BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 8 2005, pages 1403–1407
doi:10.1093/bioinformatics/bti156

Sequence analysis

A fuzzy guided genetic algorithm for operon prediction

E. Jacob[1], R. Sasikumar[2,*] and K. N. R. Nair[3]

[1]Department of Computational Modeling and Simulation, Regional Research Laboratory (CSIR), Trivandrum 695019, India, [2]Department of Computational Modeling and Simulation Regional Research Laboratory (CSIR), Trivandrum 695019, India and [3]School of Computer Science, Mahatma Gandhi University, Kottayam 686560, India

Received on May 3, 2004; revised on October 28, 2004; accepted on November 13, 2004
Advance Access publication November 25, 2004

ABSTRACT

Motivation: The operon structure of the prokaryotic genome is a critical input for the reconstruction of regulatory networks at the whole genome level. As experimental methods for the detection of operons are difficult and time-consuming, efforts are being put into developing computational methods that can use available biological information to predict operons.

Method: A genetic algorithm is developed to evolve a starting population of putative operon maps of the genome into progressively better predictions. Fuzzy scoring functions based on multiple criteria are used for assessing the 'fitness' of the newly evolved operon maps and guiding their evolution.

Results: The algorithm organizes the whole genome into operons. The fuzzy guided genetic algorithm-based approach makes it possible to use diverse biological information like genome sequence data, functional annotations and conservation across multiple genomes, to guide the organization process. This approach does not require any prior training with experimental operons. The predictions from this algorithm for *Escherchia coli* K12 and *Bacillus subtilis* are evaluated against experimentally discovered operons for these organisms. The accuracy of the method is evaluated using an ROC (receiver operating characteristic) analysis. The area under the ROC curve is around 0.9, which indicates excellent accuracy.

Contact: roschen_csir@rediffmail.com

INTRODUCTION

Bacterial genomes are organized into operons, which are clusters of adjacent genes transcribed as a single mRNA molecule. Operons can thus be considered as the fundamental units of transcription and the operon map of the genome is a key input to the reconstruction of regulatory circuits at the whole genome level. Co-transcribed genes often play related roles in the function of the organism, either interacting directly with each other, or catalyzing reactions in the same metabolic pathway. Identifying operons can thus enhance our knowledge of gene function.

Methods for computational prediction of operons have used different approaches. The most direct approach involves detection of promoter and terminator sequences at the operon boundaries (Yada *et al*., 1999) Another approach is based on the assumption that the operon structure is conserved across multiple genomes (Siefert *et al*., 1997; Overbeek *et al*., 1999). Ermolaeva *et al*. (2001) and Moreno-Hagelsieb *et al*. (2001) were able to show that conservation of gene order was indeed related to the operon structure. The assumption that genes in an operon participate in consecutive biochemical reactions has also been made use of to predict operon structure (Zheng *et al*., 2002). These approaches cannot make predictions at the whole genome level since the information available does not span the whole genome.

In order to predict the operon map of the whole genome, the simplest computational technique is based on the observation that the intergenic distance within the operons is smaller than that at the operon boundaries (Salagado *et al*., 2000; Moreno-Hagelsieb and Collado-Vides, 2002). Based on experimental data on the intergenic distance of gene pairs within operons and at operon boundaries of the *Escherchia coli* genome, Salgado *et al*. developed a log likelihood function of intergenic distance for predicting operon structure of the whole *E.coli* genome. Later Moreno-Hagelsieb *et al*. reported that the same function is applicable to other bacterial genomes also for prediction of operons.

A first level organization of the genome into operons can be made using a threshold intergenic distance value. Genes on the same strand that are closer together than the threshold value are considered as having an operon linkage. Such a first level organization can be further refined using additional criteria. In a recent paper (Strong *et al*., 2003) the distance-based gene linkages are further substantiated with linkages based on three other criteria, namely, occurrence of the gene pair in fused form in other genomes, similarity of their phylogenetic profiles and conservation of the same pair across multiple genomes. However, no effort was made to evaluate an overall score for operons or operon linkages based on the multiple criteria. Therefore the addition of more criteria reduces the number of predictions drastically since the method looks only for overlapping regions of the different criteria. Thus the additional criteria increase the specificity of the predictions but decrease the sensitivity.

