EuGene-PP: A Next Generation Automated Annotation Pipeline for Prokaryotic Genomes

E. Sallet, J. Gouzy and T. Schiex

1 Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594, INRA/CNRS, F-31320 Castanet Tolosan F-31326, France
2 INRA, Unité de Mathématiques et Informatique Appliques de Toulouse, UR 875, Castanet-Tolosan F-31326, France

ABSTRACT
Summary: It is now easy and increasingly usual to produce oriented RNA-Seq data as a prokaryotic genome is being sequenced. However, this information is usually just used for expression quantification. EuGene-PP is a fully automated pipeline for structural annotation of prokaryotic genomes integrating protein similarities, statistical information and any oriented expression information (RNA-Seq or tiling arrays) through a variety of file formats to produce a qualitatively enriched annotation including coding regions but also (possibly antisense) non-coding genes and transcription start sites.

Availability: EuGene-PP is an Open Source software based on EuGene-P (Sallet et al., 2013) integrating a Galaxy configuration. EuGene-PP can be downloaded at eugene.toulouse.inra.fr.

Contact: erika.sallet@toulouse.inra.fr

1 INTRODUCTION

Prokaryotic genome sequencing and expression quantification using RNA-Seq or tiling arrays is becoming routine. However, existing prokaryotic gene finders are either ab initio gene finders that identify only coding regions (Hyatt et al., 2010; Delcher et al., 2007) or purely RNASeq based gene finders predicting transcripts (Martin et al., 2010) and which are much less effective than their ab initio competitors for CDS prediction (Zickmann et al., 2013). Reconciling conflicting predictions is a tedious work, which is incompatible with the growing prokaryotic genome sequencing rate.

There is a need for new prokaryotic gene finders that would directly integrate all available information to produce an enriched and precise structural annotation identifying CDS but also (possibly antisense) non coding genes and transcription start sites (TSSs), avoiding tedious reconciliation. We have shown in (Sallet et al., 2013) that this can be accomplished by finely adapting Conditional Random Field based integrative eukaryotic gene finding technology (Foissac et al., 2008) to prokaryotic specificities (overlapping genes, operons...). The resulting software, EuGene-P, simultaneously exploits statistical properties of sequences, existing annotations, similarities to proteins and oriented RNASeq data to produce an enriched annotation with a better delineation of functional genomic elements, and therefore improved expression quantification.

2 APPROACH

To facilitate the application of EuGene-P, we designed a fully automatic pipeline which allows any user to directly apply EuGene-P starting just from genomic sequences and oriented sequence-based expression data (RNASeq or tiling array). The resulting Perl-based EuGene-PP pipeline has no parameter to tune (by default), accepts a variety of protein and expression data-sets of different types under most usual formats and feeds EuGene-P with:

- Markov models of coding regions trained on regions with strong similarities with a reference protein databank.
- regions of similarity with different protein databanks.
- a set of CDS predictions produced by a reliable self training ab initio gene finder. We use Prodigal (Hyatt et al., 2010).
- a set of predicted non coding RNA genes (ncRNA). We use dRNAscan-SE (Lowe and Eddy, 1997), rRNAmer (Griffiths-Jones, 2005) and RNAmmer (Lagos et al., 2007).
- a set of thresholded and rescaled profiles of measured expression on each strand along the genome (either RNASeq or tiling arrays) showing transcription.
- a set of potential transcription start sites, defined as points of sudden increase in expression identified by derivative of a smoothed version of the expression profile (Sallet et al., 2013).

EuGene-PP can run using just FASTA genomic sequences and expression data (provided as oriented Single/Pair-end reads in either FASTQ/FASTA format, mapped reads in Bam/Sam, Bed or Wig format, or tiling array data in ndf/pair files). Protein databanks for similarity detection are configurable and may include organism-specific proteomes. EuGene-PP is provided with a Galaxy configuration Goecks et al. (2010) to deploy EuGene-PP through a web interface.

