Published online 2 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D731–D737***
doi:10.1093/nar/gkn645

PhyloPat: an updated version of the phylogenetic pattern database contains gene neighborhood

Tim Hulsen[1,*], Peter M. A. Groenen[2], Jacob de Vlieg[1,2] and Wynand Alkema[2]

[1]Computational Drug Discovery (CDD), CMBI, NCMLS, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen and [2]Department of Molecular Design and Informatics, Schering-Plough, Molenstraat 110, 5340 BH Oss, The Netherlands

Received August 12, 2008; Revised and Accepted September 17, 2008

ABSTRACT

Phylogenetic patterns show the presence or absence of certain genes in a set of full genomes derived from different species. They can also be used to determine sets of genes that occur only in certain evolutionary branches. Previously, we presented a database named PhyloPat which allows the complete Ensembl gene database to be queried using phylogenetic patterns. Here, we describe an updated version of PhyloPat which can be queried by an improved web server. We used a single linkage clustering algorithm to create 241 697 phylogenetic lineages, using all the orthologies provided by Ensembl v49. PhyloPat offers the possibility of querying with binary phylogenetic patterns or regular expressions, or through a phylogenetic tree of the 39 included species. Users can also input a list of Ensembl, EMBL, EntrezGene or HGNC IDs to check which phylogenetic lineage any gene belongs to. A link to the FatiGO web interface has been incorporated in the HTML output. For each gene, the surrounding genes on the chromosome, color coded according to their phylogenetic lineage can be viewed, as well as FASTA files of the peptide sequences of each lineage. Furthermore, lists of omnipresent, polypresent, oligopresent and anticorrelating genes have been included. PhyloPat is freely available at http://www.cmbi.ru.nl/phylopat.

INTRODUCTION

Phylogenetic patterns show the presence or absence of certain genes in a set of whole genome sequences derived from different species. These patterns can be used to determine sets of genes that occur only in certain evolutionary branches. The use of phylogenetic patterns has been common practice as increasing amounts of orthology data have become available. One example is clusters of orthologous groups (COGs) (1), which included a Phylogenetic Patterns Search (PPS) and an Extended Phylogenetic Patterns Search (EPPS) (2) tool, providing the possibility of querying the phylogenetic patterns of the COG protein database using regular expressions. The ortholog database OrthoMCL-DB (3) also offers this possibility. However, PPS tools have only been available for querying proteins, and not for querying genes. The PhIGs (4), Hogenom (5) and TreeFam (6) databases all offer phylogenetic clustering of genes, but do not have the functionality of phylogenetic patterns, and do not include the full range of Ensembl (7) species. Moreover, these databases do not provide additional genomic information such as function and organization of neighboring genes. In September 2006, we introduced a database named PhyloPat (8) that offers the possibility of querying the Ensembl database using any phylogenetic pattern. Here, we show the newest version of this database, and show applications of the new functionalities that have been implemented in the web server, such as a gene neighborhood view, anticorrelating patterns, support of Entrez Gene (9) IDs and direct sequence retrieval of members of a phylogenetic lineage.

DATABASE CONTENT AND CONSTRUCTION

Content

A set of phylogenetic lineages was constructed containing all the genes in Ensembl that have orthologs in other species according to the BioMart (10) database. This set covers all of the 39 (eukaryotic) species available in Ensembl version 49 (preversions and low coverage genomes not taken into account). First, we collected the complete set of orthologies between these 39 species, consisting of 741 species pairs, 815 452 genes and 19 010 478 orthologous relationships. The orthologies within Ensembl v49 consist of 11 446 546 one-to-one relationships, 4 588 300 one-to-many relationships and 2 975 632 many-to-many relationships. These orthologies are determined by the thorough Ensembl ortholog detection pipeline

*To whom correspondence should be addressed. Tel: +31 412 668305; Fax: +31 412 662553; Email: t.hulsen@cmbi.ru.nl

(http://www.ensembl.org/info/about/docs/compara/homology_method.html). This pipeline starts with the collection of a number of best reciprocal hits [BRHs, proven to be accurate (11)] and best score ratio (BSR) values from a WU BLASTP/Smith–Waterman whole-genome comparison. These are used to create a graph of gene relations, followed by a clustering step. These clusters are then applied to build a multiple alignment using MUSCLE (12) and a phylogenetic tree using TreeBeST (http://treesoft.sourceforge.net/treebest.shtml). Finally, the above mentioned orthologous relationships are inferred from this gene tree.

Construction

After the collection of all orthologous pairs, we generated phylogenetic lineages using a single linkage algorithm. First, we determined the evolutionary order of the studied species using the NCBI Taxonomy (13) database. The phylogenetic tree [phylogram, created by TreeView (14)] of these species, together with some phylogenetic branch names, are shown in Figure 1. Second, we used this phylogenetic tree as a starting point for building our phylogenetic lineages. For each gene in the first species (*Saccharomyces cerevisiae*), we looked for orthologs in the other species. All orthologs were added to the phylogenetic lineage, and in the next round were checked for orthologs themselves, until no additional orthologies were found for any of the genes. This process was repeated for all genes in all 39 species that were not yet connected to any phylogenetic lineage yet. The complete phylogenetic lineage determination generated 241 697 lineages. Please note that the phylogenetic order that we have determined here does not affect the construction of the phylogenetic lineages in any way: changing the order only influences the numbering of the phylogenetic lineages but not the contents of the lineages. This is due to our clustering algorithm, in which each orthologous relationship is treated symmetrically. Figure 2 shows the database scheme; the phylogenetic lineages, gene neighborhood and some mapping information have been stored in six tables, and optimized for fast querying.

