

# Identification of assembly of integrated biological systems in completely sequenced genomes using Data Mining

COLOMBO T.<sup>1,2</sup>

FICHANT G.<sup>1</sup>

QUENTIN Y.<sup>1</sup>

1. Laboratoire de Chimie Bactérienne. IBSM 31, chemin Joseph Aiguier 13009 Marseille FRANCE

2. Institut de Mathématiques de Luminy. 163, avenue de Luminy 13009 Marseille FRANCE

Complex biological processes involve protein-protein interactions between a finite set of partners, here referred to as an integrated system. If the identification of the different partners can be achieved from the analysis of the genomic sequence with an appropriate combination of bioinformatic tools, their assembly in a functional system requires extra information, in particular when genes belong to families of paralogs, since several combinations of partners can be generated. A useful paradigm is the proposal that the genes encoding proteins of an integrated system frequently occurred in the same chromosomal neighborhood and thus the analysis of genomic context has been used to assemble the partners. However, such approach does not guarantee that the assembled partners are functional *in vivo* and more directly it encounters its limits when genes are not conserved in close vicinity. In order to reinforce the predictable power of an integrated system, we can use evolutionary relationships between genes encoding the partners but with the assumption that they evolve in parallel: they are vertically and horizontally transmitted as a unit. Then, if gene re-localisation occurs in one lineage, their conservation in other lineages may help to re-assemble them in the first one. The major difficulty is to unambiguously identify orthologous relationships between genes since we are dealing with families of paralogs. Therefore, we used the restricted definition of orthology proposed by Fitch, referred to as isorthology (Fitch, 2000), which allows only a one-to-one orthology link between genes belonging to two genomes.

In order to solve this problem, we have developed a new method based on **data mining** approaches (Agrawal, 1993). The genes encoding the partners are referred to as **items** and evolutionary and genomic neighborhood are referred to as **associations**. The principle of the method is to deduce functional associations between items throughout the discovery of **association rules**.

We will use the following formalism :

- $I_k^i$  stands for a gene  $k$  in a genome  $G_i$ .
- $P(I_k^i, I_l^i)$  expresses a physical relationship between two genes involved in the same integrated system in genome  $G_i$ .
- $O(I_k^i, I_m^j)$  expresses an evolutionary relationship (isorthology) between the gene  $k$  from the genome  $G_i$  and the gene  $m$  from the genome  $G_j$ .

Note that all the types of relations are bijectives.

Let two genes  $n$  and  $r$  from genome  $G_i$ , we search for association rules  $AR(I_n^i, I_r^i : j)$  between genome pair  $(G_i, G_j)$  and composed of evolutionary and physical relationships such as :

- $AR(I_n^i, I_r^i : j)$  such as :  $\exists O(I_n^i, I_m^j), P(I_m^j, I_k^j)$  and  $O(I_k^j, I_r^i)$  as shown on Fig 1.

Association rules are computed for all genome pairs  $(G_i, G_j)$  with  $1 \leq j \leq R$ , and  $i \neq j$ . The pair of genes which occurs more frequently over all comparisons is predicted as functional association. The confidence level of the prediction can be roughly estimated as i) the number of occurrences of the predicted pair over all genomes association rules supporting the predicted partner and as ii) the ratio of this number to the sum of association rules that involve one or both genes of the predicted pair. When a gene has no isorthology relationship or when a predicted pair has a support lower than the “minimum confidence threshold”, then the same formula can be applied with isoparalogy relationship, when it exists. However, in this case, the prediction will rely only upon one observation, as a consequence of the definition of the isoparalogy (Fitch, 2000). At the end of the process, the pairs of genes with non-null intersection are assembled in a putative functional integrated system.

The method has been implemented in Perl and tested with a particular family of integrated systems: the ABC transporters. These systems are retrieved from our high quality annotated database (Quentin and Fichant, 2000; Quentin et al., 2002). As a validation step, we used samples of well-known assemblies from one genome and try to reconstruct them using informations of the other genomes. Results will be presented and discussed according to the parameters used to define isorthology, the number of genomes included and the properties of the ABC families.

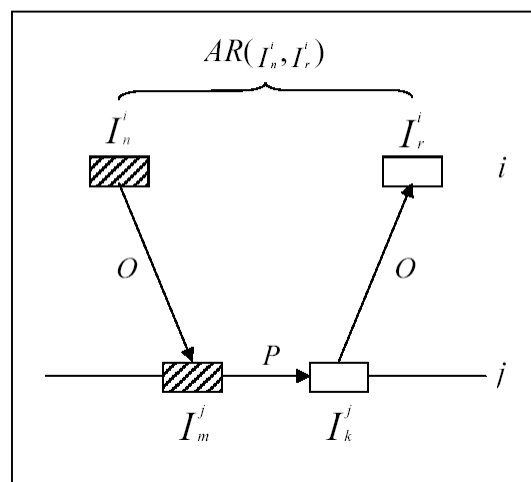


Fig 1 : Discover the association rule

### References

- AGRAWAL, R., IMIELINSKI, T., and SWANI, A. (1993) *Mining association rules between sets of items in large databases*. Proc. ACM SIGMOD, 207-216.
- FITCH, W. M. (2000) *Homology : a personal view on some of the problems*, Trends in Genetics, **16**, pages ?.
- QUENTIN, Y. and FICHANT, G. (2000) *ABCdb : an ABC transporter database. Assembly and Analysis of ABC Transport Systems in Complete Genomes*, J. Mol. Microbiol. Biotechnol., **2**, 501-504.

QUENTIN, Y., CHABALIER, J., and FICHANT, G. (2002) Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes. *Computers and Chemistry*, 26, 447-457.