Computational Analyses of Eukaryotic Gene Evolution

Sourav Chatterji

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-110
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-110.html

August 31, 2006

Computational Analyses of Eukaryotic Gene Evolution

by

Sourav Chatterji

B. Tech (Indian Institute of Technology, Kanpur) 2001

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science
with
Designated Emphasis in Computational and Genomic Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Lior Pachter, Chair
Professor Richard Karp
Professor Ian Holmes

Fall 2006

The dissertation of Sourav Chatterji is approved.

Chair	Date
	Date
	Date

University of California, Berkeley

July 2006

Computational Analyses of Eukaryotic Gene Evolution

by

Sourav Chatterji

Abstract

Computational Analyses of Eukaryotic Gene Evolution

by

Sourav Chatterji

Doctor of Philosophy in Computer Science

with

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Lior Pachter, Chair

The recent sequencing of multiple eukaryotic genomes offers an unprecedented opportunity to study the evolution of genomic elements like protein coding genes. The initial step in any such study is to obtain accurate gene annotations. Lack of sufficient experimental evidence necessitates the development of computational annotation tools. This thesis presents algorithms for genome annotation and their applications for studying gene evolution.

We first develop a Gibbs sampling approach for ab-initio identification of genes in multiple orthologous sequences. This approach leverages the evolutionary relationships between the sequences to improve the gene predictions, without explicitly aligning the sequences. We show that excellent performance can be obtained with as little as four organisms. The method overcomes a number of difficulties of previous comparison based gene finding approaches: it is robust with respect to genomic re-

arrangements, can work with draft sequence, and is fast (linear in the number and length of the sequences).

We also develop GeneMapper, a program for transferring annotations from a well annotated genome to other genomes. Drawing on high quality curated annotations, GeneMapper enables rapid and accurate annotation of newly sequenced genomes and is suitable for both finished and draft genomes. GeneMapper uses a profile based approach for mapping genes into multiple species, improving upon the standard pairwise approach.

Finally, these methods are employed to annotate the newly available fruitfly and mammalian genomic sequences. We use these annotations to study the evolution of gene structure through intron gain and loss. We test several previously proposed mechanisms of intron gain and loss. We also study the relationship between intron loss and duplication events. We find that although gene duplication is highly correlated with intron loss, structural changes in genes are not necessarily due to a loss of constraint following gene duplication as previously suggested.

Professor Lior Pachter
Dissertation Committee Chair

To the memory of my grandfather, Shri Sachis Chandra Chatterjee.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my adviser, Lior Pachter for his guidance and support during my five years at Berkeley. I learnt the basics of genomics and bioinformatics from his introductory course and since then he has been an ideal mentor to me. He always encouraged my ideas and pointed me to new directions whenever I encountered stumbling blocks in my research. I would also like to express my gratitude to my thesis committee members Professor Richard Karp and Professor Ian Holmes, for providing invaluable comments and suggestions. Their feedback played an important role in the final composition of the thesis.

I would like to thank Manikandan Narayanan for serving as a sounding board for many of my ideas. I also appreciate his willingness to read my drafts, whether they were good, bad or ugly. I thank Colin Dewey for useful discussions and for constructing innumerable Mercator maps at my request. Many of our discussions blossomed into research ideas for this thesis. I thank fellow graduate students Anat Caspi, Ariel Schwartz and Nicolas Bray for many stimulating discussions. I also thank my friends Alex Fabrikant, Animesh Kumar, Arindam Chakrabarti, Sumit Gulwani and Vinod Prabhakaran for making my stay in Berkeley a pleasant and memorable experience.

I would like to thank my parents for their constant love and encouragement. They have made many sacrifices for my education and have always supported me unconditionally in my endeavors.

My wife Neelita has been a pillar of support during this whole undertaking. I thank her for her unlimited patience and understanding. Her cheerful attitude drove away my anxieties during the ups and downs of graduate student life. Finally, a special thanks to our son Advay, whose impending arrival provided the final push for the completion of the dissertation.