The probabilistic model used by EuGene-P (Sallet et al., 2013) integrates all this information and its own RBS predictions to segment the genome in possibly overlapping coding regions,
UnTranslated Regions (UTR) and non-coding genes. The integration of expression data leads to more reliable transcripts and TSSs prediction. Prediction is performed independently on each strand allowing for the prediction of anti-sense genes.

3 RESULTS AND DISCUSSION

In (Sallet et al., 2013), we showed how EuGene-P performs when applied to the genome of the symbiont bacteria S. melloti and associated oriented RNA-Seq data. Besides its 6,308 CDS, the produced annotation contains 1,876 ncRNA genes. These ncRNA predictions, with a mean length of 107 nt, cover a large fraction of already characterized or candidate ncRNA genes. Furthermore, by looking for specific RpoE2-binding sites upstream of predicted TSSs, the S. melloti RpoE2 regulon could be extended by 3-fold, showing the added-value of predicted TSSs.

To complete this application of EuGene-P, we decided to compare the fully automated annotation produced by EuGene-PP to a recently published curated annotation of the model bacteria B. subtilis. This annotation is based on a number of condition-specific expression measures based on tiling arrays (Nicolas et al., 2012). For CDS, the two annotations were highly consistent, with more than 97% of shared CDSs (same STOP). This is consistent with the reliability of Prodigal ab initio CDS prediction. We therefore focused our evaluation on ncRNA transcription prediction. We used rfam_scan to produce a set of 207 reference ncRNA genes. We separately applied EuGene-PP using a subset of all tiling-arrays and removing all input from rfam_scan, RNAmmer or tRNAscan-SE. The comparison of the automated annotation of EuGene-PP with the curated annotation of Nicolas et al. (2012) w.r.t. to this reference gene set is given in Table 1. On three additional genomes, EuGene-PP recovers a similar fraction of the reference genes.

Table 1. For each annotation, we report the percentage of shared stops, the number of ncRNA genes (RFam and predicted), the total number of bases represented, the number of reference genes covered on more than 50% of their length by predicted genes (cover) or with a reciprocal hit covering at least 50% of both regions (recip.). Both annotations show comparable quality albeit length by predicted genes (cover) or with a reciprocal hit covering at least 50% of both regions (recip.). Both annotations show comparable quality albeit length by predicted genes (cover) or with a reciprocal hit covering at least 50% of both regions (recip.).

<table>
<thead>
<tr>
<th></th>
<th>Shared # ncRNA size &gt; 50% &gt; 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CDS Rfam pred. (kbp) cover recip.</td>
</tr>
<tr>
<td><strong>B. subtilis</strong></td>
<td></td>
</tr>
<tr>
<td>EuGene-PP</td>
<td>97% 207 2,492 817 98 55</td>
</tr>
<tr>
<td>Nicolas et al., 2012</td>
<td>95% 160 503 71 66</td>
</tr>
<tr>
<td>S. avermilitis, (Moody et al., 2013)</td>
<td></td>
</tr>
<tr>
<td>EuGenePP</td>
<td>95% 162 166 20 56 34</td>
</tr>
<tr>
<td><strong>E. coli, (Li et al., 2013)</strong></td>
<td></td>
</tr>
<tr>
<td>EuGenePP</td>
<td>96% 263 145 20 97 61</td>
</tr>
<tr>
<td><strong>S. enterica. (Kröger et al., 2013)</strong></td>
<td></td>
</tr>
<tr>
<td>EuGenePP</td>
<td>96% 290 3456 299 146 86</td>
</tr>
</tbody>
</table>

These results also show the flexibility of EuGene-PP that exploits a variety of information sources, under most usual formats, to produce an annotation comparable to a curated semi-automated structural annotation, especially on ncRNA genes which are still very difficult to predict.

ACKNOWLEDGEMENT

**Funding:** This work was supported by the ANR-08-GENO-106 grant "SYMbiMICS", in TULIP LabEx (ANR-10-LABX-41).

REFERENCES


Fig. 1. A diagram describing the input and formats accepted by EuGene-PP and how information is prepared for EuGene to produce a final GFF3 annotation.
