

# Sub-optimal predictions increase sensitivity of gene finders

Mihaela Pertea<sup>1</sup> and Steven Salzberg<sup>1</sup>

**Keywords:** gene finding, dynamic programming, alternative splicing

## 1 Introduction.

Most genome annotation projects begin by running a computational gene finding algorithm on the newly assembled DNA sequence of the target species. Most gene finders are to some extent statistical methods, usually incorporating a dynamic programming algorithm to find a gene structure that maximizes an overall score. This gene structure is then predicted as the most likely model of a gene found in a specific genomic region. While dynamic programming guarantees that the highest scoring coding region is found, it by no means guarantees that the prediction is correct. Various gene finders use different measures to score the coding regions, and therefore they capture different features of the genes. Gene finders frequently disagree on the precise gene structure of a given gene. However, if a gene finder is allowed to report multiple alternative gene structures rather than simply picking the best, the likelihood that the correct model will be contained in its output increases dramatically.

Here we present a case study on the genome of the model plant *Arabidopsis thaliana*, in which the dynamic programming algorithm of the gene finder GlimmerHMM [4] was modified to compute the top  $n$  best-scoring open reading frames. The output contains more than 90% of the correct gene models when only the top 5 best scoring predictions are reported. Note that for the sake of discussion, we use the term “gene” synonymously with “protein-coding region.”

## 2 Results.

We modified GlimmerHMM to report the 100 highest-scoring gene predictions for several organisms. Figure 1 shows the percentage of correct gene predictions included in the output of GlimmerHMM when run on a test data set of 631 non-redundant *Arabidopsis thaliana* genes, described in [3]. The genes in the training data set do not include genes with a BLASTN [2] alignment with more than 80% identity to any genes in the test set. When only the highest-score prediction is selected, GlimmerHMM predicts 372 genes exactly correct (i.e., all exon and intron boundaries are correct), for a sensitivity of ~59%. While this by itself is good performance when compared to other gene finders [3], the sensitivity jumps to 90% when the top 5 highest-score predictions are considered. Increasing the number  $n$  of reported predictions does not show significant sensitivity increases above  $n=14$ , where it reaches a sensitivity of 95%.

When the 5 top scoring set of predictions are evaluated independently of one another (i.e., the 2<sup>nd</sup>-best predictions are treated as their own set), each set of predictions achieves >98% accuracy on the test set (defined as the average of sensitivity and specificity) at the nucleotide level. At the exon level, the accuracy ranges from 90% for the best-scoring predictions to 75% for the 5<sup>th</sup>-highest. The

---

<sup>1</sup> The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland, USA, E-mail: mpertea@tigr.org and salzberg@tigr.org

results obtained for *A.thaliana* were similar to results obtained on other genomes including *Oryza sativa* (rice), *Aspergillus fumigatus*, and *Toxoplasma gondii*.

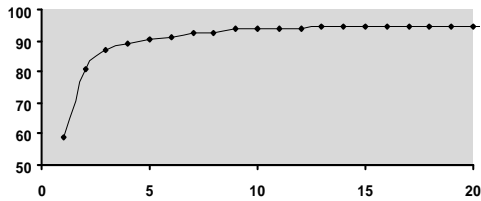


Figure 1: Correctly predicted gene models on a 631-gene *A. thaliana* data set. The x-axis represents the number of highest-scoring gene predictions, while the y-axis is the percentage of true genes correctly predicted.

Although GlimmerHMM was not trained and designed for finding alternatively spliced genes, we ran it on a set of 549 *A.thaliana* transcripts containing 1313 alternatively spliced variants validated at TIGR [6]. If only the highest scoring prediction is retained, GlimmerHMM predicts one of the correct splice variants for 323 out of the 549 genes in this set. However these predictions capture only 25% of the 1313 possible splice variants. This percentage is increased to ~45% when the top 10 best predictions are considered.

### 3 Discussion.

While there is no perfect method for gene structure prediction, a typical genome annotation pipeline will consider multiple sources of evidence including the locations of gene predictions from *ab initio* gene finders. In most cases, a gene finder must be trained specifically for the target organism [5] and very often, more than one gene finder is utilized [1] in order to achieve a reasonably accurate gene annotation, because each gene finder typically has a set of genes for which it is the only correct method. With the current state of the art in gene finding, few organisms have multiple gene finders built for them. The study presented here shows that a small range of the highest-scoring predictions contain the majority of the correct gene models. When combined with similarity approaches, one might overcome the need of several *ab initio* gene finders in order to achieve highly accurate gene annotation.

### Acknowledgements

This work was supported by NIH grant R01-LM006845.

### References

- [1] Allen JE, Pertea M, Salzberg SL. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res.* 14(1), pp 142-8.
- [2] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215(3), pp 403-10.
- [3] Korf I. Gene finding in novel genomes. 2004 *BMC Bioinformatics.* 5(1):59.
- [4] Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics.* 20(16) pp 2878-9.
- [5] Pertea M, Salzberg SL. 2002. Computational gene finding in plants. *Plant Mol Biol.* 48(1-2), pp 39-48.
- [6] The Institute for Genomic Research. Alternative donor and/or acceptor sites. [http://www.tigr.org/tdb/e2k1/ath1/altsplicing/conventional\\_splice.list.html](http://www.tigr.org/tdb/e2k1/ath1/altsplicing/conventional_splice.list.html)