Integrated Gene Prediction for Prokaryotic Genomes using EuGene

Erika SALLET[1], Jérôme GOUZY[1], Brice ROUX[1], Delphine CAPELA[1], Laurent SAUVIAC[1], Claude BRUAND[1], Pascal GAMAS[1] and Thomas SCHIEX[2]

[1] Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594, INRA/CNRS
{Erika.Sallet,Jerome.Gouzy}@toulouse.inra.fr
[2] Unité de Biométrie et d'Intelligence Artificielle UR 875, INRA, F-31320 Castanet Tolosan, France
Thomas.Schiex@toulouse.inra.fr

Abstract *With the advent of new generation sequencing, the annotation of new prokaryotic genomic sequences will occur in a data-rich context, including a variety of libraries of short reads of transcriptomic sequences. This rich context creates new potentialities in annotation. In this paper, we describe the new prokaryotic variant of the integrative gene prediction software EuGene. By leveraging RNA-Seq data, EuGene becomes capable of predicting new functional structures, including RNA genes and untranslated transcribed regions inside operons.*

Keywords Gene prediction, RNA gene, NGS, RNA-Seq.

1 Introduction

Following the initial development of gene prediction tools for prokaryotic genomes, the complexity of eukaryotic gene prediction led to the development of highly integrative gene prediction tools. Very few, if any, prokaryotic gene prediction tools have evolved along the same line, mostly because prokaryotic protein gene structures are simple and defined by open reading frames. Other dedicated tools have however been designed for the prediction of other functional (transcribed) elements such as RNA genes.

Through RNA-Seq short reads, new generation sequencing gives unprecedented access to transcriptomic data. In genomes with low gene density, such as plant or animal genomes, the availability of such transcribed sequences sampling is extremely useful to delineate gene structures. In bacterial genomes, most if not all the genome is transcribed, making such data much less easy to exploit. However, NGS technology is able to produce oriented read for which the strand of transcription is known. Such data facilitates the automatic prediction of a variety of transcribed elements, including protein genes, (possibly antisense) RNA genes and operon structures.

2 Changing EuGene Gene Model

EuGene is an eukaryotic gene finder [1,2] that can be described as a Conditional Random Field (CRF) predictor [3], a variant of random Markov fields capturing the conditional probability of structural annotations given available evidence. The default gene model of EuGene includes intergenic regions, coding exons, introns, 5'/3' untranslated terminal regions and introns within UTRs.

To be able to predict new functional elements in prokaryotes, the gene model underlying EuGene has been extensively modified. In the absence of splicing, intronic and spliced exonic states have been removed (overall 34 states removed). Conversely, new states have been introduced to capture:

- overlapping protein gene regions (on the same strand or not) on any of the 6 different coding frames.
- untranslated transcribed internal regions (UIR) between non overlapping gene appearing in the same operon on either strand. These new states complete the existing 5' and 3' UTR (untranslated terminal regions) defining operon extremities.
- and finally RNA genes on either strand.

Overall, the new prokaryotic variant of EuGene includes 30 states, compared to the 45 origianl states.

This work has been partially funded by the french "Agence nationale de la Recherche" SYMbiMICS project.

3 Integrating Evidence

In EuGene, each "feature", representing a specific type of evidence used for prediction is weighted in the CRF model and integrated through independent software plugins. We just integrated the prokaryotic translation start predictor of FrameD [4,5], based on RBS/ribosomal RNA hybridation energy, as a new plugin to get a fully functional prokaryotic gene finder capable of predicting protein genes, RNA genes and operons.

We are experimenting with the integration of oriented RNA-Seq data through existing generic plugins, either directly or following a segmentation based on the level of transcription. In the simplest variant, partial transcripts defined by oriented pair-end short reads are mapped to the genome. Their abundance at a given position is used as a weighted feature that indicates that the current region is transcribed on the corresponding strand. By integrating translation/transcription start and stop prediction, statistical models of different regions (especially coding regions) and RNA-Seq data inside a unique tool, EuGene becomes capable of discriminating protein genes (which are transcribed and follow a coding region statistical model) from RNA genes (which are transcribed but do not follow a coding model). In some sense, this is related to the QRNA [6] comparative RNA gene predictor which relies on a stochastic context free grammar model for RNA genes and a usual 3-periodic Markov model for coding regions. In a non comparative settings, we use a simple homogeneous Markov model for RNA regions instead but the integration of oriented RNA-Seq restricts the discrimination between coding and RNA genes to transcribed regions.

Similarly, a "stable" expression level inside a region, in several different conditions, identified through prior segmentation, should help delineate operons. This information can be directly injected inside EuGene as a feature informative about transcription start/stop but has not been evaluated yet.

Most, if not all, eukaryotic gene finders assume that only one strand is transcribed at a given position. To overcome this limitation, EuGene has been slightly modified to allow to perform independent gene prediction on each strand. Together with oriented RNA-Seq data, this allows to perform an automatic annotation that includes protein genes, RNA genes but also anti-sense RNA (RNA gene predicted on one strand overlapping a gene predicted on the other strand).

We are currently applying this new strand-independent prokaryotic variant of EuGene to the genome of *Sinorhizobium meliloti* using oriented RNA-seq data (representing 48Gb of reads). The results we have obtained closely match the existing genome annotation (with 6483 genes predicted compared to the 6235 annotated) and show that EuGene correctly identifies ribosomal and transfer RNA genes and many potentially new RNA genes (2040 ncRNA predicted compared to the 64 annotated ones). These new genes, predicted without any specific RNA related information (except for RNA-Seq and 3-periodic Markov coding models), needs to be experimentally evaluated.

References

[1] T. Schiex, A. Moisan, and P. Rouzé, Eugène: an eukaryotic gene finder that combines several sources of evidence. In M. Sagot, editor, *Selected papers from JOBIM'2000*, volume 2066 of *LNCS*, pages 118–133. Springer Verlag, 2001.

[2] S. Foissac, J. Gouzy, S. Rombauts, C. Mathe, J. Amselem, L. Sterck, Y. de Peer, P. Rouzé, and T. Schiex, Genome annotation in plants and fungi: EuGene as a model platform. *Current Bioinformatics*, 3(2):87–97, 2008.

[3] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Machine Learning International Workshop*, pages 282–289, 2001.

[4] T. Schiex, P. Thébault, and D. Khan, Recherche des génes et des erreurs de séquençage dans les génomes bactériens GC-riches. In O. Gascuel and M.-F. Sagot, editors, *Proc. of JOBIM'2000 (Journées Ouvertes Biologie Informatique Mathématiques)*, Montpellier, France, 2000.

[5] T. Schiex, J. Gouzy, A. Moisan, and Y. de Oliveira, FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*, 31(13):3738–41, 2003.

[6] E. Rivas and S. R. Eddy, Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.