WEB APPLICATION

Overview

We developed an intuitive web interface (Figure 3) to query the PhyloPat MySQL database containing these

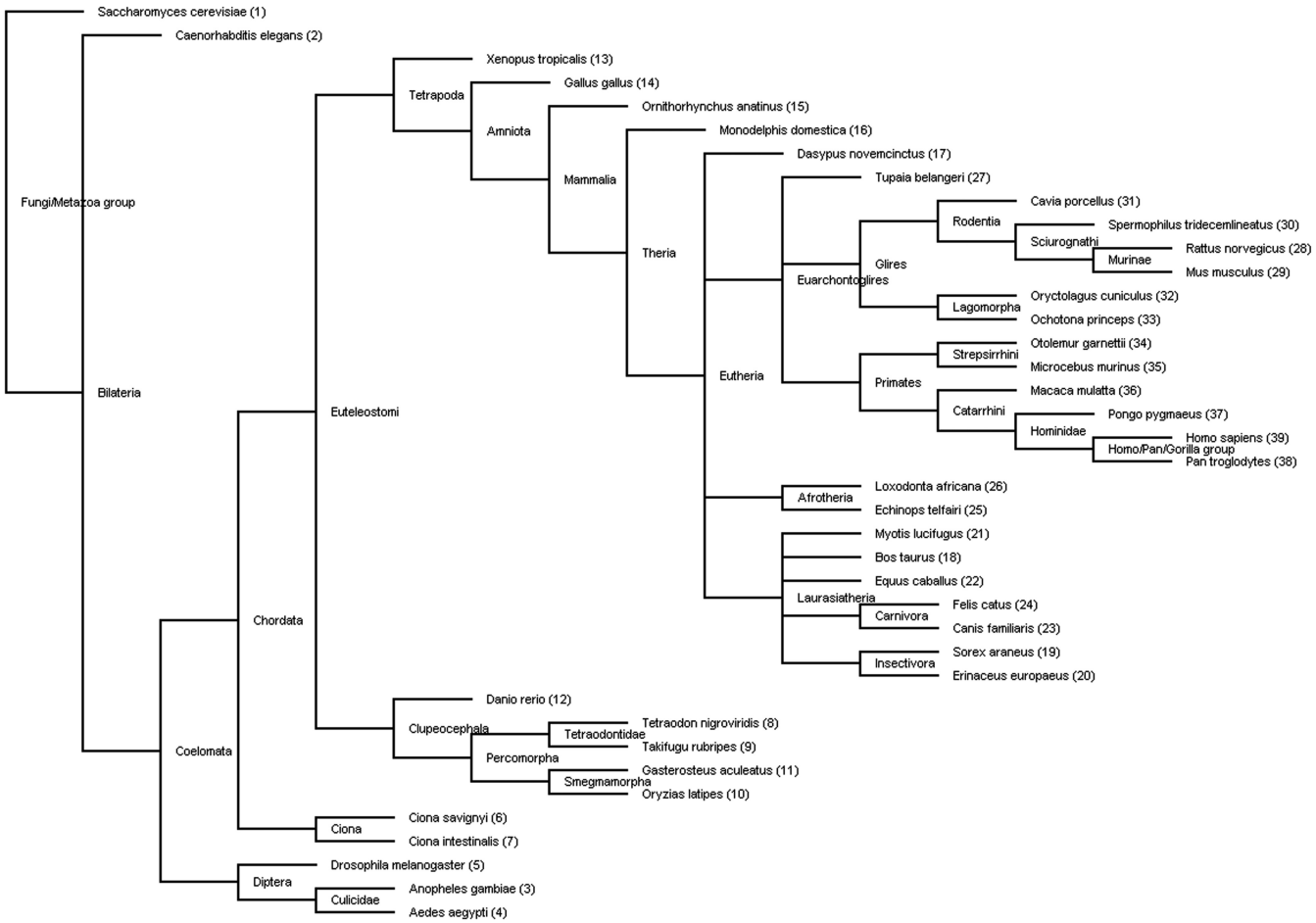


Figure 1. Phylogenetic tree of all species present in PhyloPat. This is the unrooted NCBI Taxonomy tree of all species available in Ensembl and PhyloPat. The numbers are the order in which the species are shown on the PhyloPat results pages. A phylogram version of this tree is available through the website.

phylogenetic lineages and the derived phylogenetic patterns. As input a phylogenetic pattern is used, generated by clicking a set of radio buttons or by typing a regular expression, or a list of Ensembl (7), EMBL (15), Entrez Gene (9) or HGNC symbols (16). The application of MySQL regular expressions provides enhanced querying. The output can be given in HTML, Excel or plain text format. A link to the FatiGO (17) web interface has been incorporated in the HTML output, creating easy access to functional annotation of the genes in the phylogenetic lineage. Each phylogenetic lineage can be viewed separately by clicking the PhyloPat ID (PPID). This view gives all Ensembl IDs within the phylogenetic lineage plus the HGNC (16) symbols, FASTA-format files with the proteins sequences, and the gene neighborhood. The web interface also provides some example queries, the 100 most occurring patterns, anticorrelating patterns and numerical overviews of lineages that are present in (i) all species (ii) almost all species and (iii) only one or two species. Finally, a phylogenetic tree of all included species is provided, through which each branch can be selected to view a list of branch-specific genes. This tree can be downloaded in PHYLIP (18) format.

