PhyloPat: phylogenetic pattern analysis of eukaryotic genes

Tim Hulsen[1], Jacob de Vlieg[1,2], Peter M.A. Groenen[2]

[1]CMBI/NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands
[2]MDI, NV Organon, Oss, The Netherlands

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.

Figure 1. 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 21 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.

2 --- Results

PhyloPat is an easy-to-use webserver, which can be used to query the orthologies of all complete genomes within the EnsMart database using phylogenetic patterns. This enables the determination of sets of genes that occur only in certain evolutionary branches or even single species. We found in total 446,825 genes and 3,164,088 orthologous relationships within the 21 species (figure 2) of the EnsMart v40 database. We used a single linkage clustering algorithm to create 147,922 phylogenetic lineages, using every one of the orthologies provided by Ensembl. PhyloPat provides the possibility of querying with either binary phylogenetic patterns (created by checkboxes) or regular expressions. Specific branches of a phylogenetic tree of the 21 included species can be selected to create a branch-specific phylogenetic pattern. Users can also input a list of Ensembl or EMBL IDs to check which phylogenetic lineage any gene belongs to. The output can be saved in HTML, Excel or plain text format for further analysis. A link to the FatiGO web interface has been incorporated in the HTML output, creating easy access to functional information. Finally, lists of omnipresent, polypresent and oligopresent genes have been included.

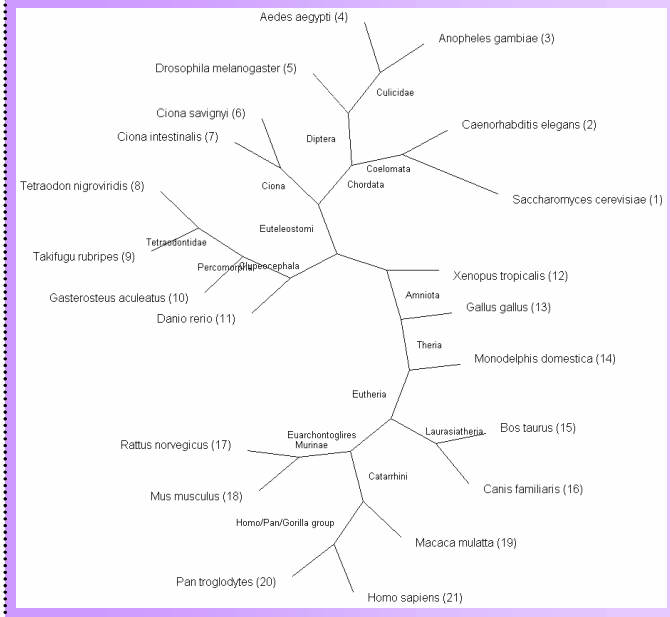

Figure 2. Phylogenetic tree of all species present in PhyloPat
This is the unrooted NCBI Taxonomy tree of all species available in Ensembl and PhyloPat. The numbers are the order in which the species are shown on the PhyloPat results pages. A phylogram version of this tree is available through the website.

3 --- Case study: Hox genes

As a case study we used the from an evolutionary point-of-view very interesting Hox genes. First, we searched the Ensembl database for human genes with the term 'hox' in the annotation. We found 44 genes, which were entered into PhyloPat. The output is shown in Table 1. 32 phylogenetic lineages were found, one of which were already present in C. elegans: PP022041. This lineage contains the Msh homeobox-like proteins. PP024984 and PP027791, containing the HOXC4 and TLX lineages, are only found in the Coelomata: A. gambiae and further. No less than 22 lineages originated in the early vertebrates, presented by T. nigroviridis. HOXD12 and HOXB13 are only present in mammals. Striking observations can be made with the fish species: all four species have significantly more Hox genes than the mammals. T. nigroviridis, for example, has 57 genes in this lineage, while M. domestica has only 35. These numbers correspond well with the fact that Teleost fish have at least seven Hox clusters, whereas mammals have only four. Mammals also have less Hox genes per cluster, demonstrating that there has been gene loss within the Hox clusters since the evolution from a vertebrate ancestor to present-day mammals.

