A new method for improving the prediction and the functional annotation of ortholog groups

Cécile Pereira, Alain Denise and Olivier Lespinet*

Institute of Genetics and Microbiology, Orsay, FRANCE

***Contact :** cecile.pereira@u-psud.fr

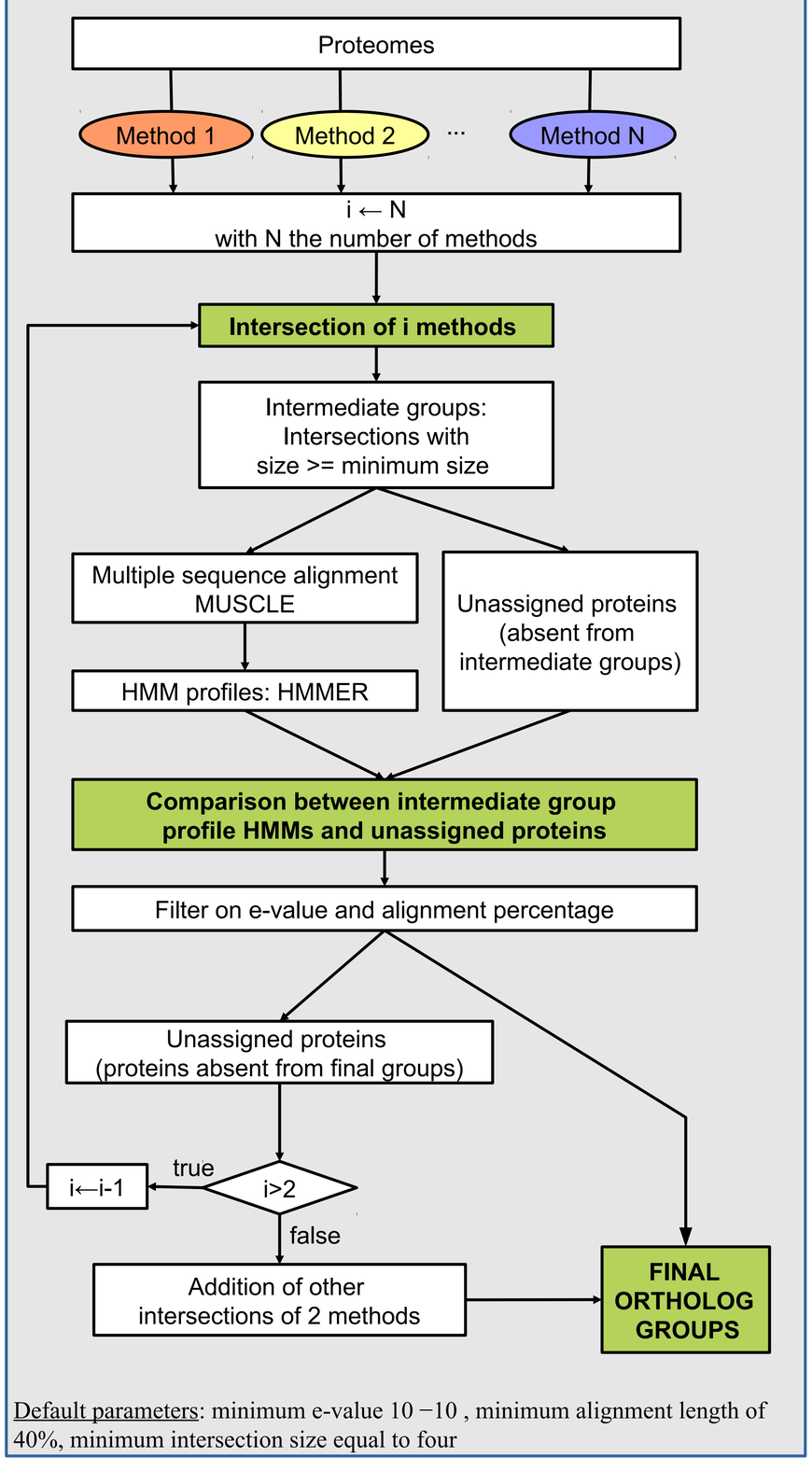
INTRODUCTION :

Orthologs are genes in different species that arose from a common ancestral gene by speciation events. Based on the 'orthology-function conjecture', the **orthologs retain the same function and thus can be used for the transfer of functional annotation from experimentally characterized genes to uncharacterized genes**. Nowadays not less than 37 databases offer prediction of orthologs. However, users should be aware that the application of different methods on the same proteomes can lead to distinct predictions.

