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

Table of Figures

List of Tables

Table of Equations