PPID	# genes per species	phylogenetic pattern	gene name(s)
PP022041	011111136562233233222	01111111111111111111	MSX1, MSX2
PP024984	00100001111100111111	00100001111001111111	HOXC4
PP027791	00111002334323333333	00111001111111111111	TLX1, TLX2, TLX3
PP049478	00000022115311232222 3	00000011111111111111	HOXB8, HOXC8, HOXD8
PP053824	00000001112001010101 1	00000001111001010101 1	HOXD11
PP053827	00000002221111111111	00000001111111111111	HOXA10
PP053828	00000002111212122222	00000001111111111111	HOXC13, HOXD13
PP053829	00000006334112222222	00000001111111111111	HOXA1, HOXB1
PP053830	00000001111001011111	00000001111001011111	HOXB4
PP053832	00000002111101111111	00000001111101111111	HOXA5
PP053833	00000002111011111011	00000001111011111011	HOXB2
PP053834	00000003110101111111	00000001110101111111	HOXD3
PP053835	00000002111011111101	00000001111011111101	HOXA9
PP053836	00000002111111111111	00000001111111111111	HOXA3
PP053838	00000002111010111111	00000001111010111111	HOXC12
PP053839	00000001111111110111	00000001111111110111	HOXD4
PP053840	00000002111120101101	00000001111110101101	HOXC11
PP053842	00000004322111111111	00000001111111111111	HOXA13
PP053844	00000003223101111111	00000001111101111111	HOXB5
PP053845	00000002111111111011	00000001111111111011	HOXB3
PP053846	00000002112111111111	00000001111111111111	HOXD10
PP053847	00000002221111111111	00000001111111111111	HOXA2
PP053849	00000003415113233333 3	00000001111111111111	HOXA6, HOXB6, HOXC6
PP053853	00000001110111111011	00000001110111111011	HOXA4
PP053854	00000003225222233133 13	00000001111111111111	HOXB9, HOXC9, HOXD9
PP053858	00000001112001111111	00000001111001111111	HOXA11
PP070659	00000000012121222222	00000000011111111111	HOXA7, HOXB7
PP075622	00000000001000111111	00000000001000111111	HOXC5
PP084287	00000000000101111111	00000000000101111111	HOXC10
PP085049	00000000000101011111	00000000000101011111	HOXD1
PP087941	00000000000011011111	00000000000011011111	HOXD12
PP089685	00000000000011111111	00000000000011111111	HOXB13

Table 1. Phylogenetic lineages containing human Hox cluster genes

Table 2 shows the further analysis of the Hox genes using the PhyloPat output. H. sapiens misses the genes HOXA8, HOXB10, HOXB11, HOXC1, HOXC2, HOXC3, HOXC7, HOXD2, HOXD5, HOXD6 and HOXD7. The absence of these 11 genes is in agreement with current knowledge of human Hox genes. Two exceptions exist: HOXC8 instead of HOXC7, and the absence of HOXA12. The HOXA12 gene cannot be found in the other mammals either.

PPID(s)	name	cl.A	cl.B	cl.C	cl.D	first sp.	position
PP053829,085049	HOX1	HOXA1	HOXB1		HOXD1	T. nigrov.	anterior
PP053847,053833	HOX2	HOXA2	HOXB2			T. nigrov.	anterior
PP053836,053845,053834	HOX3	HOXA3	HOXB3		HOXD3	T. nigrov.	PG3
PP053853,053830,024984,053839	HOX4	HOXA4	HOXB4	HOXC4	HOXD4	A. gamb.	central
PP053832,053844,075622	HOX5	HOXA5	HOXB5	HOXC5		T. nigrov.	central
PP053849	HOX6	HOXA6	HOXB6	HOXC6		T. nigrov.	central
PP070659	HOX7	HOXA7	HOXB7			G. acul.	central
PP049478	HOX8		HOXB8	HOXC8	HOXD8	C. intest.	central
PP053835,053854	HOX9	HOXA9	HOXB9	HOXC9	HOXD9	T. nigrov.	posterior
PP053827,084287,053846	HOX10	HOXA10		HOXC10	HOXD10	T. nigrov.	posterior
PP053858,053840,053824	HOX11	HOXA11		HOXC11	HOXD11	T. nigrov.	posterior
PP053838,087941	HOX12			HOXC12	HOXD12	T. nigrov.	posterior
PP053842,089685,053828	HOX13	HOXA13	HOXB13	HOXC13	HOXD13	T. nigrov.	posterior
PP027791	TLX	TLX1	TLX2	TLX3		A. gamb.	
PP022041	MSX	MSX1	MSX2			C. eleg.	

Table 2. Analysis of phylogenetic lineages containing human Hox cluster genes

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.

5 --- Acknowledgements

This work was supported financially by NV Organon. The authors like to thank Erik Franck, Nanning de Jong and Arnold Kuzniar for their suggestions. We thank Greer Wilson for critically reading this text.

6 --- Reference

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