Sabatti *et al*. (2002) calculated probability scores for operon linkages based on distance and correlation of expression patterns and performed a Bayesian classification into operons and non-operons based on these scores. It was found that the distance-based predictions could be validated and improved by correlating with expression data. They also calculated a measure of the information content of the expression data, which indicates to what extent a prediction based on expression pattern can be considered as relevant.

*To whom correspondence should be addressed.

A recent paper (Bockhorst *et al*., 2003) reports a Bayesian network approach to operon prediction at the whole genome level. This approach makes use of diverse criteria features to score candidate operons. The network is trained on a set of experimental operons and non-operons to determine the effect of the feature values on the probability of a candidate operon being a true operon. The features used for evaluation are:

(1) Promoter and terminator signals;

(2) operon length;

(3) intergenic distance;

(4) codon usage frequency;

(5) gene expression.

With this method it is possible to predict the operon structure of the whole genome with a high level of accuracy. However a training set of experimental operons is a prerequisite for the application of this method.

In this paper we propose another method by which it is possible to predict operons at the whole genome level using diverse biological criteria and no prior training. We use a genetic algorithm technique to 'evolve' an initial population of putative operon maps through progressively 'fitter' generations of operon maps obtained by crossover and mutation of the parent populations. The fitness of each member of the population is assessed using a multiple criteria scoring function. In our example implementation for *E.coli*, we use four scoring criteria:

(1) Intergenic distance;

(2) participation in the same metabolic pathway;

(3) conservation across multiple genomes;

(4) similarity of protein functions.

The main problem associated with multiple criteria is to evolve a methodology to evaluate a score that reflects their combined effect. The criteria may involve very different representations from numerical values to textual information and their relative importance; also, our level of confidence in them may vary. Therefore arriving at a combined score is not straightforward. Fuzzy logic (Kaehler, 1998, http://www.seattlerobotics.org) proves to be useful in this case. In this paper putative operons are scored using intuitive fuzzy rules. Fuzzy rules for combining the scores can be made to reflect the relative importance of the scoring criteria and the confidence we have in the data. Without any prior training on experimental data we find that our intuitive rules enable predictions of high accuracy.

SYSTEMS AND METHODS

The systems we consider in this paper are the operon structures of the *E.coli* genome and the *Bacillus subtilis* genome. The method is applicable to any prokaryotic genome limited only by the information available for that genome.

Datasets and Data Preprocessing

(1) Genome data of *E.coli* K12-MG 1655 and *B.subtilis* was downloaded from http://www.genome.ad.jp/kegg. A computer program extracts the required data, namely, gene name, strand, position and the metabolic pathways it appears in. This forms the master data file from which other input data files are created.

(2) The genes on both strands are organized into clusters based on 10 different threshold intergenic distances and stored in 10 separate files.

(3) A metabolic pathway data file is created with a record containing gene name and the pathways it is involved in.

(4) The distance file contains the gene names and the intergenic distance with the last gene on the same strand.

(5) A protein function file for *E.coli* is created using the multifunction assignment data from the database GeneProtEc (http://www.genprotec.mbl.edu/).

(6) Probability scores based on conservation of gene pairs across multiple genomes are taken from the TIGR Database (http://www.tigr.org) for *E.coli* K12.

Model overview

The method uses a hybrid soft computing approach that combines genetic algorithms (GA) with fuzzy logic to predict operons across the whole genome. While the GA attempts to evolve better clusters (putative operons) with each generation, the fuzzy logic system, applied to every cluster, weighs the contribution of different criteria to give a single crisp fitness value (in percentage) that is a measure of how good the cluster would be as an operon.

The algorithm

The main computational elements of the genetic algorithm (Goldberg, 1989) are:

(1) Representation of potential solutions;

(2) creation of an initial population;

(3) a fitness function to score the putative operons;

(4) application of the genetic operators—selection, crossover and mutation to evolve the population.