Curriculum Vitæ

Sourav Chatterji

Education

2001	Indian Institute of Technology, Kanpur, India
	B. Tech, Computer Science and Engineering
2006	University of California Berkeley
	Ph.D., Computer Science

Chapter 1

INTRODUCTION

The understanding of the biology of functional elements is one of the most important problems in molecular biology. In this thesis, we focus on the evolution of protein coding genes, the elements of the genome that contain information for biosynthesis of proteins. There has been a considerable amount of research performed on the evolution of genes at the nucleotide and amino acid level. However, in addition to coding for proteins, eukaryotic genes have an exon-intron structure in which the introns are spliced out of precursor mRNAs. This structure allows for diverse phenomena such as alternative splicing, nonsense mediated decay and regulation through untranslated regions(UTRs). In spite of such interesting features, there have been very few systematic studies on the evolution of gene structure. This is because the evolution of gene structure occurs at a much slower rate compared to the evolution of individual nucleotides in a gene and there is relatively less data to study the characteristics of gene structure evolution. Consequently, the origin and evolution of the exon intron structure of genes in eukaryotic genomes is one of the fundamental problems in evolutionary biology.

Most of the initial research about gene structure evolution involved the analysis

of small gene sets [e.g. *Tarrio et al.*, 1998]. More recent studies have been more systematic and involved the comparison of gene structure in large sets of orthologous genes in widely separated eukaryotic genomes [*Rogozin et al.*, 2003; *Roy and Gilbert*, 2005b]. Because of the very large evolutionary distances between the genomes, several questions about gene structure cannot be answered conclusively in these phylogenetically diverse studies. A denser phylogenetic sampling of genomes can help answer these questions more definitively and thus greatly improve the understanding of gene structure evolution. For example, a comparison of related genomes such as human and chimpanzee will help us find recently gained/lost introns and thus understand the mechanisms of intron gain and loss. Unfortunately, insufficient sequence data has hampered any such systematic large scale studies. The recent sequencing of multiple fly, worm and mammalian genomes offers an unprecedented opportunity to understand the evolution of gene structure. The NHGRI webpage on the status of genome sequencing [*http://www.genome.gov/10002154*] currently catalogs twenty five mammalian, twelve fruitfly and five worm genomes that have been sequenced or are being sequenced. The access to such phylogenetically dense whole genome data provides the opportunity to study the evolution of gene structure at different evolutionary timescales.

The principal objective of this thesis is to investigate the evolution of gene structure by comparison of these newly sequenced genomes. The initial step in any such study is to obtain accurate gene annotations. Therefore, the first part of the thesis concentrates on the development of computational tools for accurate annotation of protein coding genes in newly sequenced genomes. As many of the newly sequenced genomes are of draft quality, our methods are robust to sequencing errors and missing sequence. We then use these highly accurate annotations generated by our programs to study the evolution of gene structure.

1.1 Computational Gene Prediction

Automatic identification of protein coding genes is comparatively straightforward in many genomes such as *Saccharomyces cerevisiae* because most genes are intronless and a search for open reading frames (ORFs) identifies almost all genes. Consequently, the annotation of these genomes is virtually complete. On the other hand, vertebrate genes have a complex exon-intron structure which makes accurate gene prediction difficult. The problem of predicting all genes in a vertebrate genome has not been completely solved even after more than fifteen years of active research. In fact, the number of genes encoded by the human genome is still undetermined [*Pennisi*, 2003; *Glusman et al.*, 2006].

Computational genefinding methods can be broadly classified into two main categories, ab-initio methods and evidence based methods. The next two sections briefly reviews previous advances in these areas.

1.1.1 Ab-initio gene prediction

Ab-initio genefinding methods predict gene structure in DNA sequences from first principles without using external biological evidence (such as similarity to known proteins or mRNA). Most of the current generation ab-initio genefinding programs model gene structure by using hidden Markov models(HMMs). A variation of the standard genefinding HMM is shown in Figure 1.1. We only show the states for the forward strand. The HMM has states representing initial exons (E_i^j), terminal exons (E_T^i), internal exons(E_I^j) and introns ($Intron_i$). The HMM has multiple exon/intron states to represent the three possible exon frames. In addition, we have a special state representing single exon genes. The HMM ensures that the predicted gene has a consistent open reading frame(ORF) across introns. Another advantage of using

