GONNA: a Gene Ontology Nearest Neighbor Approach for the functional prediction of *P. falciparum* orphan genes

The database of the predictions

Laurent Bréhélin, Jean-François Dufayard, Olivier Gascuel

PlasmoExplore Project

Laboratoire
d'Informatique
de Robotique
et de Microélectronique
de Montpellier

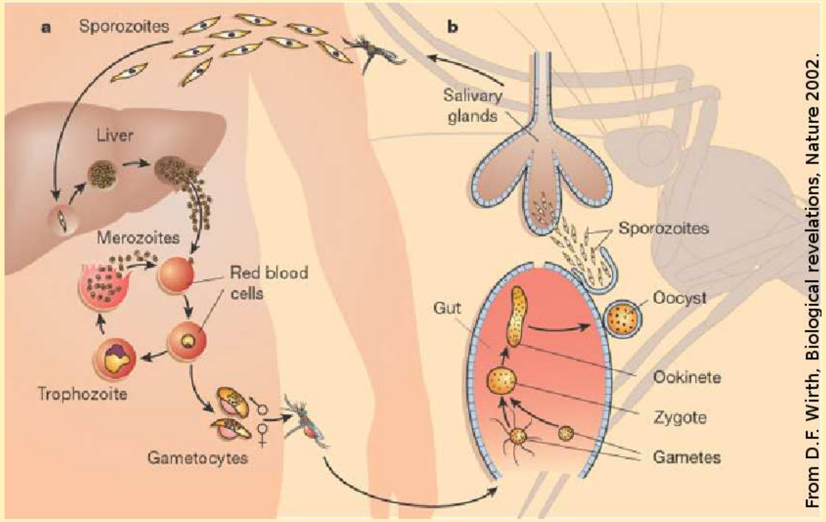
LIRMM

Méthodes
et Algorithmes
pour la Bioinformatique

ANR-06-MDCA-014

Financé par ANR

Plasmodium falciparum



From D.F. Wirth, Biological revelations, Nature 2002.

An atypical genome [Gardner et al., 2002]

- above 80% of A/T,
- only $\sim 40\%$ of the $5,300$ predicted genes can be annotated by sequence homology
  - because no homologous genes have already been characterized in other genomes
  - because standard tools fail to detect homology (sequence divergence is too large)

Non-homology based methods are needed to better characterize the $\sim 60\%$ of orphan genes

Guilt By Association (GBA) methods

Works in an intra-species way: the genes already characterized in the genome, *e.g.* by wet experiments or using sequence homology, help for the annotation of the other genes (the guilt by association principle)

Different postgenomic data can be used

- Transcriptomic data: genes with similar transcriptomic profiles are likely to share common functional roles [Eisen et al., 1998, Lockhart and Winzeler, 2000]
- Protein interaction data: proteins that share common interactors likely share common functions [Brun et al., 2003, Vazquez et al., 2003, Chen and Xu, 2004]
- Proteomic data, etc.

Two frameworks

- Unsupervised methods: unsupervised classification algorithms (clustering) + statistical test to search for over-represented functions
- Supervised methods: supervised classification algorithms to learn a gene function predictor

The Gene Ontology (GO)

