Published online 23 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D479–D482***
doi:10.1093/nar/gkn784

OperonDB: a comprehensive database of predicted operons in microbial genomes

Mihaela Pertea*, Kunmi Ayanbule, Megan Smedinghoff and Steven L. Salzberg

Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA

Received September 11, 2008; Revised October 8, 2008; Accepted October 9, 2008

ABSTRACT

The fast pace of bacterial genome sequencing and the resulting dependence on highly automated annotation methods has driven the development of many genome-wide analysis tools. OperonDB, first released in 2001, is a database containing the results of a computational algorithm for locating operon structures in microbial genomes. OperonDB has grown from 34 genomes in its initial release to more than 500 genomes today. In addition to increasing the size of the database, we have re-designed our operon finding algorithm and improved its accuracy. The new database is updated regularly as additional genomes become available in public archives. OperonDB can be accessed at: http://operondb.cbcb.umd.edu

INTRODUCTION

Large-scale comparison of complete microbial genomes reveals that numerous gene clusters are conserved across species; i.e. sets of genes occur in the same physical arrangement in two or more species. Such gene clusters often, but not always, represent co-transcribed units, or operons, which are sets of genes that are co-regulated. In general, two genes are more likely to belong to an operon if they are adjacent, have the same orientation, are positioned relatively close together and are not separated by known promoters or terminators. Genes in an operon typically have closely related functions and consequently there is a strong selective pressure for operon structure to be conserved between genomes (1). Similarly, knowledge of operon structure provides important clues to the function of genes within an operon.

Expression data from microarray experiments has previously been used to reliably identify clusters of co-transcribed genes (2,3); however, such data are not available for most bacterial and archaeal genomes, and sequencing continues to accelerate. Our operon prediction method (4) instead uses a purely computational and statistical approach, relying on conservation of gene order and orientation in two or more species to infer operon structure. We first used this method in 2001 to construct OperonDB, a database of operons for all complete microbial genomes, which at the time numbered 34 species. Although highly specific (with fewer than 2% false positives), the sensitivity of the original algorithm was estimated at 30–50% for the Escherichia coli genome. The current version of OperonDB has greater sensitivity while maintaining a similarly high level of specificity.

NEW DEVELOPMENTS

The original OperonDB algorithm defined the concept of a gene pair, and estimated the probability that genes in a conserved gene pair belonged to the same operon (4). A gene pair is defined as two adjacent genes (G1, G2) on the same strand, separated by at most 200 bp. A conserved gene pair is a gene pair that is found in two or more distinct genomes, where for pair (A, B) in genome G_1 and (C, D) in genome G_2, A and C are orthologs, and B and D are orthologs.

The probability that a conserved gene pair between two genomes belongs to an operon was estimated (in the original OperonDB algorithm) from the overall proportion of conserved gene pairs as:

$$P(\textit{gene pair in operon}) = 1 - \frac{P(\textit{conserved}|D)}{P(\textit{conserved}|S)} \times P(SN|S) - P_{\text{chance}} \quad (1)$$

where S and D are the sets of adjacent gene pairs on the same and different strands, respectively, SN is a subset of S that contains only pairs that do not belong to the same operon and P_{chance} is the probability that a conserved S pair has homologs in other genomes.

We recently re-designed the algorithm behind OperonDB to make it more efficient and to improve its accuracy. The new implementation has increased sensitivity due to a relaxation of the constraints required to assign a gene pair to an operon. For example, in the current implementation, we eliminate the adjacency requirement

*To whom correspondence should be addressed. Tel: +1 301 405 9762; Fax: +1 301 314 1341; Email: mpertea@umiacs.umd.edu

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

by requiring a gene pair to be co-linear, but allowing genes on the same strand to be separated by other genes with the same orientation. A threshold determines the maximum number of genes that are allowed to separate genes in any conserved gene pair. This relaxation allows for re-shufflings of gene order between species as well as genes inserted erroneously by misannotations. We also eliminated the 200-bp minimum separation between genes: while this limit works well for *E. coli*, other species have different distributions of intergenic lengths, and genes within operons are likely to reflect this difference (5).

