

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

A phylogenomic gene cluster resource: The phylogenetically inferred groups (PhIGs) database

Permalink

<https://escholarship.org/uc/item/1bf464fh>

Authors

Dehal, Paramvir S.
Boore, Jeffrey L.

Publication Date

2005-08-25

Peer reviewed

This Provisional PDF corresponds to the article as it appeared upon acceptance. The fully-formatted PDF version will become available shortly after the date of publication, from the URL listed below.

A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database

BMC Bioinformatics 2006, **7**:201 doi:10.1186/1471-2105-7-201

Paramvir S Dehal (psdehal@lbl.gov)
Jeffrey L Boore (jlboore@lbl.gov)

ISSN 1471-2105

Article type Database

Submission date 20 April 2005

Acceptance date 11 April 2006

Publication date 11 April 2006

Article URL <http://www.biomedcentral.com/1471-2105/7/201>

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

<http://www.biomedcentral.com/info/authors/>

A Phylogenomic Gene Cluster Resource: The Phylogenetically Inferred Groups (PhIGs) Database

Paramvir S. Dehal^{1*} and Jeffrey L. Boore^{1,2}

¹ Evolutionary Genomics Department

DOE Joint Genome Institute and Lawrence Berkeley National Laboratory

2800 Mitchell Drive

Walnut Creek, CA 94598 USA

² Department of Integrative Biology

3060 Valley Life Sciences Building

University of California

Berkeley, CA 94720 USA

*Corresponding author

Email addresses:

PD: PSDehal@lbl.gov

JLB: JLBoore@lbl.gov

Abstract

Background

We present here the PhIGs database, a phylogenomic resource for sequenced genomes. Although many methods exist for clustering gene families, very few attempt to create truly orthologous clusters sharing descent from a single ancestral gene across a range of evolutionary depths. Although these non-phylogenetic gene family clusters have been used broadly for gene annotation, errors are known to be introduced by the artifactual association of slowly evolving paralogs and lack of annotation for those more rapidly evolving. A full phylogenetic framework is necessary for accurate inference of function and for many studies that address pattern and mechanism of the evolution of the genome. The automated generation of evolutionary gene clusters, creation of gene trees, determination of orthology and paralogy relationships, and the correlation of this information with gene annotations, expression information, and genomic context is an important resource to the scientific community.

Discussion

The PhIGs database currently contains 23 completely sequenced genomes of fungi and metazoans, containing 409,653 genes that have been grouped into 42,645 gene clusters. Each gene cluster is built such that the gene sequence distances are consistent with the known organismal relationships and in so doing, maximizing the likelihood for the clusters to represent truly orthologous genes. The PhIGs website contains tools that allow the study of genes within their phylogenetic framework through keyword searches

on annotations, such as GO and InterPro assignments, and sequence similarity searches by BLAST and HMM. In addition to displaying the evolutionary relationships of the genes in each cluster, the website also allows users to view the relative physical positions of homologous genes in specified sets of genomes.

Summary

Accurate analyses of genes and genomes can only be done within their full phylogenetic context. The PhIGs database and corresponding website (<http://phigs.org>) address this problem for the scientific community. Our goal is to expand the content as more genomes are sequenced and use this framework to incorporate more analyses.

Background

The continually increasing number of whole genome sequencing projects has underscored the need for a high-throughput methodology to sort genes into orthologous sets to facilitate genome analysis. With a more robust understanding of the evolutionary history for each gene in the genome, not only can we more accurately transfer annotation across organisms, but we can also address larger biological questions regarding the evolution of genomes and species as well as the functional and biochemical processes encoded within each genome. Currently, most gene annotations rely on homologs identified by pair-wise sequence similarity to transfer the presumed function. This approach has been shown to have many drawbacks [1] which lead to annotation errors. Incorrect assignments are generally due to gene duplication events [2] giving rise to paralogs that can then acquire a new function or sub-functionalize [3, 4], accelerated rates of amino acid substitution [5] and domain shuffling [6]. Simple pair-wise comparisons cannot uncover these events.

Several approaches have been proposed to address these problems. However, most of these retain the problems associated with simply clustering genes based on sequence similarity and fail to incorporate the known evolutionary relationships of species [7-9]. Alternatively, those approaches that attempt to use some aspect of the evolutionary relationships of the species to inform the clustering process fail to then create a phylogenetic tree to uncover the relationships of the genes within the clusters [10-12].

