

Functional inference by genetic context analysis – an application to ABC transporters

Tristan COLOMBO^{†*} Belaïd BENHAMOU⁺ Cécile CAPPONI[‡] Gwennaele FICHANT[†] Yves QUENTIN[†]

[†] Laboratoire de Chimie Bactérienne. IBSM 31, chemin Joseph Aiguier 13009 Marseille

^{*} Institut de Mathématiques de Luminy. 163, avenue de Luminy 13009 Marseille

⁺ Laboratoire des Sciences de l'Information et des Systèmes. 39, avenue Joliot Curie 13013 Marseille

[‡] Laboratoire d'Informatique Fondamentale de Marseille. 39, avenue Joliot Curie 13013 Marseille

Courriel : {colombo, fichant, quentin}@ibsm.cnrs-mrs.fr, capponi@cmi.univ-mrs.fr

The availability of multiple complete genomes offers the opportunity to develop new methods to complement more traditional similarity-based methods for predicting protein function. Among them, methods that predict higher order of organisations such as metabolic pathways, regulatory networks, or protein assemblies appear to be the most promising. According to [6], these methods fall in three classes: identification of fused genes [3][9], phylogenetic profiles [12][15-16], and analysis of conservation of the local neighbourhood of genes [2][11]. In the past few years, the later received the largest attention. This approach is based on the observation that a prominent feature of bacterial chromosomes is the presence of transcriptional units (operons) composed of functionally related genes. This observation is further confirmed by statistical analysis of functionally related ORFs that have a strong tendency to be in the same strand when there are located in the same neighbourhood [11]. In this context, the question of the conservation of genomic context has been addressed throughout the analysis of genes belonging to the same transcriptional unit. Since these units are generally unknown, they are inferred as sets of genes in the same orientation and with spacer regions of less than 300 bp (referred as genes in runs in [11]).

Surprisingly, several studies have consistently shown that only a few genes exhibit highly conserved neighborhoods in many genomes, moreover most of them encode proteins in direct physical interactions. Aside models based on physical interactions and co-regulations, other hypothesis have been proposed to explain gene neighborhood conservations [8]. It is worthy of note that some of them do not rely on constraints such as orientation and spacer between genes, but just rely on genomic proximity. For example, gene clustering might be beneficial when high local concentration of protein products in the cytoplasm is necessary to perform a complex function (referred as Molarity Model in [8]). From an evolutionary perspective, genes that function together can confer a selectable phenotype and thus, their grouping in discrete region of the chromosome may facilitate their propagation to other organisms throughout horizontal transfers. In this case, the gene cluster is initially beneficial to the genes themselves, not to their host organisms (referred as Selfish Operon Model in [8]).

Therefore, there are evidences for genes functionally related to be conserved in the same neighborhood without the constraint to remain in the same operon. Hence, we decide to develop a method to delineate such conserved gene clusters from a large number of genomes, without restraining the analysis to genes in the same orientation and in a close proximity. We applied our algorithm to genes coding for proteins participating to the assembly of ABC transporters systems reconstructed from bacterian genomes. These systems, involved in the exchanges (import and export) with the external environment, are suspected be to transmitted by horizontal transfers. Genes encoding different components of these systems are known to be conserved in the same neighbourhood, but not necessary in the same putative transcriptional unit [13-14]. In addition, analysis of the proximities of these genes could reveal new partners of the integrated system, such as proteins implicated in the metabolism of the substrate.

Given a query gene in a given genome, our aim is to study the conservation of its neighbourhood in all other genomes. Thus, the analysis is anchored by the query gene in one genome and its relative in the other genome. These genes are referred as the pair of anchor genes. We assume that we are able to define relationships between genes in from two genomes (i.e. homology or orthology relationship). Another assumption of our method is that functional relationship between genes will decrease with increasing distances between them. Since our analysis do not rely on transcriptional constraints, distances can be expressed as the number of intervening genes between the related pairs of genes and without constraint on their orientation. The region to explore apart the pair of anchor genes can be either chosen by the user (fixed window size) or computed with an upper gap threshold as parameter.

If the window length is fixed, a simple way to resolve this problem is to consider neighbour conserved genes as lists in each genomes. These lists should include the anchor gene and are extended from each side until a gap greater than the accepted level is encountered. However, since conserved genes in one list are not necessarily present in the second one, an intersection of lists is performed to get the subset of common genes. At this step,

the eliminated genes may increase the gap length and reduce the list length on both sides. Hence, the procedure is iterated until convergence. At the end, we obtain the list of conserved genes around the anchor gene in both genomes, for a fixed window length and a gap threshold. Since the set of genes conserved with gap $\delta-1$ is included in the set of genes conserved with gap δ , the algorithm can start with the greater gap length, and uses the list computed at gap δ to compute the conserved genes at gap $\delta-1$. This greedy algorithm has been implemented to serve as reference. However, keeping in mind that our aim is to explore complete genomes, a more efficient algorithm will be presented. It was developed using the framework of the temporal satisfaction problem [10] belonging to the artificial intelligence field. The complexity obtained is linear and allow to analyse the large amount of biological data. This method will be implemented as a task in the ABCKB database [1].