By assuming that the conservation of a gene pair is independent of intergenic distance l, we can rewrite Equation 1 as:

$$P(\textit{gene pair in operon}) = 1 - \frac{P(\textit{conserved}|D)}{P(\textit{conserved}|S)} \times \frac{P(l|D)}{P(l|S)} \times P(SN|S) - P_{\text{chance}} \quad \textbf{2}$$

where the two probabilities $P(l|S)$ and $P(l|D)$ define the probabilities for a given S or D pair to have length l (estimated from the distributions of intergenic distances between same strand and opposite strand pairs, respectively).

Due to the enormous quantities of data involved when dealing with the numerous prokaryotic genomes available today, the most computationally intensive step in estimating the probabilities in Equation 2 is the identification of conserved pairs. To speed-up this step, we identify all orthologs between all pairs of genomes with highly parallelized BLAST (6) searches run on a grid-based system. We then find conserved gene clusters with a slightly modified version of the HomologyTeams (7) software. The algorithm implemented by HomologyTeams is a fast way to identify sets of orthologous genes. Within the conserved clusters neither the order of the genes nor their orientation need to be conserved, but a fixed threshold limits the distance between adjacent genes. We modified the HomologyTeams program to allow conserved pairs to be separated by a fixed number of non-conserved genes as well. This modification allows us to estimate both S and D conserved pairs in the context of our relaxed definition of gene pair.

Although Equation 2 is an attempt to compute the probability P that two genes belong to the same operon, it incorporates several simplifying assumptions related to both the operon model (e.g. by assuming that D pairs can never belong to operons) and the evolutionary reason for gene cluster conservation (e.g. by ignoring alternative causes such as lateral gene transfer). As shown previously (4), these assumptions serve to make our estimate of P more conservative; i.e. it is an underestimate. Therefore, for the remainder of this discussion we will use the term *confidence value* instead of probability when we refer to the estimated value of P.

The confidence value computed by Equation 2 does not take into account the evolutionary distance between the species containing the conserved gene pairs. Intuitively, the probability that a conserved gene cluster is co-transcribed should be higher if a larger evolutionary distance separates the species. To reflect this fact, we first estimate

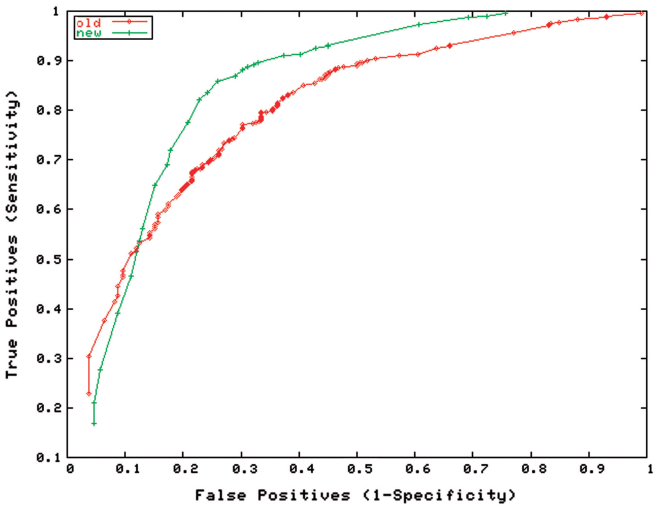

Figure 1. OperonDB ROC curves computed using the original prediction algorithm from 2001 ('old') versus the new improved algorithm ('new'). True and false positive rates are computed on a set of 602 *E. coli* experimentally confirmed operon pairs from RegulonDB.

the evolutionary distance between any two genomes G_1 and G_2 using the Jaccard distance (8):

$$d(G_1,G_2) = \frac{n(G_1) + n(G_2) - h(G_1;G_2) - h(G_2;G_1)}{n(G_1) + n(G_2)} \quad \textbf{3}$$

where $n(G_i)$ is the number of genes in G_i and $h(G_i;G_j)$ is the number of homologues from G_j in G_i. The previously computed confidence value for a conserved gene pair in two genomes G_1 and G_2 is then weighted with the ratio between $d(G_1,G_2)$ and the maximum genome distance observed in the database.