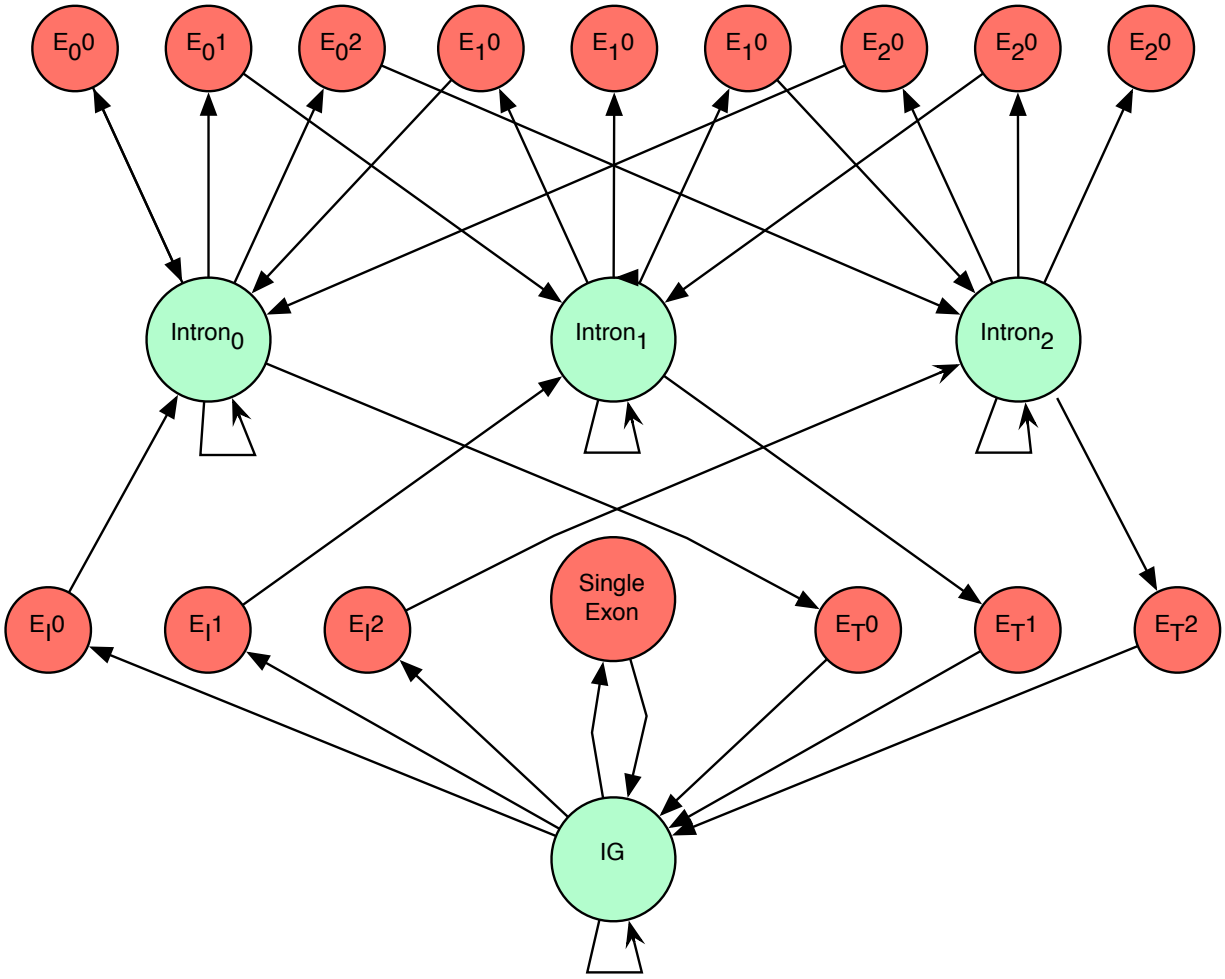


Figure 1.1. A representation of the state space and transitions of the standard genefinding HMM.

HMMs is that they are flexible and it is easy to incorporate new states to model additional features such as UTRs. The transition and emission probabilities for each state in the HMM can be learned from known genes. Given a genomic sequence, a sequence of states (or a parse) in the HMM defines a possible gene prediction and its probability. Most programs use the parse with the highest probability as the gene prediction and this optimal parse can be found efficiently by using the Viterbi algorithm. More details about the use of HMMs in ab-initio genefinding programs can be found in *Burge* [1997].

