Computational Functional Annotation using Hierarchical Orthologous Groups in OMA

Alex Warwick Vesztrocy,[1,3,5*] Adrian Altenhoff,[2,3] Henning Redestig[4] and Christophe Dessimoz[1,3,5,6,7]

[1]Department of Genetics, Evolution and Environment, University College London, London, UK [2]Department of Computer Science, ETH Zürich, Zürich, Switzerland
[3]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland [4]Bayer CropScience NV, Ghent, Belgium
[5]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland [6]Department of Computer Science, University College London, London, UK
[7]Centre for Integrative Genomics, University of Lausanne, Lausanne, Switzerland *Correspondence: alex.warwick.vesztrocy.15@ucl.ac.uk

Introduction

There are few gene-function associations that have been experimentally validated, with most having been inferred electronically with little human-validation.

The aim of this project is to propagate and infer accurate gene annotations in complex genomes, through propagation of terms to homologous genes.

The algorithm that has been developed, HOGPROP, propagates across Hierarchical Orthologous Groups (HOGs), which take into account both paralogy and orthology.

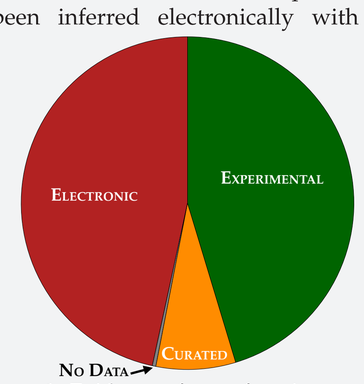

Figure 1: Evidence of gene-function annotations (GO) in *Arabidopsis thaliana* (data from UniProt-GOA). Total annotations 153,829, covering 2,848,215 terms across 25,179 genes.

What are HOGs?

A **Hierarchical Orthologous Group (HOG)** is a group of **sets of genes** arranged into a hierarchy, dependent on their location in the gene tree. Each one of these sets shares a **single common ancestor**, but genes can be a member of more than one set. This enables the comparison of highly diverged and similar species in a consistent manner.

For instance, if the following species and gene trees exist, then one could ask: *what sets exist at the taxonomic level of mammals?*

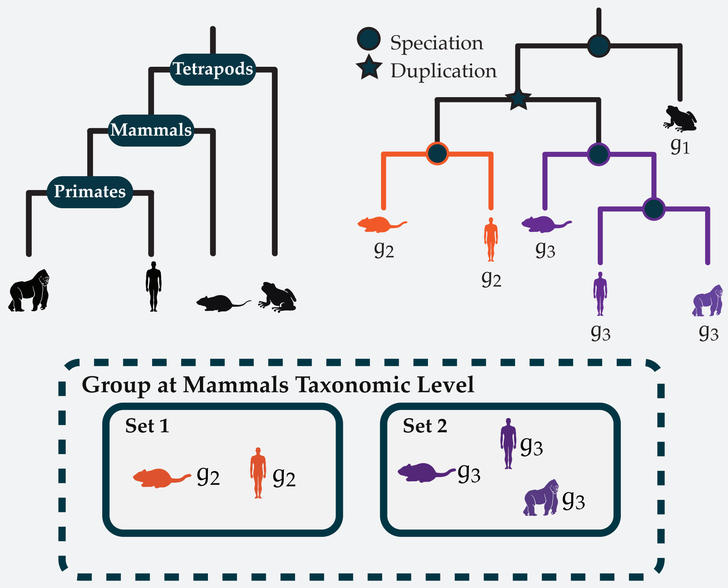

Figure 2: (Top Left) Example Species Tree; (Top Right) Example Gene Tree with Mammals gene sets highlighted; (Bottom) Gene sets at the mammals taxonomic level.

Alternatively, if one was interested in both amphibians and mammals, the HOG defines a single set of related genes. This taxonomic scoping is the advantage of using HOGs over orthologous groups.