Figure 1 shows Receiver Operating Characteristic (ROC) curves of true positive versus false positive operon predictions for the new OperonDB database. The ROC curves are computed using a set of 602 *E. coli* operon pairs from RegulonDB (9), all of which are supported by experimental evidence. In the figure, a true positive is defined as any prediction of a conserved gene pair that is contained in a documented operon, and a false positive is any predicted pair where only one of the two genes belongs to a confirmed operon from the *E. coli* data set. Our new method shows substantial improvement over the previous implementation, especially for sensitivity levels >60%. Using Figure 1, we can compute the accuracy (defined as the average of sensitivity and specificity) of the new OperonDB for all possible thresholds. The maximum accuracy achieved is 80%, which is comparable to accuracies obtained by other recent operon prediction methods (10).

IMPLEMENTATION AND WEB INTERFACE

OperonDB is accessible through a web interface at operondb.cbcb.umd.edu. The website is regularly updated and currently contains operon predictions for 550 bacterial and archaeal genomes, comprising the complete collection of finished prokaryotic genomes available from

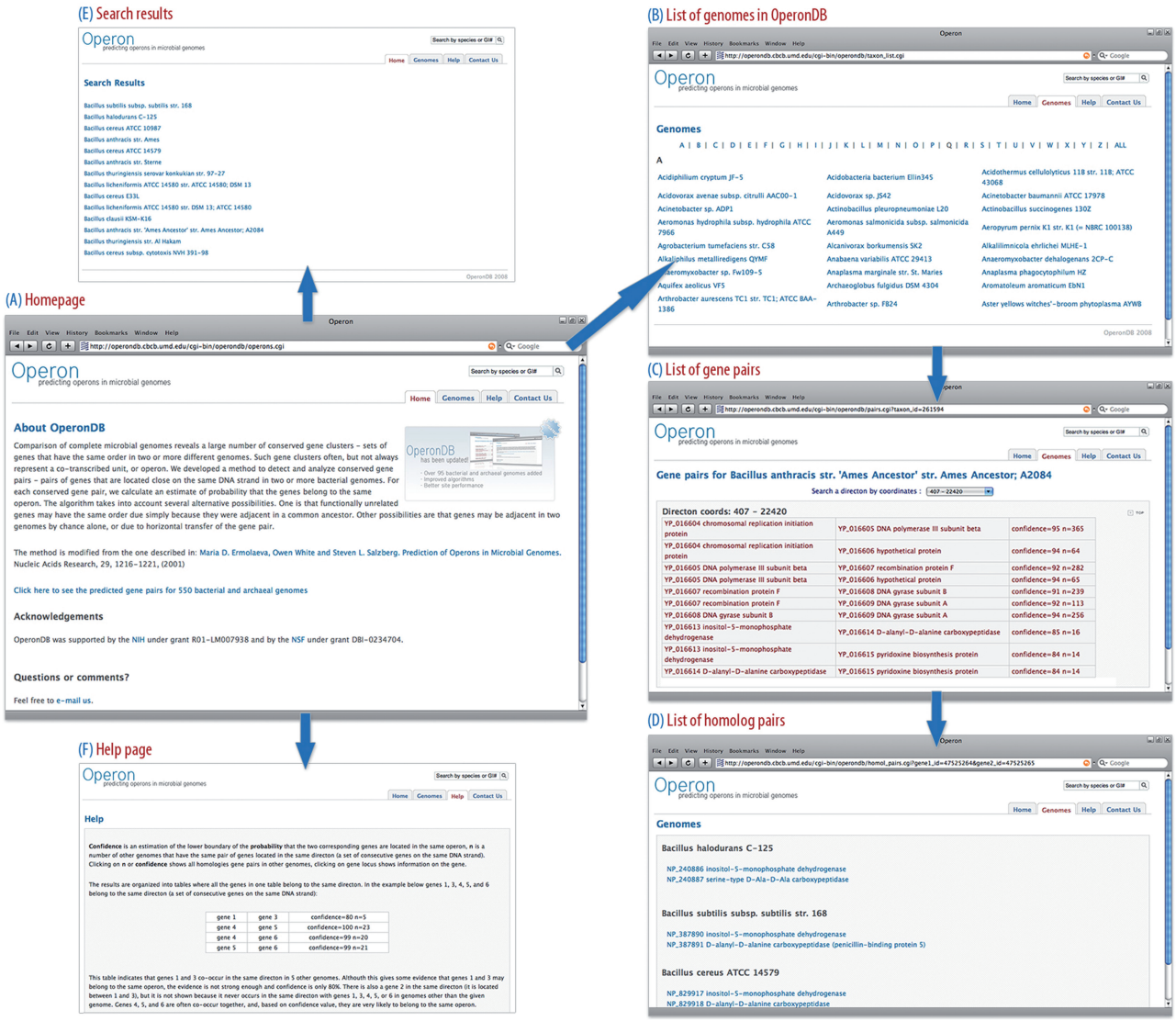

Figure 2. Site map of OperonDB. (**A**) Homepage of OperonDB. (**B**) Genomes are browsed by clicking the 'Genomes' navigation tab, which brings up an alphabetically ordered list of species. (**C**) The user can obtain a list of predicted operon gene pairs and their associated confidence values by selecting a specific organism. (**D**) Clicking on the number of genomes in which a particular pair is conserved retrieves the list of homologous gene pairs in those genomes. Each gene name is linked to the corresponding Gene Information page at NCBI. (**E**) Alternatively, a user can search for a genome by entering an organism's name or its NCBI GenInfo (GI) number in the search box located above the navigation tabs and mirrored on every page of the website. The search will return a comprehensive list of organisms that match the search query. (**F**) OperonDB's help page.

Genbank (www.ncbi.nih.gov) as of July 2008. This number will increase over time as more complete microbial genomes appear. All predictions can be downloaded in bulk, and the OperonDB software is available as free, open source software.