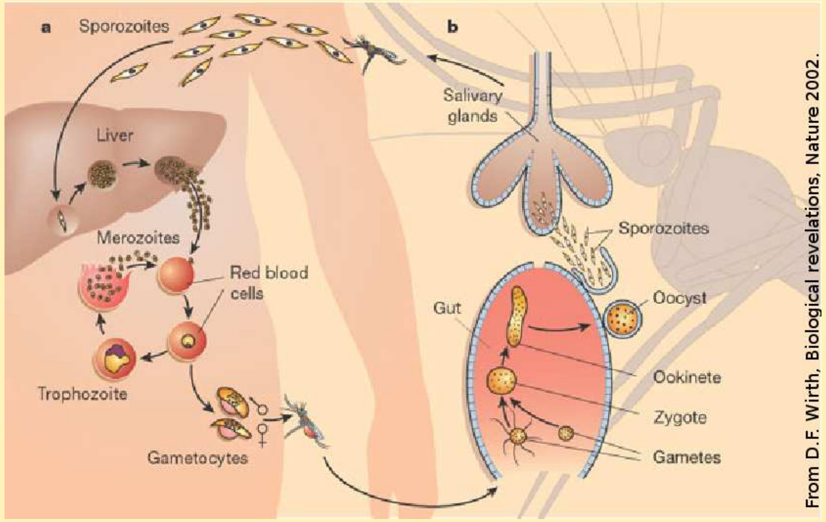
http://www.geneontology.org

- A systematic and standardized nomenclature to annotate genes in various organisms
- Three main ontologies:
  - Molecular Function
  - Biological Process
  - Cellular Component

- GO:0008150 : biological process
  - GO:0050789 : regulation of biological process
  - GO:0007582 : physiological process
    - GO:0008152 : metabolism
      - GO:0009058 : biosynthesis
        - GO:0044249 : cellular biosynthesis
          - GO:0009165 : nucleotide biosynthesis
          - GO:0016053 : organic acid biosynthesis
    - GO:0050875 : cellular physiological process
      - GO:0044237 : cellular metabolism
        - GO:0044249 : cellular biosynthesis

- Describes generalization relationships between hundreds of terms
- A gene may be annotated with several GO terms
- If a gene is annotated with a term t, then it is also annotated with all the terms that generalize t

P. falciparum: several postgenomic datasets available



From D.F. Wirth, Biological revelations, Nature 2002.

- **6 transcriptomic datasets:**
  - [Le Roch et al., 2003] 9 stages of the entire cycle of strain 3D7
    Measurements for $\sim 5,100$ genes
  - [Bozdech et al., 2003, Llinas et al., 2006] 48h intraerythrocytic developmental cycle for 3 strains: HB3, 3D7 and Dd2
    Measurements for $\sim 4,200$ genes
  - [Young et al., 2005] sexual developmental cycle (gametocytes) for 2 strains: 3D7 and NF54
    Measurements for $\sim 5,100$ genes
- **1 proteomic dataset:**
  [Florens et al., 2002, Le Roch et al., 2004] 7 stages of the entire cycle of strain 3D7
  Measurements for $\sim 2,900$ genes
- **1 protein interaction dataset:**
  [LaCount et al., 2005]
  Measurements for $\sim 1,300$ genes

GONNA - 1

Parameters

- For each postgenomic dataset d, compute a function $\mathcal{D}^d$ measuring the level of similarity $\mathcal{D}^d(g,h)$ of every pair of genes (g,h)
  - transcriptomic/proteomic data: Pearson correlation coefficient
    $\longrightarrow$ genes with correlated transcriptomic/proteomic profile have high similarity
  - protein interaction data: Czekanovski-Dice metric [Dice, 1945]
    $\longrightarrow$ genes that share many interactors have high similarity
- K and $K' \leq K$, two integers

Principle

Let g be an orphan gene

1. use the function $\mathcal{D}^d$ and the already characterized genes to search for the K nearest neighbors of g
2. for each GO term t, if at least K' of the K nearest neighbors are annotated with t, predict g to be annotated with t

GONNA - 2

Advantages

- predictions can be explained
- can be used with any present and future postgenomic dataset, as long as we have a relevant similarity measure
- consistent with the structure of the ontology
- low computing time: the confidence of the predictions can be assessed by cross-validation

Critical choices

- the similarity measure
- K: neither too large (neighbors are not similar) nor too small (sample is not representative)
- K':
  - high (close to K)
    - proportion of good predictions is high
    - few predictions on the most specific terms of the ontology
  - low
    - proportion of good predictions is lower
    - more predictions on the the most specific terms of the ontology