Omnipresent genes

An analysis of all lineages with the phylogenetic pattern '111111111111111111111111111111111111111' (or MySQL regular expression '^1+$') gives a list of 'omnipresent' genes, i.e. present in all 39 species. We found 688 omnipresent genes, which most likely have important functions, since they are present in all species. Figure 4a shows the top 15 of 5th level GO (19) molecular functions for all 2345 human genes within these omnipresent phylogenetic lineages, generated by FatiGO (17). To compare the results, we also show the top 15 of 5th level GO molecular functions for the complete set of human genes (32 584 in Ensembl v49), in Figure 4b. The GO molecular function annotation shows that omnipresent genes are more often involved in adenyl nucleotide binding compared than random human genes (17.90% versus 12.80%), and less

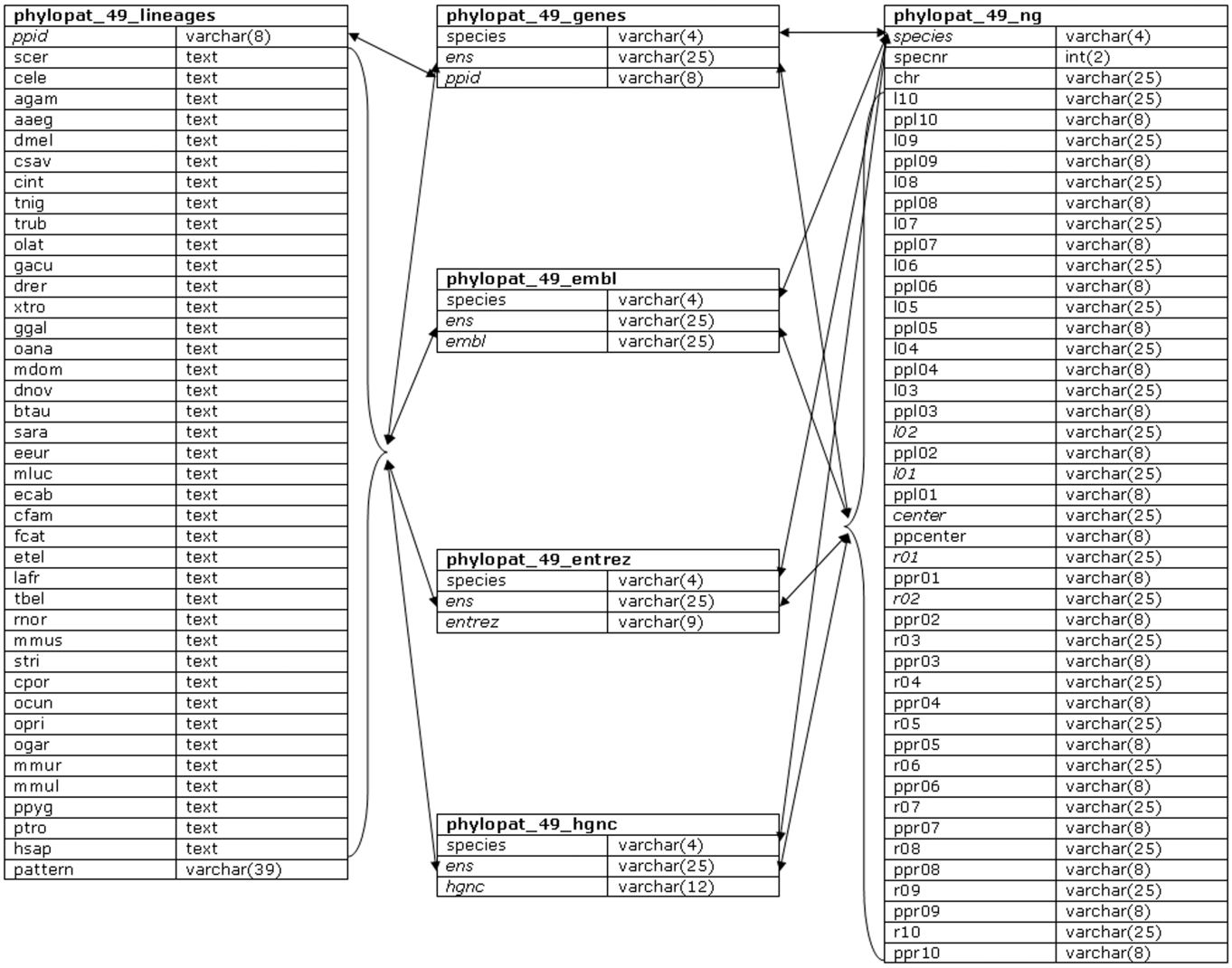


Figure 2. The PhyloPat database scheme. The database scheme shows all six tables used in the application. Table names are in bold, primary keys are in italic. Links between fields are shown with arrows. The left side of each column shows the field names, the right side shows the field types.

