Genome informatics workshop poster abstract. 2008

Incorporating tiling array expression data into a gene predictor

Benjamin Lansdell[12] **Terry Speed**[1]
lansdell@wehi.edu.au terry@wehi.edu.au

Anthony Papenfuss[1]
papenfuss@wehi.edu.au

[1] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville 3050, Australia
[2] Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia

Keywords: gene prediction, transcript mapping, tiling arrays

1 Introduction

Computational gene prediction is an important tool in the annotation of a sequenced genome. Typical *ab initio* predictors aim to identify the precise location and structure of genes, based on DNA sequence features such as codon biases, presence of start/stop codons, etc. Further improvements can be made if external evidence sources are included. To date, such external evidence has included ESTs and homologous proteins. Tiling expression arrays, which probe the non-repeat portion of the genome in an unbiased fashion, could also be used for this task, but little work has been done in this area to date. More typically, tiling arrays are analyzed by transcript mapping or 'transfragging', in which transcribed regions of the genome (exons and non-coding RNA genes) are identified.

Here we present preliminary results from a new tool, TileGene, which uses a generalized hidden Markov model (GHMM) framework to combine evidence from genomic sequence and tiling array experiments in order to improve gene prediction, or alternatively to improve transcript mapping. In addition to predicting genes, TileGene also predicts non-coding transcribed elements, thereby providing a richer annotation of a genome.

2 Method and Results

TileGene is a GHMM, similar in design and implementation to GenScan [1]. Fourth order Markov chains are used to model sequence and second order weight array matrix models are used to model signals. Signal models for start/stop codons and acceptor/donor splice sites are included. Maximum likelihood estimates of model parameters were taken from a set of known Drosophila genes. Underlying states of the model correspond to features to be inferred from input DNA sequence. Typically CDS, UTR, introns and intergenic regions are represented. In order to predict non-coding transcribed elements we also include novel states for regions showing evidence of expression but little protein-coding capability. The architecture is such that these non-coding regions can occur within intergenic and intronic regions. The addition of these states requires a number of assumptions: in the current model, we assume that these states show similar expression levels to protein-coding genes, have similar lengths to exons and show no strong nucleotide composition biases.

In order to incorporate tiling expression data into our model we use a probe-level summary statistic based on correlations. In transcribed regions, it is expected that intensities across different time points for neighbouring probes will be highly correlated, while in non-transcribed regions it is expected that intensity profiles will have lower, though typically still positive correlations. Empirical distributions of the score are measured within known expressed and non-expressed regions. This is incorporated into our model by using a dual-emission hidden Markov model. This allows for both the sequence-based models and the tiling correlation model to influence predictions. A weight may be assigned to each emission that can be discriminatively trained to maximize performance. The inclusion of data from different time points provides evidence for a larger range of transcribed elements thereby aiding in the prediction of a larger number of genes.

Table 1 shows preliminary performance data on the Adh region of Drosophila melanogaster for three versions of TileGene in order to demonstrate the improvement tiling expression provides. The models are compared to the transcript mapping method of [2] and the gene predictor Augustus [3]. Tiling array data was taken from [2], which maps transcription during Drosophila embryogenesis. Sensitivities marked with an asterisk are based on a curated set of known, expressed genes. TileGene+expression represents gene prediction with novel states and correlation score included; TileGene+expression+weighted indicates a discriminately trained weight has been used when including the tiling correlation score. Annotations include alternate transcripts, which lowers sensitivity and increases specificity at the nucleotide and exon levels.

Table 1: TileGene performance.

	Nucleotide level		Exon level		Gene level	
	Sn	Sp	Sn	Sp	Sn	Sp
Transfrag [4]	82*	45	-	-	-	-
Augustus	72	97	54	78	56	61
TileGene (*ab initio*)	77	93	51	65	35	40
TileGene+expression	65*	94	15*	32	10*	25
TileGene+expression+weighted	90*	94	57*	64	38*	40

3 Discussion

Without expression data, the performance of TileGene is comparable to older GHMM gene predictors (data not shown), but it does not perform as well as more advanced models such as Augustus. This is not surprising given the relative simplicity of the gene model implemented in TileGene.

The addition of an expression content sensor trained using a set of known, expressed genes improves the performance of TileGene at both the nucleotide and exon level based on an expressed gene test set. TileGene+expession+weighted outperforms the transfrag method, partly as a result of the model operating not at the probe-level but the nucleotide-level. The transfrag method predicts transcribed elements, not exclusively genes, resulting in the apparently poor specificity. The results show that TileGene is successful at identifying transcribed elements and classifying them as protein-coding or otherwise.

Weighting the contributions from the two content sensors improves sensitivity at all measured levels. As analysed, the expression data provide information about the location of exons but no information about the structure of the genes in which they reside, resulting in a smaller increase in sensitivity measured at the gene level than the nucleotide or exon level. Specificity remains largely unchanged, which reflects that the model is only predicting expressed genes.

GHMM gene predictors are effective at combining subtle hints of gene location but care needs to be taken when incorporating external evidence sources to ensure the GHMM does not 'misinterpret' the data and disturb the delicate balance of probabilistic models. Indeed, when expression is included without being weighted the performance is shown to decrease.

The results we present above are preliminary – optimal training and implementation of the model has yet to be determined. We hope to further investigate modeling of expressed non-coding states and the implications on gene prediction accuracy. Nonetheless, it does appear that the performance of a relatively simple gene predictor can be improved by the inclusion of tiling array data. Further, by predicting coding and non-coding features TileGene is able to provide a more detailed annotation of a genome.

References

[1] Burge, C. and Karlin, S., Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol*., 268:78-94, 1997.

[2] Manak, J., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A. and Gingeras, T., Biological function of unannotated transcription during the early development of Drosophila melanogaster, *Nature Genetics*, 38:1151-1158, 2006.

[3] Stanke, M. and Waack, S., Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, 19, Suppl. 2:ii215-ii225, 2003.

[4] http://transcriptome.affymetrix.com/publication/drosophila_development/