

EST-based detection and analysis of mammalian transcripts

Shobhit Gupta

Nov 2005

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich für Mathematik und Informatik
der Freien Universität Berlin
vorgelegte
Dissertation

1. Reviewer: Prof. Dr. Martin Vingron

2. Reviewer: Prof. Dr. Ulf Leser

Defence Date: 16-Nov-2005

For my Parents

Acknowledgments

My first days in Berlin were actually my first days outside India. That had quite an impact on my ability to get adapted to the new environment. But for that matter, the CMB department of Max Planck Institute was a very good choice. The group members were extremely friendly and easily approachable for both personal as well as work-related issues.

To start with the acknowledgments I thank my supervisor, Martin Vingron. I am most grateful to Martin for offering me an opportunity to work under his supervision. His excellence at work as well as clarity of ideas was obvious to me beforehand. Additionally he has proved to be a fantastic supervisor both in terms of project-related and personal guidance.

I express my deepest gratitude to Stefan Haas. He has been my mentor throughout. I sincerely appreciate his patience with me especially during the paper-writing sessions (read cycles). I sincerely thank my collaborators Dorothea Zink and Bernd Korn for providing a close experimental support to my computational work. Bernd has also been a critical and very helpful member of my Phd committee.

I deeply appreciate the company of all past and present CMB group members. The discussions during the weekly group meetings have been very fruitful for the advancement of my project as well as for the broadening of my knowledge-base. I would also like to thank our system administrator, Wilhelm Ruesing. He has been relentless and exceptionally prompt in cracking down on infra-structural problems.

On a personal note, special thanks goes to Birgit Pils and Johanna Holbrook for great help during my initial days. Additionally, I am thankful to my friends in Germany especially Stefanie Koeniger, Shivendra Kishore, Harindar Keer and Abha Singh. They have been very motivating during the inevitable low periods of graduate research.

Lastly but perhaps the most significantly, I appreciate the support of my mother before and during my graduate studies. I would have been nowhere without her enormous struggle for my education.

Contents

Acknowledgments	iii
1 Motivation	1
2 From simple genome to complex proteome	3
2.1 Transcription	3
2.1.1 Splicing of pre-mRNA	7
2.1.2 Alternative splicing	10
2.2 Translation	12
3 Resources for transcriptome analysis	13
3.1 DNA microarrays	13
3.1.1 Hardware	13
3.1.2 Experiment	14
3.1.3 Applications	14
3.2 Serial Analysis of Gene Expression (SAGE)	16
3.2.1 Methodology	16
3.2.2 Applications	18
3.2.3 Limitations	18
3.3 Expressed Sequence Tags (ESTs)	19
3.3.1 Generation of ESTs	19
3.3.2 Normalization of cDNA libraries	19
3.3.3 Clustering of the EST data	21
3.3.4 Assembly of the EST clusters: GeneNest	21
3.3.5 Genomic mapping of the EST data	23
3.3.6 Expression levels using EST data	29
Expression levels in GeneNest	29
Expression levels in SOURCE database	29
Tissue/tumor-specific transcripts in ASAP	31

4	Confidence-based prediction of (alternative) splicing	33
4.1	Fuzzy logic	33
4.1.1	Fuzzification of inputs	34
4.1.2	Application of fuzzy operators	34
4.1.3	Application of implication method	35
4.1.4	Aggregation of all outputs	35
4.1.5	Defuzzification of fuzzy terms	37
4.2	Fuzzy logic based prediction of (alternative) splice events	37
4.2.1	Definition of membership functions for splicing evidences	37
4.2.2	Computation of quality values for exon-intron boundaries	41
4.2.3	Evaluation of the splice signal parameter	41
4.2.4	Statistics for human data	46
4.2.5	Validation via known instances of alternative splicing	48
4.2.6	Evaluation of robustness of the model	52
4.2.7	Experimental validation	54
4.3	Alternative splicing in coding/non-coding regions	54
5	Expression patterns of (alternative) transcripts	57
5.1	Classification of cDNA libraries	57
5.1.1	Methodology	58
5.2	Tissue/tumor-specific transcripts via GeneNest and SpliceNest	58
5.2.1	Prediction approach	59
5.2.2	Experimental verification	59
5.2.3	Evaluation of tissue-specificity	62
5.2.4	Evaluation of tumor-specificity	65
5.3	Conclusions	65
6	T-STAG: An integrated portal for EST-based transcriptome analysis	69
6.1	The T-STAG database	69
6.1.1	Content	69
6.1.2	Database Design	70
6.1.3	Web-interface	71
6.2	Applications of the resource	72
6.2.1	Tissue-specific genes and splice isoforms	72
6.2.2	Rare genes/alternative isoforms and disease related genes/isoforms	72
6.2.3	Comparison of expression patterns among genes	75
6.2.4	Background definition for tissue-specific transcripts	75
6.2.5	Evolutionarily conserved expression patterns	76
7	Summary	77
A	The Genetic code	91