GENIE [*Kulp et al.*, 1996] was the first program to introduce the use of HMMs in genefinding. However, GENSCAN [*Burge and Karlin*, 1997] is the most widely used ab-initio genefinding program. A careful probabilistic modeling of splice sites, exon lengths and other features made GENSCAN significantly more accurate compared to earlier programs. Recent advances in single organism genefinding has been incremental. For example, Augustus [*Stanke and Waack*, 2003] uses a better intron sub-model, while SNAP [*Korf*, 2004] is a more flexible program that is easily adaptable for a variety of genomes. A big drawback of single organism genefinding programs is that they have a low specificity i.e. they make a lot of wrong predictions.

Comparative genefinding methods try to improve upon single organism methods by exploiting patterns of sequence homology between related genomes. They use the fact that exons are functional and are more likely to be conserved by evolution compared to introns. For example, a comparison of human and mouse orthologous genes in *Waterston et al.* [2002] show that the average sequence identity is 84.7% among orthologous exons, and 68.6% among orthologous introns (the identity among human/mouse orthologous introns is as low as 35% in some other studies such as *Batzoglou, Pachter, Mesirov, Berger, and Lander* [2000]). In addition, 91% of orthologous human mouse exon pairs have the same length, while only 1% of the orthologous introns have the same length. These conservation patterns inspired the development of a new generation of genefinding programs that use comparative analysis to distinguish exons from introns.

Pairwise genefinding programs use comparative analysis of orthologous sequences from two related species such as human and mouse to improve gene prediction. Rosetta [*Batzoglou, Pachter, Mesirov, Berger, and Lander*, 2000], the first comparison based program, uses a two step algorithm for pairwise genefinding by comparing human and mouse sequences. It first creates a global alignment of the two sequences and then uses conservation information from this alignment to make gene predictions.

Other programs such as Twinscan [*Flicek et al.*, 2003], SGP-2 [*Parra et al.*, 2003] and AGENDA [*Rinner and Morgenstern*, 2002] use conservation information from local alignments (obtained from programs such as BLAST [*Altschul et al.*, 1990] or DIALIGN [*Morgenstern*, 1999]) to make gene predictions. SLAM [*Alexandersson et al.*, 2003] and DoubleScan [*Meyer and Durbin*, 2002] use pair HMMs to simultaneously align and predict the gene structure in the two orthologous sequences. Each state of a pair HMM emits a pair of characters, but they retain many properties of HMMs (for example, an efficient Viterbi algorithm). We refer the reader to book by Durbin [*Durbin et al.*, 1998] for more details about pair HMMs.

The next logical step in comparison based genefinding methods is to use additional species to improve performance. Initial studies such as *Dewey et al.* [2004] showed that three way comparison with orthologous mouse and rat sequences helped improve gene prediction in human sequences. Recently, methods that work with an arbitrary number of species have been developed. Shadower [*Boffelli et al.*, 2003; *McAuliffe et al.*, 2004] uses phylogenetic shadowing to find exons in multiple closely related organisms such as primates. EXONIPHY [*Siepel and Haussler*, 2004] and N-SCAN [*Gross and Brent*, 2005] extends this method for more diverse genomes. However, all these methods depend upon the accuracy of the alignment. In this thesis, we develop a novel Gibbs sampling strategy [*Chatterji and Pachter*, 2004, 2005] that exploits conservation information among related species without using an alignment.

Ab-initio methods have been used in several studies [*Dewey et al.*, 2004; *Castellano et al.*, 2001; *Guigo et al.*, 2003; *Wu et al.*, 2004] to predict novel genes. However, in spite of the best efforts of several scientists for over 15 years, ab-initio genefinding programs are not very accurate. To illustrate the inaccuracies, let us examine the example in Figure 1.2 which shows a segment from human chromosome 9 in the UCSC genome browser [*Kent et al.*, 2002]. Coding exons are represented by blocks connected by horizontal lines representing introns. Untranslated regions (UTRs) are displayed as

