WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles

**Gabriel Moreno-Hagelsieb[1]* and Sarath Chandra Janga[2]**

[1] Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON N2L 3C5, Canada

[2] Program of Computational Genomics, CCG-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico

*ABSTRACT*

***Phylogenetic profiles (PPs) are one of the most promising methods for predicting functional relationships by genomic context. The idea behind PPs is that if the products of two genes have a functional interdependence, the genes should both be either present or absent across genomes. One of the main problems with PPs is that evolutionarily close organisms tend to share a higher number of genes resulting in the overscoring of PP-relatedness. The proper measure of the overscoring effect of evolutionary redundancy requires examples of both functionally related genes (positive gold standards) and functionally unrelated genes (negative gold standards). Since experimentally verified functional interactions are only available for a few model organisms, there is a need for an alternative to gold standards. The presence of operons (polycistronic transcription units formed of functionally related genes) in prokaryotic genomes offers such an alternative. Genes in operons are located next to each other in the same DNA strand, and thus their presence should result in a higher proportion of predicted functional interactions among adjacent genes in the same strand than among adjacent genes in opposite strands. Under the preceding principle, we present a confidence value (CV) designed for evaluating predictions of functional interactions obtained using PPs. We first show that the CV corresponds to a positive predictive value calculated using experimentally known operons and further validate operon predictions based on this CV in other organisms using available microarray data. Then, we use a fixed CV of 0.90 as a reference to compare PP predictions obtained using different nonredundant genome datasets filtered at varying thresholds of genomic similarity. Our results demonstrate that nonredundant genome datasets increase the number of high-quality predictions by an average of 20%. Confidence values as those presented here should help compare other strategies and scoring systems to use phylogenetic profiles and other genomic context methods for predicting functional interactions.***

Proteins 2008; 70:344–352.


***Key words: phylogenetic profiles; phylogenomics; mutual information; operons; gold standards.***

## INTRODUCTION

There are three main computational methods for the inference, or prediction, of functional relationships of gene products by genomic context: (a) gene fusions,[1,2] where genes are inferred to code for functionally interacting products if they are found to form a single gene in another organism; (b) conservation of gene order,[3,4] which relates to the expectation that operons, adjacent genes transcribed into a single messenger RNA,[5] might have a tendency to be conserved in evolutionarily distant organisms; and (c) phylogenetic profiles,[6–8] based on the idea that genes whose products have interdependent functions should both be either absent or present within a given genome, because either gene product alone would be useless without the other. Thus, the prediction of functional associations using PPs is mainly based on the similarity of vectors representing presences and absences of genes across a dataset of genomes (see Methods).

One of the most important problems in inferring functional interactions using PPs might be the presence of redundant genomes, because gene cooccurrence in redundant genomes might be more related to close ancestry than to functional interaction of gene products. The closer the two genomes are, the more genes they have in common (see for instance Ref. 9). Accordingly, Sun *et al.*[10] evaluated the effect of redundant genomes in the quality of predictions of functional associations using PPs. Sun *et al.*[10] focused their tests in *Escherichia coli* K12 because there are abundant examples of experimentally verified functional interactions in this model organism. As positive gold standards (genes known to

*Abbreviations:* PPs, phylogenetic profiles; GSS, genomic similarity score; MI, mutual information; CV, confidence value; WO pairs, pairs of adjacent genes in the same operon; TUB pairs, pairs of adjacent genes in the same strand and different transcription units (transcription unit boundaries).

Grant sponsor: Wilfrid Laurier University.
*Correspondence to: Gabriel Moreno-Hagelsieb, Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON N2L 3C5, Canada.
E-mail: gmoreno@wlu.ca
Received 19 February 2007; Revised 7 April 2007; Accepted 13 April 2007
Published online 1 August 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21564

be functionally related), the authors used the Database of Interacting Proteins[11] and EcoCyc[12] (these databases contain information on experimentally verified functional associations). Negative gold standards consisted of genes annotated with different functions. Gold standards for most other genomes are not available because most experimental evidence is limited to model organisms. The relationship among PPs depends on finding orthologs, homologous genes/proteins diverging after speciation events,[13,14] across various genomes. If we just transfer the experimentally confirmed functional annotations from a model organism to the putative orthologs in other organisms to obtain a dataset for evaluation, we would be limiting the standards to those having good similarities to genes in model organisms. The risk behind using such "orthology-annotations" is that apparent improvements in predictions of functional associations might be more related to the filtering out of genes with few orthologs than to functional interdependencies.

Alternatives to experimentally verified functional interactions might be found by using our knowledge of the functional organization of prokaryotic genomes, namely, the presence of operons.[5] Operons generally contain genes with related functions. Accordingly, previous works comparing genes in the same operon against genes at transcription unit boundaries (TUBs, the last gene in one transcription unit and the first gene in the next) have shown genes in operons to have a higher tendency to cooccur across genomes.[15,16] Experimentally verified operons are not available for most genomes. However, since genes in operons are found in the same DNA strand, the effect of the organization of genes into operons should be apparent as an increase in the proportion of adjacent genes in the same DNA strand against the proportion of adjacent genes in different strands as the scores for functional relatedness increase (Fig. 1).

