GEPETTO : An open-source Framework for Gene Prioritization

http://sourceforge.net/projects/gepetto

Vincent WALTER (walterv@igbmc.fr) , Etienne GOFFINET, Julie THOMPSON, Olivier POCH, Hoan NGUYEN (nguyen@igbmc.fr)

Laboratoire de Bioinformatique et Génomique Intégratives – Département de Biologie Structurale Intégrative
Institut de Génétique et de Biologie Moléculaire et Cellulaire – UMR7104/U964 CNRS/INSERM/UDS – 1 Rue Laurent Friès / BP 10142 / 67404 Illkirch CEDEX

KEYWORDS : *GENE PRIORITIZATION – OPEN-SOURCE – FRAMEWORK – HEREDITARY DISEASES – EVOLUTION – GENOMIC CONTEXT – MUTATION – STRUCTURE – MODULE – SYSTEM*

General overview

WHAT IS PRIORITIZATION ?

Prioritization involves the identification of the **most promising features** associated with a specific problem in dynamic systems : biological process, genetic disease, network,….

The prioritization of candidate genes (or proteins) is crucial to the study of biological processes. Such studies are now possible thanks to the availability of heterogeneous biomedical data.

WHO IS GEPETTO FOR ?

GEPETTO is designed for :

- Biologists and clinicians via the **web interface [1]**: decrypthon.igbmc.fr/sm2ph/cgi-bin/gepetto
- Bioinformaticians via the **command line** or using the **Java APIs**

WHY AN OPEN-SOURCE FRAMEWORK ?

We distribute GEPETTO as an open-source project to allow :

- **maintainability** and long-term **availability**
- **improvement** by community
- complete **customization**
- interaction with **your databases**

Thus, the whole community can benefit from your developments.

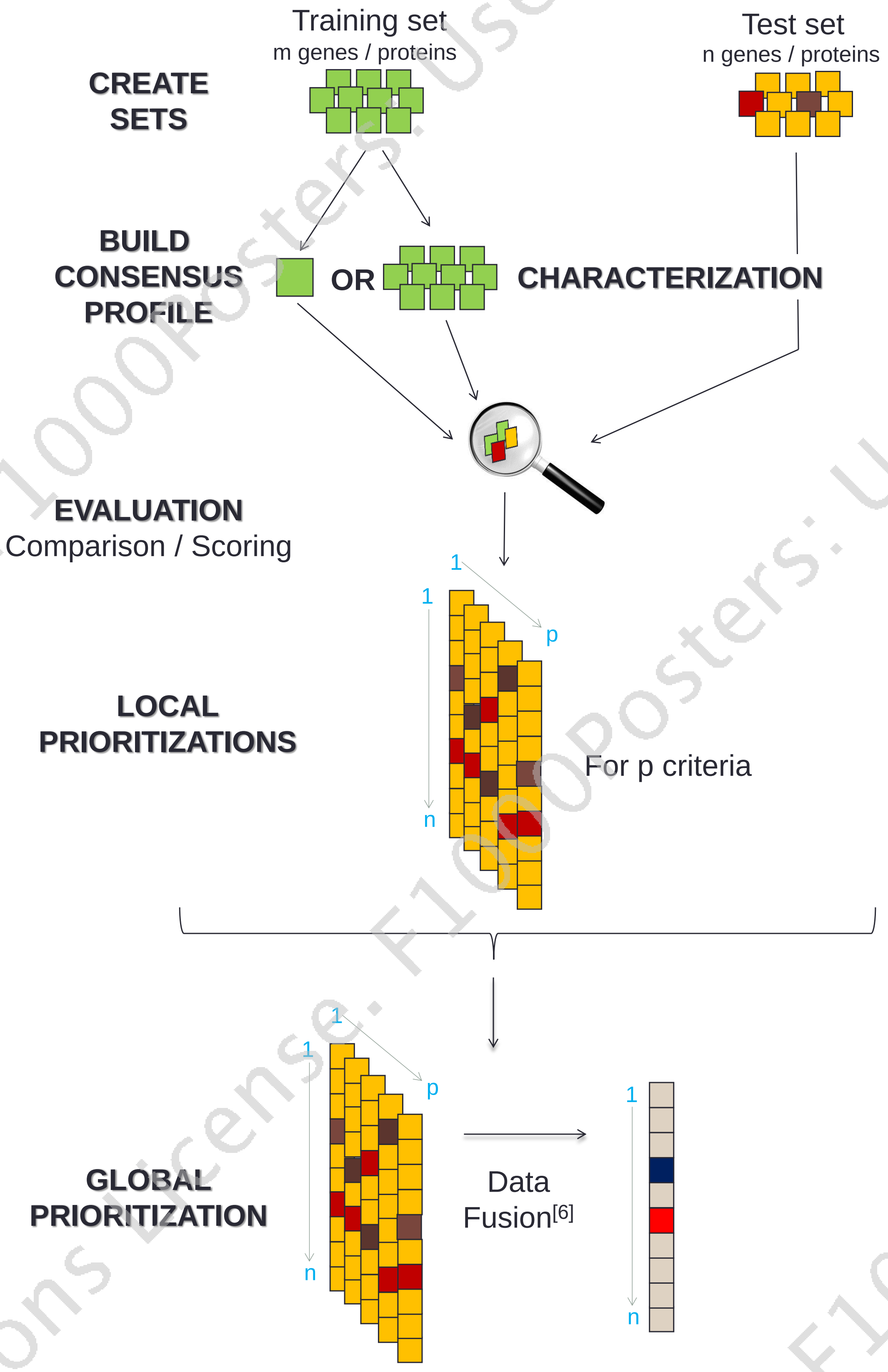
WHAT ARE THE TECHNOLOGIES USED ?