B	The RT-PCR Experiments	93
B.1	Experimental Protocol	93
B.2	List of tissues	93
C	IUPAC nucleotide ambiguity codes	95
D	Availability	97
D.1	Quality computation software	97
D.2	T-STAG software	97
D.3	T-STAG Database	97
E	List of related publications	99
F	Curriculum vitae	101

List of Figures

2.1	Central dogma of molecular biology	4
2.2	Schematic illustration of transcription	5
2.3	Regulation of transcriptional initiation	6
2.4	The mammalian consensus sequences at the 5' splice site and the 3' splice site in the pre-mRNA	8
2.5	The assembled commitment complex during pre-mRNA splicing	9
2.6	The two trans-esterification reactions during pre-mRNA splicing	10
2.7	Classification of alternative splicing	11
3.1	Illustration of a microarray experiment	15
3.2	Schematic of SAGE procedure	17
3.3	Generation of cDNA libraries from mRNAs	20
3.4	Normalization of cDNA libraries	22
3.5	The GeneNest visualization	23
3.6	SpliceNest visualization of the gene structure	25
3.7	The ASAP's geneview	26
3.8	The ASD approach: transcript confirmation for introns	27
3.9	The ASD visualization of splice isoforms for the human C2F gene	28
4.1	Linear membership functions	34
4.2	Fuzzy Logic Vs Boolean logic	35
4.3	Application of implication method	36
4.4	Membership function of common boundaries (GeneNest)	38
4.5	Membership function of common boundaries (SpliceNest)	38
4.6	Membership function of tolerance required for detection of the common boundary (SpliceNest)	39
4.7	Membership function of EST count	40
4.8	Membership function of splice signal	40
4.9	Quality values for exon/intron boundaries	42
4.10	Computation of exon quality	42
4.11	Flow-diagram of splicing confidence computation	44
4.12	Boundary quality values when the splice signal is present	45
4.13	Boundary quality values for terminal exons	46

4.14	Boundary quality values when splice signal is absent	46
4.15	Boundary quality values for gapped alignment	46
4.16	SpliceNest visualization of alternative splicing	48
4.17	Quality values of known alternative exons vs predicted alternative exons	50
4.18	Confidence values of known alternative splicing events vs predicted al- ternative splicing events	51
4.19	Quality values for the known alternative exons (AEDB) for different perturbations in the fuzzy logic model	53
5.1	Prediction approach for tissue-specific transcripts	60
5.2	RT-PCR validation experiment of a putative brain-specific isoform . . .	64
5.3	RT-PCR amplification analyzing expression pattern of gene <i>PRAME</i> .	66
6.1	T-STAG database schema	71
6.2	The T-STAG query interface	73
6.3	The hyper-linked output of T-STAG	74
A.1	The Universal Genetic Code	92

List of Tables

3.1	Tissues with specifically expressed genes (via GeneNest database) . . .	30
4.1	Rules for computing quality values for splice boundaries	43
4.2	Rules for computing confidence values for splice events	44
4.3	Perturbations applied to the fuzzy logic system	52
4.4	Experimentally verified alternative splice events	55
5.1	Tissues with predicted specific transcripts.	61
5.2	RT-PCR validation results for tissue and disease-specific splice isoforms	63
B.1	Tissues for which RT-PCR experiments were performed	94

Abbreviations

Abbreviations in alphabetical order

A	adenine
AEDB	alternative exon database
ASAP	alternative splicing annotation database
ASD	alternative splicing database
BLAST	Basic local alignment search tool
bp	base pair(s)
°C	degree celsius
C	cytosine
CAGE	cap analysis gene expression
cDNA	complementary DNA; DNA synthesized from mRNA by RT
CDS	coding sequence
CGH	Comparative genomic Hybridization
ChIP	chromatin immunoprecipitation
DNA	deoxyribonucleic acid
EST	expressed sequence tag
G	guanine
Kb	kilobase(s); 1,000 nt
Mb	megabase(s); 1,000,000 nt
mRNA	messenger RNA
nt	nucleotide
PCR	polymerase chain reaction
polyA	polyadenylation signal
RNA	ribonucleic acid
RT	reverse transcription
RTase	reverse transcriptase; an enzyme
RT-PCR	reverse transcription-polymerase chain reaction
SAGE	serial analysis of gene expression
snRNP	small nuclear ribonucleoprotein
T	thymine
U	uracil
UTR	untranslated region; part of mRNA transcripts

For a compilation of IUPAC symbols for nucleotide nomenclature see Appendix C.