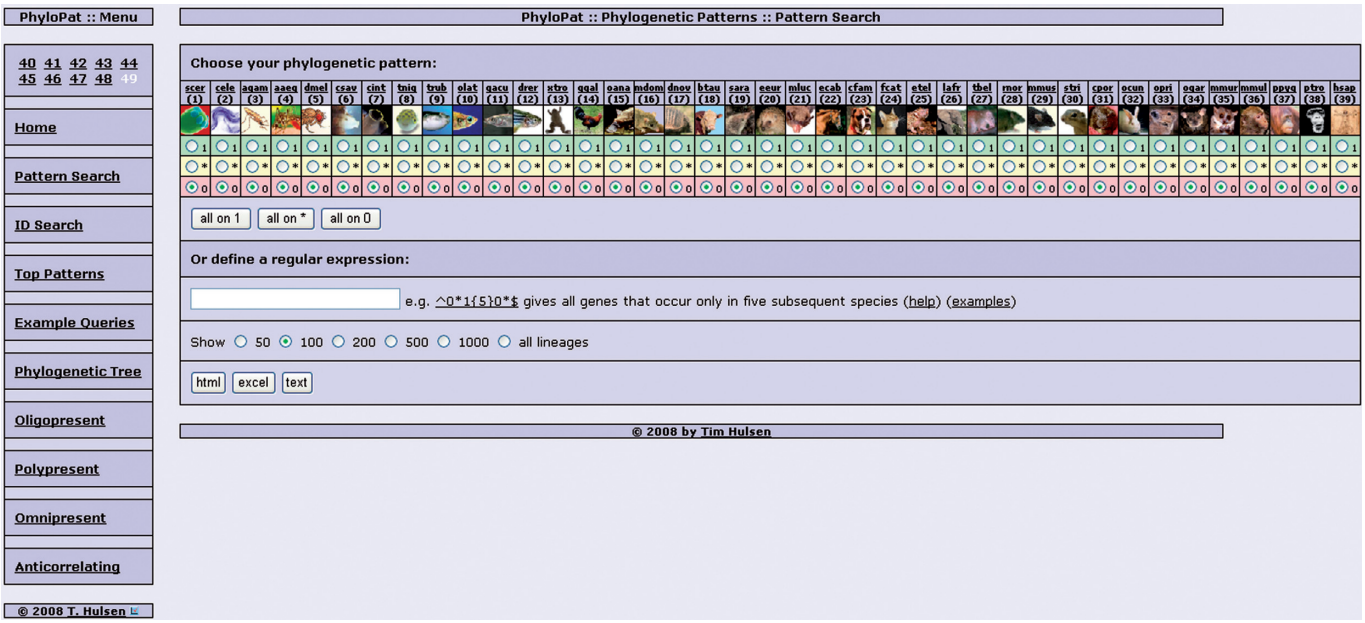


Figure 3. The PhyloPat web interface (Pattern Search tab). The web interface has the menu on the left and the input/results page on the right. On the pattern search page, the user can generate a phylogenetic pattern by clicking a radio button for each species. 1 = present, * = present/absent, 0 = absent. The buttons directly below put all 39 species on the corresponding mode. MySQL regular expressions offer the possibility of advanced querying. The user can choose to show any number of lineages and choose the output format: HTML, Excel or plain text.

often involved in transition metal ion binding (16.11% versus 23.23%). The genes with G-protein coupled receptor activity seem to be underrepresented in the omnipresent genes; whereas from the complete human genome 7.85% is involved in GPCR activity, this molecular function is not in the top 15 for the omnipresent gene set, with only 0.64%. This is likely due to the fact that GPCRs are almost absent in *S. cerevisiae* and are a class of molecules with highly specific functions in different organisms (20). However, this still needs to be proven by experimental data.

Oligopresent genes

The distribution of 'oligopresent' genes (genes that exist in only one or two species) can be used to determine which species are evolutionary most related, as the number of shared genes, that are absent in other species, can be used as a measure for the phylogenetic distance (21). It is apparent that *Ciona savignyi* and *C. intestinalis* are the closest relatives (1866 oligopresent genes), followed by *Anopheles gambiae* and *Aedes aegypti* (1206 oligopresent genes) and *Rattus norvegicus* and *Mus musculus* (557 oligopresent genes). These results correspond with the current view on the evolutionary relationships between these species. It should also be noted that the incomplete orthology information contained in the BioMart database causes the number of genes present in only one species to be very high. This will improve with each new Ensembl release, as orthology information and functional annotation are expanded and improved in each release.

Polypresent genes

A second measure for evolutionary relatedness is the distribution of 'polypresent' genes: genes that are missing in only one or two species. *Saccharomyces cerevisiae* has the highest number of missing polypresent genes: 552 polypresent genes do not occur in *S. cerevisiae* only, and 505 polypresent genes are not present in *S. cerevisiae* and a second species. When not taking into account the outlier species *S. cerevisiae*, both *Ciona* species have the highest number of missing polypresent genes: 18 lineages occur in all species except for *C. savignyi* and *C. intestinalis*.

Anticorrelating patterns

Figure 5 gives an overview of anticorrelating pattern pairs, and the numbers of lineages that have these patterns. Anticorrelating patterns are defined as patterns that are exactly opposite ('0'→'1' and '1'→'0'), and have at least five 0s and at least five 1s. Phylogenetic lineages with anticorrelating patterns can be functionally completely different, but could also be highly similar in function. For example, phylogenetic lineage PP110132 has the pattern '000000000000000001011100111001111110010' (upper line of Figure 5), while phylogenetic lineage PP004906 has the anticorrelating pattern '11111111111111110100011000011000000001101'. The PP110132 genes are all annotated by Ensembl as 'no description', but some of the PP004906 genes are annotated as 'Chromatin modifying protein 1b' (CHMP1b, in *Danio rerio*, *Gallus gallus*, *M. musculus* and *Xenopus tropicalis*). The PP110132 genes can be analogous to CHMP1b, i.e. performing a similar function to CHMP1b, without being evolutionary related.

