Computational detection of gene regulatory signals in human genome sequence

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

by

Aroul Selvam Ramadass

Trinity Hall, University of Cambridge

and

The Wellcome Trust Sanger Institute, Hinxton, Cambridge

July 2004

DECLARATION

This dissertation is my own work and contains nothing which is the outcome of the work done in collaboration with others, except as specified in the text and acknowledgements.

The work in this thesis is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other university.

Aroul Selvam Ramadass

July 2004

SUMMARY

Transcription, the first step in gene expression, is initiated from a transcription start site and terminated some distance downstream of the cleavage site. In this thesis I attempt to identify and model different regulatory signals involved in the process of transcription, towards the development of a signal based *ab initio* gene predictor.

First I attempt to identify regulatory signals in the sequence downstream of the cleavage site that may be responsible for transcription termination. Base compositional analyses reveal no significant bias in the nucleotide composition. An investigation based on free-energy minimisation Zuker algorithm indicates the possibility of a secondary structure in the sequence downstream of the cleavage site. A probabilistic machine learning algorithm based on Bayes theorem and Generalised Linear Models, Eponine, used to scan for motifs, learns a model to classify termination sites from other sequences. The model captures a few multiplex signals that might be responsible for polymerase II pause and termination. An evaluation of this termination model against annotated human chromosomes shows that the model performs better than existing methods. However a significant number of predictions also appear near the annotated start site of genes. Approximately 10% of predictions lie within genes and their density is correlated with gene length and intron size. I propose two hypotheses to explain these anomalies and discuss results from recent experiments.

Splicing is now found to be interlinked temporally and spatially with transcription and I attempt to develop a donor and acceptor site model using Eponine. Comparisons of the models with annotated sites show the models have higher positional accuracy and perform comparably with existing programs, GeneSplicer and StrataSplice.

Like transcription, translation machinery is influenced to a great extent by regulatory signals and I investigate them by scanning for motifs around translation start and stop sites using Eponine. The start model learnt only the regulatory elements and not the coding potential of exons. Despite this it performs better than the existing program NetStart, although less well than the program ATGpr.

The availability of these models creates the possibility to build an *ab initio* gene prediction program based purely on gene regulatory signals.

ACKNOWLEDGEMENTS

I thank my supervisor, Tim Hubbard, for giving an opportunity to work on this project in The Wellcome Trust Sanger Institute. I very much appreciate and am grateful to his advice and support he has given during this project. I thank Thomas Down for his help and the *Eponine* trainer that facilitated this study. Raphäel Lepläe and Mathew Pocock were grateful and patient in introducing me to the programming concepts. I thank them for this and all the help they gave me during the years I worked with them. Discussing ideas with Samiul Hasan, Yen-Hua Huang and Bernard Leong were always stimulating and I thank them for making the work interesting.

This project would not have been possible without the hard work of annotation and curation by members of Ensembl, VEGA and Sanger human Chromosome 22 group. I appreciate them for this excellent job and making the data available in the public domain.

I thank Rajkumar Sasidharan for all the support and collaboration he rendered in the 'Domain Insertion' project mentioned in the appendix. During my years of stay in Cambridge, I made numerous friends and it would be unfair if I mention only few. I thank every one of them for making my life enjoyable and happy.

I am grateful to Cambridge Commonwealth Trust and The Wellcome Trust Sanger Institute for funding and Trinity Hall College, Cambridge, for academic support.

I thank Sudhakaran Prabakaran, Jane E Swatton and James M Carr for correcting the manuscript.

Finally I would like to extend my gratitude to my parents, brother and sister for their love and support all these years.

CONTENTS

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
BP	Branch Point
BPS	Branch Point Sequence
CATH	Class Architecture Topology Homologous superfamily
CF	Cleavage Factor
CPF	Cleavage and Polyadenylation Factor
CPSF	Cleavage and Polyadenylation Specificity Factor
CstF	Cleavage stimulation Factor
CTD	Carboxy Terminal Domain
DRE	Downstream Regulatory Element
EAS	Eponine Anchored Sequence
EST	Expressed Sequence Tags
FP	False Positives
GFF	General Feature Format
GLM	Generalized Linear Model
GO	Gene Ontology
HMM	Hidden Markov Model
ISS	Intermediate Sequence Search
MAZ	Myc-Associated Zinc finger protein
NMD	Non-sense Mediated Decay
NN	Neural Networks
PABP	Poly(A) Binding Protein
PAP	Poly(A) Polymerase
PC	Position Constraint
PDB	Protein DataBase
PE	Pause Elements
RMSD	Root Mean Square Deviation
ROC	Range Operating Characteristics
RVM	Relevance Vector Machine
SCFG	Stochastic Context Free Grammar
SCOP	Structural Classification Of Proteins

SD	Shine-Dalgarno
SLBP	Stem Loop Binding Protein
snRNPs	small nuclear Ribo-Nucleo Proteins
SR	Splice Regulatory
SSP	Secondary Structure Profile
SVM	Support Vector Machine
TFBS	Transcription Factor Binding Sites
TP	True Positives
URE	Upstream Regulatory Element
UTR	Un-Translated Region
WM	Weight Matrix

LIST OF TABLES

LIST OF FIGURES