The method we present here considers *a priori* the known evolutionary relationships among the considered organisms as a guide to constructing gene clusters, then analyzes each cluster for the evolutionary relationships among the contained genes in order to reconstruct the evolutionary history of each gene family using standard analytical methods of molecular evolution. This provides a tool for the scientific community for gaining a more complete understanding of such things as evolutionary patterns of gene duplication and loss, variation in rates of amino acid substitution, and alterations in gene structure. PhIGs is the first truly comprehensive whole genome analysis phylogenetic tool allowing for accurate assessment of gene family and genome structure evolution.

Construction and Content

In this work, we develop a computational framework for the identification of sets of genes which have all descended from a single ancestral gene in the common ancestor of the lineages being examined. This collection of genes is then followed by the construction of phylogenetic trees for each set to determine relationships of the gene cluster members.

A relational database is used to store the genome annotations for each taxon. All sequence data as well as individual gene annotations, including InterPro [13] and Gene Ontology [14] assignments, intron, exon and UTR structural information, and genomic positional information are retrieved whenever available. In addition, results of analyses

such as sequence alignments, intermediate data, and trees are stored in the database.

Table 1 lists the genomes included in the current data set, which will be updated as more genome sequences become available.

The overall process involves five stages (FIGURE 1) explained in more detail below: (1) an all against all BLASTP [15] of the complete proteomes; (2) global alignment and distance calculation of the gene pairs identified by BLAST; (3) iterative, hierarchical clustering; (4) multiple sequence alignment (MSA) creation and editing; and (5) gene tree reconstruction.

All against all BLASTP and global alignment

An all-against-all BLASTP search is performed on the entire protein dataset derived from each genome. Because each BLAST only reports local alignments, a global alignment is created for each protein pair returned by BLAST with ClustalW [16]. A protein distance is then calculated using the JTT matrix and the protdist program from PHYLIP [17], hereafter referred to as the distance between genes themselves. These pair-wise protein distances and gap-free alignment lengths are then used as input for the clustering process. All alignments are stored in the PhIGs database.

Gene Clustering

Gene clustering is performed at each node of the tree, using the known evolutionary relationships of the organisms and all pair-wise protein distances as input. The objective of the clustering process is to create gene clusters at each node of the evolutionary tree

such that the genes of the descending taxa are more closely related to each other than they are to the genes from the outgroup taxa. We employ a hierarchical approach, starting at the base of the best known evolutionary tree of the organisms, and proceeding up the tree iteratively. For each bifurcating node, taxa are temporarily grouped such that those on one descending branch are labeled as clade A and those on the other as clade B. The remaining taxa, having branched earlier, are considered to be the outgroup (Figure 2). Clusters of genes are then constructed such that the included genes meet the following criteria: (1) Genes from organisms within clade A are more similar to each other than they are to genes from organisms within clade B; and (2) genes from clade A and clade B are more closely related to each other than they are to any gene in the outgroup.

Effectively, this can be achieved by first finding the top scoring alignment for each gene within any member of its sister clade, then recruiting all additional genes that have greater similarity to either one of these genes using single linkage clustering with inclusion criteria being set to the distance and alignment length of the alignment of the seed. As illustrated in Figure 2, the initial seed alignment of a pair of genes, one from each of clade A and clade B, defines an area shown in blue around representing the minimum match quality. As more genes are added to the cluster, this area grows until no more genes can be added.

Because this clustering approach is dependent on seeds, the order in which the seeds are processed will affect the clustering results. To ensure that each gene is placed in its optimal cluster, a greedy approach is used by sorting the list of seed alignments by the BLASTP score and processing the seeds by using the highest scoring seed first. In so

doing, any subsequent cluster that attempts to incorporate a gene which has already been clustered can be eliminated. It is important to note that the BLASTP score is only used to sort the seeds and clustering is based on the protein distance and alignment length. The pseudocode describing this method is available online as additional file 1: Cluster Pseudocode.

By using an iterative approach, working through the entire evolutionary tree of the organisms beginning at the base, we ensure that the most early diverging gene families create the most comprehensive clusters, with later established families properly assigned to the lineages in which they arose. Genes with a highly accelerated amino acid substitution rate, such that they are more distantly related to their sister genes than those sister genes are to a gene from the outgroup, are always excluded, since this cannot be differentiated from ancestral paralogy.

MSA and phylogenetic tree creation

A multiple sequence alignment (MSA) is created for each cluster using the ClustalW [16] program, which provides the input for phylogenetic tree reconstruction. Alignments are trimmed to remove columns that contain gap characters and the cluster is eliminated if the resulting alignment contains fewer than 100 aligned amino acid positions.