Gene neighborhood

Figure 6 shows the gene neighborhood for PhyloPat ID PP000255 (ERN1, ERN2). The human gene ENSG00000134398 has two predicted orthologs in

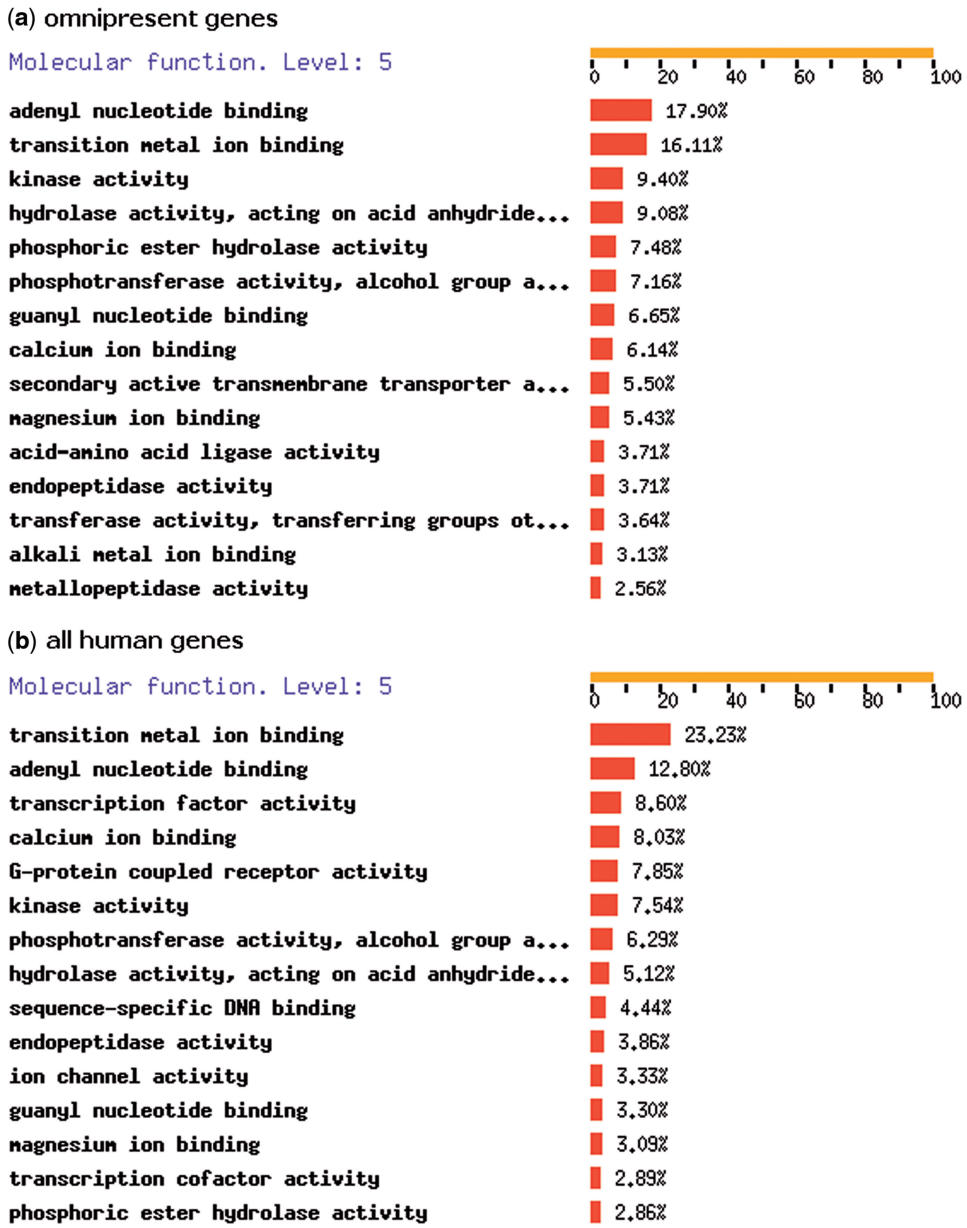


Figure 4. Gene Ontology annotations of (i) omnipresent and (ii) all human genes, created by FatiGO. (**a**) The 5th level Gene Ontology Molecular Function annotations for all 2345 human genes in omnipresent lineages. (**b**) The 5th level Gene Ontology Molecular Function annotations for all 32 584 human genes, used as a reference set.

chimpanzee: gene ENSPTRG00000007893 and gene ENSPTRG00000009535. However, only the gene neighborhoods of gene ENSPTRG00000007893 and gene ENSG00000134398 correspond, for nine of the nearest neighbors. This is called 'orthologous conservation of gene neighborhood' and it shows that the two genes involved are evolutionary related (22). In this case, we would say that the 'true' ortholog of gene ENSG00000134398 is very likely to be gene ENSPTRG00000007893. Apart from inferring 'true' orthology from the genome organization, gene neighborhoods can also be used to infer functional annotation for genes or build hypotheses about the processes or pathways that genes might be involved in.

