Bioinformatics, 34(23), 2018, 4118–4120
doi: 10.1093/bioinformatics/bty496
Advance Access Publication Date: 19 June 2018
Applications Note

Genome analysis

Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes

Blanca Taboada[1], Karel Estrada[2], Ricardo Ciria[2] and Enrique Merino[2,]*

[1]Department of Developmental Genetics and Molecular Physiology and [2]Department of Molecular Microbiology, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 4, 2018; revised on June 10, 2018; editorial decision on June 14, 2018; accepted on June 15, 2018

Abstract

Summary: Operon-mapper is a web server that accurately, easily and directly predicts the operons of any bacterial or archaeal genome sequence. The operon predictions are based on the intergenic distance of neighboring genes as well as the functional relationships of their protein-coding products. To this end, Operon-mapper finds all the ORFs within a given nucleotide sequence, along with their genomic coordinates, orthology groups and functional relationships. We believe that Operon-mapper, due to its accuracy, simplicity and speed, as well as the relevant information that it generates, will be a useful tool for annotating and characterizing genomic sequences.
Availability and implementation: http://biocomputo.ibt.unam.mx/operon_mapper/
Contact: merino@ibt.unam.mx

1 Introduction

In prokaryotes, it is common for metabolically or functionally related genes to be contiguously arranged in the genome and co-transcribed in the same polycistronic messenger RNA as a part of the same operon. As operons are biologically relevant in the regulation of gene expression, we have developed one of the most accurate algorithms for operon prediction to date (Taboada *et al.*, 2010). Our method is based on an artificial neural network (ANN) in which the inputs are the intergenic distances of contiguous genes and a score that reflects the functional relationships between the protein products. Our algorithm, when tested on a set of experimentally defined operons in *E.coli* and *B.subtilis*, reached accuracies of 94.6 and 93.3%, respectively (Taboada *et al.*, 2010). Compared to other algorithms, ours showed the highest correlations with experimentally validated operons in a recent evaluation (Zaidi and Zhang, 2017). Currently, the predicted operons of model organisms can be found in various databases (Mao *et al.*, 2009; Pertea *et al.*, 2009), including ours (Taboada *et al.*, 2012). Recent advances in sequencing technologies have made it possible for nearly any research group to determine the complete genome sequence of a particular bacterium in a fast, low-cost manner. For these newly sequenced or draft genomes, there is no easy way to predict their corresponding operons. Therefore, based on our published algorithm (Taboada *et al.*, 2010), we have developed Operon-mapper, a web server tool that can accurately and easily predict the operons of any bacterial or archaeal genome sequence.

2 Overview and implementation of the Operon-mapper web server

Operon-mapper was written in Perl. It generates HTML and JavaScript code ‘on the fly’ and integrates various sequence analysis software programs (described in the Section 3) in a Linux environment. The Operon-mapper runs on a 64 core/512 Gb of RAM server under Ubuntu Linux 16.04 LTS and is available at http://biocomputo.ibt.unam.mx/operon_mapper.

3 Results

The Operon-mapper web server, developed in Perl, consists of three main stages:

1. **Data acquisition.** This procedure is performed using a web page written in HTML and JavaScript. The only required input for

Table 1. Benchmark test of Operon-mapper using genomic sequences of different sizes and GC % contents

Organisms	Accession number	Size	GC%	Accuracy	
				NCBI ORFs	Predicting ORFs
B.subtilis	NC_000964	4216	43.5	94.1%	94.3%
C.glutamicum	NC_006958	3283	54.0	87.6%	85.3%
E.coli	NC_000913	4642	50.8	94.4%	94.4%
H.pylori	NC_00091	1668	38.9	93.1%	92.4%
L.monocytogenes	NC_003210	2944	38.0	91.6%	90.9%
L.pneumophila	NC_006368	3635	38.3	90.6%	90.1%
P.profundum	NC_006370	6403	42.0	92.1%	92.4%
S.solfataricus	NC_002754	2992	35.8	95.8%	96.1%

the operon prediction process is the genomic nucleotide sequence in FASTA format; however, the ORF genomic coordinates can also be provided by the user, either in General Feature Format (GFF) or GenBank format.

