EuGène

A Simple yet Effective Gene Finder for Eucaryotic Organisms (*Arabidopsis thaliana*)

T. Schiex, A. Moisan, L. Duret and P. Rouzé
(tschiex@toulouse.inra.fr)

Motivation

It is standard, in a thorough sequence annotation, to take into account several sources of evidence in order to try to precisely locate genes (exons/introns) in eucaryotic sequences. The sources exploited typically include:

- matches against databases (cDNA and protein databases);
- output of splice sites or translation start prediction software;
- more or less sophisticated "integrated" gene finding software, eg. GeneMark.hmm [4].

None of these sources of evidence is, alone, sufficient to decide gene locations and the manual integration of all these data is a painful and extremely slow work. The motivation of our work is, as far as possible, to automate this job using *Arabidopsis thaliana* as a first test organism.

Methods

Along this line of idea, we have designed a simple, general, efficient and yet effective graph-based approach for gene finding that allows to combine several sources of evidence. Rather than directly combining the output of existing gene finding software (as in [5]) we decided to combine the information at the lowest level in order to be able to:

1. maintain the consistency of the prediction;
2. globally assess the impact of each local choice w.r.t. all available evidence.

Given a raw DNA sequence, the basic idea is to build a directed acyclic weighted graph such that all possible consistent gene structures are represented by a path in the graph. The gene structure currently used in **EuGène** is the simplest reasonable structure, the only signals taken into account being ATG, stops and splice sites.

The directed acyclic graph used for a simple ad-hoc sequence is illustrated above. It is a series-parallel graph with 13 different tracks that correspond respectively to the 6 forward/reverse coding frames, 6 forward/reverse intronic phases and a non coding track. Each signal occurrence, between two successive nucleotides, generates one or more "switches" between two parallel tracks. Each source-sink path defines a sequence of consistent genes structures (which may be partial on the limits of the sequence). The size of the graph is in $O(n)$ where n is the sequence length.

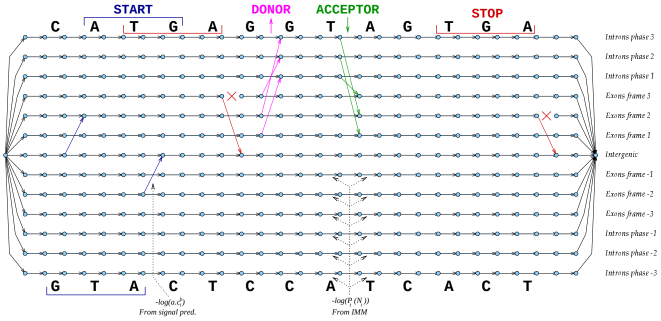

Figure 1: The DAG explored by **EuGène** for a simple sequence

To choose one path among the $O(13^n)$ paths in this graph, each edge e is weighted by a positive number w_e in such a way that shortest paths in the graph correspond to gene structures that "best respect" the available evidence. A probabilistic interpretation of this model can be given as follows: each edge e has a probability of existence P_e. Under simple independence assumptions, the reliability of a source-sink path is simply defined by the product of all the P_e in the path. A most reliable path is then a shortest path in the graph where $w_e = -log(P_e)$. The approach is comparable (although not equivalent) to an explicit state duration HMM with uniform duration densities (see [9], pp. 270) with a non-homogeneous transition matrix between hidden states (our tracks).

Given an adequate graph, a simple linear time, linear space shortest path algorithm such as Bellman's algorithm can output the best possible gene structure. We use a slightly more sophisticated algorithm that can take into account constraints on the minimum length of some gene elements (introns, single exon genes, intergenic regions). We think that this algorithm is still in $O(n)$ although a proof is needed to be affirmative.

The first version of our prototype, called **EuGène** I, integrates the following sources of information:

- output of five interpolated Markov models (IMM, [10]) for respectively frame 1, 2, 3 exons, introns and intergenic sequences. These models have been estimated on the AraClean v1.1 dataset [8]. Given the sequence, the IMMs allow to compute the probability $P_t(N_i)$ that the nucleotide N_i at position i appears on each track t. The corresponding edge is weighted $-log(P_t(N_i))$ (see Figure 1).
- output of existing signal prediction software. These software typically output a so-called "confidence" $0 \leq c_i \leq 1$ on the fact that a possible signal occurring at position i is used (i.e., the corresponding switch used). This confidence cannot decently be interpreted directly as a probability. We make the assumption that the switch's weights have the parametric form $-log(a.c_i^b)$ where the constants a and b have to be estimated for each source of evidence (see Figure 1).