In this work, we present a confidence value (CV, see Methods) that allows for the evaluation and comparison of predictions of functional interactions using PPs. The CV is possible due to the presence of operons in Prokaryotes. This value is a measure that only requires an annotated genome to find adjacent genes in the same DNA strand and adjacent genes in opposite strands. We use the CV to evaluate the effect of filtering out redundant genomes in predictions of functional interactions using PPs in all available prokaryotic genomes. Thus, the objectives of this work are (a) to provide a measure, independent of experimental datasets, that can be used to evaluate the prediction of functional interactions using phylogenetic profiles (the confidence value); (b) to validate such measure by comparing it with data derived from experiments in model organisms; and (c) to use the measure to test the effect of genome redundancy in the number of high-quality predictions of functional interactions across the prokaryotic genomes available at the time of this study (Fig. 1).

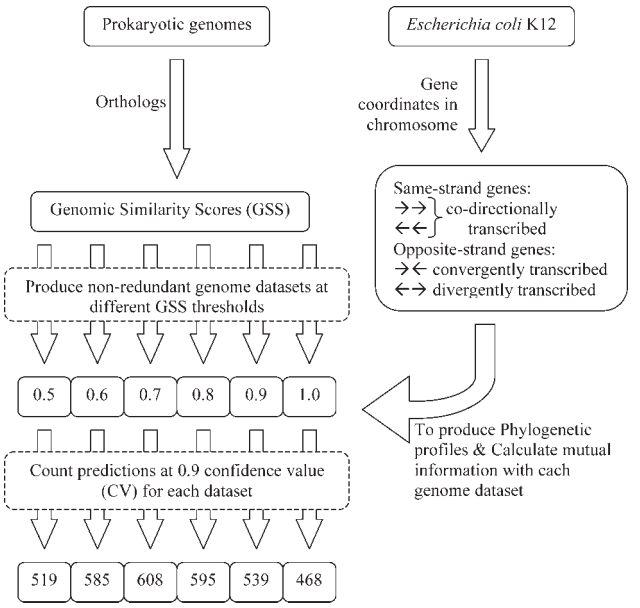


**Figure 1**

*Flow chart of this study as exemplified with the genome of* Escherichia coli *K12. Overall, we use a genomic similarity score (GSS) to filter redundancy out of the complete genome dataset. The maximum GSS is 1, meaning nearly identical genomes. Genomes were filtered at different GSS (0.5 to 1.0, where 1.0 means no filtering). The proportions of adjacent genes in the same strand and of genes in opposite strands were used to estimate a confidence value for operon predictions that translates into the relative proportion of true positives in the total number of predictions. Using nonredundant genome datasets, increases the number of predictions relative to the redundant dataset by as much as 1.30 times (608 vs. 468 predictions). In the box with same-strand and opposite-strand genes, genes are represented as arrows indicating the direction of transcription. Since genes in the same operon are functionally related, as the scores of predictions of functional association increase, the proportion of same-strand genes should increase relatively to the proportion of opposite-strand genes.*

## DATASETS AND METHODS

To build phylogenetic profiles for each gene, we used a working definition of orthology consisting of BLAST reciprocal best hits and fusions as described elsewhere.[17,18] We used the NCBI version of the program BLAST[19] to run an all-against-all comparison of all the proteins annotated within the 458 prokaryotic genomes found at the RefSeq database[20] (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) available in February 2007. The *E*-value cut-off was 1e-6 with a fixed database size (-z 5e8), soft filtering of low information content sequences (-F "m S"), and a final Smith–Waterman alignment (-s T).[21] We also required coverage of at least 50% of any of the protein sequences in the alignments.

To evaluate the similarity of phylogenetic profiles (PPs) we used mutual information (MI). The PP of a gene consists of a vector where each item represents either the presence or the absence of an ortholog to the gene in a given genome.[8] In more elaborated vectors, the

items can be numbers related to the score of the alignment of the proteins encoded by the gene and its ortholog (see for instance Ref. 22). Here, we use binary (1 and 0) profiles because we did not find a significant difference in the overall results using more complex vectors. Since the most common measure of relatedness of the PPs of two genes is mutual information (MI),[22,23] we used MI, measured in bits, to compare PPs. The MI of two vectors $I$ and $J$ is defined as:

$$\sum_{i=[0,1];j=[0,1]} P_{ij} \times \log_2 \frac{P_{ij}}{P_i \times P_j}$$

where $P_{ij}$ is the proportion of a pair $ij$ in the alignment of vectors $I$ and $J$, $P_i$ is the proportion of the value $i$ in vector $I$, and $P_j$ the proportion of $j$ in vector $J$.

To compare the similarity of genomes in the dataset, we calculated a genomic similarity score (GSS)[17,18] based on the BLAST bit-scores resulting from the comparison of all the proteins in one organism with their orthologs in another organism. The bit-score increases with the sequence similarity reaching a maximum when the two proteins compared are identical. Thus, the GSS is designed to reach a maximum of one if the proteins in one organism are identical to their corresponding orthologs in another organism[17,18]:

$$\mathrm{GSS} = \frac{\sum \mathrm{BBS_{comp}}}{\sum \mathrm{BBS_{self}}}$$

where $\mathrm{BBS_{comp}}$ is the BLAST bit-score of the alignment of a given protein with its ortholog and $\mathrm{BBS_{self}}$ is the BLAST bit-score of the alignment of the protein against itself. Because of the variable nature of alignment scoring matrices (here BLOSUM62), the GSS can vary slightly depending on which of the two genomes is used as the basis to obtain the self-scores. Thus, the final GSS is the average obtained from the perspective of both analyzed genomes. Genomes were considered redundant if their GSS was higher than a given threshold (see below). For any set of redundant genomes, we kept the one that had the highest number of orthologs in taxonomic families other than its own. We used GSS because it is a genome-based value. However, any other similarity/distance measure can be used to filter out redundant genomes.

