BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 9 2010, pages 1254–1255
doi:10.1093/bioinformatics/btq123

*Genome analysis*

# EuGène-maize: a web site for maize gene prediction

Pierre Montalent and Johann Joets*
INRA, UMR 0320 / UMR 8120 Génétique Végétale, F-91190 Gif-sur-Yvette, France
Associate Editor: Dmitrij Frishman

**ABSTRACT**

**Motivation:** A large part of the maize B73 genome sequence is now available and emerging sequencing technologies will offer cheap and easy ways to sequence areas of interest from many other maize genotypes. One of the steps required to turn these sequences into valuable information is gene content prediction. To date, there is no publicly available gene predictor specifically trained for maize sequences. To this end, we have chosen to train the EuGène software that can combine several sources of evidence into a consolidated gene model prediction.

**Availability:** http://genome.jouy.inra.fr/eugene/cgi-bin/eugene_form.pl

**Contact:** joets@moulon.inra.fr

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

Received on November 10, 2009; revised on March 17, 2010; accepted on March 18, 2010

## 1 INTRODUCTION

The B73 maize genome sequence is now available (Schnable *et al*., 2009). We can anticipate that next generation sequencing technologies will soon supply a deluge of genomic data from other maize genotypes. Therefore, in addition to the annotation provided by the maize sequence consortium for the B73 genotype, the community will also need a tool for the annotation of maize genomic sequences produced from other genotypes.

To date, the www.maizesequence.org web site provides the Filtered Gene Set including annotation of 32 540 gene models (RefGen_v1), based on biological evidence. However, no tool is provided to annotate personal sequences yet.

Fgenesh (Salamov and Solovyev, 2000) was among the first *ab initio* gene prediction softwares available for maize while it was trained for monocot species. Combiner softwares like EuGène (Foissac and Schiex, 2005) can improve their own *ab initio* prediction results by integrating information from sequence alignment software, from splice site and translation start site prediction software or from other gene finder algorithms, thereby improving prediction quality. EuGène uses probabilistic Markov models to discriminate coding sequences from non-coding ones, or genuine splice sites from false ones. Gene models generated by EuGène are associated with a score based on all the available information. In order to calculate the weight for each information source and to calibrate its *ab initio* prediction module, EuGène was trained using a maize-curated gene sequence set built in this study.

*To whom correspondence should be addressed.

## 2 METHODS

To build cognate gDNA/cDNA pairs, 6700 (4000 BAC) maize genomic sequences and 5500 full-length cDNA (FLcDNA) were extracted from the NCBI databases and filtered using an automatic pipeline designed on-site. First, cDNA sequences were trimmed using Seqclean (http://www.tigr.org/) and the Univec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Then cDNA redundancy was removed based on an 'all against all' cDNA BLAST analysis (*E*-value threshold 1e-100). Non-redundant cDNAs were then aligned onto the BAC sequences using BLAT (Kent, 2002). To avoid ambiguous alignments such as cDNA mapping to several gDNAs, a spliced alignment was computed using GenomeThreader (Gremme *et al*., 2005), and the best alignment was retained as the correct gDNA/cDNA pair. Next, for each gDNA/cDNA pair, the coding sequence was determined from an alignment with the corresponding protein from maize or rice. Each of the 247 validated pairs was manually checked before inclusion into the training set.

The third-party prediction softwares used by EuGène maize are SpliceMachine (Degroeve *et al*., 2005; donor and acceptor splicing site prediction), GenomeThreader (spliced alignment), BlastX (Altschul *et al*., 1997; protein alignment) and optionally Fgenesh (*ab initio* gene prediction). The sequence database used for prediction contains 69 306 maize FLcDNAs from GenBank, 20 508 rice mRNAs from RAPdb, 342 491 maize PlantGDB-assembled Unique Transcripts (PUT) from PlantGDB, 593 maize proteins from Uniprot/SwissProt, 19 836 rice proteins from RAPdb and 26 751 *Arabidopsis* proteins from TAIR (see the EuGène-maize web site for an updated listing of sequence resources and corresponding full references).