In this work we present a meta-approach, called MARIO, which combines several methods. The purpose is to produce better quality results by using the overlapping results obtained from several individual methods. The rationale behind our approach is that when identical results are found by several methods then they are more likely accurate. This is especially true as the prediction methods use different approaches like tree-based or graph-based methods.

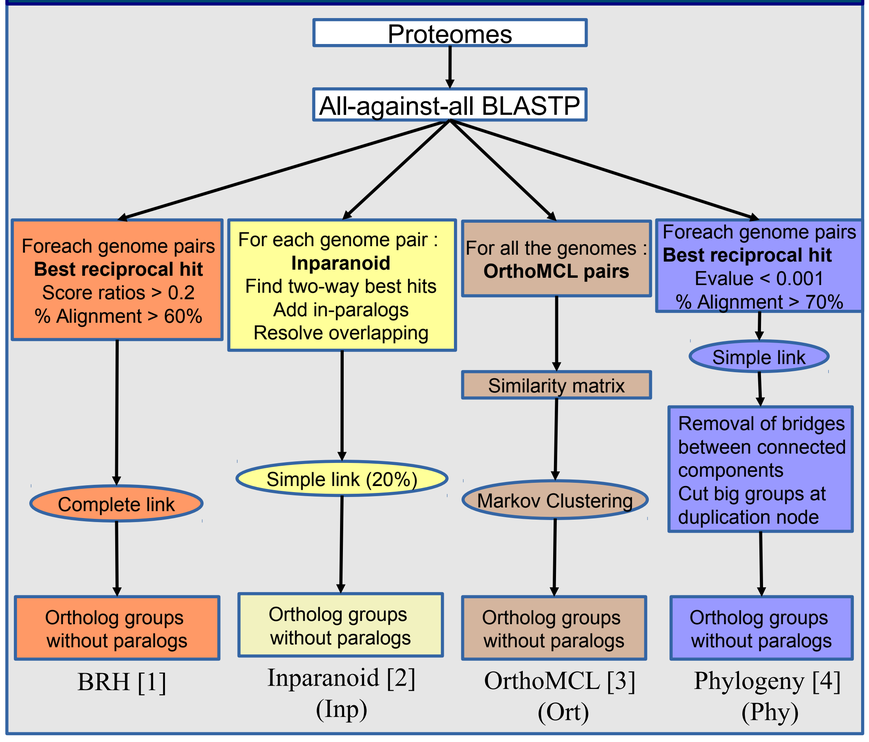
However, the overlap between multiple orthology prediction methods may lead to the loss of many true positives orthologs, especially when the number of initial methods is high. To overcome this problem the meta-approach is performed in **two steps**. An initial step **finds seeds for groups of orthologous genes** that correspond to the exact overlaps between all or at least several methods. Then we **expand these seed groups by using HMM profiles**. We report here results of MARIO using four initial approaches: BRH [1], Inparanoid [2], OrthoMCL [3] and Phylogeny [4].

1 MARIO, THE METHOD



Default parameters: minimum e-value 10 −10 , minimum alignment length of 40%, minimum intersection size equal to four

