PhyloPat version 49 – An updated version contains gene neighborhood

Tim Hulsen[1], Peter M.A. Groenen[2], Wynand Alkema[2], Jacob de Vlieg[1,2]
[1]CDD, CMBI/NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, NL
[2]MDI, Schering-Plough, Oss, NL

1 --- Introduction

Phylogenetic patterns show the presence or absence of certain genes or proteins in a set of species. They can also be used to determine sets of genes or proteins that occur only in certain evolutionary branches. Phylogenetic patterns analysis has routinely been applied to protein databases such as COG and OrthoMCL, but not upon gene databases. Here we present a tool named **PhyloPat** (figure 1) which allows the complete Ensembl gene database to be queried using phylogenetic patterns.
Address: http://www.cmbi.ru.nl/phylopat

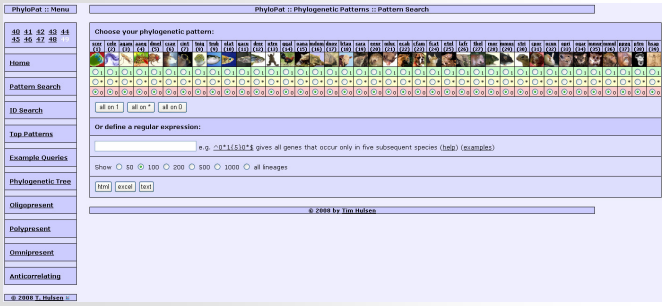

Figure 1. The PhyloPat web interface (Pattern Search tab). On the pattern search page, the user can generate a phylogenetic pattern by clicking a radio button for each species. 1 = present, * = present/absent, 0 = absent. The buttons directly below put all 39 species on the corresponding mode. MySQL regular expressions offer the possibility of advanced querying.

2 --- Results

PhyloPat is an easy-to-use webserver, which can be used to query the orthologies of all complete genomes within the BioMart database using phylogenetic patterns. This enables the determination of sets of genes that occur only in certain evolutionary branches or even single species.

Using a single linkage clustering algorithm we constructed an orthology database that currently contains:
- 39 species (see column in the middle; the species order was determined by creating an NCBI taxonomy tree and measuring the approximate evolutionary distance to man)
- 815,452 genes
- 17,332,165 orthologous relationships
- 241,697 phylogenetic lineages

Input options include:
- binary phylogenetic patterns (created by checkboxes)
- regular expressions
- branch-specific phylogenetic patterns (by clicking on a tree)
- lists of Ensembl, EMBL, EntrezGene or HUGO IDs

Output options include:
- HTML, Excel or plain text format
- links to the FatiGO web interface
- list of the top 100 patterns
- list of polypresent genes
- list of oligopresent genes
- list of omnipresent genes
- list of anticorrelating genes
- FASTA file of the peptide sequences within each each phylogenetic lineage
- gene neighborhood

5 --- Acknowledgements

This work was carried out within Biorange project SP3.2.2.

3 --- Gene Neighborhood

The new version of PhyloPat has a number of new options:
- list of anticorrelating genes, i.e. genes which have an opposite phylogenetic pattern
- EntrezGene IDs are accepted as input, next to Ensembl, EMBL and HUGO IDs
- each phylogenetic lineage contains a link to the corresponding FASTA file
- the PhyloPat ID links to all genes in that lineage, together with its 20 neighboring genes on the genome (10 on the left, 10 on the right)

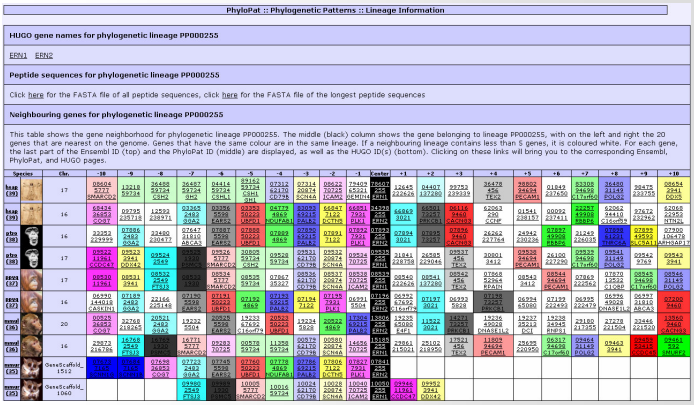

Figure 2. Lineage information page, including gene neighborhood, for PP000255. middle (black) column shows the gene belonging to lineage PP000255, with on the left and right the 20 genes that are nearest on the genome. Genes that have the same colour are in the same lineage. If a neighbouring lineage 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.

Figure 2 shows the gene neighborhood for PhyloPat ID PP000255 (ERN1, ERN2). The human gene ENSG00000134398 has two predicted orthologs in chimpanzee: gene ENSPTRG00000007893 and gene ENSPTRG00000009535. However, only the gene neighborhoods of gene ENSPTRG00000007893 and gene ENSG00000134398 correspond, for 9 of the nearest neighbors. This is called 'orthologous conservation of gene neighborhood' and it shows that the two genes involved are evolutionary related. In this case, we would say that the 'true' ortholog of gene ENSG00000134398 is very likely to be gene ENSPTRG00000007893. This is supported by the fact that these two genes have the same HUGO ID: ERN2. The HUGO ID of the other gene, ENSPTRG00000009535, is ERN1. This is just one of the many examples of the use of gene neighborhood information.

4 --- Conclusion

PhyloPat is the first tool to combine complete genome information with phylogenetic pattern querying. Since we used the orthologies generated by the accurate pipeline of Ensembl, the obtained phylogenetic lineages are reliable. The completeness and reliability of these phylogenetic lineages will further increase with the addition of newly found orthologous relationships within each new Ensembl release. The new version supports a number of new options, of which the option to show gene neighborhood is the most important.

6 --- References

1) PhyloPat: phylogenetic pattern analysis of eukaryotic genes
T. Hulsen, J. de Vlieg, P.M.A. Groenen
BMC Bioinformatics 2006, **7** (1): 398

2) PhyloPat version 49 – An updated version contains gene neighborhood
T. Hulsen, P.M.A. Groenen, W. Alkema, J. de Vlieg
Manuscript in preparation