To evaluate the confidence value as a guide to the quality of operon predictions we used the current dataset of experimentally verified transcription units of *Escherichia coli* K12 found in RegulonDB,[24] and a dataset of experimentally verified transcription units of *Bacillus subtilis* (http://odb.kuicr.kyoto-u.ac.jp/), to build datasets of experimentally verified pairs of genes in the same operon (WO pairs), and of genes at transcription unit boundaries (TUB pairs, being the last gene in one transcription unit and the first in the next) as explained previously.[25]

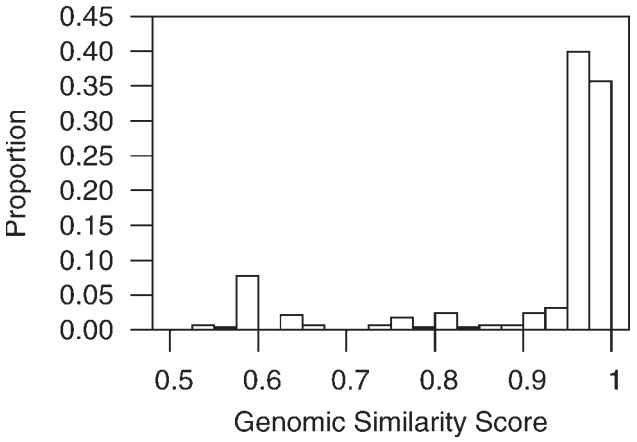


**Figure 2**

*Within-species genomic similarity score (GSS). The GSS among strains of the same species vary widely. However, most of same-species genomes have GSS above 0.90.*

## STRAINS ANNOTATED WITHIN THE SAME SPECIES VARY WIDELY IN THEIR EVOLUTIONARY RELATIONSHIP

To study the effect of redundant genomes on predictions based on the MI of PPs, we need to define genome redundancy. As a measure of evolutionary proximity, we used genomic similarity scores (GSS, see Methods). Another criterion to detect redundant genomes might be species names. However, besides the "prokaryotic species" concept is in debate,[26,27] the within-species GSS varies widely (Fig. 2). The current genome dataset contains 56 named species with two or more sequenced strains for a total of 173 genomes (Table I). The GSS between most same-species strains is above 0.90. However, *Buchnera aphidicola* (four strains) and *Prochlorococcus marinus* (nine strains) have average within-species GSS of 0.59 and 0.66, respectively. These GSS are lower than those among genomes in different taxonomic families. In other cases where the species are better defined, strains with different species names qualify as members of another species by their GSS. For a classic example, the GSS for all *E. coli* strains averages 0.97. The strains of all species of the *Shigella* genus (*S. boydii*, *S. flexneri*, and *S. sonnei*) have GSS $\geq$0.97 against most *E. coli* strains. Thus, the species names are not appropriate as filters of redundancy. A recent publication that compares genome dinucleotide signatures as a phylogenetic marker against classic phylogenetic measures, such as 16S ribosomal DNA identities, confirms this incongruence of phylogenetic distances among strains with the same species name.[28]

The result of filtering redundancy should be such that the genomes kept belong to organisms whose phyloge-


DOI 10.1002/prot

**Table I**
*Species with More Than One Strain in the RefSeq[20] Genomes Database (as of February 2007)*

| Species | Strains | Average GSS | Species | Strains | Average GSS |
|---|---|---|---|---|---|
| *Agrobacterium tumefaciens* | 2 | 0.98 | *Listeria monocytogenes* | 2 | 0.97 |
| *Bacillus anthracis* | 3 | 1.00 | *Mycobacterium avium* | 2 | 0.97 |
| *Bacillus cereus* | 3 | 0.93 | *Mycobacterium bovis* | 2 | 1.00 |
| *Bacillus licheniformis* | 2 | 0.99 | *Mycobacterium tuberculosis* | 2 | 0.99 |
| *Bacillus thuringiensis* | 2 | 0.97 | *Mycoplasma hyopneumoniae* | 3 | 0.97 |
| *Bacteroides fragilis* | 2 | 0.97 | *Neisseria meningitidis* | 3 | 0.96 |
| *Brucella melitensis* | 2 | 0.97 | *Prochlorococcus marinus* | 9 | 0.66 |
| *Buchnera aphidicola* | 4 | 0.59 | *Pseudomonas aeruginosa* | 2 | 0.99 |
| *Burkholderia cenocepacia* | 2 | 0.99 | *Pseudomonas fluorescens* | 2 | 0.82 |
| *Burkholderia mallei* | 3 | 0.98 | *Pseudomonas syringae* | 3 | 0.91 |
| *Burkholderia pseudomallei* | 2 | 0.97 | *Ralstonia eutropha* | 2 | 0.82 |
| *Campylobacter jejuni* | 3 | 0.97 | *Rhodopseudomonas palustris* | 5 | 0.79 |
| *Carsonella ruddii* | 2 | 1.00 | *Salmonella enterica* | 4 | 0.98 |
| *Chlamydia trachomatis* | 2 | 0.99 | *Shigella flexneri* | 3 | 0.99 |
| *Chlamydophila pneumoniae* | 4 | 1.00 | *Staphylococcus aureus* | 9 | 0.98 |
| *Clostridium perfringens* | 3 | 0.96 | *Staphylococcus epidermidis* | 2 | 0.99 |
| *Corynebacterium glutamicum* | 2 | 0.99 | *Streptococcus agalactiae* | 3 | 0.98 |
| *Desulfovibrio vulgaris* | 2 | 0.98 | *Streptococcus pneumoniae* | 3 | 0.98 |
| *Ehrlichia ruminantium* | 3 | 0.98 | *Streptococcus pyogenes* | 11 | 0.97 |
| *Escherichia coli* | 8 | 0.97 | *Streptococcus thermophilus* | 3 | 0.98 |
| *Francisella tularensis* | 5 | 0.97 | *Synechococcus elongatus* | 2 | 0.99 |
| *Haemophilus influenzae* | 2 | 0.97 | *Thermus thermophilus* | 2 | 0.96 |
| *Helicobacter pylori* | 3 | 0.94 | *Tropheryma whipplei* | 2 | 0.98 |
| *Lactobacillus delbrueckii* | 2 | 0.98 | *Vibrio vulnificus* | 2 | 0.97 |
| *Lactococcus lactis* | 3 | 0.92 | *Xanthomonas campestris* | 3 | 0.91 |
| *Legionella pneumophila* | 3 | 0.96 | *Xanthomonas oryzae* | 2 | 0.97 |
| *Leptospira borgpetersenii* | 2 | 1.00 | *Xylella fastidiosa* | 2 | 0.95 |
| *Leptospira interrogans* | 2 | 0.98 | *Yersinia pestis* | 5 | 0.99 |