2 SELECTED INPUT METHODS



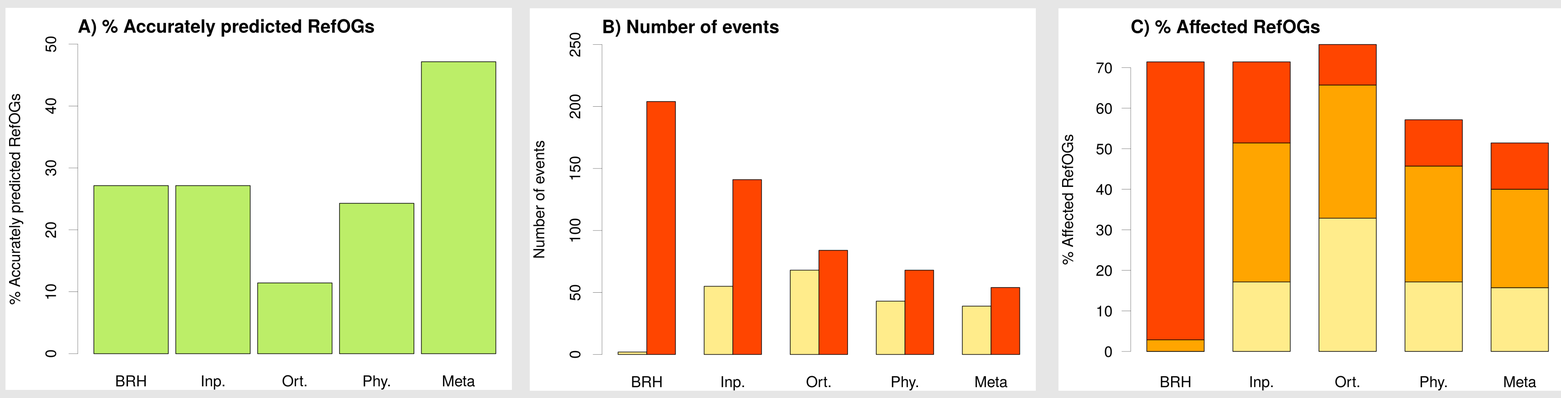
3 RESULTS

		Jaccard similarity coefficient				
		BRH	Inparanoid	OrthoMCL	Phylogeny	Meta-approach
# Identical groups	BRH	25384	0.541	0.172	0.389	0.060
	Inparanoid	7543	21342	0.248	0.340	0.093
	OrthoMCL	5272	7268	17524	0.164	0.156
	Phylogeny	4082	3260	2696	17944	0.079
	Meta-approach	3322	4705	4463	2654	14771
# Proteins		140561	163850	155984	124206	187902

Comparing the results of the four initial methods with those of the meta-approach on OrthoBENCH [5] (1519 proteins from 12 metazoan divided into 70 manually curated ortholog groups)**.**

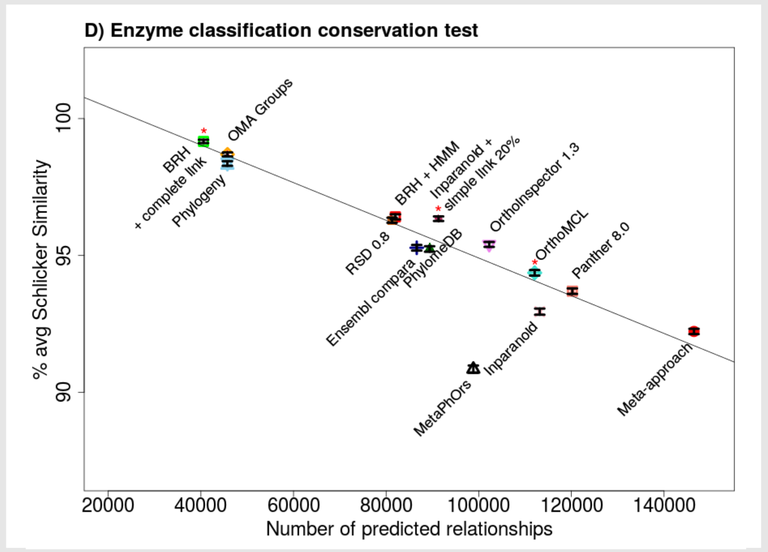
Table 1: The meta-aproach predicts fewer ortholog groups. However, the number of proteins in ortholog groups is the largest, involving biggest groups. None of the selected methods alone can explain the result of the meta-approach.

Figures A, B and C : The meta-approach increases of 73.7% in the number of accurately predicted groups compared to the highest result obtained with the four initial methods. It presents the lowest number of fissions and a number of fusions lower than three of the initial methods alone. **The meta-approach improves the results obtained with any of the initial methods.**



(A) Percentage of accurately predicted RefOGs (groups predicted without fusion or fission events), (B) Number of fusions (in dark gray) or fissions (in white), (C) Percentage of RefOGs affected by a fusion event (in dark gray), by a fission event (in white) or by the booth (in light gray). A fusion of groups corresponds to the addition of more than 3 erroneously assigned genes to a RefOG. Fissions correspond to a RefOG split in several groups: n group gives n − 1 fissions.

