**Functional annotation of orthologous groups by using hierarchical multi label classification**

Authors: Nives Škunca[1], Fran Supek[1], Panče Panov[2], Sašo Džeroski[2], Tomislav Šmuc[1]

[1] Department of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

[2] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

**Background**

The method of correlating gene occurrence patterns in selected organisms, termed phylogenetic profiling, has proven to be a useful tool in functional genomics [1]. The reasoning behind this strategy of elucidating gene function lies in the presumption that genes having been found and lost together during evolution (signalled by their occurrence pattern in the phylogenetic profile) are involved in closely related biological functions.

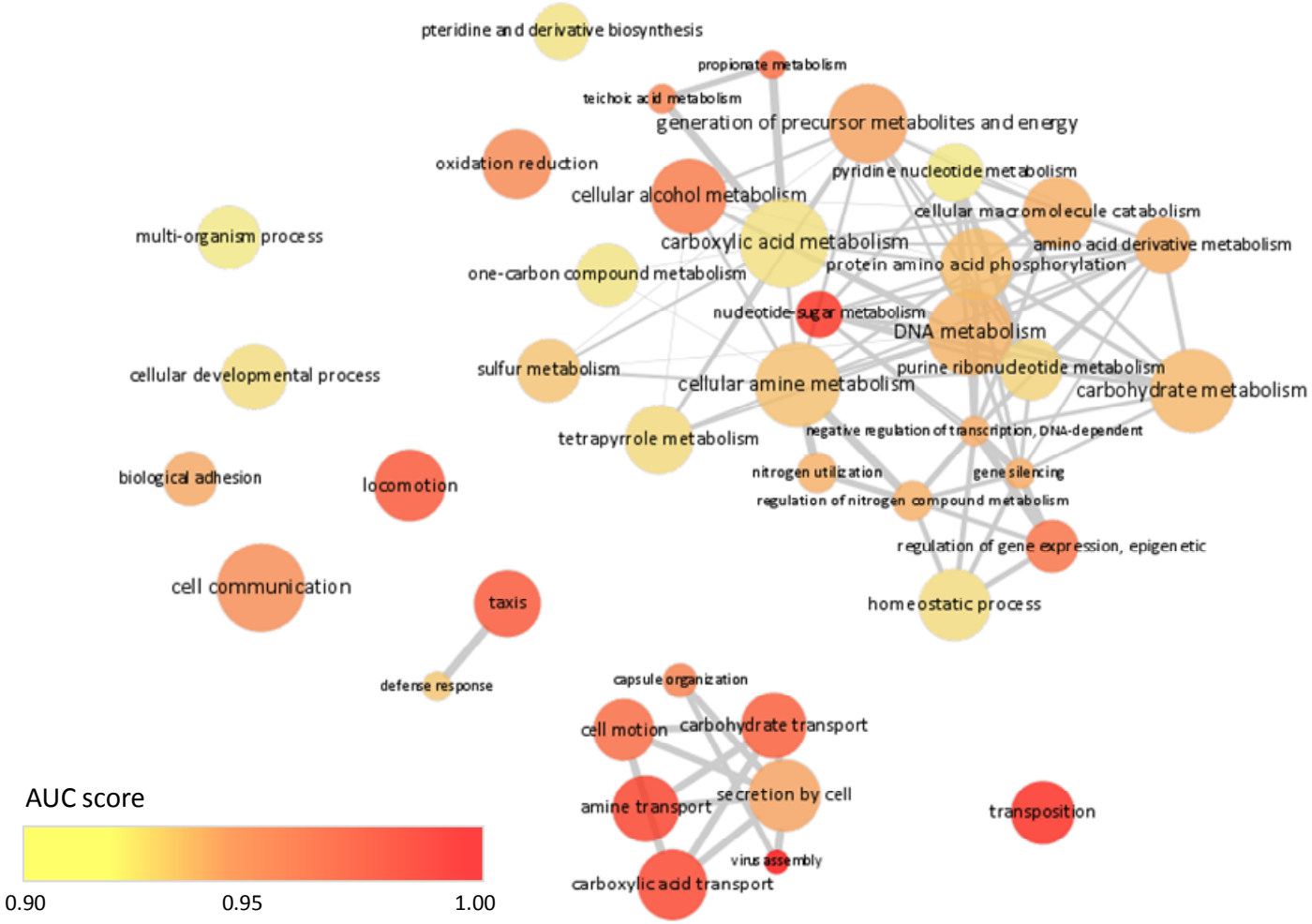
**Method**

We have focused on predicting the function of groups of orthologous proteins. Because of this, we are interested in, function-wise, highly cohesive protein groups, and have hence decided to use Orthologous Matrix Project (OMA) [2] as the basis for gene grouping, given the report on the highest specificity among the most comprehensive orthologous genes grouping methods examined [3]. We have used a machine learning algorithm based on decision trees for Hierarchical Multi-label Classification (HMC) [4] to predict Gene Ontology (GO) [5] assignments of OMA groups. The HMC extension of decision trees takes into account the directed acyclical graph layout of GO and considerably improves computational efficiency by learning to predict all classes at once.

**Results**

Results are expressed in the GO vocabulary, with the Area Under Receiver Operating Characteristic (AUC) score [6] as the performance measure. The most abundant result set with the best performance (AUC>0.90) is from the biological function ontology (Figure 1). The visualization of the results was done using Cytoscape [7].

**Figure 1**




A representation of GO categories with the best predictive performance (AUC > 0.9) in the biological process ontology. Disc size is proportional to the log number of genes in a category, and the colour gradient represents the strength of the AUC score, as shown in the legend. Grey lines represent semantic relationships between categories, per SimRel [8] method; spatial arrangement of discs approximately reflects similarity of categories. Displayed categories have been selected from a broader set by 1) eliminating very general GO assignments (assigned to >10% of OMA groups in our dataset), 2) setting AUC threshold at 0.9, 3) choosing only GO categories that have significant predictions (for each GO category, a precision > 0.8 exists for at least 5 unannotated OMA groups) and 3) by using a redundancy elimination procedure. The complete listing of results is available on request.

**Conclusion**

When predicting GO assignments for genes in OMA groups, the HMC algorithm shows convincing predictive power (Figure 1). Therefore, it can be used as 1) a powerful computation tool in narrowing the search space of potential gene candidates for a particular function and 2) annotation tool for the genes in newly sequenced genomes. The listing of function predictions together with the corresponding precision values, that is available, can be scanned: setting AUC and precision cutoff values would provide a condensed outline of target genes that are to be examined in the wet lab.

**References**


1. Kensche PR, van Noort V, Dutilh BE, Huynen MA: **Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution**. *Journal of the Royal Society Interface* 2008, **5**(19):151-170.
2. Roth AC, Gonnet GH, Dessimoz C: **Algorithm of OMA for large-scale orthology inference**. *BMC Bioinformatics* 2008, **9**:518-528.
3. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods**. *PLoS Comput Biol* 2009, **5**(1):e1000262.
4. Vens C, Struyf J, Schietgat L, Dzeroski S, Blockeel H: **Decision trees for hierarchical multi-label classification**. *Machine Learning* 2008, **73**(2):185-214.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
6. Fawcett T: **An introduction to ROC analysis**. *Pattern Recognition Letters* 2006, **27**(8):861-874.
7. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. *Genome Research* 2003, **13**(11):2498-2504.
8. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology**. *BMC Bioinformatics* 2006, **7**:302-318.