Phylogenetic trees are created using the quartet puzzling maximum likelihood method implemented in the TREE-PUZZLE [18] program using the JTT model of amino acid substitution and a gamma distribution of rates over eight rate categories with 10,000 puzzling steps to assess reliability. Quartet puzzling is chosen here as a compromise

between speed and reliability; however, the multiple sequence alignment is available for re-analysis with other tree reconstruction methods. The resulting gene tree is then reconciled with the known relationships of the organisms to determine, relative to lineage splitting, when each duplication or loss occurred, and so to determine an initial estimate of the orthology and paralogy relationships among the genes. The reconciliation process uses the most straight-forward interpretation of the tree; no alterations are made to minimize the number of duplications or gene loss events. Genes are considered orthologs if they are separated only by speciation nodes consistent with the known phylogenetic tree and considered paralogs if there is a node representing a duplication event in their shared ancestry.

The MSAs are also used to create Hidden Markov Models (HMMs) to later facilitate searching the clusters and to provide a resource for placing genes from genomes too sparsely sampled to be included in this comprehensive analysis, such as those from many EST sequencing projects.

An instructive example of this process is for the Succinyl-Coenzyme A ligase beta subunit family. In this example, considering the fungi and metazoans first as clade A and clade B, respectively, the seed alignment used is the match between a gene from *M. grisea* and one from mouse (Sucla2). The protein distance measure and gap free alignment length of this seed alignment pair is now taken to represent the maximum distance and minimum alignment length for recruiting new genes to the cluster. Any fungal or metazoan gene with a shorter distance and larger alignment length is added. In

this case the fungal gene recruits a single gene from each of the remaining fungal genomes and the mouse gene recruits two genes in each case from most of the remaining metazoan genomes and three genes from each of human, chimp, and mouse. All of these genes now included in the cluster have matches to each other that are as good or better than the initial seed alignment and do not have better matches to any other cluster. The phylogenetic tree created for this cluster ultimately shows that this gene family had a duplication at the base of the metazoan lineage, another duplication at the base of the primate lineage, and an independent duplication in the mouse lineage.

Utility

Cluster View

The PhIGs database allows users to view genes within the evolutionary context of other sequenced genomes. Because each cluster is constructed to represent the extant descendants of a single ancestral gene, the gene trees provided allow the user to see where gene duplication events have occurred and the rates of amino acid sequence change along the individual branches of the tree (Figure 3). By reconciling the gene tree with the species tree, orthology and paralogy relationships can be determined.

Comparisons of differences and similarities in annotations, such as definition line (define) gene descriptions, InterPro families, and Gene Ontology assignments, can be made with respect to the tree. The user can make a determination of whether the gene

annotations are consistent with the tree topology and whether annotations should be transferred to unannotated genes. Additionally, the genomic location and intron and exon structure of each gene is also provided, enabling analysis of such issues as whether the paralogous genes are physically clustered within a genome, indicating tandem or segmental duplication, or whether the gene family is widely dispersed. Alterations in gene intron and exon structure (and sizes) relative to other members of the cluster may be the result of biological forces acting on the genome or may simply be indicative of poor gene modeling.

The MSA for each cluster is also made available in the Cluster View. An alignment graphic, with the intron and exon structure superimposed, is shown on the page and a detailed alignment view is provided through a Jalview [19] java applet. By examining the MSA, the user can determine whether poorly aligning or missing regions of a gene contains a protein domain which may indicate the gain or loss of some function. Of course, when dealing with gene models of unknown quality, the genomic sequence should be examined for the possibility of annotation error before concluding an exon or domain loss occurred.

Gene View

All annotations related to each gene are viewable on its Gene View web page. This includes the annotations presented on the Cluster View page as well as a summary of domains found with the InterProScan [20] program (not available for all genomes) and a summary of all pair-wise alignments, including the calculated protein distance. This pair-

wise alignment information can be useful to determine whether any genes may have been left out of the cluster for failing to meet the distance and alignment length cutoffs. In some cases, this appears to be a gene model that is erroneously fragmented or merged with another, and so PhIGs provides a powerful tool for detecting these potential errors.

Searching

Searches of the database can be done by sequence similarity or by text matches to annotation fields. Text searches can be done on gene names, deflines, or InterPro annotations. Because these are associated with individual genes, the search function can be used to either return a list of genes from a selected set of taxa that contain the search term or it can return a set of clusters which contain genes matching the search term.

Because all clustering is done at the protein level, sequence similarity searches can only be performed against protein datasets. An individual sequence can be aligned against the proteins contained in the database using the BLAST program. Matches to the sequence can then be used as an entry into the cluster in which they belong. Alternatively, a similarity search can be performed directly against the Hidden Markov Models (HMMs) generated from the MSA of the clusters using the HMMER [21] program. Once a match has been made, the user can easily download either the raw fasta file of the cluster or the MSA file to create a tree incorporating the new sequence.

