A powerful non-homology method for the prediction of operons in prokaryotes

Gabriel Moreno-Hagelsieb* and Julio Collado-Vides

Program of Computational Genomics, CIFN, UNAM, A.P. 565-A, Cuernavaca, Morelos 62100, Mexico

Received on January 20, 2002; revised and accepted on March 29, 2002

ABSTRACT
Motivation: The prediction of the transcription unit organization of genomes is an important clue in the inference of functional relationships of genes, the interpretation and evaluation of transcriptome experiments, and the overall inference of the regulatory networks governing the expression of genes in response to the environment. Though several methods have been devised to predict operons, most need a high characterization of the genome analysed. Log-likelihoods derived from inter-genic distance distributions work surprisingly well to predict operons in Escherichia coli and are available for any genome as soon as the gene sets are predicted.

Results: Here we provide evidence that the very same method is applicable to any prokaryotic genome. First, the method has the same efficiency when evaluated using a collection of experimentally known operons of Bacillus subtilis. Second, operons among most if not all prokaryotes seem to have the same tendencies to keep short distances between their genes, the most frequent distances being the overlaps of four and one base pairs. The universality of this structural feature allows us to predict the organization of transcription units in all prokaryotes. Third, predicted operons contain a higher proportion of genes with related phylogenetic profiles and conservation of adjacency than predicted borders of transcription units.

Supplementary information: Additional materials and graphs, are available at: http://www.cifn.unam.mx/moreno/pub/TUpredictions/.

Contact: moreno@cifn.unam.mx

Keywords: functional genomics; comparative genomics; operon prediction; operon structure.

INTRODUCTION
Three main lines of evidence to detect functional relationships among encoded proteins within genomes by pure computational analyses have been suggested. (i) If two proteins have a related function (e.g., they take part in consecutive steps in a pathway), their encoding genes should co-occur. This would show up as a concerted appearance and disappearance of both genes among genomes in what is called a similar ‘phyletic pattern’ (Gaasterland and Ragan, 1998; Tatusov et al., 1997), or ‘phylogenetic profile’ (Pellegrini et al., 1999). (ii) Stronger evidence is the appearance of genes in conserved clusters among genomes (Dandekar et al., 1998; Overbeek et al., 1999). This assumption derives from the high tendency of operons, understood as a stretch of adjacent genes transcribed into a single messenger RNA, to contain genes with related functions. (iii) The strongest evidence of functional relationship is the finding of separate genes in a given genome that appear fused into a single gene in another genome (Enright et al., 1999; Marcotte et al., 1999). Recently, we have confirmed that genes found within experimentally characterized operons in Escherichia coli have a greater tendency to co-occur, preserve their adjacency, and appear fused, than genes found at the boundaries of transcription units (Moreno-Hagelsieb et al., 2001). The authors of a recent review have emphasized the importance of operons as a clue to predicting functional associations and to expand, via the study of their reorganization, the number of associated genes, and that operon prediction is a difficult and error-prone procedure that has never been defined in algorithmic terms (Galperin and Koonin, 2000). At the same time, we published the analysis of a collection of operons in Escherichia coli, and a transcription unit (TU) prediction method derived therein (Salgado et al., 2000). The method is based on the distances separating adjacent genes and on their functional annotations, discriminating genes within operons from those in different TUs with an accuracy of above 87% (above 82% if based only on inter-genic distances). Other procedures to predict the association of genes into operons have been proposed. One of them is based on the prediction of every aspect of a genome, from the genes to the different signals, such as promoters and terminators of transcription (Yada et al., 1999), requiring a high characterization of a genome before its application. Another method, using a probabilistic learning approach, relies on as much information as there
is available, like distances between genes, transcriptome experiments, and promoters. The procedure is also applicable mainly to highly characterized organisms (Craven et al., 2000). Finally, another group devised a tool based on the assumption that conservation of adjacency of genes in the same strand is related to the association of genes into operons (Ermolaeva et al., 2001). Using the conservation of adjacent genes in opposite strands as reference, they derive sets of genes with high probabilities of being found in operons. This has been the only method applicable to any prokaryote, but is limited to genes having orthologs in other genomes and is not aimed at predicting the complete TU organization of genomes.

