Group 7:
Ryno Lawson
Vasileios-Evripidis Kyriakidis

Comparative Genomics 2017
Practical 6 Report : Orthology Prediction

Summary

In this practical we hope to compare the ortholog databases using genes from our query genomes, while looking for differences and similarities between the web tools we used to provide us with phylogenetic information and trees

Step 1: Retrieving true IDs

From a previous practical we knew from a blast search from which species our given sequences were originally sourced. With this species name we went to Inparanoid species mapping directory to find the file name which is associated with our species name and then downloaded the corresponding proteome file for each of our species.
From url - *http://inparanoid.sbc.su.se/download/current/sequences/processed/*

05.fa.txt -> C.trachomatis -> 272561.fasta
11.fa.txt -> G.sulfurreducens ->243231.fasta
09.fa.txt -> E.coli -> 83333.fasta

To see which genes in our given genome correspond to what genes in the downloaded species genome (SG) gene, we made a database like in previous practicals with the makeblastdb on the species genome.

$makeblastdb -in 272561.fasta -dbtype prot
$makeblastdb -in 83333.fasta -dbtype prot
$makeblastdb -in 243231.fasta -dbtype prot

After which we perform a protein blast of the given genome against the SG which give us a XML file of the HSP.

$blastp -outfmt 5 -query 05.fa.txt.pfa -db 272561.fasta83333.fasta -out out_Ctrach.xml

This was further parsed through the **blastp_ResultParser.py** script (see previous practical) to pull out the best hits of our orf genes. This gave us the protein labels (from SG) from our orf genes (from our given genome proteomes), where the ID of the first hit in the XML file, is the true ID of our protein.

$python blastp_ResultParser.py out_Ctrach.xml

Step 2: Searching for orthologs in various databases

From this list of corresponding labels we went about sampling random protein sequences and testing if they was a hit when searching on Inparanoid. As the species file originated from **InParanoid** there would be a hit but not always a ortholog match. We continued until there was a ortholog match. We use the protein label to perform searches in databases. After some trial searches we chose to use only one species (*C.trachomatis*) to search for ortholog matches with the rest of our query genomes, in various databases. Such databases were **InParanoid, TreeFam, Phylomedb, Panther, OMA** and etc.

There were not always matches in all databases. For example Treefam has no prokaryotes in it would not show any prokaryotic match orthologs.

The table below describes all our query genome sequences for the practical:

File #	Organism	Size	Kingdom	# of genes
05.fa.txt	*Chlamydia trachomatis*	1042588	Bacteria	977
09.fa.txt	*Escherichia coli*	5277676	Bacteria	341
11.fa.txt	*Geobacter sulfurreducens*	4566144	Bacteria	4172
12.fa.txt	*Gloeobacter violaceus*	4659019	Bacteria	4430
28.fa.txt	*Saccharomyces Cerevisiae* (Chr VIII)	562643	Fungi	6445(Chr VIII 205)

The following four tables demonstrate the results we retrieved from our search. We focused more on three databases: InParanoid, OMA and Panther. Inside the tables we describe whether or not we got any orthologue pairs from the databases we used, the IDs of the orthologue pairs and if there was no positive result we describe our assumption on why that may have happened. Included in the tables are also the lower scoring results from the Ortholog group and the paralogues to our higher scoring ortholog.

>05.fa.txt_orf00975

***Chlamydia trachomatis* Gene : O84693**

Chlamydia trachomatis	**InParanoid Id + Cross ref(UniProt)**	**OMA OMA Id + Cross ref**	**Panther**

Geobacter sulfurreducens	Q74C08 - Q74C08_GEOSL	GEOSL01838- Q74C08 Para- GEOSK01803	No pair ***
Gloeobacter violaceus	Q7NKV3 - Q7NKV3_GLOVI	Q7NKV3 - GLOVI01361	Q7NKV3
Escherichia coli- strain K12	P77444 - SUFS_ECOLI Para - Q46925- CSDA_ECOLI	ECOLI01636 - P77444 Para ECODH02632 Para - Q46925 - ECOLI02716 Para - ECOLI02716 - Q46925 Para - ECODH01586 - SUFS_ECODH Para - ECOBW02508	P77444

Table 1. The *Geobacter sulfurreducens* gene had high score in the InParanoid database, but there was no ortholog match to the *C.trachomatis* O84693 gene in Panther, even though the *G.sulfurreducens* gene exists in the database. This could be a difference due to the algorithm.

