Comparative Genomics 2018

Practical 6: Orthology Prediction

Group number: 6

Group members: Kyle Kimler, Kajetan Juszczak

Summary

With the development of orthology search tools, scientists have begun to build up orthology databases of various taxonomies to assist in efforts to expand our understanding of computational phylogenomics. These databases can be used to predict orthologs via search/blast-like heuristics, but the databases are already set up hierarchically with only gene identifiers (with corresponding species). In this lab practical we use these databases to predict orthologs for genes we have clustered by cross-BLAST searches in the previous practicals. We have chosen OMA, Metaphors, and InParanoid. InParanoid gives results as orthologous pairs, Metaphors gives orthologous groups based on huge averages of pre-calculated phylogenetics trees, and OMA gives orthologous clusters based on sequence alignment. OMA provides the largest amount of orthology data because it clusters genes together while InParanoid and Metaphors give smaller groups more likely to exclude paralogs (they also run specific algorithms for the purpose of distinguishing paralogs and orthologs).

Key Questions to Answer

Exercise 2

Describe the used algorithms of the databases you are comparing and motivate your choice of databases [SEP]

1. **InParanoid –** Inparanoid uses BLAST of two complete proteomes to find “seed orthologs”, then trees of paralogs are built around these two orthologs by BLAST to find sequences that match more closely with the seed orthologs than with any other sequences in the two proteomes. Since it uses BLAST only between 2 species at a time, it is quick and does not assume phylogenies – instead output is presented as tables of ortholog pairs.
2. **MetaPhOrs –** MetaPhOrs and Phylomedb use phylogenetic trees to predict orthology. A search in MetaPhOrs takes every phylogenetic tree (calculated in Phylomedb) that contains the protein of interest, then calculates an overall orthology tree depending on some scoring factors for each tree from Phylomedb. It uses a “Species-overlap algorithm” where descendant orthologs are only true orthologs (not paralogs) when there is no overlap of any species below that node. This is how paralogs are separated from orthologs.
3. **OMA –** OMA, “Orthologous Matrix” is more similar to InParanoid in orthology calculation. It finds homologous sequences by Smith-Waterman local alignment and calls orthologous pairs based on sequence similarity across many proteomes, then clusters these orthologs into groups.

We first chose OMA because it provides the biggest set of results. We chose Metaphors and Inparanoid to compare to OMA because Metaphors provides a very different method of calculating orthology while InParanoid provides sort of a light-weight version of OMA.

Exercise 3

Choose 3 genes:

Tree0 - RNA polymerase sigma factor RpoE: Q8A2A9 (thetaiotamicron) in all species

Tree7 - NuoE - NADH reductase in all species

Tree9 - thioredoxin in all but thermotoga: alpha/beta hydrolase

1. **RpoE** RNA polymerase sigma factor – P0AGB6
2. **thioredoxin** "E. coli thiol disulfide reductase thioredoxin" **- trxC** in E.coli – P0AGG4
3. **NuoE -** NADH-quinone oxidoreductase subunit NuoE / NADH reductase – P0AFD1

http://betaorthology.phylomedb.org/?q=single&metaid=RPOE

http://betaorthology.phylomedb.org/?q=single&metaid=trxC

http://betaorthology.phylomedb.org/?q=single&metaid=NuoE

Pick at least 3 species for comparison of orthologs

1. **E. coli**
2. **Thermotoga maritima**
3. **Pseudomonas aeruginosa – chosen because our third reliable ortholog species spiribacter is not present in any orthology databases (discovered in ~2013)**

Discuss the achieved results with the different algorithms, especially the differences between their predictions (pairs, ortholog groups):

How do the predicted orthologs differ? Which are missing or are the same?

Predicted orthologs differ greatly between databases. In the case of thioredoxin, apparently the same protein works both as a reductase and as a redoxin, so it seems that some databases separate the functions and have a separate file for dual-function variants. In these cases some databases pick up totally different orthologs for thioredoxin than each other, even with the same protein identifier. We had to use the short protein ID for metaphors and inparanoid in some cases because it doesn't even contain entries for e. coli. RpoE was interesting because, it being an RNA-pol associated protein, you would expect it to exist in large orthology trees, and it does in OMA but is very specific in both Inparanoid and Metaphors, with only ~20 entries in each. While OMA had orthologs for each of the species we queried for each protein, except for thioredoxin-2 in thermotoga and Pseudomonas aeruginosa, OMA found all of the matches that Metahpors found, including many Pseudomonas genus orthologs for thioredoxin-2. OMA displays a lot of distant archaeal orthologs too, which seem unlikely to be orthologs in the case of at least the NADH dehydrogenase subunit since archaea diverged so long ago.

Can you find orthologs in one database that are either missing or appear as out-paralogs in another database? Why do you think this happens?

Hits may appear as out-paralogs in one database and as orthologs in another, especially in OMA, because OMA seems to cluster some protein variants together. In some cases this could be caused by the orthology trees being built up from different strains' genomes, so that

though the species may be the same, a different strain of thetaiotamicron bacteroides may have a slightly different RpoE, and RpoE itself is known to exist in 10 copies in many organisms, so what may be a paralog of one of the copies would be more likely to be an ortholog of another copy. Since you only input a gene symbol to the database you are missing information about the strain, copy, etc, and even in OMA if you write the species-specific accession number you will encounter the same problem in the case of high copy number genes – just see RpoE search in OMA - many hits in same species. High copy number genes also typically have some pseudogenes to computationally contend with. Since Metahpors relies on phylogenetic trees to build orthology groups, it could be averaging together many trees of different paralogs and spitting out a low-sequence-similarity-resolution paralog/ortholog decision. Or, since InParanoid uses BLAST, it could be missing similarity-information that OMA would find by S-W alignment.

How big are the ortholog groups for your selected genes in the databases you compare?

	RpoE	**trxC**	**NuoE**
OMA	1742	991	1137
MetaPhOrs	14	190	491
InParanoid	19	219	203

What can you say about the quality of orthology predictions with the databases you compare?

All of the databases use some quality assurance measures in their predictions. Because each of the databases predicts orthologs based on such different criteria, it's probably wise to use several of them. To me, OMA seems the most useful, since it finds all of what the other two databases find in addition to providing protein family clustering. MetaphOrs is probably the best to use alongside OMA because it uses phylogenetic analysis as part of the prediction, which OMA does not provide.

Reference

http://metaphors.phylomedb.org/

https://omabrowser.org/oma/home/

http://inparanoid.sbc.su.se/cgi-bin/index.cgi

Zvelebil and Baum – "Understanding Bioinformatics"

Roth, A., Gonnet, G. & Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9, 518 (2008).