Figures 1 and 2 give the block diagram of the algorithm.

Representation and initial population

A population of n individuals holds n alternate solutions to the problem of determining the operon map of a genome. In our case, each individual is a putative operon map of the genome. This is represented as an array of integers where each integer stands for the operon number to which the gene in that position belongs. For example, the first four genes may belong to operon number 1, the next three genes to operon number 2, etc., in which case the elements of the array will be $1, 1, 1, 1, 2, 2, 2, \ldots$ In the particular case of the *E.coli* K12 genome, this represents a putative operon map in which the genes b0001, b0002, b0003 and b0004 are considered as belonging to operon number 1; b0005, b0006 and b0007 belonging to operon number 2; and so on.

We start with an initial population in which the individuals are created by organizing the genes into operons using different threshold intergenic distance values. The population size was chosen to be 10. A computer program generates data files: initial#.dat where $\# = (1,10)$ corresponding to different threshold intergenic distances from 0 to 600 bps.

The Fuzzy Fitness Finder

The GA calls upon the Fuzzy Fitness Finder (FFF) to evaluate the solutions it generates. The FFF combines the multiple criteria of distance, metabolic pathway involvement, conservation across multiple genomes and protein function similarity of each putative operon giving a single fitness value (in percentage) for that operon.

Scoring a putative operon by multiple criteria involves combining the relative values of different criteria. Given the set of genes in a putative operon, we calculate numerical values that are a measure for the score according to each criterion as follows.

The numerical score based on intergenic distance is the average intergenic distance within the operon.

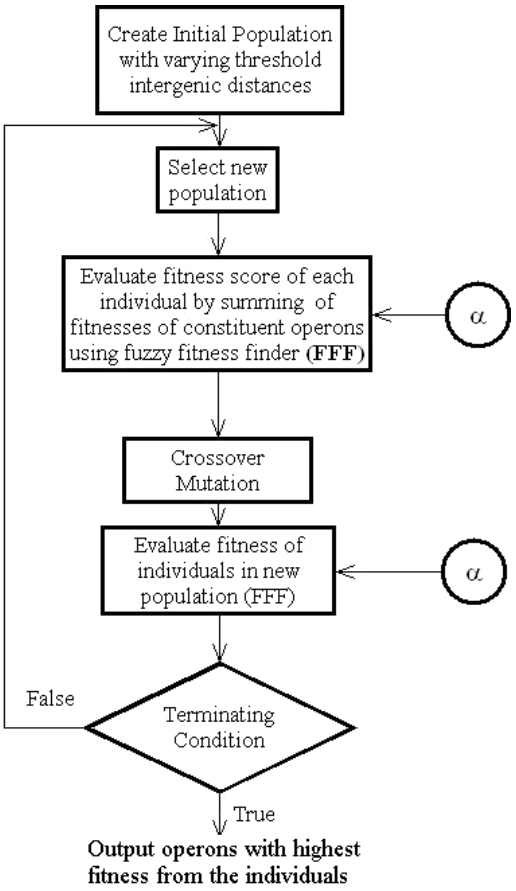

Fig. 1. Block diagram of the genetic algorithm.

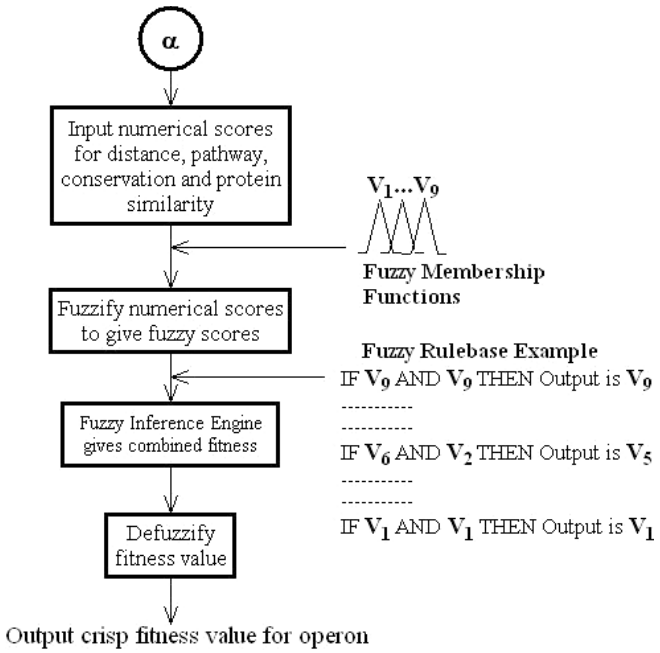

Fig. 2. Block diagram of the FFF.