To estimate these parameters, we have simply maximized the percentage of correct predictions on the same learning set (Araclean 1.1) using a simple genetic algorithm. A second version, called **EuGène** II can use, in conjunction with these basic information, results from cDNA and protein databases search:

- cDNA alignments in conjunction with splice sites are used to modify the graph as follows: matches (resp. gaps) delete intronic (resp. exonic) tracks in the graph. This forbids paths that would be incompatible with the cDNA data.
- similarly, **EuGène** II can exploit protein matches. However, one cannot be confident enough in such information to directly use matches/gaps as constraints and we therefore simply modify the $P_t(N_i)$ using a simple pseudo-count scheme.

In practice, the structure and weights of the graph can be directly modified by the user using a very simple language that allows to include information about starts, splice sites, exonic/intronic/intergenic tracks on a per nucleotide basis.

EuGène in action

In this section we show how **EuGène** works in practice by applying it to the contig 38 from the Araset dataset [6] which contains two genes with respectively 3 and 13 exons. We first collect information about splice sites and ATG by submitting the sequence to NetGene2 [11], NetPlantGene [3], SplicePredictor [1], NetStart [7]...using a dedicated Perl script. This automatically builds a file containing positions and strengths of "switches" in the graph. This file, and the Perl script can be simply modified by the user to include other sources of evidence if desired. For example, the file contains sentences like "**start r4 vrai 3.7e-02 nocheck**" which states that there is a reverse start at position 4 and the weight of the corresponding edge should be *-log*(*0.037*). The "**nocheck**" indicates that the user does not want **EuGène** to verify at the end that this ATG has been effectively used in the prediction.

We then start **EuGène** and ask for a graphical zoom from nucleotide 3001 to 7000 (region of the second gene according to Araset's annotations). On this sequence, **EuGène** I perfectly locates all exons/introns border of the 2 genes. **EuGène** outputs images in PNG or GIF format which can directly be used on Web pages. The X-axis is the sequence. The Y-axis represents successive "tracks": reverse introns, frame -3, -2, -1 exons, intergenic sequences (this includes UTR), frame 1, 2, 3 exons and forward introns.

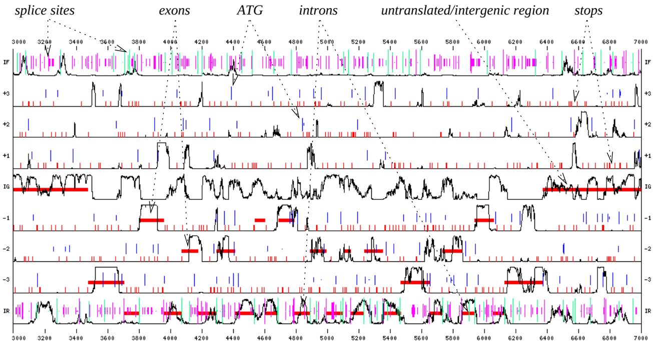

Figure 2: **EuGène** I applied to contig 38 of Araset

On each track, the black curve represents the output of the IMM models smoothed over a window of 100 nucleotides and normalized. The large red blocks represent **EuGène**'s prediction. On the exonic tracks alone, small vertical red bars represent potential stops and blue vertical bars represent potential starts (ATG), the height of the bar being representative of the quality of the ATG according to the available evidence. On the intronic tracks, green/magenta bars represent donors/acceptors. Again, the height of the bar is representative of the quality of the splice site.

EuGène I is not always as successful, eg. on the sequence of a tRNA synthetase (SYNO_ARATH), where **EuGène** I misses 2 exons and chooses wrong splice sites for 2 others. However, enough cDNA data exists so that **EuGène** II is able to unambiguously locate all exons/introns borders. On the graphic below, this additional information is provided: the intronic tracks show blue blocks that represent cDNA matches interconnected by thin lines that represent gaps (connecting splice sites). In this case, no protein database information is used.