2. **Sequence analysis**. The analysis is divided into five different tasks.
   - 2.1) ORF prediction uses *Prokka* software, which employs dynamic programming to accurately predict the 5′ and 3′ ends of all the ORFs in the given nucleotide sequence (Hyatt *et al.*, 2010; Seemann, 2014).
   - 2.2) Homology gene assignments are determined based on Hidden Markov Models (HMMs) search using the *hmmsearch* program (Eddy, 2011). This HMMs search process employs a previously constructed model set that represents each of the 4873 COGs (Taboada *et al.*, 2010; Tatusov *et al.*, 2003) and 8539 Remained Orthologous Groups (ROGs) (Taboada *et al.*, 2010).
   - 2.3) The intergenic distance evaluation is determined based on the ORF coordinates using a custom Perl program.
   - 2.4) Operon prediction is performed with an ANN implemented in R. The network inputs of our ANN are the intergenic distance between the genes and a score that reflects the functional relationships of their corresponding protein products. These scores have been defined in the STRING database (Jensen *et al.*, 2009) or in our publication (1), and they are presented for different pairs of proteins according to their associated COG or ROG. This step represents the core process of Operon-mapper, where a confidence value is evaluated for a pair of genes that might be found in the same operon. This confidence value is normalized between 0 and 1. A value greater than 0.5 indicates that the gene pair belongs to the same operon. The confidence values with the greatest accuracies are near 0 or 1, and confidence values close to 0.5 have the lowest accuracies.
   - 2.5) Gene function assignments are based on the most significant hit using DIAMOND (Buchfink *et al.*, 2015) against a core set of well-characterized proteins from the Uniprot Knowledgebase (Apweiler *et al.*, 2004).
3. **Delivery of results.** A Perl program is used to build an HTML page where the user can choose the file or set of files with the results of the different analyses performed by Operon-mapper, including the following: i) the predicted operonic gene pairs with their corresponding confidence values for being found in the same operon; ii) a list of operons with their corresponding genes; iii) the coordinates of the predicted ORFs; iv) the DNA sequences of the predicted ORFs; v) the translated protein sequences of the predicted ORFs; vi) the homology assignments of the proteins, corresponding to their COG or ROG; vii) the functional descriptions of the proteins; viii) all the above output files at once; and ix) a compressed file with all the above output files. These results are shown on the web page once the analysis is finished and are sent to the email specified by the user.

As a benchmark test, Operon-mapper was used to predict the operons of eight genomes of different sizes and nucleotide GC contents. Table 1 shows the accuracy of our predictions considering two scenarios: i) when the genomic sequence is used as the only input information, and ii) when, in addition to the nucleotide sequence, the coordinates of the genes are also provided. In these two cases, the accuracy of Operon-mapper was evaluated by comparing its predictions to experimentally determined operons; these data were recently compiled in (Zaidi and Zhang, 2017).

4 Conclusions

Operon-mapper is the first publicly available, web-based tool that is designed to predict operons in bacterial and archaebacterial genomes with only their genomic sequences as a required input. Operon-mapper has several strengths, including its accuracy, simplicity and speed. In addition to predicting operons, Operon-mapper also generates useful, relevant information that is common to most bacterial genome annotation projects, such as the identification of ORFs in a nucleotide sequence, the assignment of COGs to each of the encoded proteins, and functional annotations of proteins. For these reasons, we hope that Operon-mapper quickly becomes a reference tool in the field of bacterial genome annotation.

Acknowledgements

The authors wish to thank Shirley Ainsworth for bibliographical assistance and Juan Manuel Hurtado for computer technical support.

Funding

This work was supported by Consejo Nacional de Ciencia y Tecnologia CONACyT [grants FC 2015-2 887 and 235817] and Dirección General de Asuntos del Personal Académico [grant IN202517].

Conflict of Interest: none declared.

References

Apweiler,R. *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.

Buchfink,B. *et al*. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol*., **7**, e1002195.

Hyatt,D. *et al*. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, **11**, 119.

Jensen,L.J. *et al*. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*., **37**, D412–D416.

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

Pertea,M. *et al*. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res*., **37**, D479–D482.

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Taboada,B. *et al*. (2012) ProOpDB: prokaryotic operon database. *Nucleic Acids Res*., **40**, D627–D631.

Taboada,B. *et al*. (2010) High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res*., **38**, e130.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, **4**, 41.

Zaidi,S.S. and Zhang,X. (2017) Computational operon prediction in whole-genomes and metagenomes. *Brief. Funct. Genomic*, **16**, 181–193.