Here we first show that our distance-based method is equally efficient at discriminating genes within operons in a data set derived from experimentally determined operons of Bacillus subtilis. Then we show that operons have the same tendencies to have short distances between their genes in any prokaryote. Finally, we give evidence that distance-based predicted operons consist of functionally related genes.

DATA PREPARATION

To find orthologs, homologous genes that have diverged following speciation (Fitch, 2000), we ran all versus all comparisons of the protein-coding sequences (CDS) annotated within the genomes analysed (see below) using gapped BLASTP (Altschul et al., 1997) with an expected frequency cutoff of 0.001. Our putative orthologs were bidirectional best hits, as long as the alignments covered at least 50% of one of the sequences. Fusions were protein sequences covering a different part of the target protein in the alignment than the bidirectional best hit.

We used the 70 currently available prokaryotic genomes found at the Entrez Genome Database (Tatusova et al., 1999) (ftp://ncbi.nlm.nih.gov/genomes/Bacteria/). Since this number increases rapidly, we automated a filter to avoid repetitive genomes. The first step is to obtain the set of genes of a base genome having orthologs in another genome (e.g., E. coli K12 versus Salmonella typhimurium LT2). We then calculate the sum of the BLAST raw scores of this set of genes compared with themselves (e.g., E. coli K12 versus E. coli K12) (auto-score), and the sum of the BLAST raw scores of the same set against their orthologs in the other genome (comparison-score). We eliminate the other genome if the ratio of comparison-score and auto-score exceeds 0.7 (it is 0.87 in the example, eliminating S. typhimurium). We proceed in alphabetical order. The method left a total of 50 non-redundant genomes (see web site for a complete list with accession numbers).

OUTLINE OF THE METHOD TO PREDICT TUS

The method used to predict the TU organization of E. coli is performed in two stages (Salgado et al., 2000). We first build a database of ‘directories,’ stretches of adjacent genes in the same DNA strand with no intervening gene in the complementary strand. The procedure does not break any known operon in E. coli. Thus we assume that all directories containing a single gene correspond to TUs that require no further analysis. Then we evaluate each pair of adjacent genes within all other directories (WD pairs) to discriminate pairs of genes within operons (WO pairs) from those at TU boundaries (TUB pairs, which are the last gene in a TU and the first gene in the next one). The evaluation is based on log-likelihoods derived from the comparison, in terms of inter-genic distances and functional annotations, of experimentally known WO pairs (currently 641 pairs) versus known TUB pairs (434 pairs) of E. coli. We calculate the log-likelihood of two adjacent genes to be in an operon, at intervals of 10 base pairs (bp), as the logarithm of the fraction of WO pairs divided by the fraction of TUB pairs containing genes separated by a distance within the interval. No formal smoothing is applied. The log-likelihood of a distance interval with insufficient data is calculated as the average of the nearest neighbouring points. The functional annotations that complement the method are those developed by Monica Riley for E. coli (Riley and Labedan, 1996), currently available for only half of the genes in this organism, making log-likelihoods derived from them of limited use. Nevertheless, distance log-likelihoods account for most of the discrimination power of the method reaching an accuracy of more than 82%.

DISTANCE LOG-LIKELIHOODS ACCURATELY DISCRIMINATE WO PAIRS FROM TUB PAIRS IN BACILLUS SUBTILIS

The only list of experimentally characterized TUs we know about, besides the ones of E. coli (Salgado et al., 2001), is that of 100 operons of Bacillus subtilis (Itoh et al., 1999). We used it along with the corresponding genome sequence and annotations (Kunst et al., 1997) to build data sets of WO pairs (310 pairs) and of TUB pairs (123 pairs). We applied the method of TU prediction based on distance log-likelihoods, derived solely from the data sets of E. coli, to the B. subtilis data sets as previously described (Salgado et al., 2000). Figure 1 displays the performance of the method to discriminate WO pairs from TUB pairs in B. subtilis and E. coli. The curves of sensitivity (true positives detected per known WO pair) and specificity (true negatives per TUB pair) are almost identical in both organisms demonstrating that the method works equally well in both organisms despite their
Evolutionary distance. We also plot an estimated accuracy value. The common definition of accuracy is the sum of true positives and true negatives detected, divided by the sum of the total positives (total known WO pairs in our case) and total negatives (total known TUB pairs) in the samples used for evaluation. This measure is intended to provide an estimate of the right predictions in a given population. If the proportion of positives and negatives in the samples does not match that in the population the ‘sample’ accuracy would be misleading. Thus, whenever possible, the accuracy should be estimated taking into account such proportions in the population. Since we have calculated that half the WD pairs in *E. coli* are associated into operons (Salgado et al., 2000), we estimate the accuracy as the average of sensitivity and specificity (simple algebra will show this estimation to be equivalent to the definition of accuracy given above if the numbers of positives and of negatives are the same).

