Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W27–W32***
doi:10.1093/nar/gkm325

Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis

Jake Gunn Glanville, Dan Kirshner, Nandini Krishnamurthy and Kimmen Sjölander*

Berkeley Phylogenomics Group, University of California, Berkeley

Received January 31, 2007; Revised April 10, 2007; Accepted April 17, 2007

ABSTRACT

Phylogenomic analysis addresses the limitations of function prediction based on annotation transfer, and has been shown to enable the highest accuracy in prediction of protein molecular function. The Berkeley Phylogenomics Group provides a series of web servers for phylogenomic analysis: classification of sequences to pre-computed families and subfamilies using the *PhyloFacts* Phylogenomic Encyclopedia, *FlowerPower* clustering of proteins sharing the same domain architecture, *MUSCLE* multiple sequence alignment, *SATCHMO* simultaneous alignment and tree construction and *SCI-PHY* subfamily identification. The *PhyloBuilder* web server provides an integrated phylogenomic pipeline starting with a user-supplied protein sequence, proceeding to homolog identification, multiple alignment, phylogenetic tree construction, subfamily identification and structure prediction. The Berkeley Phylogenomics Group resources are available at http://phylogenomics.berkeley.edu.

INTRODUCTION

The standard protocol for gene function prediction involves homology-based annotation transfer (e.g. using the top BLAST hit); this approach is now known to be fraught with systematic errors (1–3). Biological processes such as gene duplication, mutation at critical residues, speciation and domain shuffling contribute to modifications of the original function that significantly complicate the process of functional annotation (1,4–6). Existing annotation errors can also be propagated by homology-based annotation transfer (7).

Phylogenomic inference of gene function is known to be the most robust and accurate method for functional annotation. This approach enables the function of a protein to be inferred in an evolutionary context, avoiding the pitfalls of simple pairwise sequence comparison based approaches, and vastly improving the accuracy of functional annotation (8–10). Phylogenomic analysis proceeds in stages, starting with homolog identification and multiple sequence alignment (MSA). The (masked) alignment is then used as input to phylogenetic tree construction. Examination of the tree topology enables biologists to discriminate between orthologs (with presumably conserved function) and paralogs (related by gene duplication, and potentially divergent in function), providing improved discrimination of specific function in instances when a protein family has evolved multiple but related distinct functions (11,12). To increase the confidence in function prediction, the source of the annotations can be examined; the Gene Ontology resource includes evidence codes for annotations for this purpose (13). The Berkeley Phylogenomics Group has developed a series of web servers for individual steps in a phylogenomic pipeline and a single web server *PhyloBuilder* that performs all the steps as shown in Figure 1. Each web server can be used individually or in combination for phylogenomic inference.

Each server includes Java applets for viewing the associated data; data can also be downloaded in standard formats. Users can bookmark a results page, or choose to receive results by email.

PHYLOFACTS PHYLOGENOMIC ENCYCLOPEDIA

PhyloFacts enables functional classification of user-submitted sequences to pre-computed families and subfamilies from across the Tree of Life (14). Hidden Markov models are provided for functional classification of novel sequences to families and subfamilies. PhyloFacts protein family 'books' include an MSA, phylogenetic trees, predicted structures and critical residues, experimental and annotation data, hidden Markov models, and links to other resources. Since the initial publication (14), the PhyloFacts resource has significantly increased in size, from ~9000 families in May 2006 to >27 000 families in April 2007. Most of this increase in size has been to expand our coverage of microbial gene families and gene families found in the human genome, including homologs in other species. New functionality included in PhyloFacts over the past year also includes super-fast classification of user-submitted sequences to global homology groups

*To whom correspondence should be addressed. Tel: 510 642 9932; Fax: 510 666 3327; Email: kimmen@berkeley.edu