Synteny Maps

These analyses produce sets of true, one-to-one orthologs, and this presentation incorporates a view of their relative physical positions across multiple genomes. As

opposed to other methods that rely on sequence similarity to create comparative genome alignments, this avoids confusion that arises from paralogy. Synteny maps are generated by selecting a genomic span from a single reference genome and one or more query genomes to align (FIGURE 4). All identified orthologous genes between the selected genome and each of the query genomes are shown.

Discussion and Conclusion

The rapidly increasing number of sequenced genomes allows us to study genes and genomes within an evolutionary context. Not only does this assist in the transfer of annotations between genes, but also allows us to uncover how the forces of evolution have shaped each genome. The PhIGs database project seeks to facilitate comparative genomic, phylogenomic, and functional genomic studies by providing a comprehensive resource for the determination of the evolutionary history for all genes from the fully sequenced genome projects. The two main properties that differentiate the PhIGs database from other clustering methods are the use of the known evolutionary relationships of the species to create gene clusters representing the descendants of a single ancestral gene and the creation of a complete phylogenetic gene tree of the cluster members using widely accepted analytic methods of molecular evolution. By combining this phylogenetic information with functional annotation, gene structure, genomic position and other datasets, the PhIGs database will prove to be a valuable resource for all fields of biology currently using genomic data.

The scientific applications of the PhIGs database are broad, extending beyond practical genome annotation and analysis. For instance, obvious applications are the use of orthologous gene clusters for: (1) organismal phylogenetic reconstruction; (2) the study of genome evolution by gene duplication; (3) gene structure evolution through the gain and loss of exons, introns, and domains; (4) the identification of gene family expansions and losses and 5) genome evolution. The PhIGs analyses have already been used to compare specifically the whole genomes of a tunicate, fish, mouse, and human, demonstrating that the relative positions in the human genome of paralogs generated by duplications at the base of vertebrates provide clear evidence in favor of the contentious hypothesis of two rounds of whole genome duplication having occurred at the base of the vertebrates, and perhaps providing the raw material for vertebrate complexity [22]. Further applications can be developed to meet other analytical needs of the scientific community.

Future development includes improvements to the underlying clustering method, incorporation of more annotation data, creation of more analysis tools and more rapid updates of newly available genomes. The functionality of the PhIGs database is currently accessible through the web interface and data files of orthology relationships for download. Our goal is to convert this into an open source project to help maintain and expand this as a resource for the scientific community.

Authors' contributions

Both authors conceived and designed the project. PSD did all of the programming and implementation. Both authors wrote and approved the final manuscript.

Acknowledgements

We thank S. Rash, W. Huang, and A. Porter for technical assistance. Funding for salary support was in part from the National Science Foundation awards MCB-0242131 and EF-0328516. This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC03-76SF00098.

References

1. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8**(3):163-167.
2. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**(2):99-113.
3. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151**(4):1531-1545.
4. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**(1):459-473.
5. Gaucher EA, Gu X, Miyamoto MM, Benner SA: **Predicting functional divergence in protein evolution by site-specific rate shifts.** *Trends Biochem Sci* 2002, **27**(6):315-321.
6. Doolittle RF: **The multiplicity of domains in proteins.** *Annu Rev Biochem* 1995, **64**:287-314.
7. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**(1):41.
8. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**(7):1575-1584.
9. Liu J, Rost B: **Domains, motifs and clusters in the protein universe.** *Curr Opin Chem Biol* 2003, **7**(1):5-11.
10. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res* 2005, **33 Database Issue**:D476-480.
11. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178-2189.
12. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J *et al*: **Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).** *Genome Res* 2002, **12**(3):493-502.
13. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD *et al*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**(1):37-40.
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.

15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
16. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.
17. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** Department of Genome Sciences, University of Washington, Seattle: Distributed by the author; 2004.
18. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**(3):502-504.
19. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20**(3):426-427.
20. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**(9):847-848.
21. **<http://hmmer.wustl.edu>**
22. Dehal P, Boore JL: **Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.** *PLoS Biology* 2005, **3**(10):e314.

Figure Legends

Figure 1: Flowchart of the PhIGs process. This is a graphical overview of the pipeline for processing gene models from many genomes into the PhIGs analysis.

Figure 2: Illustration of the clustering method. The tree shown on the left side of the figure indicates the evolutionary relationships among several hypothetical organisms, four from Clade A, two from Clade B, and one that is an outgroup. The right side of the figure illustrates a protein distance graph with circles representing proteins colored to conform to each organism, with the spatial distance of the circles proportional to their sequence distance. The cluster is created by identifying a pair of sequences (a seed) that is the shortest distance from any Clade A protein to any Clade B protein. The cluster is then grown by adding all proteins that have a shorter distance than the seed until no additions can be made. The blue cloud represents one such cluster. See text for more details.