GEPETTO is written in **Java** (core and modules), **Python** (launcher) and **R** (statistical test for local prioritizations).

GEPETTO uses other open-source frameworks :

- **Spring Framework** to standardize, simplify and reduce the source code
- **Jboss JBPM** for the workflow engine

GEPETTO - Process

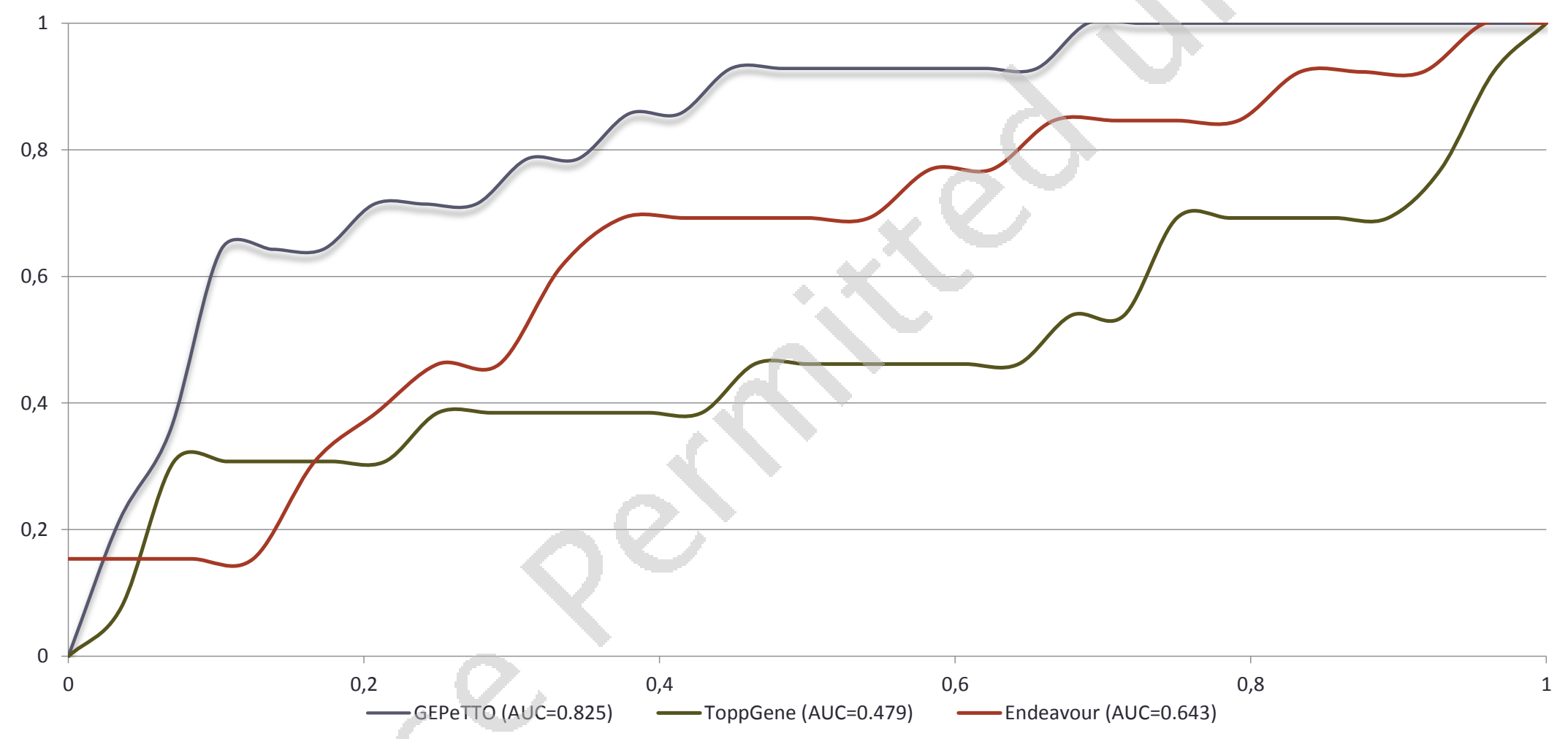


Application

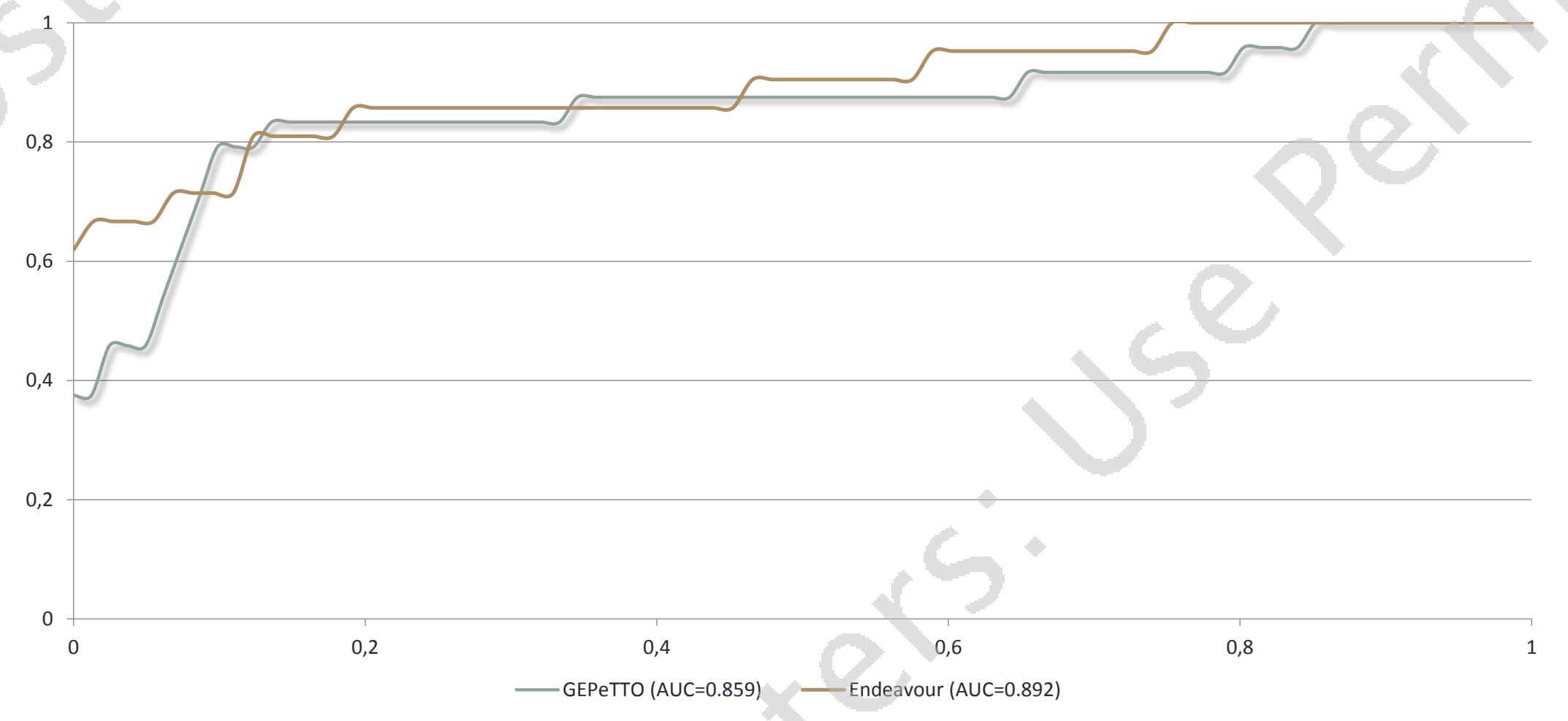
We have developed modules based on **common** criteria (sequence, protein-protein interactions, gene expression,…) and **original** ones related to **structure** and **evolution**.