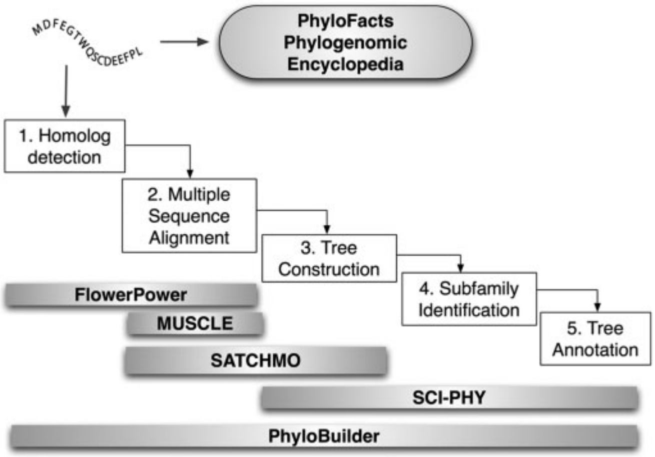


Figure 1. Berkeley Phylogenomics Group web servers for the different steps of a phylogenomic pipeline. *Top*: Users can submit sequences for classification against the PhyloFacts Phylogenomic Encyclopedia of pre-computed families and subfamilies. *Middle:* The phylogenomic pipeline. *Bottom*: Web servers for specific tasks in the pipeline. Many of these servers cover more than one step in the process, e.g. the PhyloBuilder web server, which performs all the steps of the pipeline and outputs a MSA, subfamilies, domain/3D structure predictions and phylogenetic trees overlaid with annotations.

(proteins sharing the same domain architecture) and a new protocol for functional sub-classification.

Usage: Users can submit DNA or protein sequences in FASTA format for classification to PhyloFacts families and subfamilies. PhyloFacts family books are selected for HMM scoring by a pre-processing step of BLAST search of the query sequences against the consensus sequences for each of the families in the resource; HMMs from families with BLAST E-values of 10 or better are scored against the query. An example output from PhyloFacts is shown in Figure 2. Clicking on 'View Alignment' displays the pairwise alignment between the submitted query and consensus sequence and statistics about the alignment. Clicking in the 'Search subfamilies' box for families of interest followed by clicking on the box at bottom labeled 'Search selected books for top-scoring subfamily HMMs against query' initiates the subfamily HMM-based classification; logistic regression analysis is used to differentiate sequences that can be assigned to the top-scoring subfamily and those that represent novel subtypes. Users can examine PhyloFacts protein family books by following links in the 'PhyloFacts book' column in the table of results. Super-fast classification of query sequences to families with global similarity is provided (results would otherwise include local matches). Users can bookmark a results page, or choose to receive results by email. PhyloFacts is available at http://phylogenomics.berkeley.edu/phylofacts.

