Bioinformatics, 2016, 1–2
doi: 10.1093/bioinformatics/btw330
Advance Access Publication Date: 13 June 2016
Applications Note

OXFORD

Gene expression

CONDOP: an R package for CONdition-Dependent Operon Predictions

Vittorio Fortino[1,*], Roberto Tagliaferri[2] and Dario Greco[1]

[1]Institute of Biotechnology, University of Helsinki, Helsinki, Finland and [2]NeuRoNe Lab, DISA-MIS, University of Salerno, Fisciano, Italy

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on October 12, 2015; revised on March 22, 2016; accepted on March 23, 2016

Abstract

Summary: The use of high-throughput RNA sequencing to predict dynamic operon structures in prokaryotic genomes has recently gained popularity in bioinformatics. We provide the R implementation of a novel method that uses transcriptomic features extracted from RNA-seq transcriptome profiles to develop ensemble classifiers for condition-dependent operon predictions. The CONDOP package provides a deeper insight into RNA-seq data analysis and allows scientists to highlight the operon organization in the context of transcriptional regulation with a few lines of code.
Availability and Implementation: CONDOP is implemented in R and is freely available at CRAN.
Contact: vittorio.fortino@helsinki.fi
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcriptome profiling of bacterial organisms by high-throughput RNA-seq can be used to predict operon structures. We have recently proposed a new computational strategy that integrates RNA-seq transcriptome profiles with DNA-sequence based information to accurately predict operons in a condition/transcriptome-specific manner (Fortino *et al.*, 2014). Our method is capable of highlighting significant changes occurring in the operon map annotated in the DOOR database (Mao *et al.*, 2008) in specific dynamic conditions assayed at the transcriptome level by RNA-seq. To this end, an ensemble classification system employing genomic and transcriptomic features, is trained to distinguish operon pairs (OPs) from non-operon pairs (NOPs).

The ensemble classifier is based on three different machine-learning approaches: Neural Networks (NN), Support Vector Machines (SVMs) and Random Forests (RFs). The training of each model is based on a small set of OPs and NOPs that are systematically selected from the RNA-seq transcriptome profile and considered as gene pairs with a known operon status. The trained models are finally used to predict (or re-define) the operon status of the gene pairs that are excluded from the training and validation steps, namely Door Operon Pairs (DOPs), and those that are not annotated as operon pairs in the DOOR database, namely Putative Operon Pairs (POPs). The ensemble classification system is implemented with a simple majority-voting schema that combines the classification predictions of all the trained models. After re-classifying the adjacent gene pairs, a linkage process is compiled to find adjacent genes predicted as OPs and group them into operons. This process generates a new operon map namely condition-dependent operon map. Here we present the R software package CONDOP, an implementation of this method freely available at CRAN (https://cran.r-project.org).

2 Description

CONDOP requires four data inputs, namely (i) GFF file (*.gff*) - representing gene/feature annotations, (ii) DOOR file (*.opr*) – containing non condition-dependent operon annotations retrieved from the DOOR database, (iii) the genome sequence of the target organism (*.fasta*) and, (iv), a raw count table extracted from a specific RNA-seq transcriptome profile. The raw count table must correspond to a data frame having two columns (or coverage vectors) indicating the read depth value for each genomic position on the forward and reverse strand respectively. These files are given in input to the function *pre.proc()* that accomplishes the following tasks: (i) it removes coverage depth from a user-specified type region (e.g. rRNA, tRNA, etc.), (ii) it quantifies the transcription abundance (RPKM values) for the coding

sequence and intergenic regions namely the CDSs and IGRs, (iii) it calculates cutoffs to distinguish low expressed RNA-seq data from high expressed regions on the forward and reverse strand and (iv) it identifies the start and end points in transcription.

The pre-processed information is then used as input to the function *run.CONDOP()*, which develops an ensemble classifier combining both genomic and transcriptomic features to classify operon pairs. Each gene pair is characterized by a set of features extracted from the genome sequence (the fasta file) and the RNA-seq transcriptome profile. The condition-dependent operon predictions are defined with the following tasks: (i) start and end points in transcription associated with operons annotated in DOOR are considered, they are respectively named with OSPs and OEPs, (ii) confirmed operons are then determined and the datasets of OPs and NOPs are built, (iii) operon classification using genomic and transcriptomic features of confirmed OPs and NOPs is performed, (iv) the operon status of putative operon pairs (POPs) is predicted and (v) the condition dependent operon map is compiled and exported as a tab-delimited file.

