



Annotation of bacterial genomes using improved phylogenomic profiles

F. Enault*, K. Suhre, C. Abergel, O. Poirot and J.-M. Claverie

Structural and Genomic Information, CNRS - UPR 2589, 31 chemin Joseph Aiguier, 13009 Marseille, France

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Phylogenomic profiling is a large-scale comparative genomic method used to infer protein function from evolutionary information first described in a binary form by Pellegrini *et al.* (1999). Here, we propose improvements of this approach including the use of normalized Blastp bit scores, a normalization of the matrix of profiles to take into account the evolutionary distances between bacteria, the definition of a phylogenomic neighborhood based on continuous pairwise distances between genes and an original annotation procedure including the computation of a p -value for each functional assignment.

Results: The method presented here increases the number of Ecocyc enzymes identified as being evolutionarily related by about 25% with respect to the original binary form (absent/present) method. The fraction of 'false' positives is shown to be smaller than 20%. Based on their phylogenomic relationships, genes of unknown function can then be automatically related to annotated genes. Each gene annotation predicted is associated with a p -value, i.e. its probability to be obtained by chance. The validity of this method was extensively tested on a large set of genes of known function using the MultiFun database. We find that 50% of 3122 function attributions that can be made at a p -value level of 10^{-11} correspond to the actual gene annotation. The method can be readily applied to any newly sequenced microbial genome. In contrast to earlier work on the same topic, our approach avoids the use of arbitrary cut-off values, and provides a reliability estimate of the functional predictions in form of p -values.

Contact: enault@igs.cnrs-mrs.fr

keywords: phylogenomics, functional prediction, automated annotation, bacteria, evolution.

INTRODUCTION

Phylogenomic profiling is a non-sequence-homology-based method designed to infer a likely functional relationship between proteins. It is based on the assumption that proteins involved in a common metabolic

pathway or constituting a multi-molecular complex are likely to evolve in a correlated manner. This paradigm was put to use by Pellegrini *et al.* (1999) and first applied to *Escherichia coli*. It has been further developed in other studies (Zheng *et al.*, 2002; Bilu *et al.*, 2002). According to Pellegrini *et al.* (1999), a phylogenomic profile is defined as a string of bits, each bit representing the presence or absence of an homologous gene in a given genome. Two proteins are said to have co-evolved if their profiles differ at most at one bit position, that is if their Hamming distance is at most one. The limitation of this approach is that arbitrary sequence alignment score thresholds are used to decide about the presence of orthologous genes between organisms. The method presented here avoids making such a decision, as the concept of orthologous relationships is not explicitly used. We replace the binary phylogenomic profiles introduced by Pellegrini *et al.* (1999) by continuous values consisting of the normalized scores of the best Blastp (Altschul *et al.*, 1997) alignments between each *E.coli* protein and its best matching ORFs in the other genomes. We further normalize the columns of the profile matrix so that all organisms are given a similar phylogenomic weight, independently of their global evolutionary relatedness. Different ways to compute a phylogenomic pairwise distance between genes from their conservation profiles were evaluated and compared to Pellegrini's method using the Ecocyc database of metabolic pathways as reference (Karp *et al.*, 2002). Finally, we introduce and validate an original procedure to automatically predict possible functions for anonymous genes based on the annotation of their best phylogenomic neighbors, using the MultiFun database (Serres *et al.*, 2000).

METHODS

Phylogenomic profile generation

All 4263 protein coding genes of *Escherichia coli* K-12 longer than 150 nucleotides were compared to all open reading frames (ORFs) from 71 non-redundant (only one strain per specie used) bacterial and archaeal genomes

*To whom correspondence should be addressed.

using Blastp. Let S_{ab} be the best Blastp bit score between an *E.coli* protein a and all ORFs of a bacteria b , and s_{aa} the self-score of the a protein aligned with itself. Each point of the phylogenomic profile of a protein a is computed as: $R_{ab} = S_{ab}/s_{aa}$. The profiles are as long as the number of bacterial genomes studied (here $M=71$). Note that as S_{ab} is always smaller than s_{aa} , all profile values range between zero and one. There is thus no fixed significance threshold as used in (Pellegrini *et al.*, 1999) in order to determine the presence of an orthologue to the a protein (simply, we used default Blastp parameters with bit score >50 to reduce the influence of random matches on the phylogenomic profiles). The use of the normalized Blastp scores allows each point of a profile to be weighed proportionally to the length and quality of the corresponding alignment. A second normalization procedure is then used. In order to compensate for the decreasing protein similarity (i.e. score) expected when comparing homologous genes from bacteria at increasing evolutionary distance, we normalize each column (i.e. each bacteria) by the average of the non-zero normalized scores (above the bit score threshold) obtained with this bacteria.