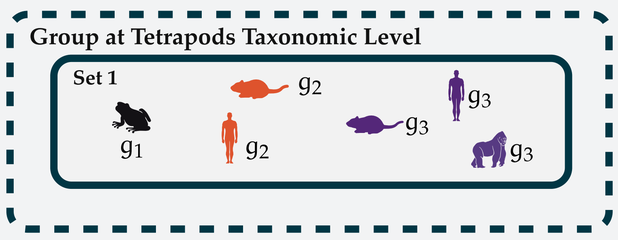

Figure 3: Gene group / set at the tetrapods taxonomic level.

The HOG structure shown here is the same as the gene tree, but many are possible. HOGs from the OMA project are available from the website* in OrthoXML format. For details on how these are generated, see Altenhoff *et al.* [Alt+13].

*http://www.omabrowser.org

Conclusions

HOGPROP, presented here, will enable rapid annotation using data from evolutionary diverse species in an evolutionary consistent manner. Work is being undertaken to show that this compares well with other state-of-the-art methods, with further improvement expected after refinement. There is also ongoing work into enabling HOGPROP to predict other types of knowledge, through the combination of different types of gene-knowledge associations.

References

[Alt+13] A. M. Altenhoff, M. Gil, G. H. Gonnet, and C. Dessimoz, "Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs", *PLoS One*, vol. 8, no. 1, e53786, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0053786.

[Jia+16] Y. Jiang *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy", *ArXiv e-prints*, Jan. 2016. arXiv: 1601.00891 [q-bio.QM].

HOGPROP – High-Precision Gene Annotation Propagation and Inference

Annotations can be propagated across a HOG. A subset of the GO annotations (experimental and some electronic annotations) are given a belief value dependent on their evidence code. These are then associated with the leaves of the HOG structure, before being pushed up and pulled down, with combination at each node.

The HOGs that are input into this method are (currently) available from the public OMA browser or from OMA standalone computations.
The aim is to release this as *free software* to enable annotation propagation locally on an individual's data.

