A Tool for Constructing EST Splice Graphs and Consensus Sequence Assembly

Ketil Malde[1], Eivind Coward[2], and Inge Jonassen[3]

Keywords: EST clustering, splice graphs, consensus sequence assembly

1 Introduction

ESTs provide an abundant and quickly growing source of genetic data, and devising efficient algorithms and tools for analysing EST data remains an important challenge for the field of Bioinformatics. We present a tool for constructing *splice graphs* from EST clusters, both for a visual rendering of the structure of cluster, and for fast assembly of high-quality consensus sequences representing the different splice variants of the gene.

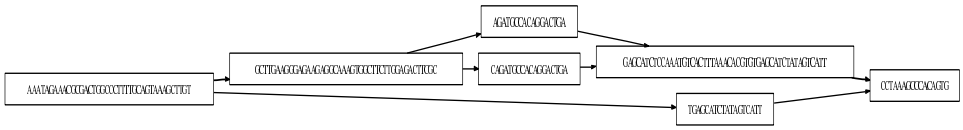

Figure 1: Visualization of a simple splice graph.

A *splice graph* [1] is an innovative way to present an EST cluster (see Figure 1 for an example). Originally based on ideas developed for "sequencing by hybridization", the splice graph is a graph where, ideally, each node in the graph represents an exon, and each edge in the graph represents a possible concantenation of exons in one or more splice variants from the gene.

It is also possible to efficiently construct consensus sequences from splice graphs. There already exist graph based methods for DNA assembly [5], but due to the different natures of mRNA and DNA, we need a slightly different approach.

2 Algorithm and Implementation

Since we do not have any a priori knowledge of the exons in the gene, we start by letting every $n-1$-word in the data set represent a potential exon. To simplify the graph, we then collapse edges where the source node has outdegree one and the destination has indegree one.

The graph is stored as dictionary of the words constituting the edges, with the nodes implicitly defined from the edges. "Lightweight" edges (with little support from the data) can then be filtered out, before visualization is done using GraphViz [6].

For consensus sequence generation, the graph is traversed using a greedy heuristic that tries at each branching point to follow the branch being supported by the most sequences consistent with previous branches.

[1]Department of Informatics, University of Bergen, Norway. E-mail: `ketil@ii.uib.no`
[2]Department of Informatics, University of Bergen, Norway. E-mail: `coward@ii.uib.no`
[3]Computational Biology Unit, BCCS, University of Bergen, Norway. E-mail: `inge@ii.uib.no`

3 Results

Both graph visualization and consensus sequence construction is very fast, on an 800MHz Sun Fire 880, a relatively large cluster consisting of 1180 sequences was assembled in less than three minutes. For comparison, Phrap, which is usually considered a fast assembler, used thirteen minutes on the same data set.

When comparing the results to six mRNAs from the same gene, our tool consistently produced contigs that more closely resembled the mRNAs, with BLAST alignment scores ranging from 2500 to 3600 bits, while Phrap's contigs were in the range 1500 to 1900.

The software is freely available [8].

4 Acknowledgements

This work was funded by the Norwegian Salmon Genome Project [7], a Norwegian Research Council program.

5 References and bibliography.

References

[1] Heber, S., et al. 2002. Splicing graphs and EST assembly problem *Bioinformatics* pp. S181-S188

[2] Malde, K. et al. 2003. Fast sequence clustering using a suffix array algorithm *Bioinformatics* vol. 19 no. 10, pp. 1221–1226.

[3] Mironov, A. et al. 1999. Frequent alternative splicing of human genes. *Genome Research* 9:1288–1293.

[4] Modrek, B. and Lee, C. 2001. A Genomic View of Alternative Splicing. *Nature Genetics* 30:13–19.

[5] Pevzner, P. A., Tang, H., and Waterman M. A. 2001. An Eulerian path approach to DNA fragment assembly *PNAS*

[6] The GraphViz web site. `http://www.research.att.com/sw/tools/graphviz/`

[7] The Salmon Genome Project. `http://www.salmongenome.no/`

[8] Further information and program downloads. `http://www.ii.uib.no/~ketil/bioinformatics`