For scoring by commonality of metabolic pathway or protein function, if two genes in an operon have a common pathway or protein function, the score for that pair is taken as 1 or else as 0. If there are m genes in an operon, all pairs of genes are scored as above and divided by ${}^{m}C_{2}$ (number of all possible gene pairs) to give a value between 0 and 1. Thus if all gene pairs have a common pathway, the pathway score for that operon will be 1.

The score based on conservation of gene pairs across multiple genomes is available (in percentage) in the TIGR Database. The score for all gene pairs in an operon is averaged over all pairs in the putative operon (i.e. summed and divided by mC2) to give a single conservation score for the operon.

The numerical values for the scores based on each criterion are then converted into fuzzy scores. The fuzzy membership functions for the criteria have been fixed using nine triangular sets (V1, . . . ,V9) where V1, V2, . . . have linguistic interpretations like extremely low, very low, etc. The fuzzy scores for the different criteria are then combined to give a single score using a fuzzy inference engine which reflects human judgment on the relative importance of each criterion. The inference engine combines the scores two at a time. It first combines the distance score with conservation score. The combined score of distance and conservation criteria is then combined with the pathway score and this combined score with the protein function score. Thus the inference engine is called three times to get the score based on all the four criteria. This procedure is easy to scale up to add more criteria. The final fitness scores are de-fuzzified using the root-sum-squares method to yield a crisp fitness value. The fuzzification schemes for the four criteria as well as the rule set comprising the fuzzy inference engine are given as Supplementary data.

The fitness of an individual is calculated as the summation of the fitness scores of all the putative operons of that individual,

$$\text{Fitness} = \sum_{i=1}^{k} \text{fitness}_i$$

where k is the number of operons in an individual.

Figure 2 gives the block diagram for the FFF. Apart from calculating fitness for a putative operon, the FFF also finds pairwise fitness values between genes. This pairwise fitness is used in guiding the mutation operations.

Selection, crossover and mutation

The selection method used is the roulette wheel method. A linear search is made through a roulette wheel with slots weighted in proportion to the fitness values of the n individuals making up that population. Individuals with higher fitness scores get more copies of themselves into the new population.

Crossover operation is performed by randomly choosing a pair of individuals from the new population, cutting at a random position on the genome and combining the fragments to form two new individuals.

The mutation operators implemented are:

(1) If two operons belong to the same strand and the pairwise fitness between the last gene of the first operon and the first gene of the second operon is very high, then the two operons are merged.

(2) If genes at the end of a putative operon have very low pair-fitness, the last gene is removed from the operon.

(3) Single gene clusters are merged if they have high pair-fitness.

The process of evaluation, selection, crossover and mutation is carried on until there is no significant difference in fitness values in successive generations. The last generation consists of the 10 best individuals as found by the algorithm. Each individual is a grouping of the genome into operons with a degree of certainty expressed in percentage. The best operons, selected from the 10 individuals so as to cover the entire genome, form the final solution.

