Computational localization of promoters and transcription start sites in mammalian genomes

Thomas Down

This dissertation is submitted for the degree of Doctor of Philosophy

Wellcome Trust Sanger Institute and Sidney Sussex College, Cambridge

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The work in this thesis has not been submitted in whole, or in part, for a degree, diploma, or any other qualification at any other university.

Thomas A. Down
May 2003, Cambridge, UK

Abstract

A number of large genomes have now been sequenced, and biologists are now faced with the challenge of identifying all the functional pieces of sequence, and understanding how they contribute to the development and life of the organism. While identification of protein coding genes, and annotating their products, has been progressing well, there are a great many open questions relating to the regulatory regions which control the expression of these genes.

Here, I investigate the question of identifying and annotating promoters, one of the most important regulatory signals in the genome, which mark the points where transcription is initiated, and regulate the transcription of genes. I present a new computational method, EponineTSS, which can predict transcription start sites in bulk genomic sequence data with excellent sensitivity and specificity. Unlike the existing methods, it gives an indication of the actual location of the transcription start site. Comparisons with available experimental data suggest that the positional accuracy of these predictions is very good. Results from this method are included as part of the Ensembl human genome annotation.

Having located transcription start sites for genes, I also discuss the use of results from comparative genomics the estimate the extent of the functional promoter region upstream of the start site. I show that the extent of promoters is very variable, and that promoter size is correlated with the function of the gene for whose regulation it is responsible. Genes associated with developmental processes tend to have particularly large, and thus presumably complex, promoters, with the homeobox transcription factors among the most extreme examples.

I also introduce sparse Bayesian learning, a recently developed approach to supervised machine learning which can be applied to the training of a wide range of model types, and embodies the principle of selecting the simplest possible model to explain the observed data. I demonstrate a new technique which makes sparse Bayesian learning much more scalable,

allowing it to be applied to very large and complex problems, and present a convenient, freely available Java library which provides a general-purpose implementation of this technique. This library was used here in the training of the transcription start site predictor, but has a wide range of applications in computational biology and beyond.

Acknowledgments

Thanks to the Wellcome Trust, and everyone at the Sanger institute who helped and encouraged me as I explored this fascinating area of research. I particularly appreciate the efforts of my project supervisor, Dr. Tim Hubbard, who has helped me along the path towards this thesis, accepted the many digressions, and added a valuable voice of caution and rigour when analyzing new results. Just because you're paranoid, it doesn't mean the data *isn't* out to get you.

All the programs I wrote during this project rest on the strong foundation of the BioJava library. I have to thank everyone who has ever contributed to this project: documentation, code, bug reports and ideas are all vital to the success of an open source effort. Right from the outset, Matthew Pocock has been a guiding light for the project. He's also a great coder, target for trying out new ideas, and friend.

Throughout this all, I've enjoyed the love, support, and good humour of Emily White, who has kept me going through the difficult moments and helped celebrate the joyous ones. Thank you.

Curated annotations for human chromosome 22 were produced by the Chromosome 22 group from the Sanger Institute and were obtained from the World Wide Web at http://www.sanger.ac.uk/HGP/Chr22/

Contents

Chapter 1. Introduction 1

1.1. Objectives of this project 6

1.2. What is a promoter 7

1.3. Resources for studying promoters 13

1.3.1. EPD: the eukaryotic promoter database 14

1.3.2. Full-length cDNA sequences 15

1.3.3. TRANSFAC 17

1.4. Existing computational methods for studying promoter 17

1.4.1. Well-known motifs and Position Weight Matrices 18

1.4.2. CpG islands 21

1.4.3. PromoterInspector 23

1.5. Other resources used in this project 24

1.5.1. BioJava 24

1.5.2. Ensembl 26

1.5.3. The Gene Ontology (GO) 28

Chapter 2. Sparse Bayesian Learning 31

2.1. Generalized Linear Models (GLMs) 32

2.2. Feature selection using pruning priors 35

2.3. A pragmatic approach to handling large spaces 42

2.4. A general-purpose Sparse Bayesian trainer 46

2.5. Sparse Bayesian Learning Discussion 48

Chapter 3. **Modeling of transcription start sites** 50

3.1. The Eponine Anchored Sequence (EAS) model 52

3.1.1. Learning EAS models 55

3.1.2. Implementation and validation of EAS 57

3.2. Training a transcription start site model 60

3.3. Model refinement using mouse cDNA sequence data 62

3.4. Validation and testing of EponineTSS 65

3.4.1. Testing on human chromosome 22: bulk genomic performance 66

3.4.2. Testing on EPD: calibration of positional accuracy 71

3.4.3. Comparison with other methods 72

3.5. Analysis of cases where promoters were not detected by EponineTSS .. 74

3.5.1. Modeling of non-detectable promoters 75

3.5.2. Correlation of promoter-detectability with gene type and function 77

3.6. EponineTSS discussion 83

Chapter 4. **Learning from comparative genomics** 88

4.1. The Eponine Windowed Sequence (EWS) model family 90

4.2. Training from non-coding homologies between human and mouse 93

4.3. Evaluating the function of mouse-similarity models as promoter predictors 100

4.4. An attempt to discover a second signal in mouse-human homologies .. 104

4.5. Comparative Genomics Discussion 109

Chapter 5. **Evolutionary conservation of promoter regions** 113

5.1. Alignment of promoter regions 115

5.2. Relationship of promoter alignments to regulatory roles 121

5.3. Discussion of promoter conservation 130

Chapter 6. **Conclusions** 134

References 139

List of Figures

1.1. Flow of biological information from the genome to the proteome. In this case, the pre-mRNA can be spliced in two possible ways, leading to different protein products. 2

