

Bioinformatics Approaches to RNA Splicing

Aaron Levine

Thesis presented for the degree of

Master of Philosophy

University of Cambridge

and

The Sanger Centre

Wellcome Trust Genome Campus

Hinxton, Cambridgeshire

August 8, 2001

Summary

With the completion or near-completion of many large genome sequencing projects, automated annotation of vertebrate genomes has become an important research priority. Yet the complex intron/exon structure and the prevalence of alternative splicing of genes in higher organisms have rendered accurate gene prediction a difficult and still unsolved problem.

In order to aid in the ongoing genome annotation projects, I have developed a splice site prediction program, StrataSplice, which predicts both canonical (GT-AG) and minor variant (GC-AG) splice sites and is designed to integrate easily into a variety of gene prediction and annotation systems. StrataSplice utilises a new splice site prediction model that combines local GC content with a standard probabilistic pattern recognition technique and shows modest but significant improvement over standard splice site prediction models. Much of this improvement occurs in gene-rich high GC regions in which previous models perform more poorly.

U12-dependent introns are a distinct class of introns found in small numbers in the genomes of most higher eukaryotes, yet they have been largely ignored in genome annotation efforts. I have conducted a computational scan for these rare introns in the draft human genome sequence and generated a new reference set of 404 U12-dependent introns, an increase of 6-fold over the number previously available in all genomes. Analysis of these introns suggested that there is a significant error rate (>0.25 percent) at the acceptor site in U12-dependent splicing and that, in contrast to U2-dependent introns, U12-dependent introns may be recognised in an exclusively exon-dependent manner.

Chromosome 22 was the first human chromosome to be sequenced and has been subject to extensive experimental and computational annotation. Combining the predictions generated by StrataSplice with expressed sequence evidence, I have generated a set of 3,199 expressed sequence confirmed introns on chromosome 22. Nearly 80 percent of these introns were previously annotated, but the remaining 20 percent (671 introns) may help identify either alternatively spliced forms of known genes or previously unidentified genes.

Acknowledgements

I feel fortunate to have spent a year working in the Informatics group at the Sanger Centre and living at Churchill College and many members of both of these communities have helped me immensely during my time in Cambridge.

I am particularly grateful for the support and guidance of my supervisor, Richard Durbin, who introduced me to the field of probabilistic sequence analysis and helped to shape my varied and only semi-connected ideas into a coherent and feasible project

I am also grateful for the useful discussions I had with and the technical assistance I received from the other members of Richard's group, namely Irmtraud Meyer, Kevin Howe and Marc Sohrmann. I would like to thank Marc in particular for the generous loan of his laptop, which was of great assistance in the preparation of this thesis.

Alex Bateman and Sam Griffiths-Jones from the pfam group have been generally helpful, as has my University supervisor, Nick Goldman from the department of Zoology. Additionally, I am particularly indebted to Matt Pocock and Thomas Down, who have patiently answered numerous Java questions, James Gilbert, who has solved several frustrating Perl and postscript problems and Michele Clamp, who has provided invaluable assistance extracting data from Ensembl.

I would also like to thank the Café Churchill crowd at Churchill College for the good times, foosball, inspiration, and many helpful discussions on the true importance of U12 introns.

I was supported by a studentship from the Winston Churchill Foundation and the Sanger Centre is supported by the Wellcome Trust.

Preface

This thesis is the result of my own work and not the product of any collaboration. As with almost any scientific endeavour the work described herein builds on the research of countless others, and they are referenced throughout the body of this thesis and in a bibliography.

This thesis is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University nor has any part of this thesis been submitted for any such degree, diploma or other qualification.

Table of Contents

1	Introduction	1
2	Introns and RNA Splicing: a short review	4
2.1	An overview of human gene structure	5
2.2	The RNA splicing process	6
2.3	U12-dependent introns and the U12 spliceosome	9
2.4	Computational analysis of RNA splicing	10
3	Identifying Splice Sites in Human Genomic DNA Sequences	13
3.1	A block dependence model for donor site identification	14
3.2	Improved splice site prediction by considering local GC content	19
3.3	Identifying non-canonical GC donor sites	30
3.4	StrataSplice: a human splice site predictor	32
4	A Computational Scan for U12-Dependent Introns in the Human Genome Sequence	34
5	A Computational Scan for U2-Dependent Introns on Human Chromosome 22	51
6	Conclusion	56
7	References	60
	Appendices	
A	Complete list of U12-dependent introns identified in chapter 4	A-1

List of Figures

2.1	Major components of a eukaryotic protein coding gene	6
2.2	Splice site consensus sequences	8
3.1	Typical log-odds score distributions for splice site prediction	16
3.2	Mutual information at donor and acceptor splice sites	17
3.3	Performance of higher order dependency model	19
3.4	Local GC content at true and false splice sites	25
3.5	Performance of stratified model at donor sites	27
3.6	Performance of stratified model at acceptor sites	28
4.1	Length distributions of U2- and U12-dependent introns	41
4.2	Branch point to acceptor site distance for U12-dependent introns	42

List of Tables

2.1	Characteristics of human genes	5
3.1	Performance of the stratified splice model on genomic DNA sequences	22
3.2	Nucleotide frequencies at stratified donor and acceptor splice sites	26
3.3	Comparison between stratified splice model and GENSCAN splice site models	29
3.4	Performance at variant GC donor sites	32
3.5	Optional parameters for customising StrataSplice	33
4.1	Diverse terminal dinucleotides of U12-dependent introns	39
4.2	Phase of U12-dependent introns	42
4.3	Putative 3' splicing errors in U12-dependent splicing	43
4.4	Alternatively spliced U12-dependent introns	45
4.5	Genes with multiple U12-dependent introns	46
4.6	Summary of expressed sequences supporting U12-dependent introns	47
5.1	Comparison of U2-dependent introns identified on chromosome 22 with pre-existing annotation	53
5.2	Alternative splicing of U2-dependent introns on chromosome 22	54