Figure 3: An example phylogenetic tree. This is one output of the PhIGs analysis that is shown on the Cluster View webpage. Instead of simply listing the members of a cluster, a phylogenetic tree is created showing the evolutionary relationships of this multigene family. In this example, we can see that this family had gene duplication events at the base of vertebrates and in the fish lineage. Because the branch lengths are proportional to the rate of amino acid substitutions, we can see how rates of evolution have varied.

Figure 4: An example Synteny Map. Genes ranging from number 205 through 301 on chicken chromosome 2 (numbered as they occur from the p-telomere to q-telomere along the chromosome) are shown as rectangles in the center of the diagram. On the left and right are the orthologs of these genes found in the human and mouse genomes as determined by the PhIGs analysis, shown as they are arranged. Black connecting lines join orthologs in the same relative transcriptional orientation whereas red lines indicate those that are inverted. Blue rectangles indicate intervening genes without identified orthologs in the genomes being compared. Cyan rectangles that do not have connecting lines, as can be seen for a portion of mouse chromosome 2, indicate that orthologs exist in chicken (the query genome), but not in the portion specified for this page.

Table 1: Taxa currently included in PhIGs

Species	Taxonomy	Genome source	URL
<i>Homo sapiens</i>	Primates	Ensembl 22.34d.1	www.ensembl.org
<i>Pan troglodytes</i>	Primates	Ensembl 22.1.1	www.ensembl.org
<i>Mus musculus</i>	Rodentia	Ensembl 22.32b.1	www.ensembl.org
<i>Rattus norvegicus</i>	Rodentia	Ensembl 22.3b.1	www.ensembl.org
<i>Gallus gallus</i>	Aves	Ensembl 22.1.1	www.ensembl.org
<i>Takifugu rubripes</i>	Pisces	Ensembl 22.2c.1	www.ensembl.org
<i>Danio rerio</i>	Pisces	Ensembl 22.3b.1	www.ensembl.org
<i>Ciona intestinalis</i>	Urochordata	JGI ciona4	genome.jgi-psf.org
<i>Drosophila melanogaster</i>	Arthropoda	Ensembl 22.3a.1	www.ensembl.org
<i>Anopheles gambiae</i>	Arthropoda	Ensembl 22.2b.1	www.ensembl.org
<i>Caenorhabditis elegans</i>	Nematoda	Ensembl 22.116a.1	www.ensembl.org
<i>Caenorhabditis briggsae</i>	Nematoda	Ensembl 24.24.1	www.ensembl.org
<i>Ustilago maydis</i>	Basidiomycota	Broad release 2	www.broad.mit.edu/annotation/fungi/fgi/
<i>Phanerochaete chrysosporium</i>	Basidiomycota	JGI whiterot1	genome.jgi-psf.org
<i>Cryptococcus neoformans</i>	Basidiomycota	TIGR CNA1	www.tigr.org/tdb/fungal/
<i>Schizosaccharomyces pombe</i>	Ascomycota	Pompep version 19	www.sanger.ac.uk/Projects/S_pombe/
<i>Aspergillus nidulans</i>	Ascomycota	Broad release 1.3	www.broad.mit.edu/annotation/fungi/fgi/
<i>Fusarium graminearum</i>	Ascomycota	Broad release 1.1	www.broad.mit.edu/annotation/fungi/fgi/
<i>Trichoderma reesei</i>	Ascomycota	JGI trire1	genome.jgi-psf.org
<i>Magnaporthe grisea</i>	Ascomycota	Broad release 2.4	www.broad.mit.edu/annotation/fungi/fgi/
<i>Neurospora crassa</i>	Ascomycota	Broad release 3	www.broad.mit.edu/annotation/fungi/fgi/
<i>Candida albicans</i>	Ascomycota	Stanford release 19	www-sequence.stanford.edu/group/candida/
<i>Saccharomyces cerevisiae</i>	Ascomycota	Saccharomyces Genome Database	www.yeastgenome.org/