Optimal distance choice

Starting from phylogenomic profiles, different pairwise distances between genes can be defined: (1) an Euclidean-like distance between any two profiles interpreted as vectors, (2) another Euclidean-like distance computed from the matrix of pairwise correlation coefficients between any two proteins, (3) a distance computed from the correlation coefficient between any two profiles and (4) a similar one with the correlation coefficients computed without the means. This last distance (4) (named $d^{(4)}$) between two genes i and j is given by:

$$d_{ij}^{(4)} = 1 - |c_{ij}| \text{ with: } c_{ij} = \frac{\sum_{k=1}^M R_{ik} \times R_{jk}}{\left[\sum_{k=1}^M R_{ik}^2 \cdot \sum_{k=1}^M R_{jk}^2 \right]^{1/2}} \quad (1)$$

Note that this distance is not a distance in a strict mathematical sense, as it is not constrained to respect the triangular inequality. However, little cases of violation of this inequality were observed, so in practice, $d^{(4)}$ was found to behave as a ‘true’ distance. In order to determine the method that is most suitable for detecting co-evolution between proteins, we use the Ecocyc database as a reference of proven association between proteins and metabolic pathways (Karp *et al.*, 2002). Ecocyc consists of 202 pathways and 676 different proteins. For validation purposes, our definition of functionally linked proteins corresponds to the enzymes involved in the same metabolic pathways, that most likely co-evolved. We

thus expect the best method to find most of such pairs of genes. Our validation protocol was as follows. For each enzyme e of Ecocyc, we determined its neighbors, i.e. the proteins of *E.coli* exhibiting the smallest phylogenomic distance with e , and identified in this set the number of proteins truly involved in the same pathways as e . This procedure was repeated for neighbor sets of increasing sizes, between 1 and 100, and computed using the different distances. In order to compare the methods, we also implemented the classical one (Pellegrini *et al.*, 1999). The base-line ‘random’ results were generated as follows. The *E.coli* gene names were randomized 1000 times, and the highest number of identified Ecocyc partner was recorded for each neighborhood size. Those numbers thus estimate the false positive identification rate for a p -value of 10^{-3} .

Independent validation of the annotation procedure

Using $d^{(4)}$ as a distance, the validity of the annotation procedure was independently tested using the MultiFun database (Serres *et al.*, 2000). In MultiFun, cellular functions have been assigned to 66% of the *E.coli* gene products. In contrast with Ecocyc, non-enzymatic proteins are included, and one gene can be associated with multiple cell functions (an average of 2–3 per gene). MultiFun can be viewed as a tree with different levels, the highest one corresponding to its most detailed annotation. In our procedure, each level was treated as an independent class. For instance, a gene belonging to the classes 1.5.1 and 6.2, was considered as being a member of the classes 1, 1.5, 1.5.1, 6 and 6.2. The annotation procedure for a target gene g is then as follows: within a given neighborhood of g (associated to a total of t annotations), we counted the number of genes annotated as of class C , say n_C , C corresponding to a total of N_C *E.coli* genes. The probability to draw at least n_C genes sharing the C annotation by chance out of the total pool of the $T=6430$ annotations is given by:

$$P(C \text{ by chance}) = \sum_{k=n_C, k < t}^{N_C} \binom{N_C}{k} \cdot \binom{T - N_C}{t - k} \quad (2)$$

A unique p -value is associated to each class C by taking the lowest value of Equation (2) for the various neighborhood sizes considered (5, 10, 20, 50, 100). The classes with the lowest p -values were retained as the best candidate annotations. This procedure was evaluated through its application to the 3069 genes annotated in MultiFun.

RESULTS

$d^{(4)}$, the best performing distance relative to Ecocyc

Figure 1 presents the results for the four distances we tested and for the binary method (Pellegrini *et al.*, 1999).

