Predicting Gene Boundary

Sridhar S. Hannenhalli **James W. Fickett**
Sridhar_Hannenhalli@sbphrd.com James_Fickett@sbphrd.com
Bioinformatics SmithKline Beecham Pharmaceuticals 709 Swedeland Road,
PO Box 1539 King of Prussia, PA-19406, USA

Keywords: computational gene prediction, gene boundary, CpG, GENSCAN, benchmark

1 Introduction

With the accelerated rate of genomic sequencing, *de novo* gene finding continues to be important. There are several gene finding programs currently in use for vertebrates. For review see [1, 2, 3] and references therein. It has been widely observed that even the best programs often combine exons from multiple genes, and it is thought that all gene finding programs tend to perform better if provided with the genomic region containing only a single gene and its immediate neighborhood. The accuracy of these programs is thus expected to deteriorate when analyzing large contigs. An independent evaluation, [4] of GRAIL, and more recently, [5] of GENSCAN confirms this.

We define the *gene boundaries* as the end points of the pre-spliced transcript of a gene on the genomic sequence, and an *anchor* as any interval on the genomic sequence internal to the gene boundaries. Given an anchor for a gene of interest, our ability to "tightly" estimate the gene boundaries will enhance the accuracy of *de novo* gene prediction. In this paper, we explore the feasibility of such a prediction using indirect evidences (not pertaining to the gene of interest) like 3' EST hits, 5' EST hits, protein hits and predicted CpG Islands.

For an anchor $[g_{5'}, g_{3'}]$ on the positive strand, let the true 5' boundary be $G_{5'}$ and let a predicted 5' outer boundary be y. If $y < G_{5'}$ we consider it a correct prediction. A trivial way to get all predictions correct is to predict the 5' end of the contig as the boundary (assuming that the gene is contained in the contig) which is clearly not desirable due to a possibly enormous amount of extra sequence 5' to the gene. We would like to maximize the fraction of correct predictions while minimizing $G_{5'} - y$. We call this latter quantity *overshoot*. For wrong predictions (when $y > G_{5'}$) we call the quantity $y - G_{5'}$ *undershoot*.

2 Results

On the set of 2362 exons, the five prime gene boundary prediction were as follows: 87% of the cases the predicted boundary included the true gene boundary with an average overshoot of 12Kb and an average undershoot of 39Kb on the remaining 13%.

The three prime gene boundary prediction were as follows: 85% of the cases the predicted boundary included the true gene boundary with an average overshoot of 17Kb and an average undershoot of 21Kb on the remaining 15%.

Improving Gene Finding via Gene boundary prediction The GENSCAN accuracy was tested on three different sets of genomic sequences. The first set, G, comprised of entire contigs with 1 or more genes. The second set G' contained, for each gene, the genomic sequence containing the gene CDS and 5Kb of flanking genomic region on either sides. The third set G'' contained, for each gene, the region within the predicted gene boundaries using the entire CDS as the anchor.

Genomic Set	Nucleotide			Exon				Gene		
	Sn	**Sp**	**CC**	**Sn**	**Sp**	**ME**	**WE**	**Similarity**	**MG**	**WG**
G	0.90	0.16	0.37	0.72	0.13	0.12	0.84	0.581	0.051	0.817
G'	0.91	0.67	0.78	0.75	0.56	0.11	0.31	0.587	0.041	0.342
G"	0.91	0.42	0.61	0.78	0.36	0.11	0.56	0.580	0.051	0.573

Table 1: GENSCAN accuracy results at nucleotide, exon and gene levels:: Sn : sensitivity, Sp : specificity, CC : correlation Coefficient. At Exon level:: ME : Missed exons, these are the real exons with no overlap with any predicted exon. WE : Wrong exons, these are the predicted exons with no overlap with any real exon. Sensitivity and Specificity include only the exactly matching exons. At gene level:: FracSimilarity : This is the fraction similarity among the overlapping real and predicted genes, calculated with respect to the true gene. MG : Missed genes, these are the real genes with no overlap with any predicted gene. WE : Wrong exons, these are the predicted genes with no overlap with any real gene.

The GENSCAN results at the nucleotide, exon and the gene levels are shown in Table 1. Specificity is the only measure that shows some improvement over the unrestricted case. Clearly, bounding the gene helps eliminate spurious exons not belonging to the gene of interest. The specificity results should be treated with some caution though. Since, in the measure used here, even the spurious exons belonging a different gene are accounted for against the specificity.

The average performance shows marginal improvement in the exon level sensitivity (72% to 78%) which is similar to the results reported in [5] (63% to 70%). There is no improvement at the level of fraction gene identity.

We analyzed the parses for the cases where there was a difference in the parses in the bound and unbound cases. There was no clear pattern in which GENSCAN (mis) behaved. In the following we discuss just a few ways in which these differences were manifested.

Among the cases where the parse of the bounded gene was better than that of the unbounded gene, a common reason was that a spurious promotor or polyA tail was picked in the unbounded case and these signals were out of bound in the bounded case. This results in the choice of a false initial/terminal exon and consequently changing the true initial/terminal exon into an internal one.

More surprising were the cases where the parse of the bounded gene was worse than that of the unbounded gene. In some of this cases, the bounding excluded essential signals and resulted in wrong parses. In some cases, even though the signals were within the bounds, a different signal and parse was chosen due to change in Isochore and hence the use of a slightly different model by GENSCAN.

References

[1] Burge, C.B. and Karlin, S., Finding the genes in genomic DNA, *Curr. Opin. Struct Biol.*, 8(3):346–354, 1998.

[2] Claverie, J.M., Computational methods for exon detection, *Mol. Biotechnol.*, 10(1):27–48, 1998.

[3] D. Haussler., Computational genefinding, *Trends Guide to Bioinformatics*, 12–15, 1998.

[4] Lopez, R., Larsen, F. and Prydz, H., Evaluation of the exon predictions of the GRAIL software, *Genomics*, 24(1):133–136, 1994.

[5] R. Guigo, M. Burset, P. Agarwal, J. Abril, R. Smith, and J. Fickett., Sequence similarity based gene prediction, In S. Suhai, editor, *Genomics and Proteomics: Functional And Computational Aspects*, chapter 1. Plenum Publishing, 1999.