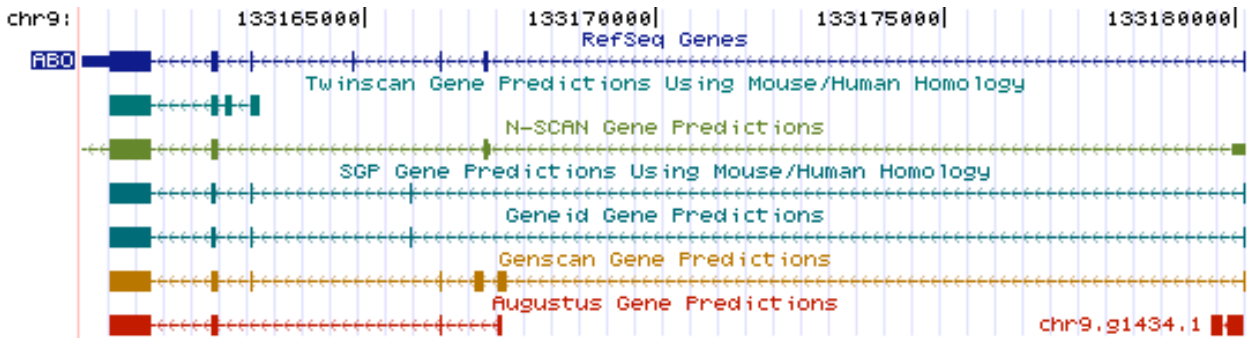


Figure 1.2. Ab-initio gene prediction tracks on a segment of the human chromosome 9 in the UCSC genome browser.

thinner blocks. The RefSeq track is based on experimental mRNA evidence and is the true gene structure. Genscan [*Burge and Karlin*, 1997], Geneid [*Parra et al.*, 2000] and Augustus [*Stanke and Waack*, 2003] are single organism ab-initio genefinding methods. N-SCAN [*Gross and Brent*, 2005], Twinscan [*Flicek et al.*, 2003] and SGP [*Parra et al.*, 2003] are comparison based methods that use conservation information with related species to improve gene predictions. The figure displays the predictions of these programs in this region. Even though the programs are able to correctly predict some of the exons, none of the methods are successful in predicting the gene structure accurately. It is also notable that many of the programs are deficient in predicting gene boundaries. We now summarize the main hurdles that impede the accuracy of ab-initio genefinding programs.

- *Pseudogenes :* Pseudogenes are sequences that are very similar to genes, but with deleterious mutations. Thus a genefinding program can wrongly annotate pseudogenes as coding sequence.

- *Long Introns :* On average, coding exons are much shorter than non-coding introns and the presence of long introns makes genefinding complicated. Furthermore, ab-initio programs are also not very good at finding very short exons (microexons).

- *Conserved non-coding sequences(CNSs) :* Comparison based genefinding programs improve upon single organism genefinding programs by using conservation information to boost exon scores (exons are much more conserved among related species compared to introns). Thus some of these methods find it hard to distinguish exons from CNSs and they can wrongly annotate CNSs as coding sequences.

- *Accurate determination of gene boundaries :* Even though genefinding programs have become fairly accurate in annotating individual exons, they are not good at assembling the exon predictions together to obtain the correct gene structure. Many gene prediction programs often split up or fuse gene boundaries. The accurate determination of gene boundaries is an important open problem and a knowledge of gene boundaries will greatly enhance the accuracy of ab-initio genefinding methods.

1.1.2 Evidence based gene prediction

Evidence based genefinding methods use information not intrinsic to the genomic DNA sequence, such as known cDNA or protein sequences to improve accuracy. Evidence based methods mostly involve aligning the evidence with the genomic sequence and they are significantly more accurate compared to ab-initio methods. However as they are dependent on external evidence, they cannot be used to make novel gene predictions.

High throughput cDNA sequencing is carried out by sequencing expressed sequence tags(ESTs). An EST is a segment of a cDNA clone obtained at random from a cDNA library. More details about cDNA sequencing can be obtained from *Adams et al.* [1991]. Gene prediction from ESTs is a two step process. The first step is EST clustering, which is the joining of overlapping ESTs into clusters to obtain

full length cDNA sequences. The second step is transcript assembly, which is the alignment of the mRNA sequence to the genome to obtain gene annotation. Unigene [*Schuler et al.*, 1996] and TGI [*Quackenbush et al.*, 2001] are the most well known EST clustering algorithms. A complementary set of methods such as EST_GENOME [*Mott*, 1997], BLAT [*Kent*, 2002] and GMAP [*Wu and Watanabe*, 2005] have been designed to align full length cDNA sequences with the genome. More recent methods such as Aceview [*Thierry-Mieg et al.*, unpublished] and ecGENE [*Kim et al.*, 2005] do EST clustering and transcript assembly simultaneously. All these methods need to account for the fact that ESTs can have a relatively high error rate (up to 3%). However, they have not been developed to align cDNA evidence with evolutionarily distant genomes. For example, they are not designed to align human cDNA with the mouse genome.