netic distances (represented here by the GSS) best avoid the overscoring effect due to evolutionary relationships, while increasing the functional-dependency signal. Since most same-species genomes have GSS above 0.90 (Fig. 2) our highest threshold was 0.90 GSS. Our lowest threshold, corresponding to a more stringent filter of redundancy, was 0.50 GSS, where all strains with the same species name would be represented by a single genome.

## OPERONS PROVIDE CONFIDENCE VALUES USEFUL FOR COMPARING PREDICTIONS USING PHYLOGENETIC PROFILES

To evaluate predictions derived from phylogenetic profiles (PPs), we devised a confidence value (CV) based on the following premises: (a) as the mutual information (MI, see Methods) increases, the similarity of PPs too increases; (b) predictions at increasingly higher PP-similarity (measured as MI) correspond to an increasing proportion of functionally related genes; (c) adjacent genes in the same-strand at high MI should mainly correspond of genes in the same operons; (d) genes in opposite-strands, divergently-transcribed and convergently-transcribed, might represent unrelated genes. Not all divergently-transcribed genes are functionally-unrelated (see for instance Ref. 29). However, the proportion of functionally related genes among divergently-transcribed genes is small compared to the same proportion among codirectionally transcribed genes.[18,29–31] Thus, we can calculate a CV that, given a MI threshold, compares the proportion of genes in opposite DNA strands against the proportion of genes in the same strand to evaluate operon predictions based on phylogenetic profiles:

$$\mathrm{CV} = 1 - 0.5\frac{P_{\mathrm{OS}}}{P_{\mathrm{SS}}}$$

This equation is similar to that previously proposed to evaluate conservation of gene order.[18,30] $P_{\mathrm{SS}}$ is the proportion of adjacent pairs of genes in the same strand, and $P_{\mathrm{OS}}$ is the proportion of adjacent genes in opposite strands. The number 0.5 is a prior probability for the genes to be in different transcription units (or to be functionally unrelated). For instance, if an organism contains 2000 total opposite-strand gene pairs and 3000 total same-strand gene pairs, while 200 opposite-strand gene pairs and 1000 same-strand gene pairs have a MI $\geq$ 0.4 bits, the CV would be $1 - 0.5 \times [(200/2000)/(1000/3000)] = 0.85$.

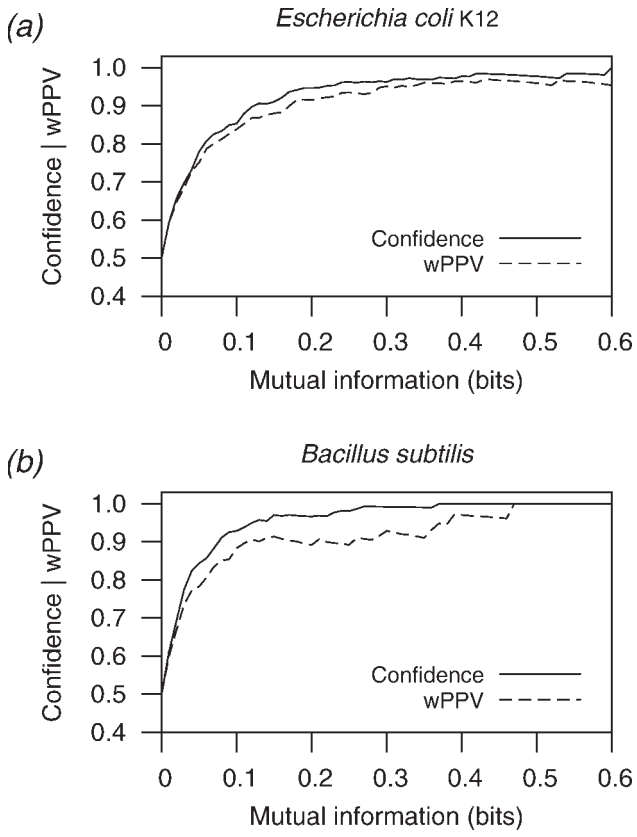


**Figure 3**

*Confidence values (CV) and weighted positive predictive values (wPPV). Confidence values calculated from the proportions of same-strand genes and opposite-strand genes are very similar to the positive predictive values calculated with known operons and known transcription unit boundaries (nonoperons) in two model organisms. The confidence value should help obtain high-quality predictions of operons across genomes.*