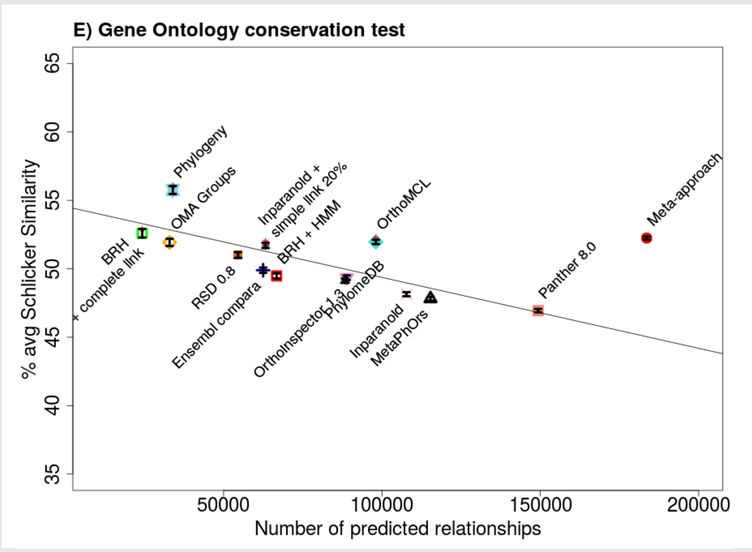
Functional similarity performance comparison on a common set of 66 species (Orthology Benchmark Service [6])



D) Enzyme classification conservation test. The linear regression curve has an intercept value of 101.8 and a regression coefficient of −6.887E−05. The black line is the linear regression obtained on all methods except the meta-approach, the metaPhors and the BRH plus HMM profiles approach. Error bars for each method are in black. Fourteen methods were compared

Figure D) Enzyme classification conservation test

The Pearson correlation test found a correlation between the number of annotated orthologs and the average Schlicker similarity obtained with the EC number annotations whether we use results of the meta-approach or not. The Pearson correlation equals −0.971 (p-value 7.436E−8) using the meta-approach and, −0.964 (p-value 8.573E−7) without the meta-approach (negative correlation hypothesis). Increasing the number of ortholog relations is correlated with a decrease in the average Schlicker similarity. All methods present a percentage of Schlicker similarity higher than 90%, revealing that **all methods, including MARIO, succeed in predicting pairs of enzymes with a similar function.**



E) Gene ontology conservation test. The linear regression curve obtained on the GO term annotation has an intercept value of 54,55 and a regression coefficient of −5.184E−05. The black lines are the linear regression obtained on all methods except the meta-approach, the metaPhors and the BRH plus HMM profiles approach. Error bars for each method are in black. Fourteen methods were compared.

Figure E) Gene Ontology conservation test

Without taking into account the meta-approach the Pearson correlation test found a correlation between the number of annotated orthologs and the average Schlicker similarity obtained on GO terms (Pearson coefficient was -0.804 (p-value 1.419E−3 with the negative correlation hypothesis). However, taking into account the meta-approach, the Pearson correlation test was not significant (−0.471 and p-value 0.06121). Furthermore, the point representing the meta-approach is above the linear regression curve showing that **the meta-approach outperforms the other methods on this dataset.**

CONCLUSION

The meta-approach appears to be a reliable method of prediction of ortholog groups. Based on the combination of existing methods, the meta-approach finds a consensus of higher quality. Both ortholog group quality and consistence of group annotation have been positively tested. The user has to be well aware that results depend of the selected input methods and on the selected parameters for the HMM profiles.

Sotware availability: The **MARIO** software which implements the meta-approach is freely available at **http://bim.igmors.u-psud.fr/mario/** .

Article in press: Cécile Pereira, Alain Denise and Olivier Lespinet: **A meta-approach for improving the prediction and the functional annotation of ortholog groups.** BMC Genomics 2014, 15 (Suppl 16)

Literature cited :

[1] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America 96(6), 2896–901 (1999)

[2] Li, L., Stoeckert, C.J., Roos, D.S.: Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Research 13(9), 2178–89 (2003). doi:10.1101/gr.1224503

[3] O'Brien, K.P., Remm, M., Sonnhammer, E.L.L.: Inparanoid: a comprehensive database of eukaryotic orthologs.Nucleic Acids Research 33(Database Issue), 476–480 (2005)

[4] Lemoine, F., Lespinet, O., Labedan, B.: Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evolutionary Biology 7(1), 237 (2007)

[5] Trachana, K., Larsson, T.a., Powell, S., Chen, W.-H., Doerks, T., Muller, J., Bork, P.: Orthology prediction methods: a quality assessment using curated protein families. BioEssays: news and reviews in molecular, cellular and developmental biology 33(10), 769–80 (2011). doi:10.1002/bies.201100062

[6] http://orthology.benchmarkservice.org/