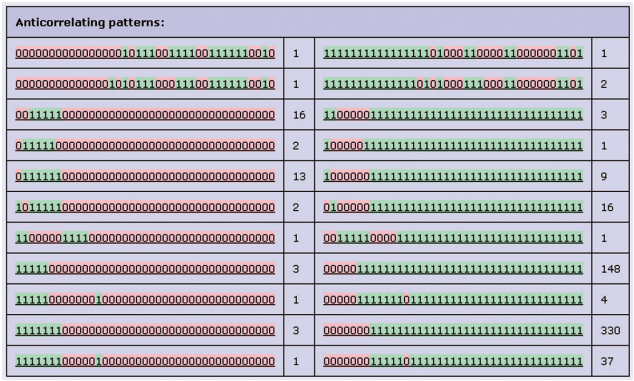
FASTA-format sequence files

Both the pattern search output and the gene neighborhood view contain links to FASTA files of the peptide sequences belonging to each phylogenetic lineage. We included two types of files: one with all peptide sequences (marked by 'A') and one with only the longest translation of each gene (marked by 'L').

DISCUSSION AND CONCLUSION

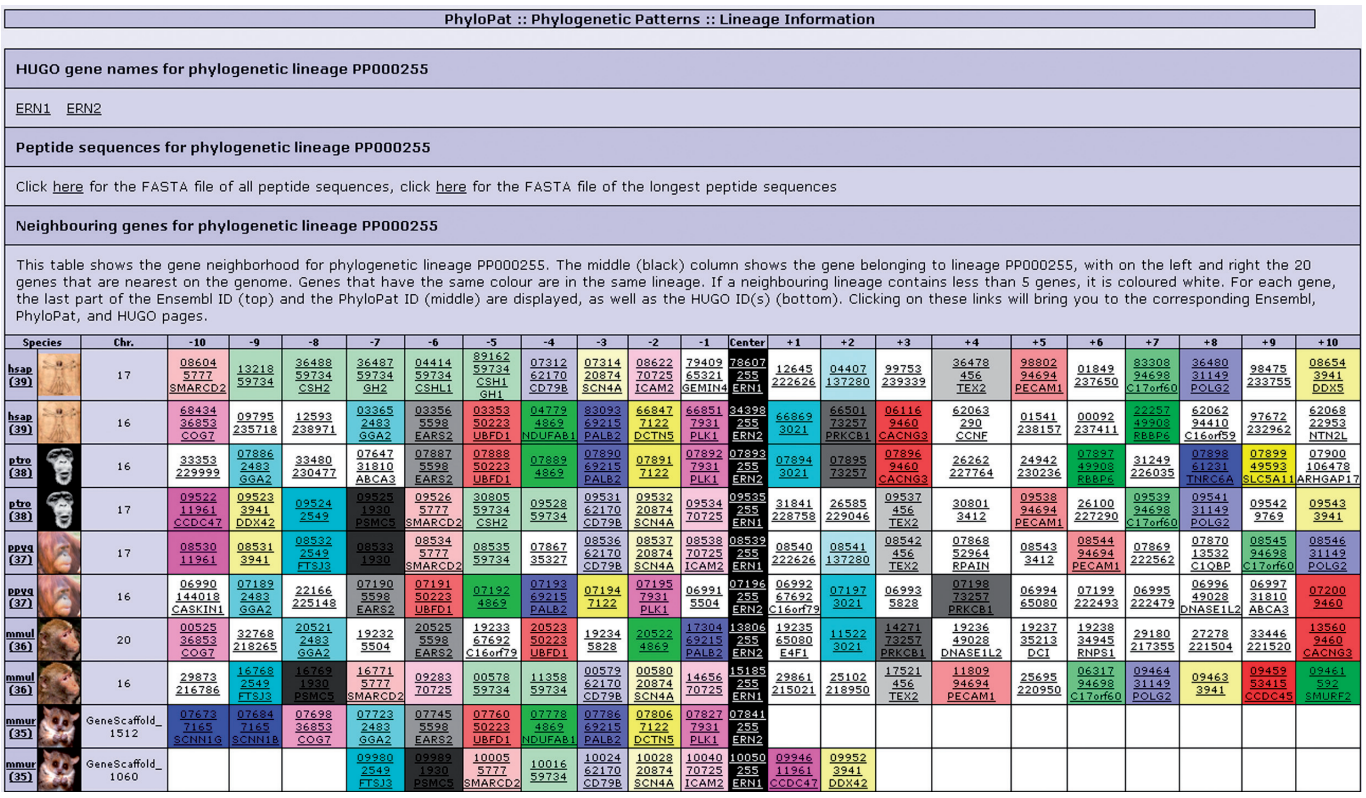
The above examples show that PhyloPat is useful in orthology detection, evolutionary studies and gene annotation. It builds on and expands the concept of