FLOWERPOWER HOMOLOGY DETECTION

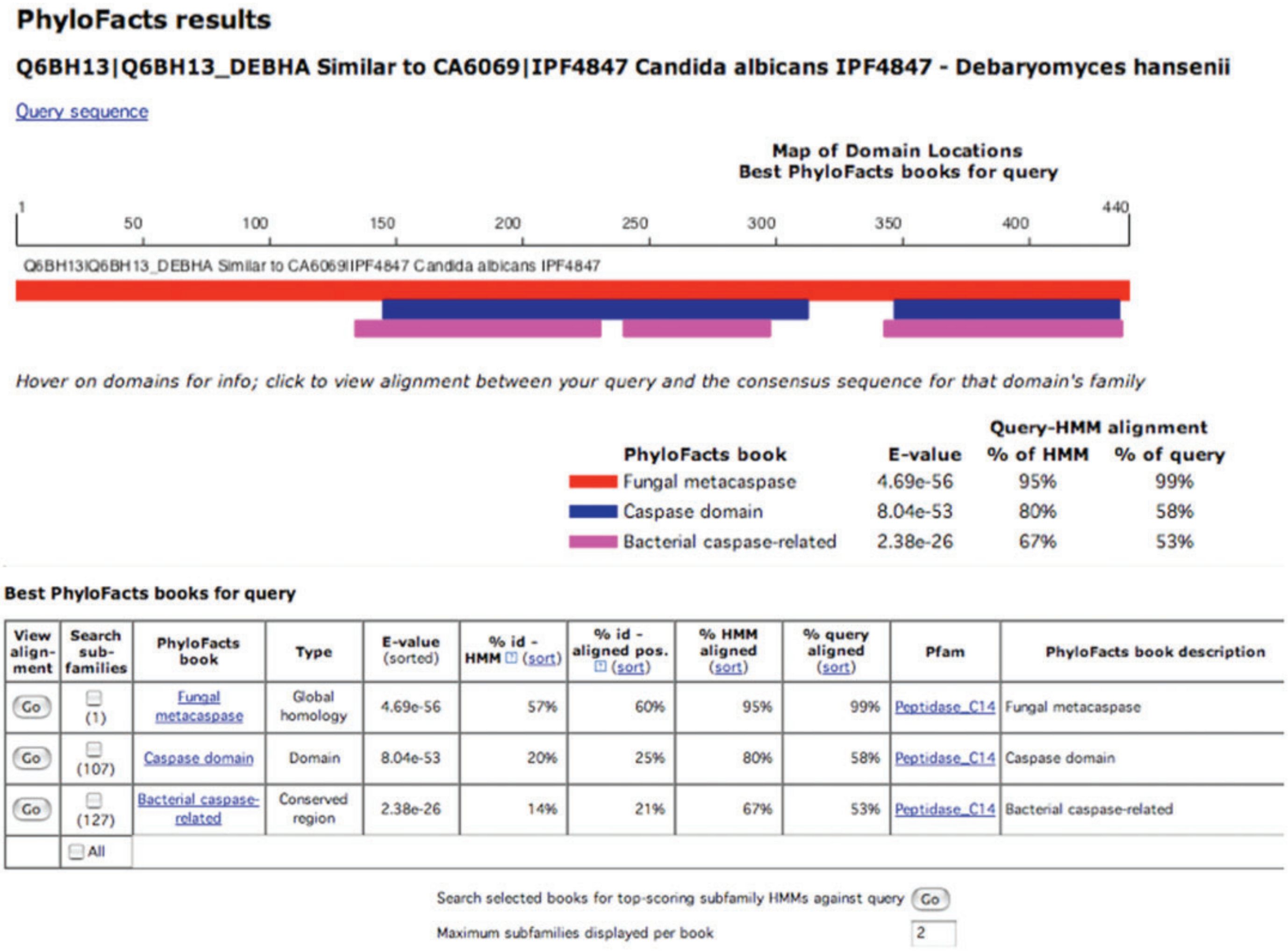
FlowerPower is an iterative homology-detection server akin to PSI-BLAST (15), but designed specifically for phylogenomic inference of function (16). FlowerPower is optimized for the retrieval of sequences sharing the same domain architecture; this prevents transfer of database annotation based on partial homology (i.e. local instead of global similarity). FlowerPower uses iterative subfamily hidden Markov model (HMM) searches against PSI-BLAST-identified homologs and alignment analysis to discriminate between partial and global homologies; this approach outperforms existing methods in gathering global homologs. *Usage*: The input to FlowerPower is a protein sequence in FASTA format; default parameters search the UniProt (17) database for proteins sharing the same domain architecture. The 'Advanced Settings' page enables users to modify the PSI-BLAST parameters for database searched, number of iterations and maximum number of hits returned. Parameters for the iterated search with subfamily HMMs can also be modified. Finally, users can choose between two homolog-selection modes: global (to both query and hit) and 'glocal' (global-local homology, retrieved sequences must align over a specific region, but can have additional structure). Results include the selected sequences, the raw FlowerPower alignment, a MUSCLE (18) re-alignment, and the results of the initial PSI-BLAST search. FlowerPower is available through http://phylogenomics.berkeley.edu/flowerpower/.

MULTIPLE SEQUENCE ALIGNMENT USING MUSCLE

The MUSCLE software produces high-accuracy multiple sequence alignments, with outstanding scores on benchmark dataset tests; it is also very fast, making it suitable for large-scale application (18). We employ MUSCLE in our internal pipeline for the PhyloFacts Phylogenomic Encyclopedia construction (14). *Usage*: The input to MUSCLE is a set of protein sequences in FASTA format. Alignments can be viewed online or downloaded in Aligned FASTA format. MUSCLE is available at http://phylogenomics.berkeley.edu/muscle.

SATCHMO

SATCHMO (Simultaneous Alignment and Tree Construction using Hidden Markov mOdels) is a progressive method of multiple sequence alignment that uses agglomerative clustering to estimate a phylogenetic tree simultaneously with the alignment. SATCHMO uses Dirichlet mixture densities (19) to construct profiles, and profile–profile scoring and alignment (20–23) to determine the phylogenetic tree topology. Each node in the tree contains a MSA and a corresponding profile. As sequences diverge in evolution, small insertions, deletions and mutations result in changes in structure and function; SATCHMO is intended to model these changes in different lineages in a family. Profiles and alignments at internal nodes in the tree represent the sequences descending from that node and may be of different lengths. The alignment at the root of the tree is an estimate of the conserved core structure defining all family members; when highly divergent sequences are input to SATCHMO this root alignment may be