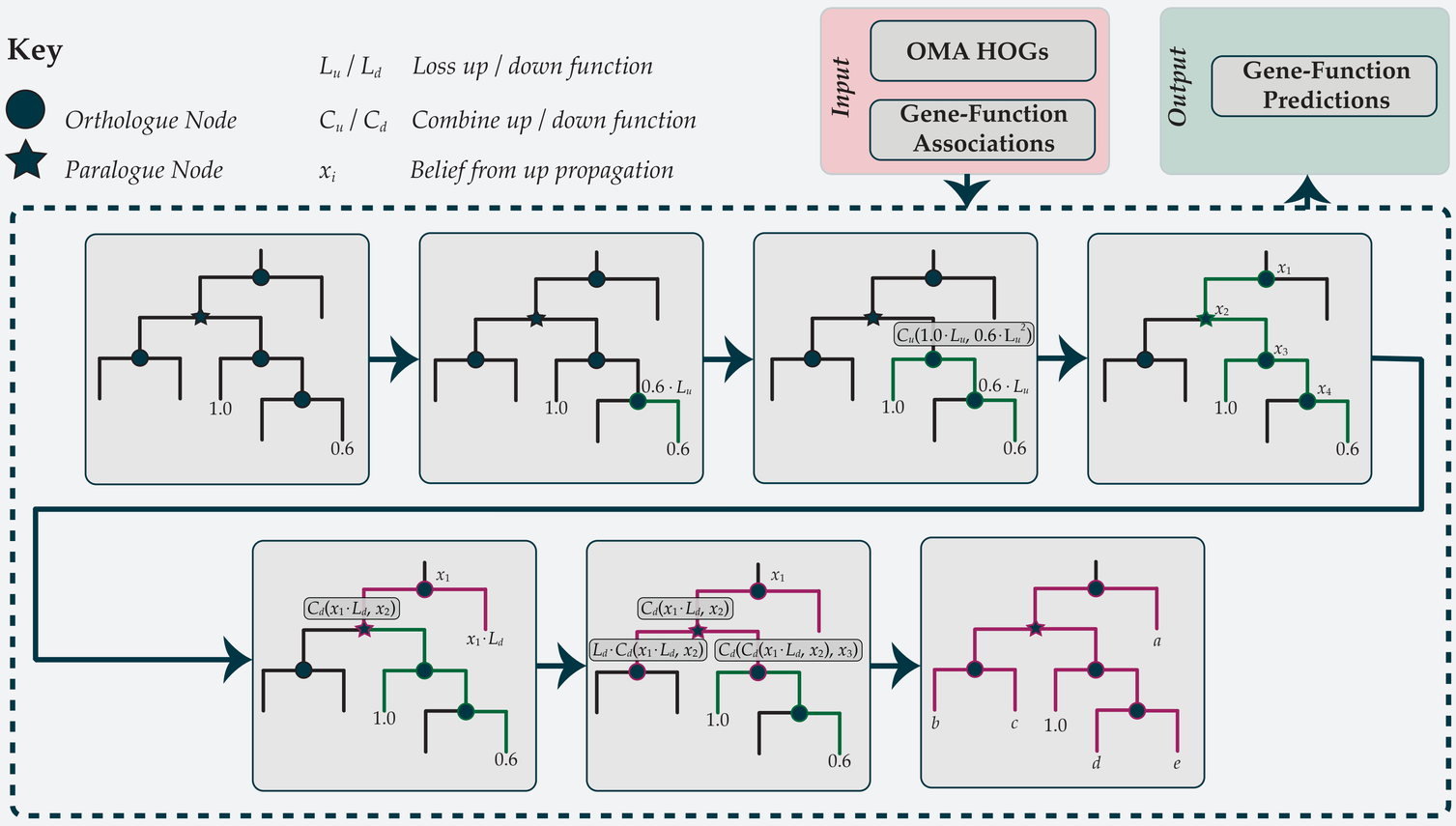

Figure 4: Overview of the method of HOG propagation, for a single term related to a gene-function association.

Results

There are promising results for HOGPROP. For instance, in the second CAFA challenge [Jia+16] it was in the top ten (of 126) methods for the overall evaluation using the minimum semantic distance (S_{min}), for two out of three of the GO namespaces. This measure is defined as

$$S_{min} = \min_{\tau} \left\{ \sqrt{ru(\tau)^2 + mi(\tau)^2} \right\},$$

where *ru* is the remaining uncertainty and *mi* is the misinformation and τ is the threshold score. A perfect predictor would be $S_{min} = 0$. For full details see [Jia+16].

Further evaluation is underway, before looking at specific applications.

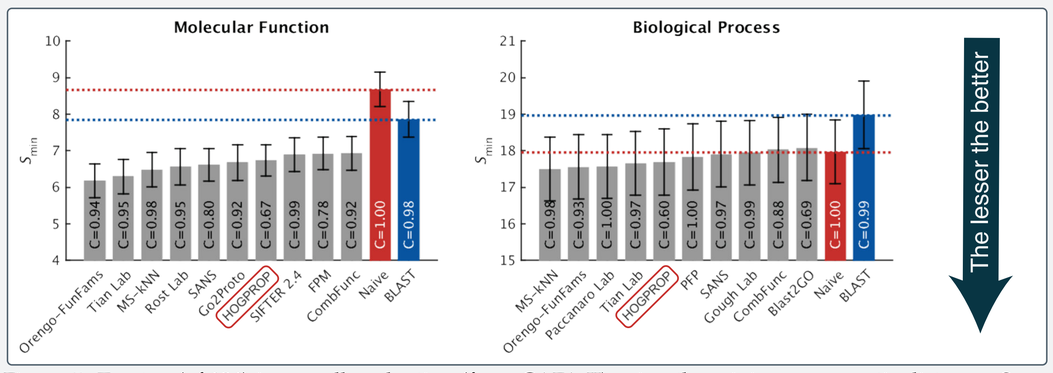

Figure 5: Top ten (of 126) in overall evaluation (from CAFA II) using the minimum semantic distance, S_{min}, for Molecular Function and Biological Process GO namespaces. The lower the distance, the higher the quality. C denotes *coverage* – the proportion of target genes for which at least one prediction was made. Plots from Figure 3 in [Jia+16].

Predictions from HOGPROP were submitted to the third CAFA experiment. Whilst making these predictions, some benchmarking was undertaken using the available data from the previous challenge [Jia+16]. Figures 7 and 8 show benchmark stats on the S_{min} measure at two stages. Figure 6 shows the F_{max} measure – the maximum of the harmonic mean of precision and recall – for the MFO at the last stage.