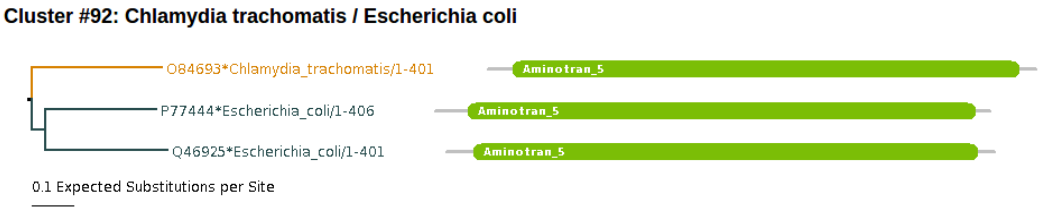

Image 1. Phylogenetic Tree for *C.trachomatis and E.Coli* clustered orthologues, from InParanoid. This shows the ortholog group as a whole plus the in paralogs

The database InParanoid, OMA and Panther were used as they consistently gave us good ortholog pairs. InParanoid and OMA are said to use Graph-based method which uses sequence similarity to form orthologous groups. These make use of pair-wise alignments in an all-against-all to generate scores. While Panther is a Tree-based method which forms sequence clusters from annotated whole genomes to make a tree and then shows orthologs and paralogs. It makes use of a HMM for families and subfamilies. The different of method between Panther and the other two graph-based methods could be the reason we don't always get a hits in Panther while we get hits is the other two databases. Plus Panther could be missing the whole genome for that particular species and shows results for alternative strains.

>05.fa.txt_orf01120_rev

***Chlamydia trachomatis* Gene : O84797**

Chlamydia trachomatis	**InParanoid Id + Cross ref(UniProt)**	**OMA OMA Id + Cross ref**	**Panther**

Geobacter sulfurreducens	P61667 - MUTS_GEOSL	GEOSL01795-MUTS_GEOSL	P61667
Gloeobacter violaceus	Q7NLT8 - MUTS_GLOVI	GLOVI01024 - MUTS_GLOVI	Q7NLT8
Escherichia coli- strain K12	P23909 - MUTS_ECOLI	ECOLI02640 - P23909 Para = ECODH02558-MUTS_ECODH Para = ECOBW02432-MUTS_ECOBW	P23909

Table 2. All databases found orthologue pairs for the 084797 gene of *C.trachomatis*

>05.fa.txt_orf00354_rev
***Chlamydia trachomatis* Gene : O84260**

Chlamydia trachomatis	**InParanoid Id + Cross ref (UniProt)**	**OMA OMA Id + Cross ref**	**Panther**
Geobacter sulfurreducens	Q74BN0 Para -Q74A21 Para-Q749F7	GEOSL01981-Q74BN0 Para - GEOSL02528 - Q74A21 Para - GEOSL02739 - Q749F7	No Q74BN0 match Para - Q749F7 Para - Q74A21
Gloeobacter violaceus	Q7ND52 - Q7ND52_GLOVI	GLOVI04361 - Q7ND52	Q7ND52
Escherichia coli- strain K12	P0A6B7-ISCS_ECOLI	ECOLI02457- P0A6B7	P0A6B7

Table 3. The *Geobacter sulfurreducens* gene was not an ortholog to *C.trachomatis* O84260 gene in Panther, but it exists in the Panther database.