PhyloFacts results

Q6BH13|Q6BH13_DEBHA Similar to CA6069|IPF4847 Candida albicans IPF4847 - Debaryomyces hansenii

Query sequence

Map of Domain Locations
Best PhyloFacts books for query

1
50
100
150
200
250
300
350
400
440
Q6BH13IQ6BH13_DEBHA Similar to CA6069IIPF4847 Candida albicans IPF4847

Hover on domains for info; click to view alignment between your query and the consensus sequence for that domain's family

		Query-HMM alignment	
PhyloFacts book	E-value	% of HMM	% of query
🟥 Fungal metacaspase	4.69e-56	95%	99%
🟦 Caspase domain	8.04e-53	80%	58%
🟪 Bacterial caspase-related	2.38e-26	67%	53%

Best PhyloFacts books for query

View align-ment	Search sub-families	PhyloFacts book	Type	E-value (sorted)	% id - HMM (sort)	% id - aligned pos. (sort)	% HMM aligned (sort)	% query aligned (sort)	Pfam	PhyloFacts book description
Go	□ (1)	Fungal metacaspase	Global homology	4.69e-56	57%	60%	95%	99%	Peptidase_C14	Fungal metacaspase
Go	□ (107)	Caspase domain	Domain	8.04e-53	20%	25%	80%	58%	Peptidase_C14	Caspase domain
Go	□ (127)	Bacterial caspase-related	Conserved region	2.38e-26	14%	21%	67%	53%	Peptidase_C14	Bacterial caspase-related
	□ All									

Search selected books for top-scoring subfamily HMMs against query Go

Maximum subfamilies displayed per book 2

Figure 2. Result of functional classification against PhyloFacts. The figure shows HMM scoring results for the UniProt sequence Q6BH13 from *Debaryomyces hansenii*. The search retrieves protein family books constructed using three different protocols: global homology, conserved region and domain. Subfamily classification is enabled by selecting books (clicking in boxes at left side of table, under 'Search subfamilies') followed by clicking the 'Go' button at bottom. See text for details.

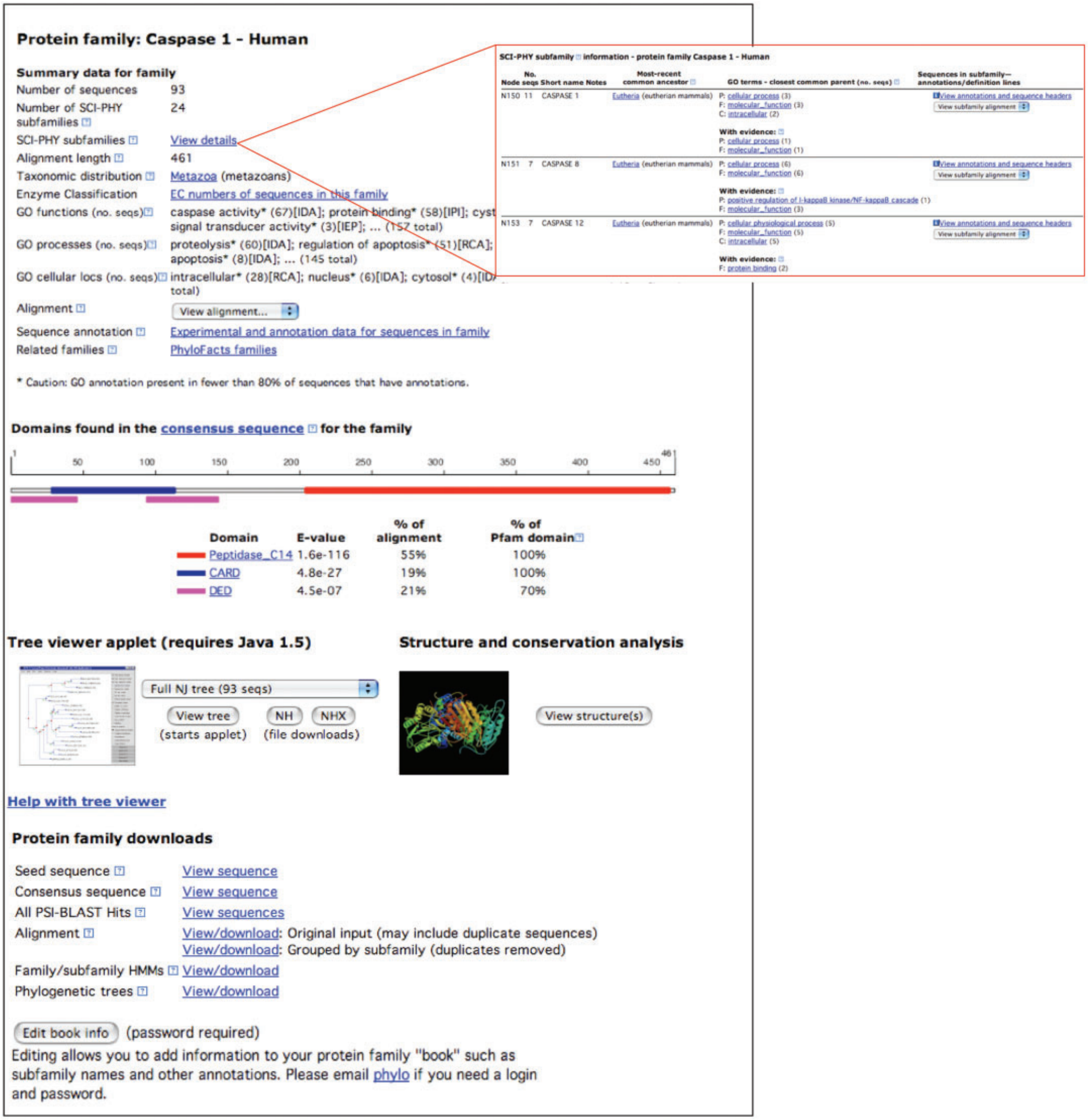
a small fraction of the average sequence length. Tree topologies produced using SATCHMO are consistent with expert-defined subtypes; alignment accuracy is also high (20). *Usage*: The input to SATCHMO is a set of unaligned protein sequences, in FASTA format. The SATCHMO root alignment can be viewed online using a Java applet or downloaded from the website. Special SATCHMO tree-alignment viewing software is available online (currently for PCs only) enabling the different alignments descending from each internal node of the tree to be examined separately. SATCHMO is available at http://phylogenomics.berkeley.edu/satchmo.