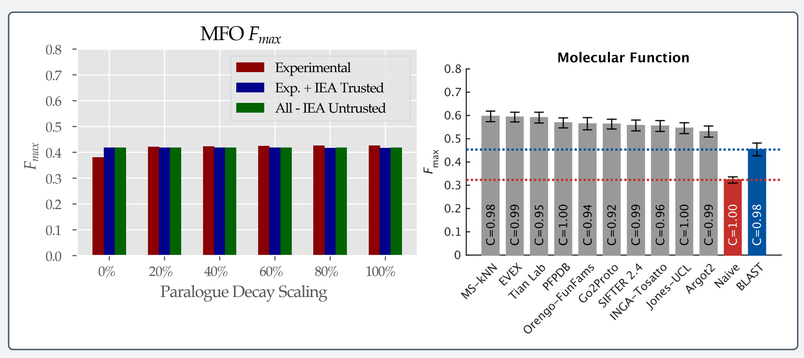

Figure 6: F_{max} measure for the MF ontology. Extra input through evidence code filtering, using max-combination. Coverage increased by ~ 18.3% with trusted IEA.

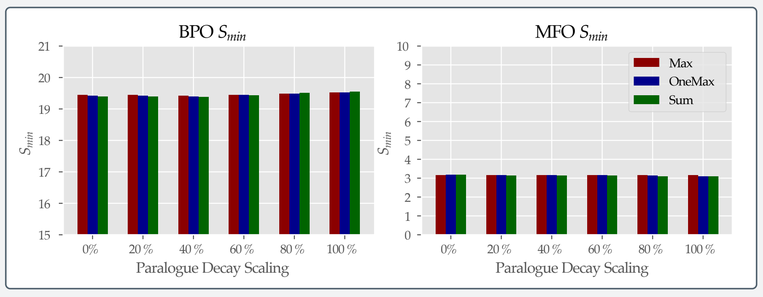

Figure 7: S_{min} measure for BP / MF ontologies. Experimental input only, testing three different combination methods. Coverage ~ 69% / ~ 67%.

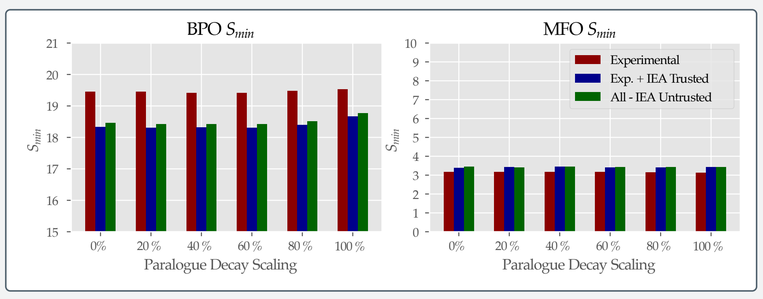

Figure 8: S_{min} measure for BP / MF ontologies. Extra input through evidence code filtering, using max-combination.
Coverage increased by ~ 16.8% / ~ 18.3% with trusted IEA.

Ongoing Work

Due to the nature of the HOGPROP algorithm, it lends itself well to other forms of data.

The overall aim of HOGPROP is to create a general gene knowledge propagation algorithm, using OMA HOGs.

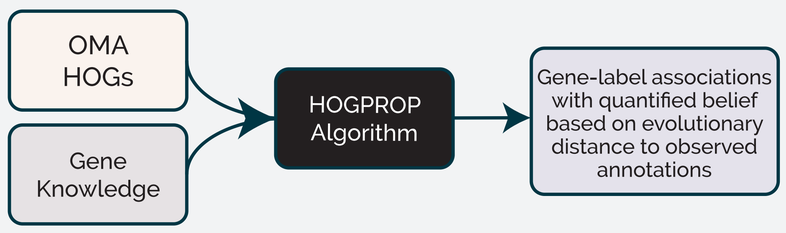

Figure 9: Overall aim of HOGPROP – to be a general gene knowledge propagation algorithm.