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 possible orthologs for genes we have clustered by cross-BLAST searches in the previous practicals. We have chosen OMA, Metaphors, and InParanoid since TreeFam only has very few genomes.

**Key Questions to Answer**

**Ex. 2**

1. **OMA - describe the used algorithms of the databases you are comparing and motivate your choice of databases**
2. **InParanoid -** [**http://inparanoid.sbc.su.se/cgi-bin/faq.cgi#how**](http://inparanoid.sbc.su.se/cgi-bin/faq.cgi#how) **– 3.**
3. **MetaPhOrs - http://metaphors.phylomedb.org/?q=help#Methodology**

**Ex. 3**

**Choose 3 genes:**

**Tree0 - RNA polymerase sigma factor RpoE ALL: Q8A2A9 (thetaiotamicron)**

**Tree7 - NuoE - NADH reductase**

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

**1. RpoE RNA polymerase sigma factor**

**2. thioredoxin "E. coli thiol disulfide reductase thioredoxin" - trxC in E.coli**

**3. NuoE - NADH-quinone oxidoreductase subunit NuoE / NADH reductase.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RpoE** | **trxC** | **NuoE** |
| **OMA** | 742 | 991 | 1137 |
| **MetaPhOrs** | 14 | 190 | **491** |
| **InParanoid** | 19 | 219 | 203 |

[**http://betaorthology.phylomedb.org/?q=single&metaid=RPOE**](http://betaorthology.phylomedb.org/?q=single&metaid=RPOE)**,** [**http://betaorthology.phylomedb.org/?q=single&metaid=trxC**](http://betaorthology.phylomedb.org/?q=single&metaid=trxC)**,**

[**http://betaorthology.phylomedb.org/?q=single&metaid=NuoE**](http://betaorthology.phylomedb.org/?q=single&metaid=NuoE)

**Pick at least 3 species for comparison of orthologs – is this 3 species to search out of our 4 or 3 species from each gene’s result list to compare? Assuming latter.**

**You can do an online BLAST to find gene ID’s. Write up how you got gene ID**

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

* 1. **How do the predicted orthologs differ? Which are missing or are the same?**
  2. 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!
  3. **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?**
  4. Another reason that hits may appear as out-paralogs in one database and as orthologs in another is that the databases themselves are built up from different genomes. 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 I think if you write the accession number you will encounter the same problem in the case of high copy number genes. High copy number genes also typically have some pseudogenes to computationally contend with.
  5. **How big are the ortholog groups for your selected genes in the databases you compare?**
  6. tabled.
  7. **What can you say about the quality of orthology predictions with the databases you compare?**

OMA is better than InParanoid because it has a greater database to search; MetaPhOrs is better than OMA because it uses OMA's database in addition to others like PhylomeDB. As a side-note, PhylomeDB caused my PC to crash when I tried to search, and Metaphors caused a similar reaction when attempting to browse. Need to write a little bit about the match, and maybe compare results to a famous tree of life!

**Reference**

[**http://metaphors.phylomedb.org/**](http://metaphors.phylomedb.org/)

[**https://omabrowser.org/oma/home/**](https://omabrowser.org/oma/home/)

[**http://inparanoid.sbc.su.se/cgi-bin/index.cgi**](http://inparanoid.sbc.su.se/cgi-bin/index.cgi)

Zvelebil and Baum – “Understanding Bioinformatics”