We have compared the results of GEPETTO with the most popular gene prioritization tools : **Endeavour [8]** and **ToppGene [9]** using data related to AMD (Age-related Macular Degeneration).

A) Estimation of the capabilities to prioritize AMD known genes at the top of the ranked list

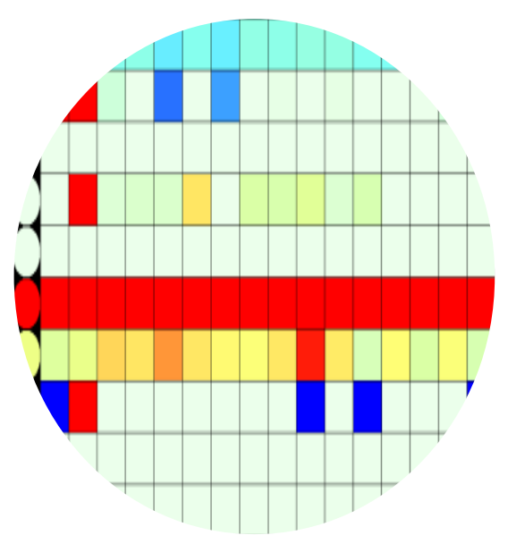


B) Comparison of phenotypically close diseases, Retinitis Pigmentosa (RP) and AMD, which do not affect the same genes



GEPETTO can **identify** and **discriminate** AMD/RP genes as well as Endeavour or ToppGene. Results are hopeful and suggest that **genomic context** and **evolution** may be good parameters for gene prioritization.

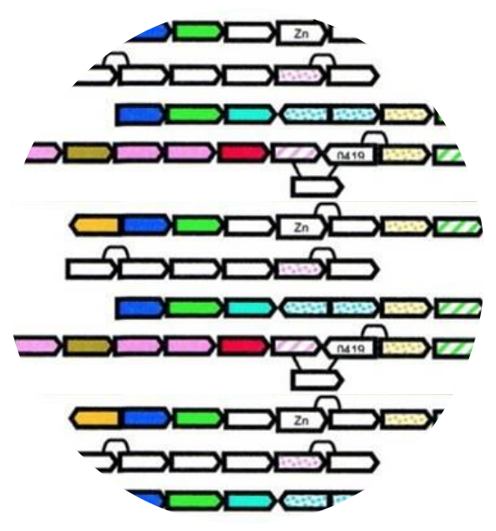
Local prioritizations



EVOLUTION EVOLUCODE

Compare evolutionary barcodes generated by Evolucode **[2]**. This is a representation that highlights similar evolutionary behavior.

- 10 parameters
- 16 vertebrates



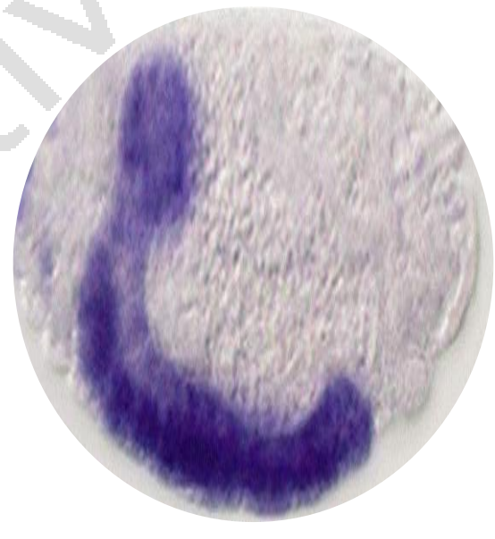
GENOMIC CONTEXT

Compare environment of genes using distance and occupancy profiles for proximal genetic elements (CpG islands, histones modifications, SNP) or epigenetic features (PolII binding sites, open chromatin, nested repeats) .

