Mathematical Methods for Comparative *Ab* Initio Gene Prediction

Irmtraud Margret Meyer
Trinity College
Cambridge

August 2002

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

Preface

The work presented in this dissertation was carried out at the Sanger Institute in Cambridge between January 2000 and August 2002. 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. No part of this dissertation nor anything substantially the same has been or is being submitted for any qualification at any other university.

Summary

This dissertation introduces two novel methods for the comparative prediction of protein coding genes in eukaryotic genomes. The first method, implemented in a program called **DOUBLESCAN,** is an ab initio method which simultaneously predicts the gene structures and the alignment of two evolutionarily related input **DNA** sequences from the sequence of their **A**, **C**, G, T bases only. The second method, implemented in a program called **PROJECTOR,** is a homology based method which predicts gene structures in one **DNA** sequence according to the known gene structures of a related **DNA** sequence and which simultaneously aligns the two **DNA** sequences. Both methods employ a probabilistic pair Hidden Markov model and are capable of predicting partial, complete and multiple genes **as** well **as** pairs of genes which are related by events of exon-fusion or exon-splitting. Predictions are generated using two different algorithms: the Hirschberg algorithm whose predictions are generated in linear memory and quadratic time and a new algorithm, called the Stepping Stone algorithm, whose memory and time requirements scale both linearly with the length of the input sequence. This work describes the theoretical concepts underlying the two novel methods and their implementation into computer programs and demonstrates the validity and generality of the approach by evaluating the performance of the gene prediction on a test set of mouse (Mus *musculus*) and human (*Homo* sapiens) **as** well **as Caenorhabditis *elegans*** and Caenorhabditis briggsae **DNA** sequence pairs.

Acknowledgements

First of all, I would like to thank my supervisor, Richard Durbin, for his advice, support and encouragement. I thank Kevin Howe, Matthew Pocock, Raphael Leplae, Aaron Levine, Marc Sohrmann, Ashwin Hajarnavis, Lachlan Coin, Thomas Down and all the other members of the Wellcome Trust Genome Campus who have made both science and life on the campus interesting and enjoyable.

I am grateful to Trinity College, Cambridge, for an External Research Studentship and to the Wellcome Trust for a Prize Studentship.

Contents