SCI-PHY AND SUBFAMILY HMM CONSTRUCTION

SCI-PHY (Subfamily Classification in PHYlogenomics) uses Bayesian and information-theoretic approaches to construct a hierarchical tree and cut the tree into subtrees to identify functional subfamilies (24). Subfamily hidden Markov models are constructed using Dirichlet mixture densities to derive a position- and subfamily-specific weighting scheme to share information across subfamilies; this has been shown to increase the separation between homologous and unrelated sequences and to provide high specificity of classification (25). *Usage*: The input to SCI-PHY is a MSA in either Aligned FASTA or the UCSC A2M format. Outputs include the MSA divided into subfamilies, the SCI-PHY tree, and subfamily and family HMMs in both HMMER and UCSC SAM formats. The SCI-PHY tree can be downloaded or viewed online using the Java ATV applet (26). SCI-PHY is available at: http://phylogenomics.berkeley.edu/SCI-PHY.

PHYLOBUILDER

PhyloBuilder is an automated computational pipeline for phylogenomic analysis, starting from an input protein sequence. PhyloBuilder is a modified version of the



Protein family: Caspase 1 - Human

Summary data for family

Number of sequences	93
Number of SCI-PHY subfamilies	24
SCI-PHY subfamilies	View details
Alignment length	461
Taxonomic distribution	Metazoa (metazoans)
Enzyme Classification	EC numbers of sequences in this family
GO functions (no. seqs)	caspase activity* (67)[IDA]; protein binding* (58)[IPI]; cyst signal transducer activity* (3)[IEP]; ... (157 total)
GO processes (no. seqs)	proteolysis* (60)[IDA]; regulation of apoptosis* (51)[RCA]; apoptosis* (8)[IDA]; ... (145 total)
GO cellular locs (no. seqs)	intracellular* (28)[RCA]; nucleus* (6)[IDA]; cytosol* (4)[IDA total)
Alignment	View alignment...
Sequence annotation	Experimental and annotation data for sequences in family
Related families	PhyloFacts families

* Caution: GO annotation present in fewer than 80% of sequences that have annotations.

