Progress Report

03.08.2017

# HamFAS-only KO annotation analysis

The comparison between the performance of HamFAS against BlastKOALA and KAAS has shown, that there were 165 KO-unannotated yeast proteins that have been assigned KOs only by HamFAS approach.

1. *Origin of annotations*

17 out of 165 KO annotations that have been predicted only by HamFAS were originated from reference fungi. A large number of KOs (100/165) come from other eukaryote taxa. Other 29% annotations have been obtained from reference archaea and bacteria species.

=> how to explain ko from eukaryotes:

(1) no fungal orthologs - only 6 fungal ref => so many independent losses or just few?

or (2) fungal orthologs have no KO => YES (92/100 have fungal orthologs)

29% (archaea and bacteria) could be false pos => if they do not complement the pathways or don't make any sense

1. *Number of Pfam annotated domains*

Although a large fraction of yeast proteins (38%) have only one Pfam domain annotated, the others still have been assigned with at least 2 domains.

=> to answer the question how informative the proteins are (if only one domain drives the fas score for those proteins => uninformative! But here there are still large number of proteins that have more than 1 domain).

CÁCH TIẾP CẬN KHÁC (can be used as a master thesis): using FAS filter to estimate the "wrong identified/predicted gene" (the one which have significantly low fas score in comparison to the large fraction (the average) of other genes)

# PhyloProfile app

## Supporting of OrthoXML files

I have 2 scripts (*data/orthoxmlParser.pl* and *data/orthoxmlParser.py*) for parsing OrthoMCL files.

The Python script depends on 2 Python libraries *beautifulSoup* and *xml* (so those have to be installed in order to use this Python parser script). Previous problem of OMA's XML files was due to the parser library. It has been fixed now.

The alternative Perl script is based only on regular expression matching. Currently it works if the format of input XML files is similar to the test case (OrthoMCL files created by InParanoid and OMA). It can be used as a backup, when there is any change in the Python parser libraries that makes that script crash.

## Reformatting the profile plotMacintosh HD:Users:trvinh:Desktop:Screen Shot 2017-08-03 at 11.22.22.png

The smallest dots for *absent orthologs* will not be present on the profile plot.

The legend can dynamically adapt to the input information (e.g. it will not be shown if only the presence/absence information is given).

# Next tasks:

- Collect pathways for LCA and extant microsporidia species (Done! Data are stored in our mysql-database ***pathways\_kegg***) & compare them.

- Summarize the comparison analysis of HamFAS against BlastKOALA and KAAS.

- Package PhyloProfile and write draft manuscript.