

ABSTRACT: Prediction of protein coding genes is most commonly done with a probabilistic model or heuristic incorporation of alignments from several sources. The former method tends to be less precise than the latter, but the latter requires a large amount of expensive data. Newer genome sequencing projects are often limited to early-phase, whole-genome shotgun assembly and ESTs.

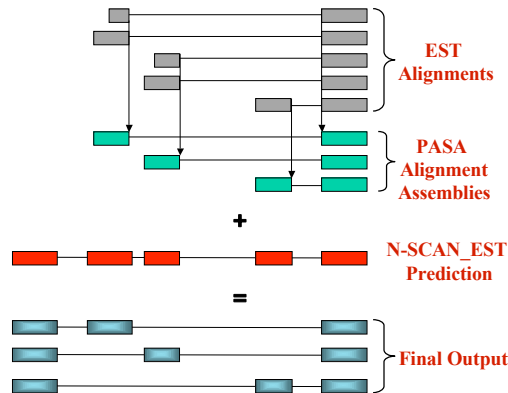
Many gene predictors combine the use of EST alignments and genomic sequence evidence, however, they do so inefficiently, throwing away most of the clues left by the ESTs. We show that the combination of a state-of-the-art gene predictor, N-SCAN_EST, with alignment assemblies from ESTs can predict a significant amount of transcripts correctly with an order of magnitude fewer ESTs than required by N-SCAN_EST alone. We present a pipeline, which uses N-SCAN_EST and the alignment assembler, PASA, for generating these predictions from newly-sequenced genomes.

INTRODUCTION

N-SCAN_EST uses a model which incorporates EST evidence (Wei, et. al., 2006), built upon a Genscan-like (Burge, et. al, 1997) DNA hidden Markov model and a phylogenetic conservation model (Gross, et. al, 2006). N-SCAN_EST takes clues from EST alignments by shadowing their genomic locations onto an EST sequence. This “shadow” method serves to prevent the long running time that would occur if all ESTs were considered independently. However, it ignores the distinctiveness of overlapping alignments, many of which imply alternate splice forms.

PASA (Haas, et. al, 2003) can filter and reduce redundant ESTs by combining them into maximal alignment assemblies, representing putative gene structures. This drastically reduces the volume of EST evidence, making it feasible to process in a gene predictor. We harness the data produced by PASA to improve N-SCAN_EST predictions.

COMBINING PREDICTIONS WITH ALIGNMENT ASSEMBLIES

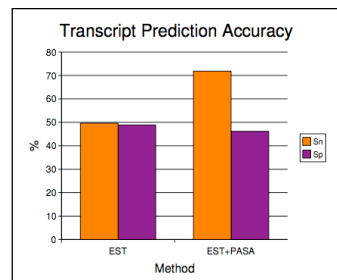


1. ESTs are aligned to the genome and assigned genomic coordinates.
2. PASA clusters all alignments with consistent internal splice site coordinates into **alignment assemblies**. Each distinctive assembly are used as evidence for transcription.
3. N-SCAN_EST makes its gene prediction based on DNA, conservation and EST evidence.
4. The **alignment assemblies** are combined with the gene prediction and filtered for consistency.

PERFORMANCE ANALYSIS

We performed four-fold cross validation using the FlyBase annotation of the *D.melanogaster* genome. Performance was measured in terms of the total number of transcripts correctly predicted. The chart to the right shows a substantial increase in transcript sensitivity, from near 0.5 to above 0.7. A modest decrease of <.05 was seen in specificity.

$$S_n = \frac{TP}{FN + TP} \quad S_p = \frac{TP}{FP + TP}$$



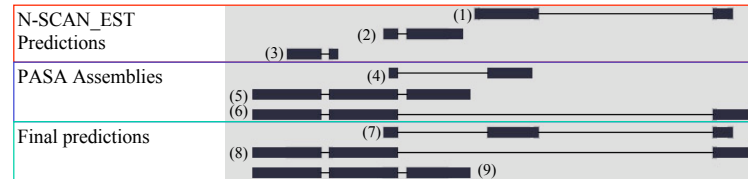
EXAMPLES

EST gives evidence for a skipped exon



A new intron is predicted. The PASA assembly that gives evidence that the 6th exon is sometimes skipped. Other assemblies (not shown) support the N-SCAN_EST prediction.

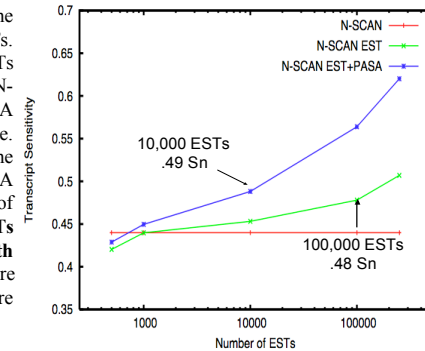
ESTs merge incorrectly split genes



Three predicted genes are merged into three overlapping isoforms of the same gene. N-SCAN_EST predicts several exons correctly, however since it can only predict one isoform per locus, it splits the gene into three separate genes. PASA assembly (4) with N-SCAN_EST prediction (2) give (7), (2), (3) and (5) give evidence for (9), and N-SCAN_EST predictions (1) and (3), combined with PASA assembly (6) give evidence for (8).

EFFECT OF REDUCING EST EVIDENCE

Here we attempt to determine if the system will survive given fewer ESTs. We reduced the number of ESTs introduced to the system at all steps (N-SCAN_EST training, alignment, PASA updates) and measured the performance. The graph to the right shows that the using N-SCAN_EST with PASA predicts approx. the same number of transcripts correctly with **10,000 ESTs as N-SCAN_EST alone does with 100,000 ESTs**. The ESTs were randomly pared and all larger sets were a superset of smaller sets.



CITED WORK

- Burge, C. and Karlin, S. *Journal of Molecular Biology*, 268(1), 1997.
 Gross, S. and Brent, M.R. *Journal of Computational Biology*, 13(2), 2006.
 Haas, B.J. et. al. *Nucleic Acids Research*, 1(19), 2003.
 Wei, C. and Brent, M.R. *BMC Bioinformatics*, July 2006.