Positive predictive values (PPV, true positives divided by the sum of false positives and true positives) are one of the most common statistical measures of prediction quality as evaluated against known positive and negative sample datasets. The confidence value is not a PPV (it is not derived from known datasets), but it is intended to represent one. Accordingly, the CV is somewhat equivalent to a weighted positive predictive value for operon predictions (wPPV, Fig. 3). The wPPV, given a MI threshold, is calculated as:

$$\mathrm{wPPV} = \frac{P_{WO}}{P_{WO} + P_{TUB}}$$

where $P_{WO}$ is the proportion of WO pairs, and $P_{TUB}$ is the proportion of same-strand TUB pairs. This PPV is weighted in the sense that it is equivalent to a PPV where the number of known positives and the number of negatives are the same.

## CONFIDENCE VALUES REPRESENT POSITIVE PREDICTIVE VALUES FOR OPERON PREDICTIONS

To further evaluate the CV as a point of reference to select high-quality operon predictions, we obtained the data on correlation of expression of adjacent genes for six genomes previously presented by Price *et al.*[32] We found that pairs of genes predicted to be in the same operon at a CV of 0.90 have a tendency toward higher correlation values than pairs of genes in opposite strands (Fig. 4). The two smaller genomes, *Chlamydia trachomatis* (1 Mbp) and *Helicobacter pylori* J99 (1.6 Mbp), show more noise as a consequence of the low number of genes, and thus of predictions (59 and 76 predicted pairs, respectively). It is also important to note that in these two genomes the opposite-strand gene pairs also have a tendency toward high correlation of expression. Perhaps, the reduced genomes of parasites are left with a higher proportion of functionally related genes, or perhaps most of the genes are expressed at the same time as a consequence of their limited environment and scarcity of genes coding for transcription factors.[33]

Previously, operon predictions in the cyanobacterium *Synechocystis* PCC6803 have been problematic due to its unusual high spacing between genes and some annotation issues.[17] Our PP-based operon predictions show obvious higher correlations of expression in this organism than those of the predictions based on intergenic distance. The correlations also seem to surpass those of the operons predicted by Price *et al.*[32] (Fig. 4 within the reference). Since our PP-based predictions have better correlation of expression, they might represent samples that should help better study the properties necessary for overall operon predictions in this and other organisms that do not follow the intergenic distance tendencies prevailing in most other Prokaryotes.

The CV allows for the selection of high-quality operon predictions and provides a point of comparison for different versions of PPs. However, an operon-based CV cannot measure the quality of overall predictions of functional interactions using PPs. We are assuming that the proportions of genes in operons and of genes at different transcription units are equal, which is quite reasonable.[30,34,35] To properly calculate a CV for overall predictions, we would need to know the proportion (the prior probability) of functionally interacting genes among all possible pairs of genes and adjust the CV accordingly. For instance, there are 4242 annotated protein-coding genes in the genome of *E. coli* K12 (GenBank version: NC_000913.2). This number solves to $(4{,}242 - 1) \times (4{,}242/2) = 8{,}995{,}161$ total possible pairs. To calculate a CV we would need to know how many of these ~9 million possible pairs functionally interact. We will explore