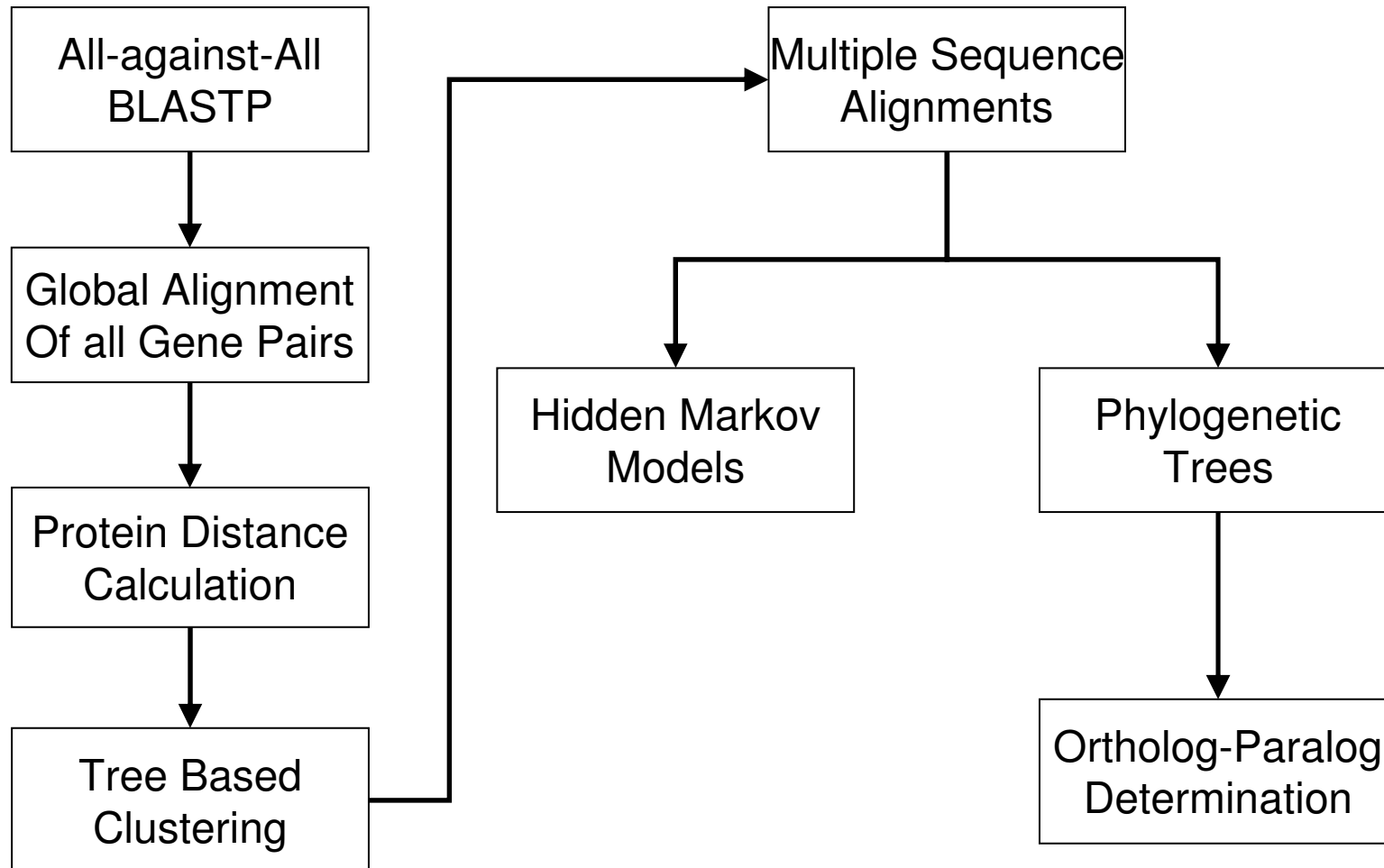
Additional file

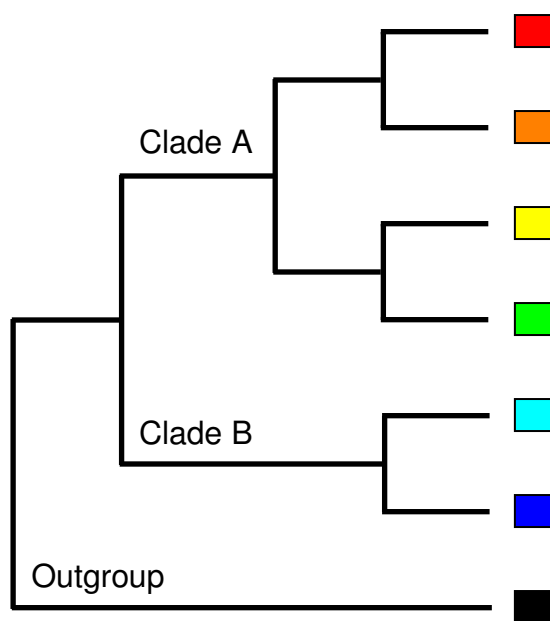
File name: Pseudocode.pdf

File format: pdf

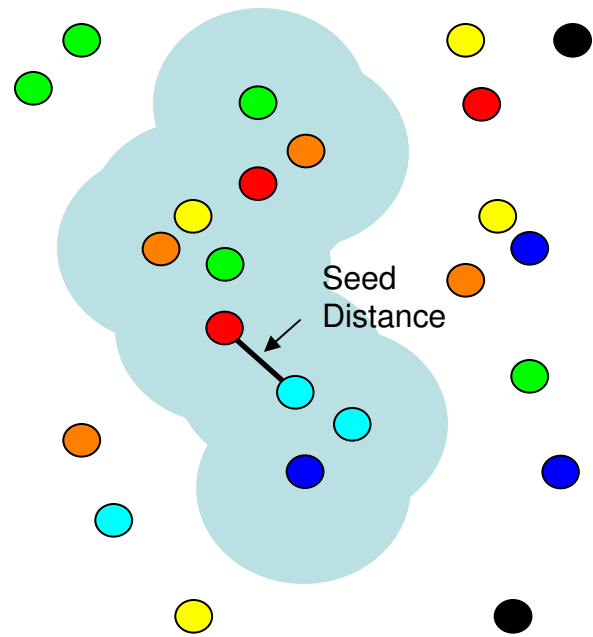