RESULTS AND DISCUSSION

Our method predicts whole operons and calculates an overall score for each operon. The algorithm was evaluated by comparison of its predictions with the experimentally discovered operons included in the *E.coli* database RegulonDB (http://www.cifn.

unam.mx/Computational_Genomics/regulondb). However because the predicted and experimental operons could be of different sizes, it is more convenient to compare the pairwise linkages implied by the predicted and experimental operons. Therefore at the termination of the algorithm, we convert our operon scores to pairwise linkage scores, giving each adjacent gene pair in the predicted operon the overall score of the operon. Similarly we deduce 807 gene pairs with operon linkages from the multigene transcriptional units given in RegulonDB. We use this as the positive experimental test set for evaluating the sensitivity of our method. In order to test the specificity of the method we also need a set of gene pairs that are known to be not linked; i.e. genes that occur at the operon borders. Obtaining a set of gene pairs that are known to be unlinked is not straightforward because experimentally, it is easier to detect co-transcription than it is to detect the absence of co-transcription. In this work, as done by earlier authors (Ermolaeva *et al*., 2001; Salgado *et al*., 2002), we took gene pairs that occur at the borders of the experimental operons as unlinked if they occurred on the same strand. For example if A, B, C, D and E are consecutive genes on the same strand and RegulonDB records BCD as an operon, A–B and D–E were considered as border pairs. We selected as our negative test set, 312 gene pairs at the (same strand) borders of the experimental operons that we use as our positive test set.

We also implemented the method with *B.subtilis* using only two criteria, namely intergenic distance and metabolic pathway. We used the same rules as for *E.coli* for scoring the operons in *B.subtilis*. For evaluating the predictions we used experimental data received in personal communication from Shujiro Okuda of Kyoto University (Okudo, 2004, personal communication). A set of 703 operon pairs and 194 border pairs were obtained from this dataset.

Evaluating the accuracy of prediction

We use an ROC curve to evaluate the overall accuracy and predictive value of the method. The ROC analysis is a standard approach to evaluate the sensitivity and specificity of diagnostic procedures. It estimates a curve, which describes the inherent tradeoff between sensitivity and specificity of a diagnostic test. Each point on the ROC curve is associated with a specific diagnostic criterion—in this case it is the cut-off score value of an operon linkage above which we diagnose the operon linkage as true. The ROC curve is obtained by plotting the True Positive rate (fraction of experimental linkages that are predicted by the method) against the False Positive rate (1 – fraction of experimental borders predicted by the method), for different values of the cut-off score. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test because it indicates nearly 100% sensitivity with almost zero false positive rate. The closer the curve comes to the 45° diagonal of the ROC space, the less accurate the test because the 45° diagonal represents a random guess situation. The area under the ROC curve is a measure of the predictive value of the method.

Evaluation of the effect of the different criteria

Figure 3 shows the ROC curves for predictions using each criterion individually and all the four criteria together for *E.coli*. The areas under the ROC curves for the different criteria are also given in the figure along with the legends. Taken individually, intergenic distance and participation in the metabolic pathway are the most effective prediction criteria. Intergenic distance criterion fails to detect some of the experimental operons because some linked genes have large

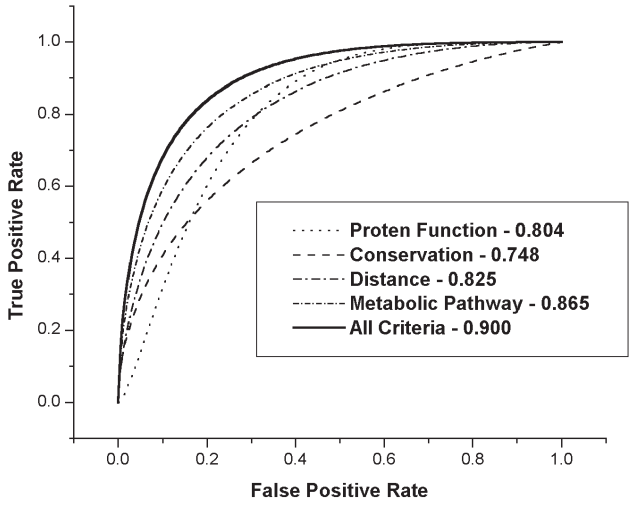

Fig. 3. ROC curves for *E.coli* obtained using different criteria individually and together. Areas under the curves are given in the legend.