DOI 10.1002/prot

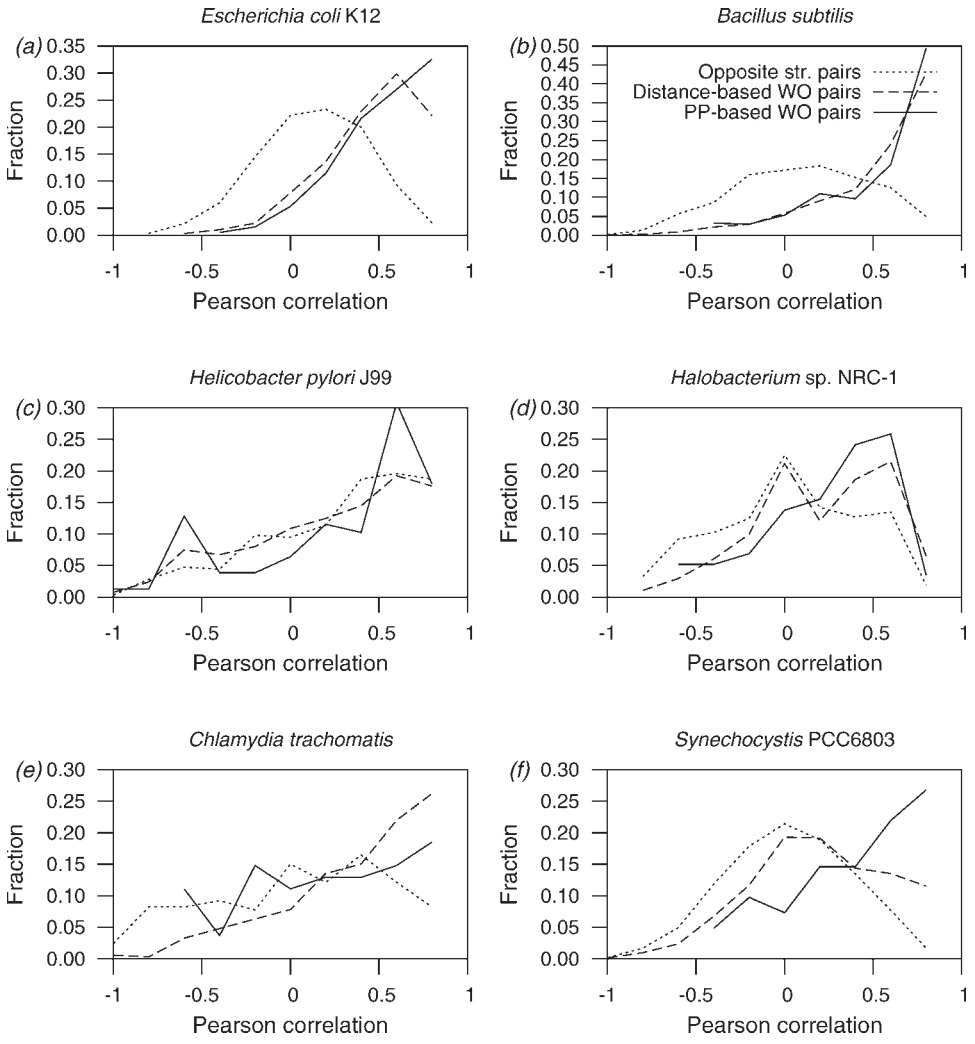


**Figure 4**

*Correlation of expression of predicted operons. The figure presents the Pearson correlation of expression in microarrays as calculated by Price et al.[32] Pairs of genes predicted to be in operons using a confidence value of 0.90 tend to have higher correlations than opposite strand genes. Predictions of operons by intergenic distances were performed as published previously.[17]*

approaches to solve this problem in another work (Moreno-Hagelsieb and Janga, in preparation).

## NONREDUNDANT GENOMES INCREASE THE NUMBER OF HIGH QUALITY PREDICTIONS USING PHYLOGENETIC PROFILES

To test the effect of filtering out redundant genomes, we predicted functional interactions at a fixed CV of 0.90. A fixed CV ensures the same quality for all predictions (here operon predictions), and improvements should be measurable as increases in the number of predictions obtained. We tested nonredundant genome datasets obtained at different thresholds of GSS, from 0.50 to 0.90, to build phylogenetic profiles. We required at least 50 predictions to proceed with further analyses. To also avoid redundancy in the presentation of results, we used a genome dataset filtered at 0.90 GSS to show the effects of genome redundancy in PP-based operon predictions.

Nonredundant genome datasets result in an increase in the number of operon predictions at a fixed CV (Table II). The number of predictions seems to increase as the GSS decreases, with the datasets filtered at a GSS of 0.50 showing the maximum average improvement (1.27 times as many predictions as those obtained using a redundant

**Table II**
*Increase in Number of Predictions Using Nonredundant Genome Datasets*

| | All genomes | | Genomes ≥ 2.5 Mbp | |
|---|---|---|---|---|
| GSS | Average increase | Proportion | Average increase | Proportion |
| 50 | 1.27 | 0.72 | 1.26 | 0.72 |
| 60 | 1.25 | 0.83 | 1.24 | 0.84 |
| 70 | 1.23 | 0.84 | 1.23 | 0.86 |
| 80 | 1.20 | 0.83 | 1.20 | 0.87 |
| 90 | 1.18 | 0.83 | 1.18 | 0.88 |

The genomic similarity score (GSS) of 0.70 produces the best combination of improvement and a high proportion of genomes increasing their number of predictions.

dataset). However, the same dataset is the one showing the smallest proportion of genomes showing improvements in number of predictions (Table II).