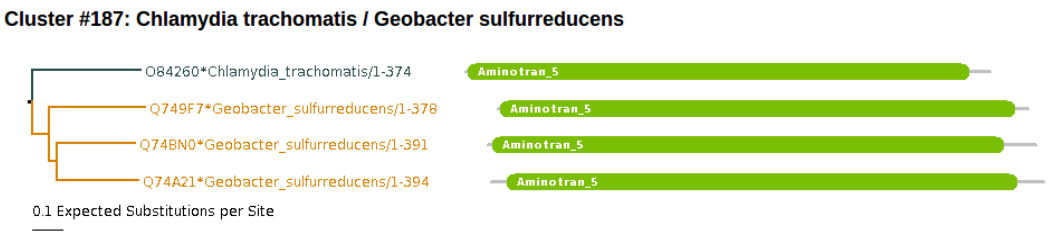

Image 2. Phylogenetic Tree for *C.trachomatis and G.sulfurreducens* clustered orthologues, from InParanoid. This shows the ortholog group as a whole plus the in paralogs

>05.fa.txt_orf00525
***Chlamydia trachomatis* Gene : O84379**

Chlamydia trachomatis	**InParanoid Id + Cross ref(UniProt)**	**OMA OMA Id + Cross ref**	**Panther**
Geobacter sulfurreducens	No pair	No pair	No pair
Gloeobacter violaceus	Q7NI34 Q7NK28 - Q7NK28_GLOVI Q7NK14 - Q7NK14_GLOVI	No Q7NI34 match Para - GLOVI01651 Q7NK14	No pair
Escherichia coli- strain K12	P0AAE5	ECOLI01561 - P0AAE5	P0AAE5

Table 4. *G.sulfurreducens* gave no ortholog hits in all three databases *G. violaceus,* but exists in all databases. Also for *G. violaceus* OMA database gave different orthologue pairs than InParanoid and we got no pairs at all from Panther. The *Geobacter sulfurreducens is not well characterized in most databases.*

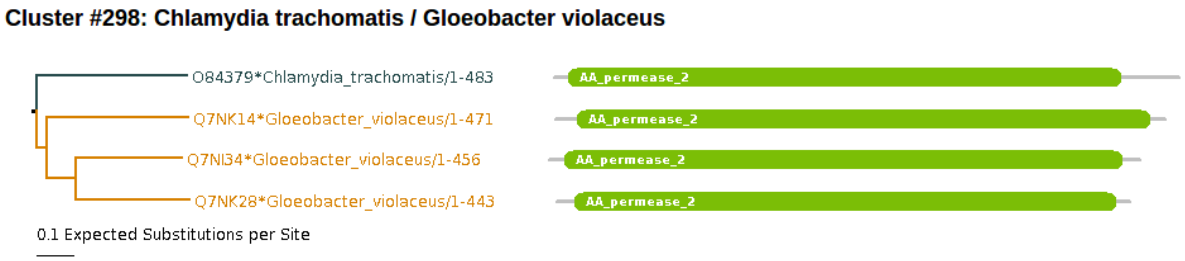

Image 3. The tre for orthologs group for *Gloeobacter violaceus*.

There were no good results which show orthology between our species when we searched for all three genes in Phylomedb and metaPhOrs. We didn't use treeFam at all because all our genome sequences procaryotic and treeFam only accepts Eukaryotic genomes. PhylomeDB uses a phylogeny-based method (tree-based method) and typically accepts only Ensembl, Swissprot, Trembl identifiers. Which were not always available from Uniprot, however some other database did offer alternative identifiers but switching between formats in cumbersome. PhylomedDB also requires at least 3 significant hits in the genomes to make a tree. We would have thought we should get good results from MetaPhOrs as it uses seven popular homology prediction services however *Geobacter sulfurreducens and Gloeobacter violaceus* didn't show up is the orthologs for our Chlamydia trachomatis genes, as they were either not in the database or the identifiers didn't get a hit in the search.

References:

PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements.

Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas
Nucl. Acids Res. (2016) doi: 10.1093/nar/gkw1138

Altenhoff A et al., *The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements*, Nucleic Acids Research, 2015, 43 (D1): D240-D249 (doi:10.1093/nar/gku1158).

"InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic"
Erik L.L.Sonnhammer and Gabriel Östlund
Nucleic Acids Res. 43:D234-D239 (2015)

PhylomeDB - http://phylomedb.org/help3

MetaPhOrs - http://orthology.phylomedb.org/?q=faq