## 3 RESULTS

### 3.1 Gene prediction assessment

The training gene set was compared with 330 curated genes (Haberer *et al*., 2005) and was found to be representative of maize genes (Table 1). To assess EuGène-maize, we performed

**Table 1.** Comparison of several maize gene set statistics

| | Training set[a] | Curated set[b] | Maize all[c] | Maize cDNA[d] |
|---|---|---|---|---|
| Genes | 247 | 330 | 32 540 | 20 867 |
| Av. gene size (kb) | 3.5 | 4 | 3.7 | 3.5 |
| Exons | 1321 | 1520 | – | – |
| Av. no. of exon/gene | 5.4 | 4.6 | 5.3 | 4.7 |
| Av. exon size (kb) | 0.22 | 0.25 | 0.3 | 0.3 |
| Av. intron size (kb) | 0.52 | 0.6 | 0.52 | 0.58 |
| G + C gene (%) | 47.6 | – | 47.1 | 47.1 |
| G + C exon (%) | 53.4 | 55.4 | 52.7 | 53.4 |
| G + C intron (%) | 42.3 | 42.3 | 42.1 | 42.5 |

[a]The curated maize gene training set built in this study.
[b]A maize set of curated genes from Haberer *et al*. (2005).
[c]All maize genes in the B73 RefGen_v1 filtered set.
[d]Maize gene models supported by FLcDNAs are from Schnable *et al*. (2009).

**Table 2.** EuGène-maize and GeneBuilder assessment comparison

| | Missed Loci | Exon | |
|---|---|---|---|
| | | Se (%) | Sp (%) |
| Genebuilder B73 RefGen_v1 | 4 | 79 | 74 |
| EuGène-maize | 1 | 73 | 84 |

Se, sensitivity (fraction of actual exons predicted among total actual exons); Sp, specificity (fraction of actual exons predicted among total predicted exons).

gene prediction on eight BACs (AC211245, AC190915, AC204601, AC186187, AC211225, AC193983, AC200414 and AC194325) for which manually curated annotations of 42 genes are available (Liu *et al*., 2007). We compared these results (Table 2) with predictions from GeneBuilder (Liang *et al*., 2009) B73 RefGen_v1 (Schnable *et al*., 2009). Nearly all loci are detected by both predictors; however, GeneBuilder missed several mono-exonic genes. Exon-level assessment shows that the GeneBuilder is more sensitive yet less specific than EuGène. A gene containing 18 exons was incorrectly split by both tools (GRMZM2G119544_E01 and GRMZM2G119496_E01). Another gene (GRMZMM2G086779) containing four exons was split by EuGène only. In two other instances (GRMZM2G520535 and GRMZM2G177098) GeneBuilder incorrectly merged two adjacent genes, whereas EuGène failed only once.

### 3.2 Web access

EuGène-maize is available online (see Availability section). Genomic sequences can be masked prior to the prediction step. Masking is computed by RepeatMasker (A.F.A. Smit *et al.*, unpublished data) using the mips Repeat Element Database (Redat 4.3) (Spannagl *et al*., 2007). RepeatMasker low complexity masking option is disabled. The user may also submit, if available, the output file from the Fgenesh software (version 2.4). The gene prediction computation takes $<5$ min for a 200 kb genomic sequence and $>1$ h if the RepeatMasker option is enabled.

The results are compressed into an archive file and e-mailed to the user. The archive contains the parameters and options used for prediction, the submitted sequence, the masked sequence (if relevant), the annotation file (gff, gff3 and fasta format) and a HTML file that allows results to be displayed by a web browser.

## ACKNOWLEDGEMENTS

The authors greatly appreciate the advice of Thomas Schiex and Jérôme Gouzy, technical support from Christophe Caron and Veronique Martin from MIGALE Plateforme and careful reading by Delphine Vincent. The curated annotation of the 42 maize gene set was kindly provided by Dr Clémentine Vitte.

*Funding*: French 'Agence nationale de la recherche' (ANR); Génoplante BIEP program.

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Degroeve,S. *et al*. (2005) SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics*, **21**, 1332–1338.

Foissac,S. and Schiex,T. (2005) Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics*, **6**, 25.

Gremme,G. *et al*. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Soft. Technol.*, **47**, 965–978.

Haberer,G. *et al*. (2005) Structure and architecture of the maize genome. *Plant Physiol*., **139**, 1612–1624.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.

Liang,C. *et al*. (2009) Evidence-based gene predictions in plant genomes. *Genome Res*., **19**, 1912–1923.

Liu,R. *et al*. (2007) A GeneTrek analysis of the maize genome. *Proc. Natl Acad. Sci. USA*, **104**, 11844–11849.

Salamov,A.A. and Solovyev,V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res*., **10**, 516–522.

Schnable,P.S. *et al*. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Spannagl,M. *et al*. (2007) MIPS plant genome information resources. *Methods Mol. Biol.*, **406**, 137–159.