Protein based methods make use of alignments of known protein sequences with genomic sequences, and form an important component of pipelines such as ENSEMBL [*Birney et al.*, 2004b]. Such programs include DPS [*Huang*, 1996], Procrustes [*Gelfand et al.*, 1996], Genomescan [*Yeh et al.*, 2001] and GeneWise [*Birney et al.*, 2004a]. To some extent, these programs are designed to work with proteins from related species. Although they work quite well with highly conserved proteins, they are not as accurate for diverged protein sequences.

Reference based genefinding methods use gene annotations from a reference species as evidence to predict the gene structure in a target species. In analogy to cDNA based methods, these *reference based genefinding* methods align mRNA from a reference gene to a target sequence, however they exploit additional information about splice sites. Projector [*Meyer and Durbin*, 2004], one of the first reference based methods uses a pair HMM to transfer annotations from the reference species to the target sequence. In this thesis, we introduce a new reference based algorithm called GeneMapper which is significantly more accurate compared to existing reference based

and protein based methods. We use GeneMapper as the main component of our whole genome annotation pipeline.

Another class of evidence based methods such as JIGSAW [*Allen and Salzberg*, 2005] and Exonhunter [*Brejova et al.*, 2005] use multiple sources of evidence to predict gene structure. Theoretically, multiple source methods should outperform single source methods, but we believe that computational genefinding based on multiple sources is still an area of active research. It is also interesting to note that methods such as GeneComber [*Shah et al.*, 2003] improve ab-initio gene prediction by combining predictions from several ab-initio genefinding programs.

1.2 Evolution of Gene Structure

There is enormous diversity in the structure and organization of genes in living organisms. Prokaryotic genes are intronless while most eukaryotic genes have introns that are spliced from precursor mRNA. There is lot of variation in gene structure even within eukaryotic genomes. For example, most yeast genes are intronless whereas human genes have an average of about 8 introns per gene. Such diversity in gene structure has been used to explain the evolution of genomic complexity and determination of population size[*Lynch and Conery*, 2003]. In addition, introns are believed to play a role in natural selection [*Comeron and Kreitman*, 2000], but their exact role is not well understood. Therefore, the origin and evolution of exon-intron structure of genes in eukaryotic genomes remains one of the most important unanswered questions in evolution. We now review the current understanding of gene structure evolution.

1.2.1 Origin of Exon Intron Structure of Genes

Introns are classified according to the mechanisms by which they are removed from the precursor mRNA. The mRNAs that contain Group I and Group II introns are self splicing, i.e. their structure facilitates the removal of introns from the RNA transcripts. The difference between these two intron types is in the mechanism of the self catalytic splicing process. These two intron types have been found in bacterial and organellar genomes. In this thesis, we mostly deal with *spliceosomal introns*. Spliceosomal introns are found in eukaryotic nuclear genomes and a spliceosomal complex is used to splice out the intron from the precursor mRNA. The splicing mechanism of spliceosomal introns resembles the mechanism in Group II introns and this similarity suggests that spliceosomal introns evolved from Group II introns [*Cavalier-Smith*, 1991]. It has been recently hypothesized that the formation of the nucleus coincided with the evolution of exon-intron structure and the incipient function of the nuclear envelope was to allow mRNA splicing [*Martin and Koonin*, 2006].

There are two theories about the origins of exon-intron structure in eukaryotic genomes. The exon theory of genes or "Introns Early" theory [*Gilbert*, 1978; *Gilbert et al.*, 1997; *Roy*, 2003] hypothesizes that genes in the original cell were assembled by exon shuffling. The theory proposes that spliceosomal introns have been lost in prokaryotic lineages and they continue to exist in eukaryotic genomes. The alternative "Introns Late" theory [*Palmer and Logsdon*, 1991] postulates that introns were invented during eukaryotic evolution and they were spread by insertion into unsplit, pre-existing genes.

Spliceosomal introns are absent in all prokaryotic genomes whereas they are widely distributed in eukaryotic lineages. This phylogenetic distribution of introns strongly supports the introns late theory. On the other hand, the introns early theory is supported by intron phase and protein structure correlations among related genomes.

