Gene Prediction using a configurable system for the integration of data by Dynamic Programming

Thesis by

Kevin Howe

submitted for the degree of

Doctor of Philosophy

University of Cambridge

St. John's College

and

The Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge

(Submitted on February 20, 2003)

Summary

A new approach to the computational identification of protein-coding gene structures in genomic DNA sequence is described. It overcomes rigidities inherent in most existing gene prediction methods, for example those based on Hidden Markov Models (HMMs), by supporting a flexible computational model of how sequence signal signals fit together into complete gene structures.

The primary result of the work is a gene prediction tool for the assembly of evidence for individual gene components (features) into complete gene structures. The system is completely configurable in that both the features themselves, and the model of gene structure against which candidate assemblies are validated and scored, are external to the system and supplied by the user. The gene prediction process is therefore tied neither to any specific techniques for the recognition of sequence signals, nor any specific underlying model of gene structure.

The methodology is implemented in a piece of software called "GAZE" which uses a dynamic programming algorithm to obtain (i) the highest scoring gene structure consistent with the user-supplied features and gene-structure model, and (ii) posterior probabilities that each feature is part of a gene. The algorithm includes a novel pruning strategy, ensuring that it has a run-time effectively linear in the length of the sequence without compromising accuracy. The effectiveness of the approach is explored by applying it to the prediction of gene structures in sequences of the nematode worm *C. elegans*.

GAZE allows the integration of gene prediction data from multiple, arbitrary sources. It is important for the accuracy of the system that the various pieces of evidence are weighted appropriately with respect to each other. A novel strategy for the automatic determination of optimal values for these weights is described. The method uses numerical analysis and dynamic programming to maximise a probabilistic accuracy function with respect to the weights. Its effectiveness is demonstrated in the context of the development a gene prediction system for vertebrate sequences using GAZE.

Contents

List of Tables

List of Figures

Preface

Too many people have helped me during my time at the Sanger Centre to name individually. I feel it appropriate however to give some people a special mention.

I first came to the Sanger in October 1998 to work on the Pfam database, under the guidance of Alex Bateman and Ewan Birney. Being the young and impressionable newcomer to bioinformatics that I was, Alex and Ewan have to take a degree of credit/blame for the way I now approach problems in this field. Although my involvement with Pfam has diminished in recent years due to other commitments, Alex in particular has continued to to take an active interest in my scientific development, and for that I thank him.

This thesis represents the result of these other commitments, work which I began in October 1999. During this time, my primary source of guidance has been my supervisor, Richard Durbin. I thank him for ideas, direction, encouragement and not least for tolerating my (what must be sometimes infuriating) indecisiveness.

Declaration of originality

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.

Introduction

The working draft of the human genome is now nearly two years old [112], with announcement of the finished article expected later this year. The near-completion of this effort has seen a redirection of resources, resulting in an acceleration in the genome-sequencing of other organisms studied in experimental biology, such as mouse and zebrafish. According to the National Centre for Biotechnology Information, nearly 900 genomes are either finished or currently being sequenced[1]. The fact that such large scale sequencing is possible represents an incredible achievement, both in technology/engineering, and sheer organisation. However, genomes only become useful resource for science through biological interpretation, i.e. *annotation* of the role of the different parts of the sequence in cellular processes. Without annotation, genome sequencing is, to paraphrase Ernest Rutherford, nothing more than stamp collecting.

The specific problem addressed by this thesis is that of the annotation of the *gene structures* in a genome. Annotation of a genome in terms of its constituent genes and their intron-exon structure allows us to infer a set of proteins for an organism. Furthermore, the genomic context of the genes can provide insight into the regulatory mechanisms that determine where and under what conditions the corresponding proteins are expressed, as well as being a useful resource for experimental biology.

Gene structure annotation of genomic sequence is still most accurately performed by trained experts, combining the results of a number of computational and experimental analyses with biological knowledge and heuristics. This is naturally a slow

[1] http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html, 3rd February 2003

process, and the huge volume of sequence data being generated places an unrealistic demand on the number of experts required to perform this skilled activity. In addition, for the annotation of a large vertebrate genome to be completed in any reasonable amount of time, it is necessary to divide the sequence amongst up to a hundred annotators. This can have the undesirable result that different sections of a genome can be annotated with different standards and procedures. Reliable, completely automated methods for gene structure annotation would therefore firstly cope with the rate at which genome sequence data is being generated and secondly provide gene structure annotation that is *consistent*.

This thesis describes a new approach to the automated prediction of gene structures in genomic DNA sequences. Despite progressive improvement in the accuracy of computational methods in the last fifteen years, they remain imperfect. The problem therefore still attracts considerable research interest both into the biological processes of transcription, RNA processing and translation that determine the gene structure of a genome, and into methods for the recognition of the sequence signals involved in these processes. The integration of new knowledge and methods into complete gene prediction systems is however often inhibited by rigidities of design, such as a fixed assumed underlying model of the compositional and structural properties of genes.

The primary motivation for my research has been to accelerate the integration of new and possibly disparate knowledge and techniques into the gene prediction process. To this end, I have developed a structured framework for the assembly of gene prediction evidence from multiple, arbitrary sources into complete gene structure predictions. Careful design and certain assumptions allow the system to make probabilistic statements about its predictions, and this in turn facilitates a principled approach to the problem of determining an optimal weighting strategy for the various types of evidence employed.

The organisation of the dissertation is as follows. Chapter 1 discusses some of the issues and techniques of computational gene-structure prediction. The aim is to provide an introduction and broad survey, as many of the issues that are directly relevant to the work presented in the remainder of the thesis are expanded upon where appropriate.

After this short review, the dissertation can be viewed as comprising of two parts. The first part (chapters 2 and 3) describes a framework for the integration of arbitrary gene prediction data, and its application to the development of a gene finder for *C. elegans* sequences; the second part (chapters 4 and 5) describes a new approach to probabilistic parameter estimation and its application to the performance-tuning of a gene prediction system for vertebrate sequences.

Chapter 2 describes the details of the framework, implemented in a program called “GAZE”. I briefly explain the elements of the system, with focus on the *configuration file* that controls the assembly of the external evidence into complete gene predictions. I then go on to describe the dynamic programming algorithms used by GAZE for the calculation of the optimal gene structure and posterior probabilities for parts of gene structures, including a novel search-space pruning strategy. To end, I contrast GAZE with other, similar approaches to computational gene prediction.

Chapter 3 describes the application of GAZE to gene prediction in *C. elegans* sequences. I outline the stepwise development of an initial configuration, and explore the effects of extending the model in two ways, first to account for a worm-specific peculiarity of gene structure, and second to make use of sequence similarity information. I also examine the probabilistic aspects of the system and explore some of their potential applications.

Chapter 4 addresses the problem of identifying an optimal weighting for the scores attached to the different types of evidence employed in an integrated gene prediction system. Two methods for estimating optimal weights for the elements of a GAZE configuration are described. The first is based on a classical maximum likelihood approach; the second is a novel method which I call *Maximal Feature Discrimination* (MFD). I contrast these with other similar techniques, particularly those used for Hidden Markov Models.

Chapter 5 describes the application of Maximal Feature Discrimination to the training of a simple GAZE model for gene finding in vertebrate sequences, and compares the results with those obtained using the classical maximum likelihood method. I extend the simple model with each of three types of additional evidence and demonstrate the effectiveness with which MFD is able to determine weights for the new model elements.

Finally chapter 6 concludes the dissertation by briefly summarising the important aspects of the work, and suggests possible areas for further research.