BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 24 2010, pages 3127–3128
doi:10.1093/bioinformatics/btq601

Sequence analysis

Advance Access publication October 28, 2010

PPO: Predictor for Prokaryotic Operons

Li-Yeh Chuang[1], Jui-Hung Tsai[2] and Cheng-Hong Yang[3,4,*]

[1]Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University, [2]Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, [3]Department of Network Systems, Toko University, Chiayi and [4]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: We present an operon predictor for prokaryotic operons (PPO), which can predict operons in the entire prokaryotic genome. The prediction algorithm used in PPO allows the user to select binary particle swarm optimization (BPSO), a genetic algorithm (GA) or some other methods introduced in the literature to predict operons. The operon predictor on our web server and the provided database are easy to access and use. The main features offered are: (i) selection of the prediction algorithm; (ii) adjustable parameter settings of the prediction algorithm; (iii) graphic visualization of results; (iv) integrated database queries; (v) listing of experimentally verified operons; and (vi) related tools.

Availability and implementation: PPO is freely available at http://bio.kuas.edu.tw/PPO/.

Contact: chyang@cc.kuas.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 7, 2010; revised on October 15, 2010; accepted on October 19, 2010

1 INTRODUCTION

An operon consists of one or more consecutive genes on the same DNA strand that share a common promoter and terminator and are therefore co-transcribed into a single-strand mRNA sequence. Genes belonging to the same operon tend to work together. Operon prediction can be used to infer the functions of putative proteins (Yada *et al.*, 1999), and also provides valuable information for drug design and determining a protein's functions. However, the number of operons known to date is relatively small, and experimental methods to verify operons cannot keep up with genome sequencing in price and speed. Improving the operon prediction process has thus become a very important issue.

Many operon databases have been constructed and made available. RegulonDB (Pertea *et al.*, 2008) and DBTBS (Sierro *et al.*, 2008) collect experimentally validated operons on the *Escherichia coli* and *Bacillus subtilis* genome, respectively. DOOR (Database of prOkaryotic OpeRons) (Mao *et al.*, 2009) and MicrobesOnline (Price *et al.*, 2005) collect and sort the predicted operons based on the method that was used to identify them. ODB (Operon DataBase) collects the operons of 203 genomes and contains both predicted and experimentally verified operons (Okuda *et al.*, 2006).

*To whom correspondence should be addressed.

In this study, we present a web server called PPO (Predictor for Prokaryotic Operons), which integrates a database with a predictor. The PPO database collects the predicted operons of 592 genomes, and provides a search function for operons and genes. Furthermore, PPO contains references to some literature methods and improves on their optimization algorithms, thus resulting in better prediction performance. Users can freely select the necessary parameters. The ease of use and the data provided make it a very useful operon platform for bioinformatics research.

2 METHODS

Of the databases currently in use, only DOOR and ODB provide operon prediction functions. The method used by the ODB employs four types of associations to predict operons, whereas the DOOR database utilizes a decision tree-based classifier for prediction. However, the DOOR database only allows users to upload the gene location information files, protein sequence files or genome sequence files. The database does not provide a parameter setting interface for the user. ODB only allows users to set up biological property parameters, whereas algorithm parameters can not be set. In order to allow users to freely select related parameters and use different algorithms, we provide an operon prediction platform ideally suited for these needs. In PPO, parameters can be freely adapted. The structure of the PPO system and its flowchart is shown in Supplementary Figure 1. A comparison of PPO and other currently used platforms is shown in Supplementary Table 1.

2.1 Optimization algorithms

The PPO system currently offers a binary particle swarm optimization (BPSO), a genetic algorithm (GA), and the ODB and DOOR predictors to predict operons (Supplementary Fig. 1). The particle swarm optimization (PSO) technique consists of a population-based evolutionary algorithm (Kennedy and Eberhart, 1995). PSO was developed through simulation of the social behavior of organisms, such as birds flocking and fish schooling. Each particle from a swarm represents a candidate solution to the problem. The individual best value ($pbest_i$) is the position of the i-th particle with the highest fitness at a given iteration; the best position of all *pbest* particles is called global best (*gbest*). Particles use their individual memory (*pbest*) and knowledge gained by the swarm as a whole (*gbest*) to move around a multidimensional search space until the stop condition is reached.

A standard GA has three main operators, namely a selection, crossover and mutation operator (John, 1975). In a first step, the chromosomes in a population are randomly initialized. Parents with a high fitness value have a greater probability of being selected by the selection operator. Two selected parents create two offsprings through the crossover operator. Then the mutation operator is applied to change the dimension randomly based on the mutation rate. Thus, each chromosome is updated at each generation

until an optimal chromosome is obtained or the computational limitations are reached.