Assessing the confidence of the predictions made with a dataset

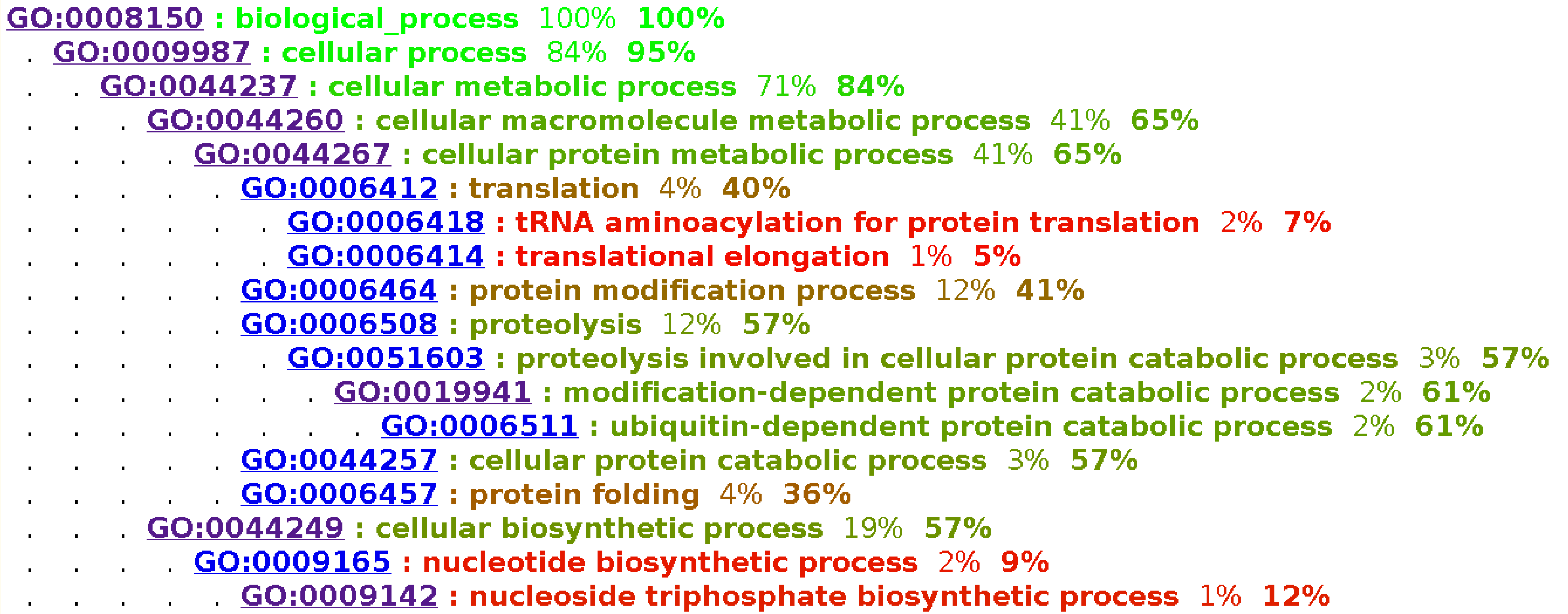
Leave-one-out Cross-validation (CV) [Hastie et al., 2001]

1. run GONNA on each characterized gene as it was an orphan gene
2. for each GO term t, compute the proportion of times GONNA is right when predicting that a gene has annotation t:
   True Discovery Rate (TDR) associated with t

Features

- confidence of the predictions can be estimated for each GO term
- highlights the parts of the ontology that are more suitable to apply a GBA approach with the considered dataset

An extract achieved with [Le Roch et al., 2003]



```
GO:0008150 : biological_process  100%  100%
 . GO:0009987 : cellular process  84%  95%
 .   . GO:0044237 : cellular metabolic process  71%  84%
 .   .   . GO:0044260 : cellular macromolecule metabolic process  41%  65%
 .   .   .   . GO:0044267 : cellular protein metabolic process  41%  65%
 .   .   .   .   . GO:0006412 : translation  4%  40%
 .   .   .   .   .   . GO:0006418 : tRNA aminoacylation for protein translation  2%  7%
 .   .   .   .   .   . GO:0006414 : translational elongation  1%  5%
 .   .   .   .   . GO:0006464 : protein modification process  12%  41%
 .   .   .   .   . GO:0006508 : proteolysis  12%  57%
 .   .   .   .   .   . GO:0051603 : proteolysis involved in cellular protein catabolic process  3%  57%
 .   .   .   .   .   .   . GO:0019941 : modification-dependent protein catabolic process  2%  61%
 .   .   .   .   .   .   .   . GO:0006511 : ubiquitin-dependent protein catabolic process  2%  61%
 .   .   .   .   . GO:0044257 : cellular protein catabolic process  3%  57%
 .   .   .   .   . GO:0006457 : protein folding  4%  36%
 .   .   . GO:0044249 : cellular biosynthetic process  19%  57%
 .   .   .   . GO:0009165 : nucleotide biosynthetic process  2%  9%
 .   .   .   .   . GO:0009142 : nucleoside triphosphate biosynthetic process  1%  12%
```

