

ExonHunter: A Comprehensive Approach to Gene Finding

Broňa Brejová, Daniel G. Brown, Ming Li, Tomáš Vinař*

Keywords: gene finding, genome comparison, homology search, hidden Markov models

1 Introduction.

We present ExonHunter, a new and comprehensive gene finder system that outperforms existing systems, featuring several new ideas and approaches. Our system combines numerous sources of information (genomic sequences, ESTs, and protein databases of related species) into a gene finder based on a hidden Markov model in a novel and systematic way. In our framework, various sources of information are expressed as *advisors*—partial probabilistic statements about positions in the sequence and their annotation. We then combine these into the final prediction via a quadratic programming method, which we show is an extension of existing methods. Allowing only partial statements is key to our transparent handling of missing information and coping with the heterogeneous character of individual sources of information. As well, we give a new method for modeling length distribution of intergenic regions in hidden Markov models.

On a commonly used test set, ExonHunter performs significantly better than the existing gene finders ROSETTA, SLAM, or TWINSCAN, with more than two thirds of genes predicted completely correctly. Supplementary material is available at <http://www.bioinformatics.uwaterloo.ca/supplements/05eh/>

2 Results.

Table 1 shows the comparison of ExonHunter with other gene finding programs evaluated by [Alexandersson et al., 2003] on the ROSETTA data set. ExonHunter used advisors based on human and mouse ESTs, human, mouse, and chicken protein alignments, and mouse and *Drosophila* genome-genome comparison. We have outperformed all other tested programs at both exon and nucleotide levels, except for nucleotide specificity. At the gene level, ExonHunter identifies more than two thirds of genes in the data set completely correctly.

One could object to this test since many of the genes in the ROSETTA set are also found in the database of human ESTs or proteins. Therefore, we also evaluated the program without advisors based on human information. We still maintain the highest sensitivity on both exon and nucleotide levels, with only a 2% drop in exon specificity; the change mostly affects the gene statistics.

*School of Computer Science, University of Waterloo, Waterloo, ON N2L3G1, Canada. E-mail: bbrejova@cs.uwaterloo.ca

	GN	RA	SM	TN	TN.p	SGP-1	EH	EH-nh
Gene Sn	44%	—	—	—	—	—	74%	68%
Gene Sp	41%	—	—	—	—	—	66%	62%
Exon Sn	82%	83%	78%	84%	86%	70%	91%	89%
Exon Sp	73%	83%	76%	77%	82%	76%	83%	81%
Nucl. Sn	98%	94%	95%	98%	96%	94%	99%	99%
Nucl. Sp	88%	98%	98%	89%	94%	96%	93%	93%

Table 1: **Comparison on ROSETTA set.** Results for GENSCAN (GN), ROSETTA (RA), SLAM (SM), TWINSCAN (TN), TWINSCAN.p (TN.p; alignments from known orthologs only), and SGP-1 are from [Alexandersson et al., 2003] (the authors did not report gene statistics). The EH column gives the results achieved by ExonHunter with all advisors. The EH-nh column corresponds to ExonHunter results without advisors originating in human datasets. We used standard definitions of sensitivity (Sn) and specificity (Sp), see, for example, [Alexandersson et al., 2003].

3 Methods.

Gene finding seeks to label each position in a given DNA sequence as intergenic, intron, exon (in six different reading frames), donor site, acceptor site, start codon, or stop codon. An HMM for gene finding defines a conditional probability distribution over all possible *annotations* (sequences of labels) of a specific sequence. To predict genes using an HMM, we find the annotation A^* that maximizes $\Pr(A^*|sequence)$.

To combine an HMM with other supplementary sources of evidence (genome-genome sequence comparison, EST or protein alignments, etc.), we express each source of evidence as one or several *advisors*. An advisor provides a partial information about probabilities of individual labels. The granularity of this information is determined by the nature of the source of evidence. For example, EST match does not help to predict reading frame of an exon.

Definition 1 (Advice of an advisor). *Let Σ be the set of labels. The advice of advisor a at position i is a partition π_a of the set Σ and a probability distribution $p_a(S)$ over all partition elements $S \in \pi_a$. The value $p_a(S)$ is an estimate of the probability that the correct label at position i is in set S , given the information available to advisor a .*

Advice of all advisors is combined using quadratic programming, extending linear opinion pool method [Tax et al., 2000] commonly used to combine predictions without making independence assumptions about the sources of the information. The resulting probability distribution $\Pr(A|evidence)$ is combined with $\Pr(A|sequence)$ with Bayesian principles, and we find the annotation A^* maximizing $\Pr(A^*|sequence, evidence)$.

References

- [Alexandersson et al., 2003] Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research*, 13(3):496–502.
- [Tax et al., 2000] Tax, D. M. J., van Breukelen, M., Duin, R. P. W., and Kittler, J. (2000). Combining multiple classifiers by averaging or multiplying? *Pattern Recognition*, 33:1475–1485.