The debate is further muddled by the fact that there has been extensive intron gain and intron loss in eukaryotic kingdoms [*Rogozin et al.*, 2003]. However, there is increasing evidence for more nuanced views of the competing models. For example, it is clear that most introns do not predate the eukaryotic-prokaryotic ancestors and are fairly new. Phylogenetic studies also suggest that the spliceosome was present in the ancestor of all extant eukaryotes [*Collins and Penny*, 2005]. But the central question of whether the eukaryotic-prokaryotic ancestor had any introns is unresolved and is still a subject of vigorous debate.

1.2.2 Mechanisms of Gene Structure Evolution

The evolution of gene structure is a slow process and the exon-intron structure of a gene is highly conserved in related species. For example, a comparison of orthologous genes in the human and mouse genome showed that 86% of the genes have identical numbers of exons [*Waterston et al.*, 2002]. In fact, the exon-intron structure of some genes is highly conserved over very large evolutionary timescales even when the sequence homology of the proteins coded by the orthologous genes is very low [*Betts et al.*, 2001; *Yoshihama et al.*, 2002]. Over these smaller timescales, most of the changes in the gene structure occur through changes in intron length. This is because introns are not under any evolutionary constraints and they are prone to rampant insertions and deletions. Introns can also serve as sinks for transposable elements such as Alus [*McNaughton et al.*, 1997]. In fact, insertions and deletions in introns can be used to infer phylogenies [*Ogurtsov et al.*, 2004]. A more detailed analysis of the evolution of the lengths of orthologous introns has been carried out by *Yandell et al.* [2006].

The exon-intron structure of a gene can also evolve through gain and loss of coding sequence. *Exon shuffling* is one of the most widely studied mechanisms by which

existing exons recombine or duplicate to develop new exon-intron gene structures. Exon shuffling can occur by *exon duplication*, *exon insertion* or *exon deletion*. *Patthy* [1999] catalogs genes that are known to have been formed by exon shuffling. Genes are also known to generate new functions by integrating transposable elements into the coding sequence of a host gene[*Nekrutenko and Li*, 2001; *Lorenc and Makalowski*, 2003]. A gene can also develop a new structure by gain and loss of alternative spliced forms. For example, a recent comparison of orthologous genes in two dipteran species by *Malko et al.* [2006] found that only 80% of *alternative exons* are conserved. Therefore, the remaining 20% of the alternative exons must have been gained or lost during evolution. Lastly, the structure of a gene can be altered through gene fusion and gene fission [*Snel et al.*, 2000].

In this thesis, we concentrate on the evolution of gene structure by gain and loss of introns. These events do not affect the protein coded by the gene but the deciphering of the mechanisms can help us understand the origin and attributes of the exon-intron structure of genes. The loss and gain of introns occur at a much lower rate compared to nucleotide/amino acid substitutions [*Roy and Gilbert*, 2005b]. Even though these events are comparatively rare, introns are known to have been gained and lost in diverse eukaryotic lineages [*Rogozin et al.*, 2003]. We now briefly review the mechanisms of intron gain and intron loss in eukaryotic genomes. For a more detailed review, the reader is referred to *Roy and Gilbert* [2006].

1.2.3 Mechanisms of Intron Gain and Loss

The most popular theory about intron gain postulates that new introns are formed by duplication of previously present introns [*Tarrio et al.*, 1998]. According to this theory, a previously spliced intron can be inserted into an mRNA, which is reverse transcribed to cDNA that recombines with the genome. Group II introns are known

to propagate through a similar mechanism. There is also evidence that some of the recently gained introns have been inserted by this mechanism [*Tarrio et al.*, 1998; *Coghlan and Wolfe*, 2004]. However, all evidence for this theory seems to be indirect and there is no direct proof of insertion of introns by this mechanism.

Another theory for intron gain first espoused by Francis Crick [*Crick*, 1979], hypothesizes that novel introns arise by insertion of transposons. There are multiple examples of recently inserted transposons are spliced out of precursor mRNA [*Giroux et al.*, 1994]. These examples clearly prove that new introns can be formed by insertion of transposable elements. However a recent study of comparatively recent introns in worms in *Coghlan and Wolfe* [2004] suggests that very few novel introns are repeat elements. The role of transposons in the evolution of gene structure is also muddled by the fact that many genes are known to gain new functions by recruiting transposons as coding sequence [*Nekrutenko and Li*, 2001].