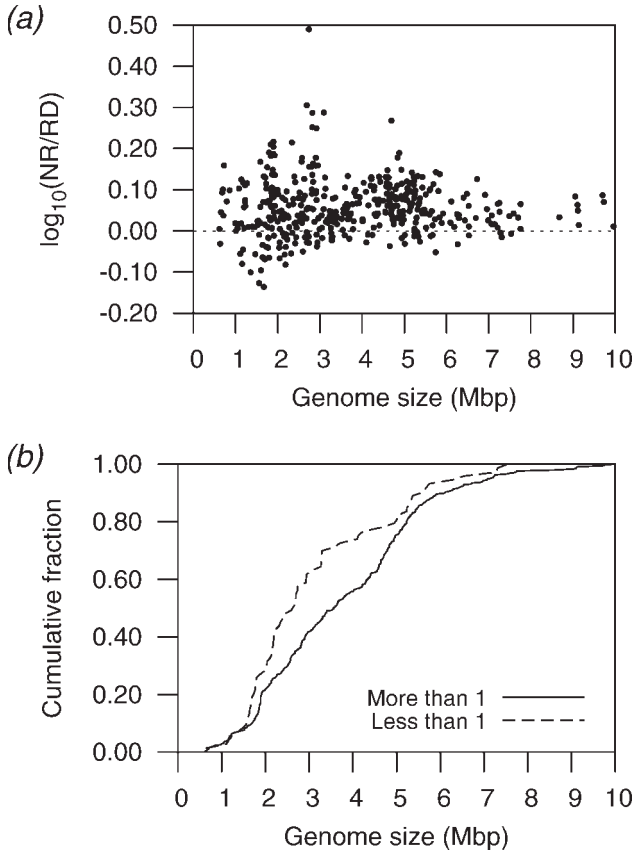


**Figure 5**
*Prediction improvement using nonredundant genome datasets. All predictions were produced at a confidence value of 0.90. (**a**) The improvement is shown as the logarithm of the ratio of the number of predictions obtained using a nonredundant genome dataset versus the number of predictions using a complete [redundant] genome dataset. A log-ratio above 0 (dotted line) indicates an increase in the number of predictions. (b) The cumulative frequencies of genomes showing increases in the number of predictions (ratio above 1), and of genomes showing decrements in number of predictions (ratio below 1), show that genomes where the nonredundant dataset did not improve the number of predictions tend to be smaller.*

In *E. coli* K12, the maximum increase in number of predictions occurs with a genome dataset filtered at 0.70 GSS, with an increase from 468 pairs of genes in operons using the complete genome dataset, to 608 (1.30 as many predictions) using a nonredundant genome dataset filtered at 0.70 GSS. The CV of 0.90 is attained at 0.23 bits of MI using the complete genome dataset, and at 0.12 bits using the nonredundant genome dataset. The higher MI necessary to attain the same CV evidences the overscoring effect of redundancy.

**Table III**
*Organisms Always Showing Decrements in Number of Predictions Using Nonredundant Genome Datasets*

| Genome | Size (Mbp) | GSS to next redundant group | Overannotation |
|---|---|---|---|
| *Aeropyrum pernix* K1 | 1.67 | 0.23 | 24.00 |
| *Agrobacterium tumefaciens* C58 | 5.67 | 0.46 | 20.10 |
| *Buchnera aphidicola* Sg | 0.64 | 0.61 | 0.37 |
| *Burkholderia mallei* NCTC 10229 | 5.74 | 0.44 | 31.41 |
| *Burkholderia mallei* SAVP1 | 5.23 | 0.43 | 33.81 |
| *Burkholderia pseudomallei* 1710b | 7.31 | 0.44 | 15.44 |
| *Campylobacter jejuni jejuni* NCTC 11168 | 1.64 | 0.31 | 14.88 |
| *Campylobacter jejuni jejuni* 81–176 | 1.70 | 0.31 | 20.58 |
| *Campylobacter jejuni* RM1221 | 1.78 | 0.31 | 25.03 |
| *Dehalococcoides ethenogenes* 195 | 1.47 | 0.33 | 33.56 |
| *Deinococcus radiodurans* R1 | 3.28 | 0.31 | 23.53 |
| *Helicobacter acinonychis* Sheeba | 1.56 | 0.31 | 23.61 |
| *Helicobacter hepaticus* ATCC 51449 | 1.80 | 0.30 | 24.01 |
| *Helicobacter pylori* 26695 | 1.67 | 0.31 | 15.37 |
| *Mannheimia succiniciproducens* MBEL55E | 2.31 | 0.55 | 19.78 |
| *Neisseria gonorrhoeae* FA 1090 | 2.15 | 0.45 | 34.09 |
| *Nitrobacter hamburgensis* X14 | 5.01 | 0.39 | 33.40 |
| *Neisseria meningitidis* FAM18 | 2.19 | 0.44 | 22.02 |
| *Neisseria meningitidis* Z2491 | 2.18 | 0.44 | 30.26 |
| *Pseudomonas aeruginosa* PAO1 | 6.26 | 0.45 | 14.83 |
| *Pelobacter propionicus* DSM 2379 | 4.24 | 0.34 | 22.12 |
| *Synechococcus elongatus* PCC 7942 | 2.74 | 0.32 | 23.76 |
| *Streptococcus pneumoniae* D39 | 2.05 | 0.66 | 23.80 |
| *Streptococcus pneumoniae* R6 | 2.04 | 0.65 | 27.29 |
| *Synechococcus* JA-3-3Ab | 2.93 | 0.32 | 28.49 |
| *Synechococcus* WH 8102 | 2.43 | 0.32 | 33.14 |
| *Thiomicrospira denitrificans* ATCC 33889 | 2.20 | 0.30 | 17.41 |
| *Xylella fastidiosa* 9a5c | 2.73 | 0.40 | 55.86 |

It is not easy to understand why these organisms fail to improve their number of predictions when genome redundancy decreases. The main reasons seem to be small genome size (see also Fig. 5), and low similarity to any groups of redundant genomes. Overannotations (the annotation of genes that do not exist) can add to the problem. The GSS reported is the genomic similarity score to the closest group of redundant genomes other than its own. We calculated overannotation by the "SwissProt method" published by Skovgaard *et al.*[36] Overannotations tend to be smaller than those published by Skovgaard *et al.*,[36] due to the cleanup of annotations performed to produce the RefSeq database.[20] Gray cells indicate values that might help understand the lack of improvement in these genomes.


