potential sugar transporters of a type represented by the family-type model PF00083 are present while only one of them is proximal to genes associated with rhamnose catabolism (Table 1). In other cases, even though the model is not an equivalog, only one protein is identified in the genome. In the best case, the single candidate gene is near other genes associated with the biological process as well (rhamnose epimerase in Table 1, for instance), and that arrangement is repeated across multiple genomes.

The identification of genes with such auxiliary evidence can be fed back into the model building process, supporting the accurate definition of conserved-function family boundaries. A significant number of TIGRFAMs equivalogs have been constructed and/or validated in this manner over the past 2 years. Having constructed such models, a new round of Genome Properties evaluation may promote the assertion of state for that property from 'some evidence' to 'YES' and may further clarify the proper function of ambiguously assigned genes. Iteration of this cycle of improvements to Genome Property assertions and improvements to the underlying TIGRFAMs models has proved a remarkably robust method of pathway and system reconstruction.

Even with this context-driven approach available, there are many families for which the construction of equivalogs remains impractical or impossible. In such cases, the Genome Properties system can single out a protein according to both its membership in some relatively broad protein family and the proximity of its gene to other genes identified as hallmarks of the property. By these mechanisms, some 200 such TIGRFAMs and Pfam models of types other than equivalog have been incorporated so far into the sets of rules that drive Genome Properties. These links made by Genome Properties can be highly informative for annotation of those proteins (Table 1). On occasion they may serve to identify essential components, which can be added to the requirements for asserting the YES state.

Phylogenetic profiling using TIGRFAMs and Genome Properties

Phylogenetic profiling is the process of inferring links between protein families and biological process based on patterns of co-occurrence with other protein families involved in the same biological processes (6). In practice, the phylogenetic distribution for some biological process may differ from the pattern of the individual protein families that contribute essential parts of that process. This can happen, of course, for several reasons. Members of one protein family may substitute occasionally for those of another. An enzyme that performs a specific function may participate in different processes in different species. The set of parameters used to discriminate all true members of a protein family from all other proteins may be imprecise, especially prior to manual review. Missed gene calls, poor start-site predictions and sequencing errors may result in profiles with missing members.

Genome Properties, by creating composite objects that represent biological processes as opposed to molecular functions, can be used to generate phylogenetic profiles of much higher fidelity than those based on individual protein

Table 1. An example of a gene cluster in *Mycobacterium smegmatis* MC2 identified as being responsible for rhamnose catabolism by Genome Properties analysis (http://cmr.tigr.org/tigr-scripts/CMR/shared/GenomePropDefinition.cgi?prop_acc=GenProp0457)

Locus	MSMEG_0577	MSMEG_0578	MSMEG_0579	MSMEG_0580		MSMEG_0581	MSMEG_0582
Direction	→	→	→	→		→	→
Original annotated name	*Bacillus subtilis* YulD protein homolog lin2978	*cis, cis*-muconate transport protein MucK, putative	Sugar isomerase	Oxidoreductase, short-chain dehydrogenase/reductase family, putative		Sugar kinase	Ribose operon repressor
Post Genome Properties annotated name	Putative L-rhamnose-1-epimerase (RhaU)	Putative rhamnose-specific major facilitator family transporter	L-rhamnose isomerase (RhaI)	Rhamnulose-1-phosphate aldolase/alcohol dehydrogenase (RhaD)		Rhamnulo-kinase (RhaB)	Putative rhamnose catabolism operon transcriptional regulator
HMM evidence	PF05336: Protein of unknown function (DUF718)	PF00083: Sugar (and other) transporter	TIGR02635: l-rhamnose isomerase	TIGR02632: rhamnulose-1-phosphate aldolase/alcohol dehydrogenase		PF00370: FGGY family of carbohydrate kinases, N-terminal domain	PF00532: Periplasmic binding proteins and sugar binding domain of the LacI family
HMM type	Family	Family	Hypothetical equivalog	EQUIVALOG		Domain	Domain
Number of hits in genome	1	58	1	1		10	25
Additional evidence	Apparent operon; only candidate in genome	Apparent operon	Apparent operon; only candidate in genome	Only candidate in genome		Apparent operon	Apparent operon
Genome Properties component	Rhamnose epimerase	Rhamnose transporter	L-rhamnose isomerase (RhaA/RhaI)	Rhamnulose-1-p aldolase (rhad)	ADH	Rhamnulo kinase (RhaB/RhuK)	Rhamnose catabolism regulator
GP-required	No	Yes	Yes	Yes	No	Yes	Yes
GO process term	GO:0019301: rhamnose catabolism	GO:0015762: rhamnose transport	GO:0019301: rhamnose catabolism	GO:0019301: rhamnose catabolism		GO:0019301: rhamnose catabolism	GO:0043463: regulation of rhamnose catabolism

families. Profiles created from Genome Properties data tend to smooth out the errors associated with individual protein family models. Considering the rhamnose catabolism property illustrated in Table 1, eight genomes containing apparently extraneous hits to certain of the components of the system can be removed from the profile, while seven genomes containing all but one component may be included within the profile. Additionally, some of the components, such as the aldolase and the isomerase are captured by two independent, non-homologous equivalogs. Profiles associated with each of these four models would represent only a subset of the genomes expressing this rhamnose catabolism pathway.

This phenomenon of non-orthologous displacement within pathways is particularly common in prokaryotic genomes. It is frequently the case that, when reviewing Genome Properties data across many genomes, a number of genomes are identified lacking only a single component, and that this cannot be resolved by attempts to broaden the associated equivalog TIGRFAMs model or fix gene-calling errors. Often successful phylogenetic profiling may be accomplished in these cases by constructing a profile consisting of those genomes lacking that hits to that TIGRFAMs model, but otherwise having all required components of the Genome Property (5).

Phylogenetic profiling using the methods originally proposed and most frequently used requires profiles calculated globally over a comprehensive set of protein families, something that TIGRFAMs and Genome Properties with their more limited scope cannot provide. Nevertheless, a recently published method, called Partial Phylogenetic Profiling (7), is independent of such global calculations and requires only one high-fidelity query profile, something these databases are well-suited to provide.

Availability of TIGRFAMs and Genome Properties

The TIGRFAMs database is available at www.tigr.org/TIGRFAMs. The Genome Properties database is available at www.tigr.org/Genome_Properties. In addition to access to TIGRFAMs and Genome Properties data through their own home pages, both systems have been thoroughly integrated into the CMR (cmr.tigr.org). All genes which are members of TIGRFAMs or Pfam families are linked to the appropriate models. Similarly, all genes which have been mapped to Genome Properties processes are linked to the appropriate Genome Properties' web pages. All Genome Properties assertions can be downloaded from the CMR.

ACKNOWLEDGEMENTS

This work has been supported by grants from the United States National Science Foundation (DBI-0445826), the United States Department of Energy (DE-FG02-04ER63935) and the United States National Institute of Allergy and Infectious Disease (NIH-NIAID-DMID-04-34, HHSN266200400038C). Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
2. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
3. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
5. Haft,D.H., Selengut,J.D., Brinkac,L.M., Zafar,N. and White,O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, **21**, 293–306.
6. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
7. Haft,D.H., Paulsen,I.T., Ward,N. and Selengut,J.D. (2006) Exopolysaccharide-associated protein sorting in environmental organisms: the PEP-CTERM/EpsH system. Application of a novel phylogenetic profiling heuristic. *BMC Biol.*, **4**, 29.