When we look for conserved neighbourhood, we first assume that it is possible to identify conserved genes between genomes. This is generally achieved by the delineation of orthologous relationships. This notion of orthology received prominent attention because it is generally admitted that orthologous genes encode the same function in different genomes. However, if the notion of orthologous genes can be clearly established at the evolutionary level [4-5], its implementation is almost impossible since evolutionary pathway of the genes cannot be reconstructed with enough confident level. However, one may ask whether orthologous relationships are an absolute requirement when we are looking for conserved gene clusters. Indeed, co-occurrence of homologous genes could be as well as informative to suspect functional links. An illustration of this idea is the identification of unexpectedly frequent associations of genes encoding two-component systems with genes encoding ABC transporter in *Bacillus Clostridium* group [7] which were found functionally linked, at least at the transcriptional level.

References

- [1] CAPPONI (C.), CHABALIER (J.), QUENTIN (Y.) and FICHANT (G.), *A Knowledge Base for Integrated Biological Systems*. IEEE Intelligent Systems, 16, 52-60, 2001.
- [2] DANDEKAR (T.), SNEL (B.), HUYNEN (M.) and BORK (P.), *Conservation of gene order: a fingerprint of proteins that physically interact*. Trends Biochem Sci. 23, 324-328, 1998.
- [3] ENRIGHT (A. J.), ILIOPOULOS (I.), KYRPIDES (N. C.) and OUZOUNIS (C. A.), *Protein interaction maps for complete genomes based on gene fusion events*. Nature 402, 86-90, 1999.
- [4] FITCH (W. M.) *Distinguishing homologous from analogous proteins*. Syst. Zool., 19, 99-113, 1970.
- [5] FITCH (W. M.), *Homology : a personal view on some of the problems*. Trends in Genetics, 16, pages ?, 2000.
- [6] HUYNEN (M. A.), and SNEL (B.), *Gene and context: integrative approaches to genome analysis*. Adv Protein Chem., 54, 345-79, 2000.
- [7] JOSEPH (P.), FICHANT (G.), QUENTIN (Y.) and DENIZOT (F.), *Genetic link between ABC permease and regulatory systems in the Bacillus/Clostridium group suggests an involvement in a common physiological process*, . J. Mol. Microbiol. Biotechnol., (in press).
- [8] LAWRENCE (J. G.), *Gene transfer, speciation, and the evolution of bacterial genomes*. Curr Opin Microbiol. 2, 519-523, 1999.
- [9] MARCOTTE (E. M.) , PELLEGRINI (M.), Ng (H. L.), RICE (D. W.), YEATES (T. O.) and EISENBERG (D.), *Detecting protein function and protein-protein interactions from genome sequences*. Science. 285, 751-753, 1999.
- [10] MONTANARI (U.), *Networks of constraints : fundamental properties and application to picture processing*. Information Sciences, 7, 95-132, 1974.
- [11] OVERBEEK (R.), FONSTEIN (M.), D'SOUZA (M.), PUSCH (D. G.) and MALTSEV (N.), *The use of gene clusters to infer functional coupling*. Proc. Natl. Acad. Sci. USA, 96, 2896-2901, 1999.
- [12] PELLEGRINI (M.), MARCOTTE (E. M.), THOMPSON (M. J.), EISENBERG (D.) and YEATES (T. O.), *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles*. Proc Natl Acad Sci USA. 96, 4285-4288, 1999.
- [13] QUENTIN (Y.) and FICHANT (G.), *ABCdb : an ABC transporter database. Assembly and Analysis of ABC Transport Systems in Complete Genomes*. J. Mol. Microbiol. Biotechnol. 2, 501-504. 2000.
- [14] QUENTIN (Y.), CHABALIER (J.) and FICHANT, (G.), *Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes*. Computers and Chemistry, sous presse, 2002.
- [15] SNEL (B.), BORK (P.) and HUYNEN (M. A.), *Genome phylogeny based on gene content*. Nat Genet. 21, 108-110, 1999.
- [16] SNEL (B.), BORK (P.) and HUYNEN (M. A.), *Genome evolution. Gene fusion versus gene fission*. Trends Genet. 16, 9-11, 2000.