phylogenetic pattern tools like EPPS (2), and on gene databases like PhiGs (4), Hogenom (5) and TreeFam (6). The originality of PhyloPat lies in the combination of these two aspects: phylogenetic pattern querying and gene family databases. In PhyloPat, it is possible to determine a species set that should be included (1), a species set that should be excluded (0) and a species set which presence is indifferent (∗). This, and the use of regular expression queries, enables quite complex PPSs and clustering. Furthermore, we aim to provide an easy-to-use web interface in which the Ensembl database can be queried using phylogenetic patterns. Users can see which gene families are present in a certain species set but missing in another species set. The output of PPSs can be easily analyzed by the FatiGO tool, like we demonstrated in Figure 4. Another advantage of PhyloPat is that it relies on the Ensembl database only. Other gene databases use a wide range of gene and protein data sets, each with their own standards and methodologies. By using only the popular Ensembl database as input, we create a nonredundant database, through which it is possible to easily study lineage-specific expansions of gene families. Finally, the new options of the web application of PhyloPat make it easier to query the database and to retrieve the sequences from the lineage of interest. The gene neighborhood view adds a new level of information: genomic context can help in locating evolutionary-related genomic clusters of genes, and in detecting the 'true orthologs' within large sets of predicted orthologs as well as in functional annotating less well known genes. PhyloPat will be updated with each major Ensembl release to ensure up-to-date and reliable phylogenetic lineages. Older versions of PhyloPat (starting with version 40) are maintained and linked to the corresponding Ensembl archive pages. Future versions of



Anticorrelating patterns:			
00000000000000001011100111001111110010	1	11111111111111101000100001000000001101	1
00000000000000101011100011100111110010	1	11111111111110101000111000100000001101	2
00111110000000000000000000000000000000	16	11000001111111111111111111111111111111	3
01111100000000000000000000000000000000	2	10000011111111111111111111111111111111	1
01111110000000000000000000000000000000	13	10000001111111111111111111111111111111	9
10111110000000000000000000000000000000	2	01000001111111111111111111111111111111	16
11000001111000000000000000000000000000	1	00111110000111111111111111111111111111	1
11111000000000000000000000000000000000	3	00000111111111111111111111111111111111	148
11111000000010000000000000000000000000	1	00000111111101111111111111111111111111	4
11111110000000000000000000000000000000	3	00000001111111111111111111111111111111	330
11111110000010000000000000000000000000	1	00000001111101111111111111111111111111	37

Figure 5. Anticorrelating patterns. The anticorrelating pattern page of PhyloPat version 49. Columns 1 and 3 show the anticorrelating phylogenetic patterns, columns 2 and 4 the numbers of phylogenetic lineages that have these patterns.



PhyloPat :: Phylogenetic Patterns :: Lineage Information

HUGO gene names for phylogenetic lineage PP000255

ERN1 ERN2

Peptide sequences for phylogenetic lineage PP000255

Click here for the FASTA file of all peptide sequences, click here for the FASTA file of the longest peptide sequences

Neighbouring genes for phylogenetic lineage PP000255

This table shows the gene neighborhood for phylogenetic lineage PP000255. The middle (black) column shows the gene belonging to lineage PP000255, with on the left and right the 20 genes that are nearest on the genome. Genes that have the same colour are in the same lineage. If a neighbouring lineage contains less than 5 genes, it is coloured white. For each gene, the last part of the Ensembl ID (top) and the PhyloPat ID (middle) are displayed, as well as the HUGO ID(s) (bottom). Clicking on these links will bring you to the corresponding Ensembl, PhyloPat, and HUGO pages.