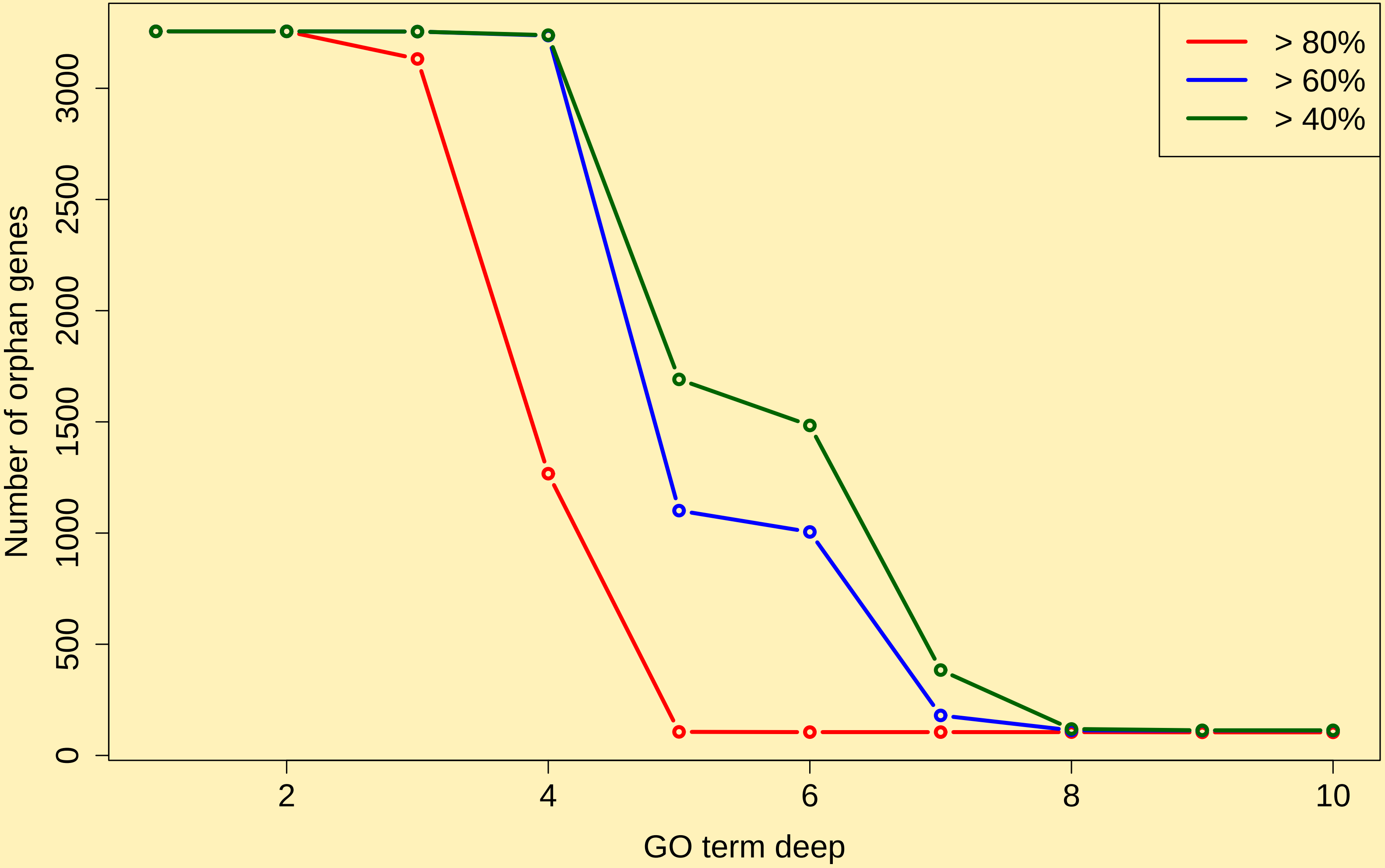
The database of the predictions

```
http://atgc.lirmm.fr/plasmo_draft/
```