Fig. 1. Comparison of the adequacy of the method to discriminate between genes in operons and genes at the borders of transcription units. Performance was measured against data sets of adjacent genes in the same DNA strand known to be in the same operon or known to be at transcription unit boundaries in (a) *Escherichia coli*, and (b) *Bacillus subtilis*. Sensitivity is the fraction of true positives detected and specificity is the fraction of true negatives. Accuracy was estimated as the average of sensitivity and specificity.

Fig. 2. Inter-genic distance frequency distributions of adjacent genes transcribed in the same direction. Almost all genomes show the characteristic peak between −20 and 30 bp due to the presence of operons at the same location. The examples here are the ones where we found the highest (*Thermotoga maritima*) and lowest (*Synechocystis* sp.) peaks (see also Figure 3).

### THE INTER-GENIC DISTANCE DISTRIBUTIONS OF ALL PROKARYOTES REVEAL A CONSERVED PEAK RELATED TO WO PAIRS

Eyre-Walker (1995) noticed before that the most frequent distances between adjacent genes in the same DNA strand in *E. coli* are the overlaps of 4 and 1 bp. Our previous analyses demonstrate these to be the prevalent distances between adjacent genes within operons in the same organism (Salgado et al., 2000). Genes at TU boundaries have an average distance of about 190 bp with a slight preponderance at 120 bases. As a consequence, the inter-genic distance distribution (IDD) of all *E. coli* WD pairs shows a characteristic peak reflecting the abundance of operons in this organism. The IDD of WO pairs of all other prokaryotic genomes have a peak in the very same position, suggesting that their operons have the same characteristic distances separating their genes (see two examples in Figure 2). The prevalent distance is also the overlap of 4 bp in most genomes, except in *Haemophilus influenzae* and the three organisms of the genus *Mycoplasma*, where it is the overlap of 1 bp, and *Ureaplasma urealyticum* with 2 bp of separation. In all these exceptions the next most frequent distance is the overlap of 4 bp. The differences in the size of the peaks are generally small, and might reflect a different abundance of operons, a different tendency of WO pairs to lie close to each other, or the variation in the qualities of genome annotations. Assuming that the differences reflect the abundance of WO pairs of each genome, we can estimate the fractions of WD pairs that belong to operons (Figure 3) by dividing the fraction of WD pairs at the peak (from −20 to 30 bp of distance) in the problem genome, by...
Fig. 3. Estimated fraction of pairs within operons per pair within directon (WO pairs/WD pairs). We calculate this fraction by comparing the fraction of WD pairs under the peak (from −20 to 30 bp of inter-genic distance) against that of WD pairs of *E. coli* K12, and from previous knowledge indicating that the distribution of *E. coli* WD pairs is consistent with a combination of an equal number of WO pairs and TUB pairs (Salgado et al., 2000). The species are sorted in taxonomic groups. For a full list with complete species names with accession numbers see [http://www.cifn.unam.mx/moreno/pub/TUpredictions/REDUNDANCY.html](http://www.cifn.unam.mx/moreno/pub/TUpredictions/REDUNDANCY.html).

the fraction of WD pairs at the peak in *E. coli*. Then we multiply the result by 0.5, which is the ratio of WD pairs per WD pair we have previously calculated for *E. coli* (Salgado et al., 2000).

CONSERVED PAIRS BETWEEN EVOLUTIONARY DISTANT SPECIES SHOW IDDS SIMILAR TO THOSE OF EXPERIMENTALLY CONFIRMED WO PAIRS

As mentioned above, another group devised a method to estimate the probability of two adjacent genes being in the same operon (Ermolaeva et al., 2001). We downloaded the sets of genes with high probabilities to be found in operons (TIGR WO pairs) from their web site (http://www.tigr.org/tigr-scripts/operons/operons.cgi). In several cases, the number of pairs is too small. Thus, we mixed all TIGR WO pairs within some taxa to examine their IDDs (Figure 4). The IDDs of TIGR WO pairs are very similar to those of *E. coli* WO pairs. This is consistent with our previous results showing that conserved adjacent pairs crossing the eubacterial and archaeal domains display IDDs very similar to those of *E. coli* WO pairs (Moreno-Hagelsieb et al., 2001). From this and the results above, we can infer that WO pairs share the same tendency to keep short inter-genic distances in every prokaryote, even in Archaea.