1.2. Illustrative diagram showing the basic structure of a eukaryotic protein-coding gene. 3

1.3. X-ray structure of transcription factor Pu.1 (shown in red) bound to a synthetic DNA substrate (green). 10

1.4. A (simplistic) overview of keys steps in transcription initiation at a *polII* promoter. 11

1.5. X-ray structure of DNA (shown in green) coiled around an octamer of core histone proteins (red). 12

1.6. A protocol for oligonucleotide-labeling of mRNAs with intact cap structures, used to preferentially clone full-length mRNA sequences. 16

1.7. Hidden Markov Model of a motif (states m1 to m5) embedded in a longer sequence. 20

1.8. Logo view of a Position Weight Matrix model of the TATA box (redrawn from data on the EPD website) 21

1.9. Example of a directed acyclic graph of Gene Ontology terms. 29

2.1. Plot of the logistic function (equation 2.1.2). 34

2.2. Plots of the Gaussian distribution, $\mathcal{G}(\beta_i \mid 0, \alpha_i^{-1})$, for various values of α. 39

2.3. Example of sparse Bayesian learning in Cartesian 2-space. The data points chosen as centers for the final set of basis functions are highlighted in green. 41

2.4. Example of a data set used for testing of training speed. 44

2.5. Training time vs. problem size for full-set and incremental sparse Bayesian learning. 44

3.1. Example schematic architecture of an Eponine Anchored Sequence model. 54

3.2. Schematic of an EAS model learned from the synthetic dataset, showing the three spiked motifs. 58

3.3. Learning curves for the training of an EAS model on the synthetic dataset. 59

3.4. Accuracy vs. coverage (ROC) curve for a model trained on the synthetic dataset. 60

3.5. The EponineTSS_1 model, trained from 389 mammalian sequences from EPD. .. 62

3.6. Histogram showing offsets of DBTSS start sites relative to those of corresponding EPD entries. 63

3.7. The EponineTSS_2 model, trained from 599 mouse sequences. 65

3.8. Selection of regions to include when the pseudochromosome sequence. 68

3.9. Accuracy vs. coverage for two EponineTSS models on the pseudochromosome. 69

3.10. Density of prediction from the EponineTSS_2 model relative to annotated gene starts. 70

3.11. Accuracy vs. coverage for the EponineTSS_2 model on pseudochromosomes based on old (2.3) and new (3.1b) curated annotation of chromosome 22. 71

3.12. Density of EponineTSS_2 predictions relative to the annotated TSS of EPD entries. 72

3.13. Intersection of “correct” predictions of promoters by EponineTSS, PromoterInspector, and CpG islands. 75

3.14. Model trained on negative-selected FANTOM data. 76

3.15. Accuracy vs. coverage for model trained on negative-selected FANTOM data, with random predictions for comparison. 77

3.16. Example GO term lineage. 79

4.1. Length distribution of sequence regions aligned between human and mouse by ensembl-compara version 5. 94

4.2. Scatter of scores from two different EWS models. 97

4.3. Ensembl contigview displays for selected portions of human chromosome 22, showing windows with high scores for one of the homology models (labeled “ews1_0.95”), and predictions from the EponineTSS_2 model (labeled “Eponine”). 98

4.4. ROCs for EWS models trained from sets of 2000 or 4000 human-mouse homologies. 100

4.5. Intersection of transcription start sites correctly predicted by the EponineTSS, EponineHomol, and CpG methods. 103

4.6. A set of basis functions learned by training the EWS-scaffold system on human-mouse homologous sequences. Each cell of this table shows an individual basis function, made up of between one and three sequence motifs. 105

4.7. Scores for a model trained from the reverse-selected dataset vs. EponineHomol_1. 106

4.8. Contigview displays showing both EponineHomol_1 and EponineHomol_2 predictions. 108

4.9. Intersection of “correct” promoter predictions from EponineTSS, EponineHomol_1, and EponineHomol_2. In each compartment, the first figure indicates the number of start sites which were also detected by a CpG island predictor, while the second figure gives the total number. 109

5.1. Agreement of promoter orthology with protein orthology. Panel a shows coverage (proportion of sequences correctly paired) at high levels of accuracy, with and without additional repeat-masking using **etandem.** Panel b shows a wider range of coverages, and only includes the results with **etandem** masking. 119

5.2. Comparison of orthology in the windows [-2000:0] and [-4000:-2000] 121

5.3. Histograms of sequences binned by number of bases of promoter sequence included in blastn alignments to the orthologous promoter. 125

5.4. Aligned regions from 250 transcription-associated genes, sorted by number of aligning bases. 126

5.5. Dot plot between human and mouse upstream regions with low similarity. 127

5.6. Dot plot between human and mouse upstream regions with moderate similarity. 128

5.7. Dot plot between human and mouse upstream regions with strong similarity. .. 129

List of Tables

3.1. The four position weight matrices used in the EponineTSS_2 model. 66

3.2. Sensitivity and selectivity of various promoter-prediction mechanisms on the human pseudochromosome. 74

3.3. GO terms applied preferentially to **found** genes. 80

3.4. GO terms applied preferentially to **unfound** genes. 81

4.1. Words learned from three EWS models trained from mouse-human homologies. 96

4.2. Logo view of the motifs used in the EponineHomol_1 model. 102

4.3. Logo view of the motifs in the EponineHomol_2 model. 107

5.1. Default **blastn** parameters used 117

5.2. GO terms which are overrepresented in the long-aligning promoter set. 123

5.3. GO terms which are overrepresented in the short-aligning promoter set. 124