3 Results

Here we show how to use CONDOP to identify changes in the operon organization of *Escherichia coli* under two different growth conditions (Fig. 1).

```
library(CONDOP)
data(ct1)
ga = system.file("extdata", "NC_000913.gff", package="CONDOP")
da = system.file("extdata", "1944.opr", package="CONDOP")
in.objs = pre.proc(gff.file = ga, door.op.file = da, "NC_000913",
            list.cov.dat = list(ct1))
res = run.CONDOP(data.in = in.objs,
                 bkgExprCDS = 0.2, bkgExprIGR = 0.2,
                 cl.run = 30, nfolds = 5, cons = 2,
                 find.ext = TRUE, return.all = FALSE,
                 save.TAB.file = "condop")
```

Escherichia coli K12 wild-type cells (MG1655) were grown in LB medium with and without α-methylglucoside (McClure *et al.*, 2013). BOWTIE (Langmead *et al.*, 2009) and SAMtools (Li *et al.*, 2009) were used for the RNA-seq read alignment and the generation of the count table data files respectively. The necessary inputs are included into the CONDOP R package. The *pre.proc()* function compiles the input data structures for the main function *run.CONDOP()*, which is then used to build the condition dependent operon map. An important step of the *run.CONDOP()* function is the selection of the 'confirmed operons' (Fortino *et al.*, 2014), upon which the set of OPs and NOPs is built. The users can influence this selection by properly modifying the thresholds (bkgExprCDS and bkgExprIGR) used for defining the detectable expression above the background (Supplementary Material S1). The set of OPs and NOPs is used for the training and validation of three different classification models, which predictions are combined by using a simple voting schema. The users can exploit the parameters *cons* to specify the minimum number of positive votes to declare a gene pair as operon pair. For instance, with *cons* equals to 2, the voting system tags a gene pair as an OP when at least two classifiers have predicted that gene pair as an OP. All the models are trained and evaluated by using cross-validation and bootstrap. Users can indicate the number of folds and bootstrapping iterations by setting the *nfolds* and *cl.run* parameters respectively. Further details about the parameter setting and the description of the data inputs/output of the CONDOP functions are provided as supplementary materials (Supplementary Material S1).

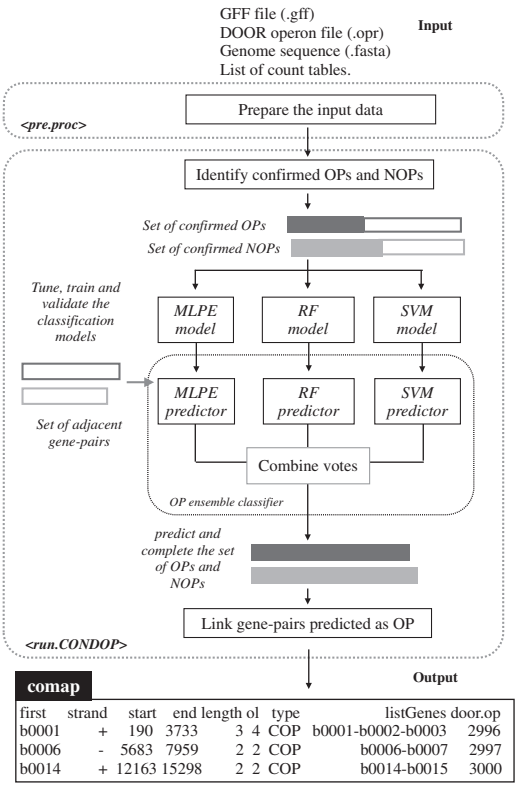

Fig. 1. CONDOP workflow – the function *pre.proc()* provides the necessary data structures for the main function *run.CONDOP()*. It develops an ensemble operon pair classifier that combines genomic and transcriptomic features. The ensemble classifier consists of three machine-learning models that are trained on a small set of confirmed operon pairs (OPs) and non-operon pairs (NOPs). The OPs and NOPs are extracted from 'confirmed' operons annotated in the DOOR database. The confirmed operons are systematically found by searching for start and end points in transcription grouping consecutive, active coding-sequence and intergenic regions, indicated with CDSs and IGR respectively. The trained ensemble classifier is used to predict the operon status of all gene-pairs including DOOR-based operon pairs, namely DOPs, and putative operon pairs (POPs). Finally, a linkage process is exploited to combine consecutive predicted operon-pairs and, so, build the map of condition-dependent operons namely comap

Funding

This work has been supported by the European Commission, under grant agreement FP7-309329(NANOSOLUTIONS) and by the Academy of Finland, under grant agreements 275151 and 292307.

Conflict of Interest: none declared.

References

Fortino,V. *et al.* (2014) Transcriptome dynamics-based operon prediction in prokaryotes. *BMC Bioinformatics*, **15**, 145.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*,**25**, 1–10.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mao,F. *et al.* (2008) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, 459–463.

McClure,R. *et al.* (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.*, **41**, e140.