Fig. 4. Inter-genic distance distributions of genes with high probabilities to be in operons. As shown, pairs of genes predicted to be in operons by Ermolaeva et al. (2001), labelled as Proteobacteria, Firmicutes and Archaea, show inter-genic distance distributions very similar to those of experimentally determined pairs in operons of *E. coli*.

PREDICTED WO PAIRS HAVE A HIGHER PROPORTION OF FUNCTIONALLY RELATED GENES THAN PREDICTED TUB PAIRS

We have shown before that co-occurrence, measured as the number of co-occurring orthologous gene pairs divided by the sum of co-occurring pairs and orphans found in another genome, and conservation of adjacency, measured as the number of pairs found adjacent or fused per co-
Fig. 5. Consistency in the conservation of predicted genes within operons versus that of predicted transcription unit boundaries. After applying our distance-based prediction of transcription units to each of the non-redundant set of prokaryotic genomes, we calculated the number of genomes where the predicted WO pairs have co-occurrences or conservation of adjacencies superior to those of predicted TUB pairs, divided by the total number of non-redundant genomes. For instance, predicted WO pairs in *A. fulgidus* have superior co-occurrences than predicted TUB pairs in 48 out of 49 genomes, giving a consistency of \(\sim 0.98\). Most predictions show consistencies above 0.8 in co-occurrence, and above 0.9 in conservation of adjacency. The species are sorted in taxonomic groups. For a full list with complete species names and accession numbers see http://www.cifn.unam.mx/moreno/pub/TUpredictions/REDUNDANCY.html.

A CLOSER LOOK AT THE EXCEPTIONAL CASES

Though the IDDs of WD pairs of most genomes show peaks similar to that of *E. coli*, there are a few extreme examples. *Thermotoga maritima* has an IDD of WD pairs very similar to that of WO pairs of *E. coli*, suggesting that most of its WD pairs are found in operons (Figure 3). The contrary holds for the Cyanobacteria *Nostoc* and *Synechocystis*, implying, in a simple interpretation, that these genomes have very few operons. We tested this assumption by finding the overlap between our distance-based predictions and the TIGR WO pairs (Figure 6). Only *Synechocystis* is included in the TIGR WO pairs. We might conclude that operons have a different tendency to be close together in *Synechocystis* since the distance-based method predicts the lowest number of WO pairs in the corresponding TIGR WO pairs. However, except for co-occurrence in *Synechocystis*, our predicted WO pairs and TUB pairs in both Cyanobacteria are among the well behaved in terms of the above-mentioned tests (see Figure 5). The remaining problems might be related to genome annotation.
Fig. 6. Fraction of the pairs of genes with high probability of being in operons as calculated by Ermolaeva et al. (2001), that are also predicted to be in operons by our distance-based method. The species are sorted in alphabetic order.