- Run GONNA on all available datasets using two sets of parameters (K, K'):
  - a stringent set $(K = 6, K' = 4)$ to achieve high TDRs in the most suitable GO terms
  - a non-stringent set $(K = 6, K' = 2)$ to allow predictions in the more “difficult” GO terms
- Pool all the predictions made with the different postgenomic datasets
- The database can be
  - browsed through the Gene Ontology
  - queried by GO terms and genes

Assessing the global performances

Achieved with the transcriptomic dataset of [Le Roch et al., 2003]



References

[Bozdech et al., 2003] Bozdech, Z., Llinas, M., Pulliam, B., Wong, E., Zhu, J., and DeRisi, J. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, 1(1).

[Brun et al., 2003] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1).

[Chen and Xu, 2004] Chen, Y. and Xu, D. (2004). Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, 32(21):6414–6424.

[Dice, 1945] Dice, L. (1945). Measure of the amount of ecologic association between species. *Ecology*, 26:297–302.

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868.

[Florens et al., 2002] Florens, L., Washburn, M., Raine, J., Anthony, R., Grainger, M., Haynes, J., Moch, J., Muster, N., Sacci, J., Tabb, D., Witney, A., Wolters, D., Wu, Y., Gardner, M., Holder, A., Sinden, R., Yates, J., and Carucci, D. (2002). A proteomic view of the plasmodium falciparum life cycle. *Nature*, 419(6906):520–526.

[Gardner et al., 2002] Gardner, M., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R., Carlton, J., Pain, A., Nelson, K., Bowman, S., Paulsen, I., James, K., Eisen, J., Rutherford, K., Salzberg, S., Craig, A., Kyes, S., Chan, M., Nene, V., Shallom, S., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M., Vaidya, A., Martin, D., Fairlamb, A., Fraunholz, M., Roos, D., Ralph, S., McFadden, G., Cummings, L., Subramanian, G., Mungall, C., Venter, J., Carucci, D., Hoffman, S., Newbold, C., Davis, R., Fraser, C., and Barrell, B. (2002). Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer.

[LaCount et al., 2005] LaCount, D., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J., Schoenfeld, L., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. (2005). A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, 438(7064):103–107.

[Le Roch et al., 2004] Le Roch, K., Johnson, J., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S., Williamson, K., Holder, A., Carucci, D., 3rd Yates, J., and Winzeler, E. (2004). Global analysis of transcript and protein levels across the plasmodium falciparum life cycle. *Genome Res*, 14(11):2308–2318.

[Le Roch et al., 2003] Le Roch, K., Zhou, Y., Blair, P., Grainger, M., Moch, J., Haynes, J., Vega, P. D. L., Holder, A., Batalov, S., Carucci, D., and Winzeler, E. (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508.

[Llinas et al., 2006] Llinas, M., Bozdech, Z., Wong, E., Adai, A., and DeRisi, J. (2006). Comparative whole genome transcriptome analysis of three plasmodium falciparum strains. *Nucleic Acids Res*, 34(4):1166–1173.

[Lockhart and Winzeler, 2000] Lockhart, D. and Winzeler, E. (2000). Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–836.

[Vazquez et al., 2003] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, 21(6):697–700.

[Young et al., 2005] Young, J., Fivelman, Q., Blair, P., P, P. d., Le Roch, K., Zhou, Y., Carucci, D., Baker, D., and Winzeler, E. (2005). The plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol*, 143(1):67–79.