Title of data: Pseudocode

Description of data: The pseudocode for PhIGs database construction





Species Tree



Protein Distance Graph

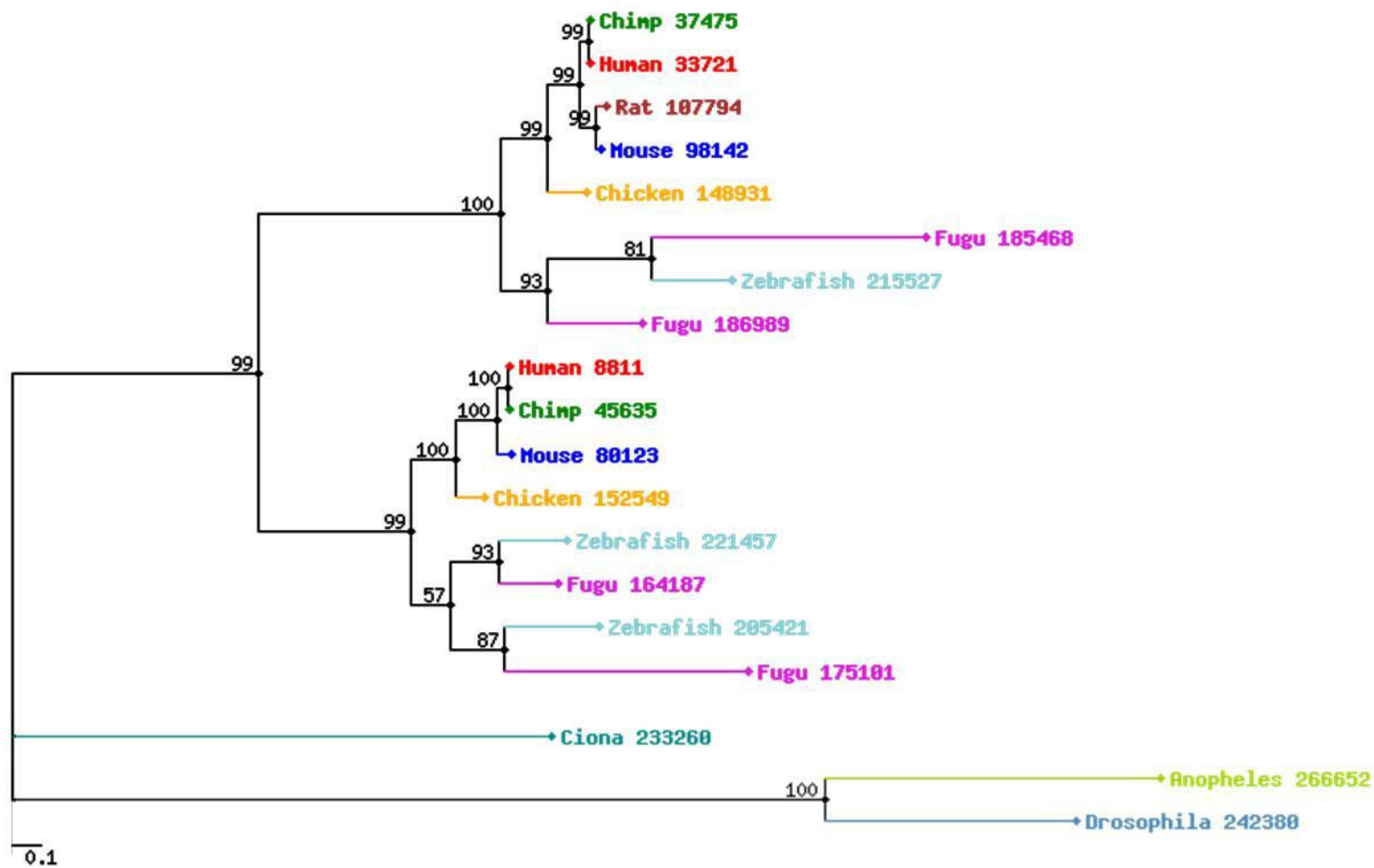


Figure 3

