Gene finding integrating external evidence

Thomas Schiex[1], Catherine Mathé[2], Annick Moisan[1], and Pierre Rouzé[2]

[1] INRA, Toulouse, France,
http://www-bia.inra.fr/T/schiex
[2] INRA, Gand, Belgique

Abstract

EuGène, a gene finder for *Arabidopsis thaliana*, combines the output of several information sources, including output of other software or user information. To achieve this, a weighted DAG is built in such a way that a shortest feasible path in this graph represents the most likely gene structure of the underlying DNA sequence. The usual simple Bellman linear time shortest path algorithm for DAG has been replaced by a shortest path with constraints algorithm which remains linear both in time and space. EuGène effectiveness has been recently assessed on Araset, a recent dataset of *Arabidopsis thaliana* sequences used to evaluate several existing gene finding software. It appears that, despite its simplicity, EuGène gives results which compare very favorably to existing software.

For locating genes (exons/introns) in eukaryotic sequences, it is important to take into account several sources of evidence. The sources exploited typically include matches against databases (cDNA, EST, protein databases...), output of signal prediction software, "integrated" gene finding software, experimental evidence or human expertise. The motivation of our work is, as far as possible, to automate this job for *Arabidopsis thaliana*. Several integrated gene finders exist that integrate protein or cDNA similarities in their prediction [5, 2]... EuGène is naturally closely related to these tools. It most striking peculiarity is it's parasitic behavior: EuGène has been designed to exploit other tools or information sources, including human expertise or other integrated gene finding programs.

To be able to integratec information, we combine the information at the lowest level in order to be able to maintain the consistency of the prediction and globally assess the impact of each choice w.r.t. all evidence. Given a DNA sequence, we define a DAG such that all possible consistent gene structures are represented by a path in the graph. 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. To weight the edges, output of interpolated Markov models (IMM, [6]) is used for content analysis, while the output of existing signal prediction software is used for signals (for each software, parameters are estimated by optimization on a learning set). A second version, called EuGène+ can use, in conjunction with these basic informations, results from cDNA/EST and protein databases search. In practice, the structure and weights of the graph can also be directly modified by the user using a very simple language. Bellman space/time linear shortest path algorithm [1] has been sophisticated to take into account constraints on the minimum length of gene elements (introns, single exon genes, intergenic regions) and remains linear. Globally, the approach is comparable to explicit state duration HMM with uniform duration densities.

The “Araset” dataset [4] has been used to assess recent gene finding software: GeneMark.HMM [3], GlimmerA [6], EuGène and EuGène+, FGenesH and FGenesHG (Salamov and Solovyev, unpublished, a variant that predicts non standard AG/GC splice sites). In this evaluation, EuGène+ uses SPTR, PIR and TrEMBL as protein databases and a cDNA/EST database built from EMBL using SRS.

We report only gene level prediction quality: a gene is considered as correctly predicted if all its exons are correctly predicted. Il all exons are not correctly predicted, the gene can be completely missing, partial, wrong, split or fused. In each case, the overall performance is abstracted in the two usual *specificity* (S_p) and *sensitivity* (S_n) measures. See [7] for further information.

Program	pred	correct	missed	partial	wrong	split	fused	S_n	S_p
GeneMark.hmm	187	69	1	98	24	2	12	41%	37%
GlimmerA	265	50	2	116	58	35	0	30%	19%
EuGène	184	108	1	59	14	4	2	64%	59%
EuGène+	186	131	0	37	15	4	2	78%	70%
FGenesH	167	81	12	75	15	0	8	48%	49%
FGenesHG	166	83	12	73	15	0	10	49%	50%

References

[1] Bellman, R. *Dynamic Programming.* Princeton Univ. Press, Princeton, New Jersey, 1957.

[2] Kulp, D., Haussler, D., Reese, M., and Eeckman, F. Integrating database homology in a probabilistic gene structure model. In *Pacific Symp. Biocomputing* (1997), pp. 232–44.

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

[4] Pavy, N., Rombauts, S., Déhais, P., Mathé, C., Ramana, D., Leroy, P., and Rouzé, P. Evaluation of gene prediction software using a genomic data set: application to arabidopsis thaliana sequences. *Bioinformatics 15*, 11 (Nov. 1999), 887–99.

[5] Rogozin, I., Milanesi, L., and Kolchanov, N. Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci. 12*, 3 (Jun 1996), 161–70.

[6] Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., and Tettelin, H. Interpolated markov models for eukaryotic gene finding. *Genomics 59*, 1 (1999), 24–31.

[7] Schiex, T., Moisan, A., and Rouzé, P. Eugène, an eukaryotic gene finder that combines several type of evidence. In *Selected papers from JOBIM'2000*, no. 2066 in LNCS. Springer Verlag, 2001, pp. 118–133.