DOI 10.1002/prot

Organisms with decreased numbers of predictions tend to have small genomes (Fig. 5). Seventeen of the 28 genomes displaying the highest number of predictions using the redundant genome dataset have genome sizes below 2.5 Mbp, while seven have genome sizes above 4.0 Mbp (Table III). Size might affect these results because the numbers of adjacent genes in the same-strand and in opposite strands might not be enough to properly calculate a CV. Two more characteristics seem to affect these results: the GSS to the closest group of redundant genomes and genome overannotation (the annotation of genes that do not exist). If a genome is very distant to the redundant groups of genomes, the probability that the PP will be overscored is lower because fewer genes will find orthologs within such redundant groups. In this regard, the GSS to redundant groups is generally low for most of the genomes having a maximum number of predictions with the complete genome dataset (Table III). Overannotation might result in nonexistent same-strand and opposite-strand genes, deviating the results. We measured overannotation by the SwissProt method suggested by Skovgaard *et al.*[36] and found that a few of the genomes with highest numbers of predictions with redundant genomes have overannotation above the average of 20% (Table III). Overall, the reasons for failure to increase the number of predictions in some genomes are hard to evaluate. These might range from genome size (insufficient samples for proper calculation of a CV, see also Fig. 5), to the presence of genes in groups of redundant genomes. Some groups of genes might work together only within an evolutionarily limited clade. Such groups might not be detected by generic phylogenetic profiles. Designing other strategies might be needed to further improve predictions and deal with limited evolutionary scopes.

These results show that the elimination of redundant genomes reduces overscoring and increases the signal-to-noise ratio in most genomes. Thus, our simple method to obtain a nonredundant genome dataset has a positive impact in the predictive quality of PPs in most Prokaryotes.

## CONCLUDING REMARKS

Jansen and Gerstein[37] have insisted on the importance of using positive and negative gold standards to train and evaluate large-scale experimentally-determined functional interactions, as well as computational predictions of functional associations. Our results here suggest that our knowledge of genome organization can help obviate such standards in most Prokaryotes, and that the use of appropriate contrasting datasets can help better understand the effects of different parameters, such as redundancy in the genome dataset, on predictions of functional associations.

Nonredundant genome datasets and operon predictions using PPs can be found at: http://popolvuh.wlu.ca/Phyl_Profiles/.

## ACKNOWLEDGMENTS

GMH acknowledges Gary Molenkamp for computer assistance and the Shared Hierarchical Academic Research Computing Network (SHARCNET: http://www.sharcnet.ca/) for computational facilities. We thank Warren F. Lamboy for critical reading of the manuscript. SCJ acknowledges support from Julio Collado-Vides.

## REFERENCES

1. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
2. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
3. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998;23:324–328.
4. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999;96:2896–2901.
5. Jacob F, Perrin D, Sanchez C, Monod J, Edelstein S. [ The operon: a group of genes with expression coordinated by an operator. C R Acad Sci Paris 1960;250:1727–1729]. Comptes rendus biologies 2005;328:514–520.
6. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997;278:631–637.
7. Gaasterland T, Ragan MA. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 1998;3:199–217.
8. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.
9. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet 1999;21:108–110.
10. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, Li Y. Refined phylogenetic profiles method for predicting protein–protein interactions. Bioinformatics 2005;21:3409–3415.
11. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 2004;32:D449–D451. Database Issue.
12. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 2005;33:D334–D337. Database Issue.
13. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool 1970;19:99–113.
14. Fitch WM. Homology a personal view on some of the problems. Trends Genet 2000;16:227–231.
15. Moreno-Hagelsieb G, Trevino V, Perez-Rueda E, Smith TF, Collado-Vides J. Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. Trends Genet 2001;17:175–177.
16. Moreno-Hagelsieb G, Collado-Vides J. Operon conservation from the point of view of *Escherichia coli*, and inference of functional

interdependence of gene products from genome context. In Silico Biol 2002;2:87–95.
17. Moreno-Hagelsieb G, Collado-Vides J. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 2002;18 (Suppl. 1):S329–S336.
18. Janga SC, Moreno-Hagelsieb G. Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res 2004;32:5392–5397.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
20. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005;33:D501–D504. Database Issue.
21. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
22. Date SV, Marcotte EM. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol 2003;21:1055–1062.
23. Huynen M, Snel B, Lathe W, 3rd, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 2000;10:1204–1210.
24. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 2006;34:D394–D397. Database Issue.
25. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci USA 2000;97:6652–6657.
26. Cohan FM. What are bacterial species? Annu Rev Microbiol 2002; 56:457–487.
27. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J. Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol 2005;3:733–739.
28. van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T. The reach of the genome signature in prokaryotes. BMC Evol Biol 2006; 6:84.
29. Korbel JO, Jensen LJ, von Mering C, Bork P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 2004;22:911–917.
30. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. Nucleic Acids Res 2001;29:1216–1221.
31. Rogozin IB, Makarova KS, Wolf YI, Koonin EV. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Brief Bioinf 2004;5:131–149.
32. Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 2005;33:880–892.
33. Ranea JA, Buchan DW, Thornton JM, Orengo CA. Evolution of protein superfamilies and bacterial genome size. J Mol Biol 2004; 336:871–887.
34. Cherry JL. Genome size and operon content. J Theor Biol 2003; 221:401–410.
35. Moreno-Hagelsieb G. Operons across prokaryotes: genomic analyses and predictions 300+ genomes later. Curr Genom 2006;7:163–170.
36. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A. On the total number of genes and their length distribution in complete microbial genomes. Trends Genet 2001;17:425–428.
37. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 2004;7:535–545.

 DOI 10.1002/prot