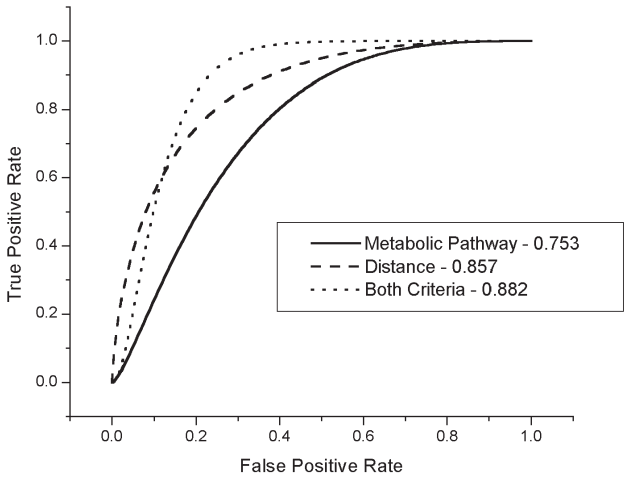

Fig. 4. ROC curves for *B.subtilis* obtained using two different criteria individually and together. Areas under the curves are given in the legend.

intergenic distances between them. The criterion of participation in the same metabolic pathway is a very 'sure' criterion in that its predictions are almost always correct (high specificity). However this data applies only to genes that code for enzymes and therefore cannot predict all the operon linkages. The conservation data does not cover the whole genome, but a high score by this criterion is highly indicative of operon linkage. The protein function criterion, on the other hand, is useful for detecting most of the experimental linkages, but the number of false positives it predicts is very high. Our fuzzy inference engine reflects these observations. For example a high score according to the protein function criterion is not considered as indicating high probability of showing linkage unless it is supported by a high score in one of the other criteria.

When all the four criteria are used the area under the curve is 0.9, which demonstrates the excellent accuracy of the predictive method.

The ROC curves for *B.subtilis* using the distance and metabolic pathway criteria individually and together are shown in Figure 4. The metabolic pathway data being scantier and probably less reliable

for *B.subtilis* compared to *E.coli*, this criterion makes only a small improvement to predictions based on distance alone. Unlike in the case of *E.coli* inclusion of the metabolic pathway criterion seems to increase the number of false positives. Using both criteria, the area under the ROC curve is 0.882, again showing good accuracy of prediction.

Predicted operons

The method predicts 745 multigene operons for *E.coli* and 844 multigene operons for *B.subtilis* with varying levels of confidence. The complete lists of transcriptional units, with their scores, are given as Supplementary Material.

CONCLUSIONS

- This method proves to be a useful computational tool for prediction of operons at the whole genome level with high level of accuracy.
- A methodology has been developed for using different kinds of biological information in combination for scoring the predictions. The method can be easily scaled up to add more scoring criteria.
- Using information from diverse biological features improves the predictive value of the method because when information on one of the features is missing or unreliable, information on another feature can contribute to the score.
- The method does not need prior training. Intuitive rules are sufficient to make predictions of high accuracy.

ACKNOWLEDGEMENTS

We wish to express our thanks to Ms Yamuna Devi and Mr Jijoy Joseph for helping with the preparation of the data files. Our sincere gratitude to Prof. Javed Iqbal for initiating us into the work and to Prof. T.K. Chandrasekhar for his encouragement and support.

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.

Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA, http://www.seattlerobotics.org

Kaehler,S.D. (1998) *Fuzzy Logic—An Introduction*, http://www.seattlerobotics.org

Moreno-Hagelsieb,G., Trevino,V., Perez-Rueda,E., Smith,T.F. and Collado-Vides,J. (2001) Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. *Trends Genet*, **17**, 175–177.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) Operon conservation from the point of view of *Escherichia coli* and inference of functional interdependence of gene products from genome context. *In Silico Biol.*, **2**, 87–95.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Sabatti,C., Rohlin,L., Oh,M. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–2893.

Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.

Siefert,J.L., Martin,K.A., Abdi,F., Widger,W.R. and Fox,G.E. (1997) Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J. Mol. Evol.*, **45**, 467–472.

Strong.M., Mallick,P., Pellegrini,M., Thompson,M.J. and Eisenberg,D. (2003) Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. *Genome Biol.*, **4**, R59.

Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.

Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221–1230.