

Computational Analysis of Genomes

Matthew R. Pocock

This dissertation is submitted for the degree of Doctor of
Philosophy.

April 2003

Supervisors: Dr T. J. P. Hubbard, Dr N. Goldman

The Sanger Centre, Cambridge; Darwin College, Cambridge

This dissertation is the result of my own work and includes nothing which is the
outcome of work done in collaboration.

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.

Matthew R. Pocock,

April 2003, Cambridge, United Kingdom

Dedication

I would like to thank all of those who have supported me through the process of producing this dissertation. Particular thanks must go to Tim Hubbard, who has been a source of great help and provided direction where needed without smothering me with micro-management. Nick Goldman and Ed Griffiths have both been valuable sounding boards throughout. Thomas Down has been my staunchest ally as we developed BioJava and also the DAS protocol from embryonic beginnings to the well-respected projects they are now. It would be unfair not to thank all of those who have helped with BioJava, as coders, testers and users. In particular, a mention must go to Chris Dagdigian for managing the hardware. I have enjoyed my time here, and this is in no small part due to the friendliness of those I meet daily in the Sanger Centre, the EBI and from the Ensembl project. Lastly, I must dedicate this work with heartfelt gratitude to Caroline. Without her love all academic achievements would be worthless.

*Fate is unmoved by one's pitiful hopes; what changes, bowing to fate, is
what one hopes for.*

(Liza Dalby, The Tale of Murasaki p239)

Abstract

Recently we have been blessed with a simultaneous rise in the volume of biological data and the power of computers. This has necessarily led to the emergence of the field of Bioinformatics, where the study of entire genomes rather than individual genes is the norm.

This dissertation describes the development and application of the software framework BioJava, designed from the outset to provide a strong foundation for the implementation of different machine learning algorithms. BioJava allows genomic size datasets to be efficiently manipulated in a range of hardware environments.

A variety of supervised and unsupervised learning techniques were applied to data sets on the scale of whole genomes taking advantage of the BioJava framework.

Firstly, unsupervised learning was used to look for underlying structure in the genome sequence of whole Malaria chromosomes. Time-reversible 1st order Hidden Markov Models (HMMs) learned signals based on sequence composition that appear to correlate closely with biological units, such as exons, introns, repeats and non-coding genomic regions. This demonstrates the ability of unsupervised methods to discover biologically meaningful information within genomic sequence.

Secondly, supervised learning was used to develop a regression method able to predict recombination rate within human chromosomes. Support Vector Machines (SVMs) using suffix tree kernels were trained on human chromosome 22 sequence and were able to learn a signal reproducibly, although it was not clear how well this models recombination rate.

Finally, supervised learning was used to develop a classification method able to detect subtle signals in noisy and small sets of micro-array expression data. A Bayesian technique for training linear models was applied to learn sparse models. These were able to distinguish between tumour samples that had been treated with a drug and those that had not. The models produced by this method can be readily interpreted in terms of individual genes, and in this case made good biological sense.

This dissertation illustrates how a framework of modular and reusable software components can be used together with advances in artificial intelligence to help us interpret the data flowing from high throughput projects in the post genomic era.

Table of Contents

Dedication	ii
Abstract	iii
Table of Contents	v
Table of Figures	ix
List of Tables	xi
Table of Equations	xii
Chapter 1 Introduction.....	1
1.1 Existing Software Development Frameworks for Bioinformatics.....	2
1.1.1 The NCBI Toolkit.....	3
1.1.2 Bioperl.....	4
1.1.3 EMBOSS.....	5
1.2 BioJava.....	6
1.3 Machine Learning	8
1.3.1 Clustering, Classification and Regression for Single Items.....	9
1.3.2 Signal Analysis with Hidden Markov Models.....	18
1.4 Implementation and Use of BioJava.....	24
Chapter 2 The BioJava Core Interfaces	24
2.1 Java as a Language for Bioinformatics	24
2.2 Nested Exceptions and Assertions	24

2.3	Changeability	24
2.4	Symbols, Alphabets and SymbolList.....	24
2.5	Locations, Sequences and Features.....	24
2.6	Probability Distributions and Hidden Markov Models.....	24
2.7	Query.....	24
2.7.1	Motivations	24
2.7.2	Initial Implementation.....	24
2.7.3	Limitations of This System.....	24
2.8	Recent Developments	24
2.8.1	The Tag-Value Parser	24
2.8.2	Flat File Indexing.....	24
2.8.3	Annotation Types.....	24
2.8.4	Enhanced Feature Filters.....	24
2.8.5	Change Hubs.....	24
2.8.6	Bit Packed Sequences	24
2.9	Conclusions.....	24
Chapter 3	HMMs for whole <i>Plasmodium Falciparum</i> Chromosomes.....	24
3.1	Introduction.....	24
3.2	Simple HMM Architectures.....	24
3.2.1	Methods.....	24
3.2.2	Results.....	24