In an effort to make OperonDB faster and more portable, we have made several changes that have dramatically improved its performance and user interface. The most significant performance upgrade resulted from a switch of the database management system from Sybase to MySQL (5.0.22). The database also received a performance boost from a new database schema, better optimization tuning and an improved caching system. In consequence, OperonDB is easily scalable to a large number of genomes, and we routinely update it with little or no system downtime.

OperonDB's web interface, which uses an Apache web server and a collection of php and perl-cgi scripts, was redesigned with a more intuitive system for data browsing. Figure 2 shows a site map of OperonDB.

FUTURE PLANS

Accurate knowledge of promoters and terminators should allow us to improve further our predictions of transcription boundaries. Although promoters are notoriously difficult to predict, the TransTermHP system (11) does an excellent job—for at least some species—of finding rho-independent terminators. And as mentioned above, when expression data are available, operon predictions can be even more accurate. We will continue to explore these and

other methods for improving the accuracy of our operon finding algorithm.

FUNDING

National Institutes of Health (R01-LM006845 and R01-GM083873). Funding for open access charge: National Institutes of Health (R01-LM006845 and R01-GM083873).

Conflict of interest statement. None declared.

REFERENCES

1. Lawrence,J.G. and Roth,J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.
2. Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–2893.
3. De Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput.*, 276–287.
4. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
5. Ermolaeva,M.D. (2005) Operon finding in bacteria. In Subramaniam, S. (ed), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. John Wiley & Sons, New York, pp. 2886–2891.
6. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
7. He,X. and Goldwasser,M.H. (2005) Identifying conserved gene clusters in the presence of homology families. *J. Comput. Biol.*, **12**, 638–656.
8. Wolf,Y.I., Rogozin,I.B., Grishin,N.V. and Koonin,E.V. (2002) Genome trees and the tree of life. *Trends Genet.*, **18**, 472–479.
9. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
10. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
11. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.