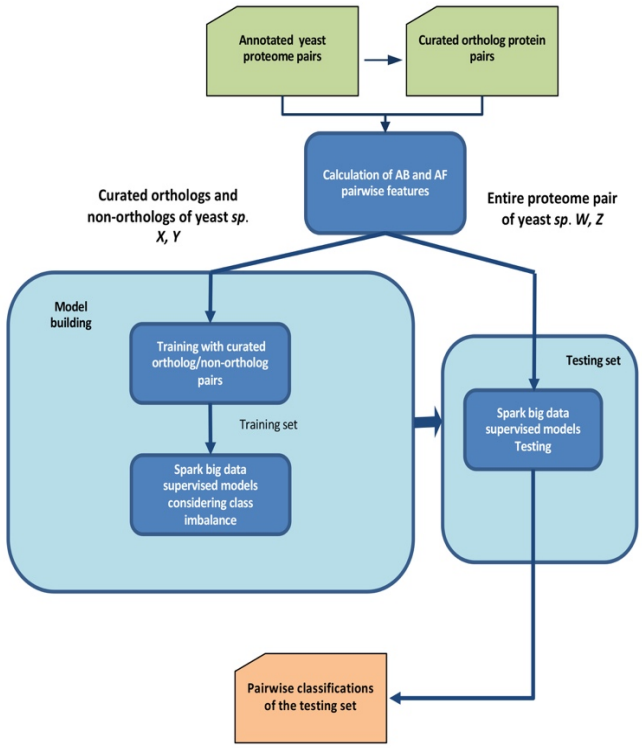
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-04

Surveying Alignment-free Features for Ortholog Detection in Related Yeast Proteomes by using Supervised Big Data Classifiers

Deborah Galpert Cañizares (deborah@uclv.edu.cu) [a], Alberto Fernández (alberto@decsai.ugr.es) [b], Francisco Herrera (herrera@decsai.ugr.es) [b], Agostinho Antunes (aantunes@ciimar.up.pt) [c,d], Reinaldo Molina-Ruiz (reymolina@uclv.edu.cu) [e], Guillermin Agüero-Chapin (gchapin@ciimar.up.pt) [c,d*]

[a]*Departamento de Ciencia de la Computación. Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba.*
[b]*Department of Computer Science and Artificial Intelligence, Research Center on Information and Communications Technology (CITIC-UGR), University of Granada, 18071 Granada, Spain.*
[c]*CIIMAR/CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208 Matosinhos, Porto, Portugal.*
[d]*Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.*
[e]*Centro de Bioactivos Químicos (CBQ), Universidad Central ¨Marta Abreu¨ de Las Villas (UCLV), Santa Clara, 54830, Cuba.*

Graphical Abstract



Abstract

Methods for pairwise ortholog detection (POD) usually relies on alignment-based (AB) similarity measures. However, AB algorithms are still limited in POD since they may fail in the presence of certain evolutionary and genetic events. In this sense, POD is an open field in bioinformatics demanding either constant improvements in existing methods or new effective scaling algorithms to deal with Big Data.

In a previous paper, we developed a Big Data supervised POD approach considering several AB pairwise gene features and the low ortholog pair ratios found between two proteo-

mes. (Galpert, del Río et al. 2015). Although the higher sensitivity achieved for our supervised POD models in relation to classical POD methodologies, when were comparatively evaluated on the *Saccharomycete* yeast benchmark dataset (Salichos and Rokas 2011); they were implemented in MapReduce framework and tested on a single yeast genome pair.

In (Galpert, Fernández et al. 2018), we propose some improvements to our supervised POD approach by i) surveying the incorporation of alignment-free pairwise similarity measures ii) evaluating other classifiers under the Big Data Spark platform and iii) extending the test set to other related *Saccharomycete* yeast proteomes https://doi.org/10.1186/s12859-018-2148-8.

Research Highlights

- The Big Data Spark architecture improved both the quality of supervised POD models and the execution time in relation to the previous ones built on the Hadoop MapReduce framework.
- The incorporation of AF pairwise similarity measures into the supervised POD models did not significantly improve the classification rates, however it was especially useful for POD within the "twilight zone" of sequence similarity
- The success of our supervised Big Data POD approach was confirmed by extending the test set to other yeast proteomes

References

- Galpert, D., S. del Río, F. Herrera, E. Ancede-Gallardo, A. Antunes and G. Agüero-Chapin (2015). "An Effective Big Data Supervised Imbalanced Classification Approach for Ortholog Detection in Related Yeast Species." BioMed research international **2015**.

- Salichos, L. and A. Rokas (2011). "Evaluating ortholog prediction algorithms in a yeast model clade." PloS one 6(4): e18755.

- Galpert, D., Fernández, A., Herrera, F., Antunes, A., Molina-Ruiz, R. and Agüero-Chapin, G. (2018). Surveying alignment-free features for Ortholog detection in related yeast proteomes by using supervised big data classifiers. *BMC bioinformatics*, *19*(1), p.166.