In PPO, the BPSO parameters and the GA parameters can be freely set based on experimental requirements. In ODB, users can set the property parameters, such as intergenic distance, intergenic step, reaction step and others. The DOOR predictor, however, requests uploading of three datasets to predict operons. The prediction results are then sent to a user mailbox.

2.2 Biological properties

Several properties have been proposed to predict operons on prokaryotic organisms. These properties can be classified into the following five categories: intergenic distance, conserved gene clusters, functional relations, genome sequence and experimental evidence. In this study, the intergenic distance, metabolic pathway, cluster of orthologous groups (COG) gene function, gene length ratio and operon length property were employed to predict operons. Of these five properties, researchers can choose any or all based on their research requirements.

2.3 Visualization of prediction results

In order to easily evaluate the distribution of the prediction results, PPO provides graphic visualization of the results after the operon prediction process is completed. Five types of operon-related information are graphically depicted in as many charts, namely the operon size, the operon length, the intergenic distance in an adjacent genes within operons (WO) pair, the intergenic distance in a transcription units borders (TUB) pair and the intergenic distance in all gene pairs. Supplementary Figure 2 illustrates the relationship between the parameter settings and the prediction results. The graphic visualization allows researchers to set related parameters accurately and obtain better prediction results.

2.4 Tools supported

The PPO system also adds a motif search (Bailey *et al.*, 2006) and a Protein Basic Local Alignment Search Tool (PBLAST) interface. Users can set related parameters to predict the structure of similar operons. If an operon is similar to the one identified, it can be found by the conserved sequence motifs across the promoter sequences of the operons. In addition, genes within operons are often conserved between different genomes. Users can obtain the conserved pairs by applying PBLAST to sequences of specific genes. In the PBLAST interface, the related parameters, i.e, program and database parameters can also be freely set.

2.5 Database

We recently proposed a prediction algorithm called BPSO for operon prediction (Chuang *et al.*, 2010), which reached predicting accuracies of 92.1%, 93.3% and 95.9% on *Bacillus subtilis*, *Pseudomonas aeruginosa* PA01 and *Staphylococcus aureus*, respectively. The experimental results show that BPSO not only increased the prediction accuracy on the three genome datasets tested, but also obtained a good balance between sensitivity and specificity.

The PPO system uses MySQL (http://www.mysql.com/) as the database management system, and employs Apache on its web server to store and manage all operon information. Additionally, JSP and Java script were used to construct the web page and dynamics, respectively. The *Organism View* page shows the gene location, strand, length, PID, gene, synonym, code, COG and product of each organism. In addition, NCBI can be accessed via the PID hyperlink. This allows researchers to obtain detailed data for a specific gene. The *Operon View* page shows the operon ID, operon size, strand, PID, synonym and other data. Genes or operons of interest can be searched based on various conditions. Multiple conditions can be selected via AND or OR operations. After a target gene is found, the system shows information related to the gene on the web page. Operons can also be found in PPO based on different operon conditions, such as the operon ID, operon size, etc.

3 CONCLUSION

A novel platform for operon prediction called PPO is proposed, which provides a highly reliable predictor and a database for bioinformatics research. In this platform, the superior BPSO and the GA method are applied to predict operons, but two other methods from the literature can also be selected. The used prediction algorithm, biological properties and related parameters can be freely selected. The database collects information related to highly likely putative operons on 592 prokaryotic genomes and provides search functions for users. Ongoing efforts are made to enhance the algorithms, properties and literature-based operon data for PPO, thus providing a perfect operon prediction platform for researchers.

ACKNOWLEDGEMENTS

Funding: National Science Council in Taiwan (grant NSC96-2221-E-214-050-MY3, NSC96-2622-E-214-004-CC3 and NSC97-2622-E-151-008-CC2, in parts).

Conflict of Interest: none declared.

REFERENCES

Bailey,T. *et al*. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369.

Chuang,L. *et al.* (2010) Binary particle swarm optimization for operon prediction. *Nucleic Acids Res.*, **38**, e128.

Holland,J. (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI.

Kennedy,J. and Eberhart,R. (1995) Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 4, pp. 1942–1948.

Mao,F. *et al.* (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–D463.

Okuda,S. *et al.* (2006) ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res.*, **34**, D358–D362.

Pertea,M. *et al.* (2008) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, **37**, D479–D482.

Price,M.N. *et al.* (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.

Sierro,N. *et al.* (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.

Yada,T. *et al.* (1999) Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.