Studies in Probabilistic Sequence Alignment and Evolution

Thesis by

Ian Holmes

In Partial Fulfilment of the Requirements
for the Degree of
Doctor of Philosophy

University of Cambridge

Queens' College

and

The Sanger Centre
Wellcome Trust Genome Campus
Hinxton Hall, Hinxton
Cambridge

(Submitted on December 18, 1998)

Acknowledgements

During the past three years I have been very fortunate in having had the opportunity to discuss computational biology and Bayesian methods with Alex Bateman, Ewan Birney, Bill Bruno, Richard Durbin, Sean Eddy, David Haussler, Des Higgins, Tim Hubbard, Anders Krogh, Chip Lawrence, David MacKay, Graeme Mitchison, Richard Mott, Kimmen Sjölander, Guy St.Clair Slater, Victor Solovyev, Erik Sonnhammer and many other workers at the Wellcome Trust Genome Campus in Hinxton.

I especially thank my supervisor, Richard Durbin, for his thoughtful encouragement, guidance and advice, for being a fair and consistent critic and for his tireless patience and support.

I am very grateful for the assistance I have received from the Sanger Centre. I would particularly like to thank Sharon Thornton and Sheila Skingley, the Informatics group secretaries during the course of my PhD, and the Systems Support team for their ever-rapid response.

I would also like to thank my proofreaders, whose comments were vital: Alex Bateman, Ewan Birney, Tony Cox, Richard Durbin and Guy Slater.

Thank you to the people who have supported me most during the last year: Andrew Holmes; Jennifer Holmes; and Claire Hibbitt.

This work was supported by a grant from the Medical Research Council.

Summary

The complete sequencing of whole genomes presents opportunities for detailed study of molecular evolution. This thesis combines theoretical developments of Bayesian approaches in bioinformatics with analysis of duplications in the recently completed *C.elegans* genome.

Developments in the Bayesian probabilistic framework for sequence analysis using hidden Markov models (HMMs) are described. The principal HMM algorithms are reviewed including alignment, training and model comparison. Theory is developed for prediction of alignment accuracy and tested using simulations. Software to provide accuracy measures for multiple alignments, based on the popular HMMER suite of profile-based alignment algorithms, is presented and evaluated with reference to the Pfam database of multiple alignments.

Several of these statistical techniques are applied to an analysis of genomic duplications in the *C.elegans* genome. The completion of this - the first animal genome - offers an opportunity to study the random duplication that are believed to be the first step in the evolution of a new gene. The construction of a database of non-coding duplications is described and measurements of molecular evolutionary parameters in *C.elegans* are calculated from the data and reported. A method of dating gene duplications using alignments between conserved introns is presented and compared to existing methods using Bayesian techniques developed earlier in the dissertation. Amongst the principal agents involved in creating genomic duplications are transposons; one of the simplest families of transposon is the Tc1-*mariner* family, of which two distinct active subfamilies are well-known in *C.elegans*. Using HMM profiles, six new subfamilies of *mariner*-like transposon have been identified in the *C.elegans* genome. Several of the new subfamilies display interesting homologies to one another, suggestive of common mechanisms of transpositional catalysis.

Finally, the software tools developed during this project are described and made available for public retrieval from the Sanger Centre web site.

Contents

List of Figures

List of Tables