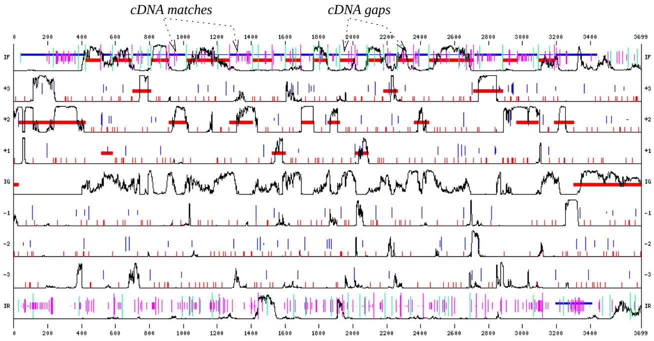

Figure 3: **EuGène** II applied to a tRNA synthetase with cDNA data

Evaluation

The "Araset" dataset has originally been designed and used to assess several existing gene or signal finding software (full paper presented at this conference). The clear winner of this evaluation is GeneMark.hmm [4]. We therefore decided to compare **EuGène** to GeneMark.hmm on this dataset (which does not share any sequence with the Araclean dataset used for **EuGène** parameter estimation). The results presented below are "gene model" level results (we refer the reader to [6, 2] for a precise definition of these measures). Naturally, similarly improved results are obtained at the nucleotide, exon or protein level. For **EuGène** II, we used SPTR and a cDNA database built using EMBL, and cleaned from documented partially or alternatively spliced cDNA.

# genes	actual	predicted	correct	missing	partial	wrong	split	fused	sensit.	specif.
GeneMark	168	208	67	1	100	27	18	12	40%	32 %
EuGène I	168	196	102	1	65	21	8	2	61%	52 %
EuGène II	168	198	125	1	42	22	9	0	74%	63 %

Although an in-depth analysis is needed to be more conclusive, we think that the strength of **EuGène** I lies in the quality of its basic components (IMM, NetGene2...), the existence of an intergenic Markov model, and in the fact that, except for the IMM, its parameters have been estimated by maximum of "successful recognition" rather than maximum likelihood. This probably tends to compensate for possible weaknesses in the global model.

This report is very preliminary and we expect to significantly enhance **EuGène**'s effectiveness in a near future (and apply it to other organisms). Actually, compared to other gene finding algorithms, **EuGène** is relatively simple: it uses a single Markov model set independently of GC%, does not take into account signals such as polyA or promoters. This should leave room for improvements.

Acknowledgements : this work has been supported by a grant from the « Programme prioritaire génome & fonction » from the *Institut national de la recherche agronomique* (INRA). We would also like to thank N. Pavie and C. Mathé (INRA, Gand) for their help with Araset. S. Rombauts (INRA, Gand), B. Lescure ,H. Chiapello and I. Small (INRA, Versailles) for sharing their experience de l'arabette.

References

[1] V. Brendel and J. Kleffe, *Prediction of locally optimal splice sites in plant pre-mRNA with application to gene identification in Arabidopsis thaliana genomic DNA*, Nucleic Acids Res., 26 (1998), pp. 4749–4757.

[2] M. Burset and R. Guigo, *Evaluation of gene structure prediction programs*, Genomics, 34 (1996), pp. 353–367.

[3] S. Hebsgaard, P. Korning, N. Tolstrup, N. Engelbrecht, P. Rouzé, and S. Brunak, *Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information*, Nucleic Acids Res., 24 (1996), pp. 3439–3452.

[4] A. V. Lukashin and M. Borodovsky, *GeneMark.hmm: new solutions for gene finding*, Nucleic Acids Res., 26 (1998), pp. 1107–1115.

[5] K. Murakami and T. Takagi, *Gene recognition by combination of several gene-finding programs*, BioInformatics, 14 (1998), pp. 665–675.

[6] N. Pavy, S. Rombauts, P. Déhais, C. Mathé, D. Ramana, P. Leroy, and P. Rouzé, *Evaluation of gene prediction software using a genomic dataset: application to Arabidopsis thaliana sequences*, in Proc. of 2^{d} Georgia Tech conference on BioInformatics, Atlanta, Nov. 1999.

[7] A. Pedersen and H. Nielsen, *Neural network prediction of translation initiation sites in eukaryotes: prespectives for EST and genome analysis*, in Proc. of ISMB'97, AAAI Press, 1997, pp. 226–233.

[8] P. R. P.G. Korning, S.M. Hebsgaard and S. Brunak, *Cleaning the genbank arabidopsis thaliana data set*, Nucleic Acids Res., 24 (1996), pp. 316–320.

[9] L. Rabiner, *A tutorial on hidden markov models and selected application in speech recognition*, Proc. IEEE, 77 (1989), pp. 257–286.

[10] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, *Microbial gene identification using interpolated markov models*, Nucleic Acids Res., 26 (1998), pp. 544–548.

[11] N. Tolstrup et al., *A branch-point consensus from Arabidopsis found by non circular analysis allows for better prediction of acceptor sites*, Nucleic Acids Res., 25 (1997), pp. 3159–3163.