SCI-PHY subfamily information - protein family Caspase 1 - Human

Node	No. seqs	Short name	Notes	Most-recent common ancestor	GO terms - closest common parent (no. seqs)	Sequences in subfamily—annotations/definition lines
N150	11	CASPASE 1		Eutheria (eutherian mammals)	P: cellular process (3) F: molecular_function (3) C: intracellular (2) With evidence: P: cellular process (1) F: molecular_function (1)	View annotations and sequence headers View subfamily alignment
N151	7	CASPASE 8		Eutheria (eutherian mammals)	P: cellular process (6) F: molecular_function (6) With evidence: P: positive regulation of I-kappaB kinase/NF-kappaB cascade (1) F: molecular_function (3)	View annotations and sequence headers View subfamily alignment
N153	7	CASPASE 12		Eutheria (eutherian mammals)	P: cellular physiological process (5) F: molecular_function (5) C: intracellular (5) With evidence: F: protein binding (2)	View annotations and sequence headers View subfamily alignment

Domains found in the consensus sequence for the family

	Domain	E-value	% of alignment	% of Pfam domain
🟥	Peptidase_C14	1.6e-116	55%	100%
🟦	CARD	4.8e-27	19%	100%
🟪	DED	4.5e-07	21%	70%

Tree viewer applet (requires Java 1.5)

Structure and conservation analysis

Help with tree viewer

Protein family downloads

Seed sequence	View sequence
Consensus sequence	View sequence
All PSI-BLAST Hits	View sequences
Alignment	View/download: Original input (may include duplicate sequences) View/download: Grouped by subfamily (duplicates removed)
Family/subfamily HMMs	View/download
Phylogenetic trees	View/download

Edit book info (password required)

Editing allows you to add information to your protein family "book" such as subfamily names and other annotations. Please email phylo if you need a login and password.

Figure 3. Result of PhyloBuilder run for human Caspase-1. PhyloBuilder takes an input protein sequence and outputs a web page containing a cluster of homologous proteins, multiple sequence alignment, neighbor-joining tree, predicted subfamilies, PFAM domains, transmembrane domains and signal peptides, and retrieval of Gene Ontology (GO) and Enzyme Classification (EC) data. *Top*: Summary data include the number of homologs retrieved, taxonomic distribution, EC numbers and GO annotations and evidence codes. SCI-PHY subfamilies can be viewed by clicking on the link labeled 'View details' (see inset at top right). The multiple sequence alignment can be viewed using JalView or hypertext. A spreadsheet with annotations for all sequences is available under 'Experimental and annotation data for sequences in family'. *Middle*: PFAM and transmembrane domain/signal peptide predictions are displayed. Neighbor-joining and SCI-PHY trees can be viewed using ATV. Homologous 3D structures can be viewed using JMOL; residues predicted to be critical using evolutionary conservation analysis are displayed on the structure. Catalytic Site Atlas data are included. *Bottom*: Various downloads are available, including a multiple sequence alignment for the family and individual subfamilies, a FASTA file for all PSI-BLAST hits, NJ tree and HMMs for the family and SCI-PHY subfamilies in HMMER and SAM formats. An 'Edit book info' button enables users to add descriptive labels to families and subfamilies (as shown in the inset at top right). See text for details.

pipeline we use to populate the PhyloFacts Phylogenomic Encyclopedias with protein family books (14). The *PhyloBuilder* pipeline has multiple stages, as shown in Figure 1. In stage 1, FlowerPower is used to retrieve global homologs for the user-supplied sequence. Program parameters for this stage are set by default to maximize the retrieval of proteins sharing a common domain architecture; alternative settings are provided to enable users to request the selection of *glocal* homologs (sequences sharing a common domain but which may have different overall folds). If fewer than three sequences matching user criteria are identified, the program skips stages 2 and 3 and jumps directly to stage 4. In stage 2, the FlowerPower cluster is aligned using MUSCLE, followed by alignment masking in preparation for phylogenetic analysis (removing columns containing >70% gap characters). In stage 3, the masked alignment is used as the basis for neighbor-joining tree construction using the PHYLIP software (27), and submitted to the SCI-PHY software for subfamily identification. In stage 4, Gene Ontology annotations and evidence codes (13), Enzyme Classification data, and other data are retrieved for sequences in the cluster, and put into a spreadsheet, separated into SCI-PHY subfamilies. The species of origin, accession and definition lines are overlaid on the neighbor-joining tree, and can be viewed using the ATV tree-viewer/editor Java applet. In stage 5, domain and 3D-structure predictions for the family as a whole are performed based on analysis of the consensus sequence for the family: PFAM domains (28) are predicted (using the PFAM gathering threshold), transmembrane domains and signal peptides are predicted using the Phobius server (29), and homologous 3D structures are identified using BLAST analysis against the Protein Data Bank (PDB) (30). The phylogenetic trees produced by the *PhyloBuilder* web server can be used to identify orthologs manually; users can also download these trees and alignments for input to automated ortholog identification programs such as Orthostrapper (11) and RIO (12). *Usage*: Users paste in (or upload) a protein sequence for analysis. Results are stored for ten days; users can request long-term storage of these results. *PhyloBuilder* program outputs include a multiple sequence alignment, phylogenetic tree, subfamily identification, predicted domain/3D structure, and experimental and annotation data (see Figure 3). PhyloBuilder is available at http://phylogenomics.berkeley.edu/phylobuilder.