Species	Chr.	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	Center	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
hsap (39)	17	08604 5777 SMARCD2	13218 59734	36488 59734 CSH2	36487 59734 GH2	04414 59734 CSHL1	89162 59734 CSH1 GH1	07312 62170 CD79B	07314 20874 SCN4A	08622 70725 ICAM2	79409 65321 GEMIN4	78607 255 ERN1	12645 222626	04407 137280	99753 239339	36478 456 TEX2	98802 94694 PECAM1	01849 237650	83308 94698 C17orf60	36480 31143 POLG2	98475 233755	08654 3941 DDX5
hsap (39)	16	68434 36853 COG7	09795 235718	12593 238971	03365 2483 GGA2	03356 5598 EARS2	03353 50223 UBFD1	04779 4869 NDUFAB1	83093 69215 PALB2	66847 7122 DCTN5	66851 7931 PLK1	34398 255 ERN2	66869 3021	66501 73257 PRKCB1	06116 9460 CACNG3	62063 290 CCNF	01541 238157	00092 237411	22257 49908 FBBP6	62062 94410 C16orf59	97672 232962	62068 22953 NTN2L
ptro (38)	16	33353 229999	07886 2483 GGA2	33480 230477	07647 31810 ABCA3	07887 5598 EARS2	07888 50223 UBFD1	07889 4869	07890 69215 PALB2	07891 7122	07892 7931 PLK1	07893 255 ERN2	07894 3021	07895 73257	07896 9460 CACNG3	26262 227764	24942 230236	07897 49908 FBBP6	31249 226035	07898 61231 TNRC6A	07899 49593 SLC5A11	07900 106478 ARHGAP17
ptro (38)	17	09522 11961 CCDC47	09523 3941 DDX42	09524 2549	09525 1930 PSMC5	09526 5777 SMARCD2	30805 59734 CSH2	09528 59734	09531 62170 CD79B	09532 20874 SCN4A	09534 70725	09535 255 ERN1	31841 228758	26585 229046	09537 456 TEX2	30801 3412	09538 94694 PECAM1	26100 227290	09539 94698 C17orf60	09541 31143 POLG2	09542 9769	09543 3941
ppyg (37)	17	08530 11961	08531 3941	08532 2549 FTSJ3	08533 1930	08534 5777 SMARCD2	08535 59734	07867 35327	08536 62170 CD79B	08537 20874 SCN4A	08538 70725 ICAM2	08539 255 ERN1	08540 222626	08541 137280	08542 456 TEX2	07868 52964 RPAIN	08543 3412	08544 94694 PECAM1	07869 222562	07870 13532 C1QBP	08545 94698 C17orf60	08546 31149 POLG2
ppyg (37)	16	06990 144018 CASKIN1	07189 2483 GGA2	22166 225148	07190 5598 EARS2	07191 50223 UBFD1	07192 4869	07193 69215 PALB2	07194 7122	07195 7931 PLK1	06991 5504	07196 255 ERN2	06992 67692 C16orf79	07197 3021	06993 5828	07198 73257 PRKCB1	06994 65080	07199 222493	06995 222479	06996 49028 DNASE1L2	06997 31810 ABCA3	07200 9460
mmul (36)	20	00525 36853 COG7	32768 218265	20521 2483 GGA2	19232 5504	20525 5598 EARS2	19233 67692 C16orf79	20523 50223 UBFD1	19234 5828	20522 4863	17304 69215 PALB2	13806 255 ERN2	19235 65080 E4F1	11522 3021	14271 73257 PRKCB1	19236 49028 DNASE1L2	19237 35213 DCI	19238 34945 RNPS1	29180 217355	27278 221504	33446 221520	13560 9460 CACNG3
mmul (36)	16	29873 216786	16768 2549 FTSJ3	16769 1930 PSMC5	16771 5777 SMARCD2	09283 70725	00578 59734	11358 59734	00579 62170 CD79B	00580 20874 SCN4A	14656 70725	15185 255 ERN1	29861 215021	25102 218950	17521 456 TEX2	11809 94694 PECAM1	25695 220950	06317 94698 C17orf60	09464 31149 POLG2	09463 3941	09459 53415 CCDC45	09461 592 SMURF2
mmur (35)	GeneScaffold_ 1512	07673 7165 SCNN1G	07684 7165 SCNN1B	07698 36853 COG7	07723 2483 GGA2	07745 5598 EARS2	07760 50223 UBFD1	07778 4869 NDUFAB1	07786 69215 PALB2	07806 7122 DCTN5	07827 7931 PLK1	07841 255 ERN2										
mmur (35)	GeneScaffold_ 1060				09980 2549 FTSJ3	09989 1930 PSMC5	10005 5777 SMARCD2	10016 59734	10024 62170 CD79B	10028 20874 SCN4A	10040 70725 ICAM2	10050 255 ERN1	09946 11961 CCDC47	09952 3941 DDX42								

Figure 6. Lineage information of PP000255. Lineage information page, including gene neighborhood, for PP000255 (ERN1/ERN2). The middle (black) column shows the gene belonging to lineage PP000255, with on the left and right the 20 genes that are nearest on the genome. Genes that have the same color are in the same lineage. If a neighboring lineage contains less than five genes, it is colored white. For each gene, the last part of the Ensembl ID (top) and the PhyloPat ID (middle) are displayed, as well as the HGNC symbol(s) (bottom), linking to the corresponding Ensembl, PhyloPat and HGNC pages.

PhyloPat might contain more features such as a statistical significance measure for the comparison of multiple phylogenetic patterns, and user-defined species sets for the calculation of orthologous groups.

ACKNOWLEDGEMENTS

This work was part of BioRange SP3.2.2 project of the Netherlands Bioinformatics Centre (NBIC), and was supported financially by Schering-Plough corporation.

Conflict of interest statement. None declared.

REFERENCES

1. Natale,D.A., Galperin,M.Y., Tatusov,R.L. and Koonin,E.V. (2000) Using the COG database to improve gene recognition in complete genomes. *Genetica*, **108**, 9–17.
2. Reichard,K. and Kaufmann,M. (2003) EPPS: mining the COG database by an extended phylogenetic patterns search. *Bioinformatics*, **19**, 784–785.
3. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
4. Dehal,P.S. and Boore,J.L. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, **7**, 201.
5. Dufayard,J.F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perriere,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
6. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
7. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
8. Hulsen,T., de Vlieg,J. and Groenen,P.M. (2006) PhyloPat: phylogenetic pattern analysis of eukaryotic genes. *BMC Bioinformatics*, **7**, 398.
9. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
10. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
11. Hulsen,T., Huynen,M.A., de Vlieg,J. and Groenen,P.M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.
12. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
13. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
14. Page,R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 357–358.
15. Cochrane,G., Akhtar,R., Aldebert,P., Althorpe,N., Baldwin,A., Bates,K., Bhattacharyya,S., Bonfield,J., Bower,L., Browne,P. *et al.* (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **36**, D5–D12.
16. Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.*, **34**, D319–D321.
17. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool .for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
18. Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
19. 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. *Nat. Genet.*, **25**, 25–29.
20. Perez,D.M. (2005) From plants to man: the GPCR ''tree of life''. *Mol. Pharmacol.*, **67**, 1383–1384.
21. Korbel,J.O., Snel,B., Huynen,M.A. and Bork,P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.*, **18**, 158–162.
22. Notebaart,R.A., Huynen,M.A., Teusink,B., Siezen,R.J. and Snel,B. (2005) Correlation between sequence conservation and the genomic context after gene duplication. *Nucleic Acids Res.*, **33**, 6164–6171.