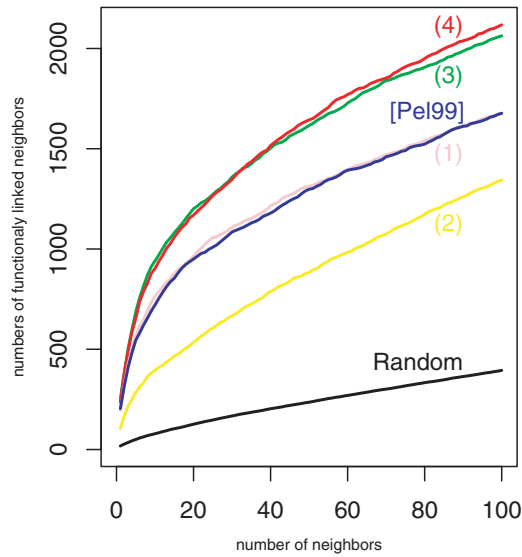


Fig. 1. Comparison between our method computed with the four different distances (1-4), that of Pellegrini *et al.* (1999) [Pel99] and the best of the 1000 random simulations. For each neighborhood size from 1 to 100, the number of correctly identified functionally related proteins (belonging to the same Ecocyc pathway) is indicated.

$d^{(4)}$, the best performing distance, allows to identify 2118 pairs of enzymes acting in the same metabolic pathway, while only 1677, 1344 and 2063 for respectively $d^{(1)}$, $d^{(2)}$ and $d^{(3)}$. Our method applied with $d^{(3)}$ and $d^{(4)}$ also outperformed the binary one, only predicting 1677 pairs of functionally linked enzymes. Using randomized datasets, we find that at most 395 gene pairs would be associated by chance at a p -value of 10^{-3} . This increases the number of genes identified as being evolutionarily related by about 25% with respect to the original method, whereas the fraction of ‘false’ positives is smaller than 20%.

Sensitivity/Specificity of the annotation procedure

Using the $d^{(4)}$ distance at a 10^{-11} p -value threshold, our method produced 3122 predictions with 50% of them appearing in the 6433 annotations of the 3069 genes of MultiFun (Fig. 2). Out of the total annotations to be retrieved, 1555 are correctly generated. At this p -value threshold, phylogenomic information alone thus appears to generate already known annotations with a sensitivity of 25%. In addition, the 200 best predictions’ (the smallest p -values) correspond to 160 already known annotations. At this very high stringency, the specificity of the method can thus reach—at least—80%. A more precise evaluation of the specificity is difficult to make given that annotated

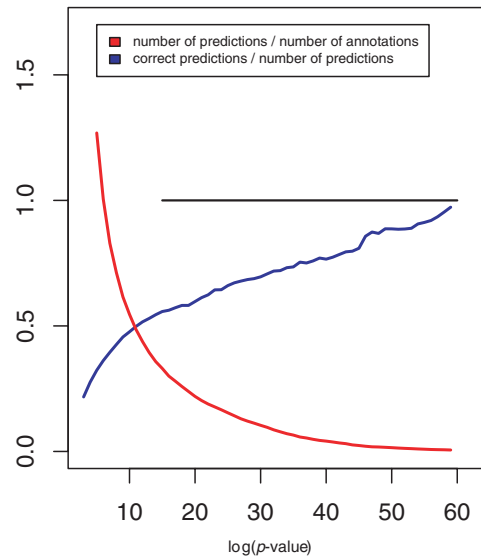


Fig. 2. Comparison of the predictions made for the MultiFun genes and their actual annotations at various p -value thresholds.

genes may have additional—yet unknown—functions. Thus, a prediction not corresponding to the MultiFun annotation should not be automatically counted as being wrong. A detailed annotation of bacterial genomes of biomedical interest is now in progress and will be published elsewhere upon completion. The annotations and associated phylogenomic profile visualization are available at <http://igs-server.cnrs-mrs.fr/phydbac/>.

REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bilu,Y. and Linia,M. (2002) Algorithms in Bioinformatics. *Second International Workshop, WABI 2002. Rome, Italy, September, 2002*. Springer, pp. 121–129.
- Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
- Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenomic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
- Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
- Zheng,Y., Roberts,R.J. and Kasif,S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.*, **3**, 121–129.