THE PROBLEM OF GENOME ANNOTATION

The two main problems in genome annotations that might affect our predictions are incorrect prediction of gene start codons and inclusion of nonexistent genes (over-annotation). To evaluate the effect of these problems in our predictions we sought sources of corrected annotations. The HAMAP project (High quality Automated Microbial Annotation of Proteomes) is aimed at using the high quality of data included in SwissProt (Bairoch and Apweiler, 2000), to devise computational methods to automatically perform corrections on the annotation of genomes. HAMAP reports that the genome annotations of E. coli as originally published (Blattner et al., 1997) contained about 750 wrong start codons. We downloaded all the HAMAP data files (http://www.expasy.ch/sprot/hamap/), and performed comparisons with the Entrez data in order to find the corrected genes. We were able to find 507 corrections most of which move the start codon only three bp shorter. If the coordinates of the genes are corrected using this information the accuracy of our predictions increases slightly to 84%. Unfortunately, the database contains very few corrections for other genomes. However, there is also a report that if the program GeneMark is set to keep the longest probable gene, erroneous start codons contain an average of ~20 more bp (Hayes and Borodovsky, 1998). In any case we need to discuss our exceptional cases. The genome of T. maritima (Nelson et al., 1999) was sequenced and annotated at The Institute for Genomic Research (TIGR). Other TIGR microbial genomes do not show such exceptional peaks, and, thus, the highest peak observed must be a genome tendency. In the case of Synechocystis (Kaneko et al., 1996), however, the conclusions are different. This genome appears by all accounts to contain genes, even at independently predicted operons (TIGR WO pairs), separated by longer distances than any other genome analysed. However, there is experimental evidence that this organism uses a very unusual start codon, TTG (Sazuka and Ohara, 1996), which does not appear in any of the genes annotated in the corresponding Entrez files. It is also compelling that HAMAP corrected start codons for this organism result in longer sequences more frequently than in other organisms. In this regard, an improvement in the GeneMark program to predict genes within prokaryotes was performed to increment the accuracy of start codon predictions (Besemer et al., 2001). While their results confirm that the most frequent distances between genes are those we report here, the authors also mention that they have developed a specific program, in preliminary testing, to predict the start codons for Synechocystis (http://opal.biology.gatech.edu/GeneMark/GeneMarkS/index.html). Thus either Synechocystis is exceptional at many levels, or corrections in the annotation process will show that the IDD of WD pairs of this organism is similar to that of E. coli. The recently released genome of Nostoc (Kaneko et al., 2001) was produced by the same group that released the genome of Synechocystis, and thus its exceptional behaviour might be due to the very same reasons: the necessity of a special method to detect the start codons, a true difference in their operon organization tendencies, and/or a lower proportion of operons. However, It seems that Cyanobacteria have a lower proportion of operons, and that the inconsistencies between our predictions and the TIGR WO pairs might be solved once the special method to predict start codons is developed.

Gene over-annotations might have more profound implications, as evidenced in the case of Aeropyrum pernix (Kawarabayasi et al., 1999), where the extreme over-annotation results in a short peak and an abundance of long overlaps. The curated Entrez file of this organism (NC_000854) contains less over-predicted genes after their elimination during the construction of the Clusters of Orthologous Groups (COGs) (Tatusov et al., 2001). The IDD of this curated genome looks more like that of E. coli (Figure 7). The clean up includes only very hypothetical genes that overlap genes within COGs, and so further improvements can be expected if other overlapping genes are eliminated. Again, such extreme over-annotations are rare (the other evident case being that of Pyrococcus horikoshii (Kawarabayasi et al., 1998)), and thus our method will perform well when used to predict the TU organization of most prokaryotic genomes. Clean up of start codons and over-predictions are beyond the scope of this report, but it is clear that such problems should be taken into account in any genome analyses and predictions derived therein.
ACKNOWLEDGEMENTS

This work was supported by grant number NC028 from: Consejo Nacional de Ciencia y Tecnología to J.C.-V.
We appreciate fruitful discussions with Temple F. Smith and technical support from Víctor del Moral and César Bonavides.

REFERENCES


CONCLUDING REMARKS

The tendency of genes within operons to be slightly overlapped or close together might derive from an economy effect combined with a lack of necessity for space between genes. Another reason for little spacing might be to protect the resulting mRNA from degradation by association with ribosomes (Schneider et al., 1978). Our results suggest that the tendency to keep such distances is universal and strong enough to stand out in the form of a peak in frequency distribution plots, even when diluted against the distances between all genes in TUB pairs. Such universality might also be explained in terms of the theory of the selfish operon (Lawrence and Roth, 1996). Put in the simplest terms, the short distances might reduce chances of physical separation of the associated genes, thus ensuring a concerted transfer. Whatever the reasons, the consequences permit the use of the same distance derived log-likelihoods to predict operons with high accuracy in most if not all prokaryotes. Some refinements can be applied. For instance, we are currently cleaning up false negatives appearing as fusions in other genomes, or showing conservation of adjacency within evolutionarily distant species. As for false positives, we expect to be able to add corrections by the prediction of transcriptional promoters. Predicted operons represent key information that will improve functional inferences useful not only in the interpretation of transcriptome and proteome experiments, but also in the study of the whole regulatory system of prokaryotic genomes. Since the method requires no further information but the coordinates of the genes within a genome, it is applicable as soon as the CDSs of a given genome are mapped.