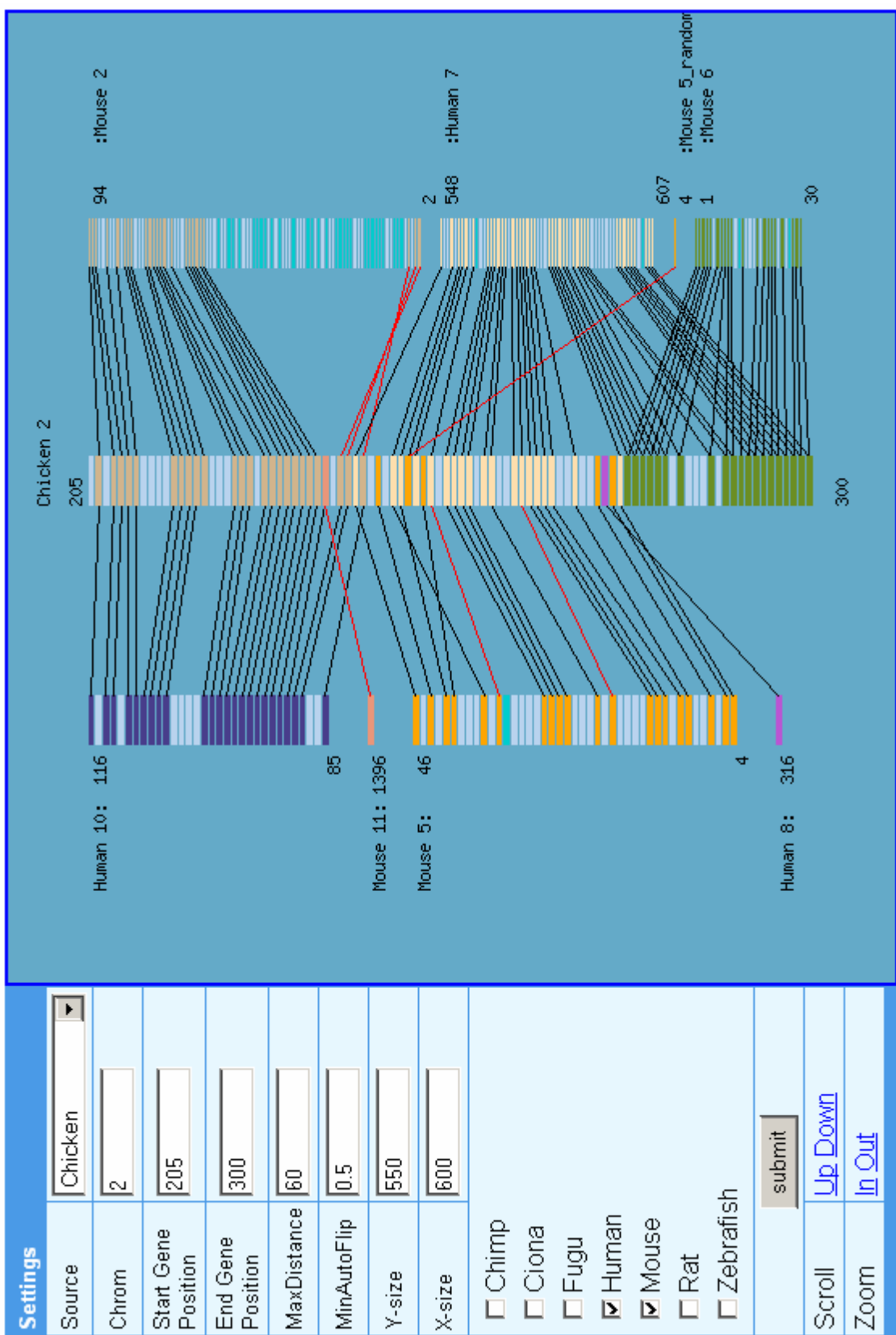


Figure 4

Additional files provided with this submission:

Additional file 1 : pseudocode.pdf : 20Kb

<http://www.biomedcentral.com/imedia/1442830240779874/sup1.PDF>