SEQUENCE

Compare sequences of proteins in the test set to a collection of reference sequences (training set) :

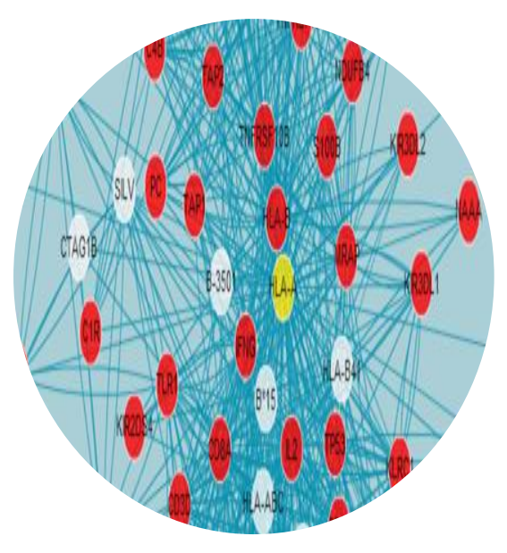
- Local alignment (BlastP **[3]**)
- Global alignment (Needlman-Wunsch)



TISSULAR EXPRESSION

Use trranscriptomic data from GxDb expression data obtained by microarray analysis :

- 79 human tissues
- 61 mouse tissues
- 6 normalization methods (gcRMA, MAS5,…)

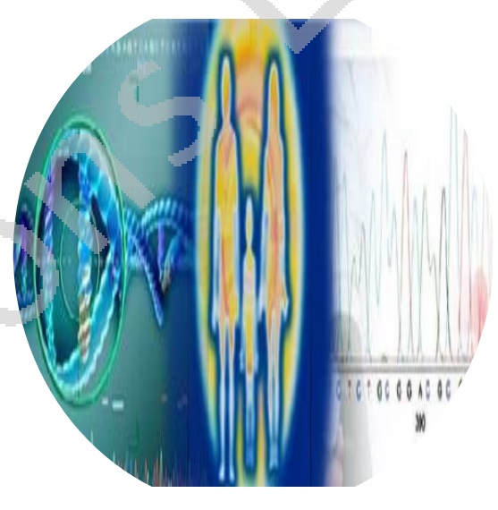


STRING-DB INTERACTION

Compare the overlap between all interactors of candidate elements using data provided by String-DB **[4]** (with a cut-off of 0.7 for the confidence score).

DISEASE PROBABILITY

Integrate the probability of involvment in a dominant or a recessive hereditary disease using data provided by IDGP **[5]**.



AMD SPECIFIC GWAS

Use the double-GC from confidential GWAS data of Age-related Macular Degeneration (AMD) study provided by the Institut de la Vision (Paris)

CUSTOMIZED MODULE

Related to your own database with custom criteria related to gene or associated protein.

Global prioritization

The current global prioritization method uses **order statistics** described by Aerts et al **[6]**. We want to implement other methods based on ranks to improve the prioritization :

- **Robust Rank Aggregation**
- **Mallows' Model**
- **GPSy [7] Model**

Perspectives

The future developments will focus on :

- Implementation of **new global** prioritization **methods**
- Creation of **modules** dedicated to **structural** data
- Integration of **SNP** parameters from MSV3d **[10]** using **Bayesian networks** for prioritization
- Integration of **orthology** aspects (mice, rat,…)
- Extension to **other organisms** and **ncRNA**
- **Pattern** extraction for hereditary causing-genes

References

[1] Friedrich, A., et al. (2010). **SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases**. Hum. Mutat. 31, 127–135.

[2] Linard, B., et al. (2011). **EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data**. Evolutionary Bioinformatics 61.

[3] Altschul, S.F., et al. (1990). **Basic local alignment search tool**. J. Mol. Biol. 215, 403–410

[4] Von Mering, C., et al. (2003). **STRING: a database of predicted functional associations between proteins.** Nucleic Acids Res. 31, 258–261

[5] Calvo, B., López-Bigas, N., et al. (2007). **A partially supervised classification approach to dominant and recessive human disease gene prediction**. Comput Methods Programs Biomed 85, 229–237.

[6] Aerts, S., et al. (2006). **Gene prioritization through genomic data fusion**. Nat. Biotechnol. 24, 537–544.

[7] Britto, R., et al. (2012). **GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development.** Nucleic Acids Research.

[8] Tranchevent, L.-C., et al. (2008). **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** Nucleic Acids Res. 36, W377–384

[9] Chen, J., et al. (2009). **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization**. Nucleic Acids Res. 37, W305–311.

[10] Luu, T.-D., Rusu, A.-M., Walter, V., et al. (2012b). **MSV3d: database of human MisSense Variants mapped to 3D protein structure**. Database (Oxford) 2012, bas018.

Acknowledgments

This work was funded by Université de Strasbourg (UDS), Institut de Génétique et de BIologie Moléculaire et Cellulaire (IGBMC), Strasbourg and Centre National de la Recherche Scientifique (CNRS), France. The data sets used to develop GEPETTO were provided by the Institut de la Vision (IdV), Paris.