FUTURE WORK

The PhyloFacts Phylogenomic Encyclopedia is under continuous expansion; we plan to continue our development of this resource to cover all protein families across the Tree of Life. The conservative parameterization of homology clustering component of the PhyloBuilder server occasionally results in a somewhat restrictive set of homologs when global homology is enforced. We plan to explore PhyloBuilder parameter settings that retain selectivity while optimizing sensitivity, and to allow users to input a multiple sequence alignment constructed independently instead of being dependent on the FlowerPower clustering used in PhyloBuilder. Computational efficiency remains a significant challenge in phylogenomic inference. Many of the steps in a phylogenomic pipeline are computationally intensive; this causes us to limit the size of inputs and the number of jobs submitted per day (see individual web server pages for guidelines). We plan to improve the computational efficiency of these servers and also increase the size of our compute cluster in order to overcome this limitation.

ACKNOWLEDGEMENTS

The authors wish to thank the numerous developers of bioinformatics web servers and databases providing methods or data included in the PhyloFacts resource and web servers. The PhyloFacts resource is supported by a Presidential Early Career Award for Scientists and Engineers (PECASE) from the National Science Foundation (DBI-0238311), and by a grant from the National Human Genome Research Institute of the NIH (R01HG02769). We thank Ian Tegebo and Graham Melcher for assistance in quality control tests of the web servers described here. Funding to pay the Open Access publication charges for this article was provided by NIH (R01HG02769).

Conflict of interest statement. None declared.

REFERENCES

1. Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences – where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.
2. Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 55–67.
3. Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, REVIEWS0005.
4. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
5. Kaessmann,H., Zollner,S., Nekrutenko,A. and Li,W.H. (2002) Signatures of domain shuffling in the human genome. *Genome Res.*, **12**, 1642–1650.
6. Rajalingam,R., Parham,P. and Abi-Rached,L. (2004) Domain shuffling has been the main mechanism forming new hominoid killer cell Ig-like receptors. *J. Immunol.*, **172**, 356–369.
7. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.
8. Brown,D. and Sjölander,K. (2006) Functional classification using phylogenomic inference. *PLoS Comput. Biol.*, **2**, e77.
9. Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
10. Sjölander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.
11. Storm,C.E. and Sonnhammer,E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, **18**, 92–99.
12. Zmasek,C.M. and Eddy,S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3**, 14.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

14. Krishnamurthy,N., Brown,D.P., Kirshner,D. and Sjölander,K. (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol.*, **7**, R83.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Krishnamurthy,N., Brown,D. and Sjölander,K. (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol. Biol.*, **7**(Suppl. 1), S12.
17. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
18. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
19. Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I.S. and Haussler,D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–345.
20. Edgar,R.C. and Sjölander,K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.
21. Edgar,R.C. and Sjölander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, **20**, 1301–1308.
22. Edgar,R.C. and Sjölander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309–1318.
23. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
24. Sjölander,K. (1998) Phylogenetic inference in protein superfamilies: analysis of SH2 domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 165–174.
25. Brown,D., Krishnamurthy,N., Dale,J.M., Christopher,W. and Sjölander,K. (2005) Subfamily hmms in functional genomics. *Pac. Symp. Biocomput.*, **10**, 322–333.
26. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
27. Felsenstein, J. (2005). 3.6a2.1 ed. Distributed by the author. Department of Genome Science, University of Washingtion, Seattle.
28. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
29. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
30. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.