3.3	HMM Architectures with Complementary Emission Distributions	24
3.3.1	Methods.....	24
3.3.2	Results.....	24
3.4	First Order HMMs with Time-Reversible Transition Probabilities.....	24
3.4.1	Methods.....	24
3.4.2	Results.....	24
3.5	Discussion.....	24
3.6	Future Directions	24
Chapter 4	Investigation of Recombination Rates Using SVMs	24
4.1	Introduction.....	24
4.1.1	Support Vector Machines	24
4.1.2	BioJava APIs for Support Vector Machines.....	24
4.2	Methods.....	24
4.2.1	Searching for a Signal Affecting Recombination Rates Using a Word-Frequency Kernel Function	24
4.2.2	Construction and Training of an SVM for Predicting Recombination Rate.....	24
4.3	Results.....	24
4.3.1	Recombination Rates Predictions	24
4.3.2	Cross-Validation	24
4.4	Discussion.....	24

Chapter 5	RVMs for Classification of Expression Data.....	24
5.1	Introduction.....	24
5.2	Cellular Responses to Doxorubicin	24
5.3	Generalized Linear Models.....	24
5.4	Micro-array Classification Using a Support Vector Machine Implemented as a Linear Kernel RVM.....	24
5.4.1	Framework for Generalised-Linear-Models amenable to Expression Arrays.....	24
5.4.2	RVM Analysis Using the Small Working Set Heuristic.....	24
5.4.3	Function of Genes Identified by GLM Models.....	24
5.5	Conclusions, Applications and Future Work.....	24
	Concluding Remarks.....	24
	References.....	24

Table of Figures

Figure 3-1 Emission probabilities for the four pair-state model.....	24
Figure 3-2 Diagram of the <i>P. Falciparum</i> chromosome 3 and the state paths through three models	24
Figure 3-3 Diagram of the <i>P. Falciparum</i> chromosome 2 and the state paths through three models	24
Figure 3-4 Emission Spectrums for all Pair-State Models.....	24
Figure 3-5 Diagram of the alignments of the 3,4 and 5 state-pair models to Malaria chromosome 3	24
Figure 3-6 Diagram of the alignments of the 3,4 and 5 state-pair models to Malaria chromosome 2	24
Figure 3-7 Counts for Biological Feature and States for the 2-5 Pair-State Models ...	24
Figure 3-8 Normalized Counts of States for Biological Features.....	24
Figure 3-9 Normalized counts of Biological Features for States.....	24
Figure 4-1 Comparison of physical and genetic distances along chromosome 22	24
Figure 4-2 Total Results of Training the SVM using Uniform Counts	24
Figure 4-3 Moving Average for Uniform Counts models of Depth 4-6.....	24
Figure 4-4 Moving Average for Uniform Counts models of Depth 7-9.....	24
Figure 4-5 Total Results of Training the SVM using Normalized Rates	24
Figure 4-6 Moving Average for Normalized Rates: Depths 4-6	24
Figure 4-7 Moving Average for Normalized Rates: Depths 7-9	24

Figure 4-8 Accuracy for Recombination SVMs Under 3-Way Jack-knifing24

Figure 4-9 Predictions Across the Entire Chromosome from the 3 Jack-knife Models
for Depth of 524

Figure 5-1 Scatter Plot of the Two Topoisomerase II Probes Used.24

Figure 5-2 Expression Levels for Each Probe Used24

Figure 5-3 Average Weights Across Relevant Models.....24

Figure 5-4 Average Weights Across All Models.....24

List of Tables

Table 3-1 Forward-strand and reverse-strand counts.....	24
Table 3-2 State-transitions and their reverse-complements.....	24
Table 5-1 GLM for all before-after pairs (to 4 s.f.)	24
Table 5-2 Genes used by cross-validation models.....	24

Table of Equations

Equation 1-1 A Hypothesis Function.....	10
Equation 1-2 Error of a Hypothesis	11
Equation 1-3 Some Error Functions	12
Equation 1-4 Dot Products for Items Decomposable into Sub-Spaces with Dot- products Defined.....	14
Equation 1-5 Definition of Kernel Functions	15
Equation 1-6 A Polynomial From a Two-dimensional Coordinate to a Coordinate Containing One Component for each Possible Product Involving up to Two Dimensions	15
Equation 1-7 Dot products between two polynomial mappings reduced to terms involving the dot product of the unmapped variables.....	16
Equation 1-8 Polynomial Kernel Function	16
Equation 1-9 Definition of a Probabilistic Hidden Markov Model	21
Equation 1-10 Emission and Transition Probabilities	22
Equation 1-11 Definition of All Legal State-Sequences.....	23
Equation 1-12 Likelihood of Observing a Given Sequence and Labelling	23
Equation 1-13 Common Dynamic Programming Recursions as Applied to Probabilistic Hidden Markov Models.....	24
Equation 4-1 Equation of a Plane	24
Equation 4-2 Normal to a Plane as a Weighted Sum of Vectors	24

Equation 4-3 Definition of a Support Vector Machine.....	24
Equation 4-4 Basis Functions for Kernel Functions and Data Points.....	24
Equation 4-5 SVMs in Terms of Basis Functions	24
Equation 4-6 The Normalizing Kernel	24
Equation 4-7 SuffixTree Kernel.....	24
Equation 5-1 Bayes Theorem.....	24
Equation 5-2 Rearrangement of Bayes Theorem.....	24
Equation 5-3 Bayes Theorem in Words.....	24