The classical theory of intron loss states that introns are lost by recombination of a reverse transcribed mRNA transcript with the genome, resulting in the loss of introns [*Bernstein et al.*, 1983]. As reverse transcriptase works from the 3' to the 5' end and is often incomplete, this theory predicts that more introns should be lost from the 3' end compared to the 5' end. Because of this mechanism of reverse transcriptase, this theory also predicts that many introns should be lost in tandem. An alternative theory of intron loss hypothesizes that introns are lost by genomic deletion [*Kent and Zahler*, 2000; *Cho et al.*, 2004]. This theory predicts that intron loss is inexact. It also predicts that intron loss is random and not biased towards the 3' ends of genes. Large scale comparative studies of orthologous genes have resulted in mixed results in validating the predictions of these competing theories of intron loss.

It seems that none of the theories can explain the evolution of introns. However, scientists studying intron evolution have been constrained by limited data and a novel

mechanism might explain all observed intron gains. We believe that a more dense phylogenetic sampling will help us in answering these questions more accurately.

1.3 Overview of the thesis

The goals of this thesis are twofold. The first goal is to develop a comprehensive computational system for accurate annotation of protein coding genes in a newly sequenced genome. As we have discussed in the previous section, ab-initio and evidence based methods are complementary to each other. Ab-initio methods are useful in finding novel genes, while evidence based methods are much more accurate in predicting gene structure. Therefore, we have developed novel ab-initio and evidence based methods as a part of this thesis. We also use these tools to develop a gene prediction pipeline for annotating newly sequenced genomes. The second goal of this thesis is to study the evolution of gene structure in a wide range of eukaryotic genomes. We have mainly concentrated on the gain and loss of introns in evolution.

Our research in ab-initio genefinding has mostly concentrated on developing comparative methods, as they are much more accurate compared to ab-initio methods. As described earlier, comparative genefinding methods exploit the fact that exons are functional and are more likely to be conserved compared to introns. Most comparative methods such as Rosetta [*Batzoglou et al.*, 2000] and Twinscan [*Flicek et al.*, 2003] obtain this conservation information from either global or local alignments. However alignments of orthologous genes are not always reliable, especially because conserved exons are much shorter than non-conserved introns. In Chapter 2, we develop a novel Gibbs sampling strategy that exploits conservation information without using alignments. We believe that this strategy will be an important approach for improving accuracy of genefinding programs in the badly aligned regions of the genomes.

In Chapter 3, we describe GeneMapper, a program for annotating newly sequenced genomes by transferring gene annotations from well annotated reference genomes (such as *D. melanogaster* and *H. sapiens*). The rationale behind developing GeneMapper is that a lot of resources have been invested in annotating genomes of these model organisms and and it is unreasonable to expect similar efforts to be expended for the myriad of genomes that are now being sequenced. GeneMapper provides an alternative way to accurately annotate these genomes by transferring annotations from reference genomes. For example, we have used *D. melanogaster* FlyBase annotations to annotate the newly sequenced fruitfly genomes. If a gene is to be mapped into multiple species, GeneMapper uses a novel profile based approach that is an improvement over the standard pairwise approach. We show that GeneMapper is much more accurate compared to existing programs such as Projector and GeneWise. GeneMapper is designed to be robust to missing sequence and sequencing errors, so that it is suitable for both finished and draft genomes.

GeneMapper maps known genes from a reference genome to newly sequenced genomes and thus implicitly creates a data set of orthologous genes for studying the evolution of genes. In Chapter 4, we use this data set to study the evolution of genes in multicellular eukaryotes. We are particularly interested in the evolution of gene structure. The sequencing of multiple dipteran and mammalian genomes opens new vistas as the more dense phylogenetic sampling of species can help us answer many unresolved questions about evolution of gene structure. We use GeneMapper annotations to study the gain and loss of introns in mammalian and dipteran (fruitfly) genomes. We test previous theories of intron gain and loss. In addition, our findings in Diptera also provide an explanation for the 5' bias in the position of introns in eukaryotic genomes. We also study the relationship between intron loss and duplication events. We find structural changes in genes are not necessarily due to a loss of constraint following gene duplication as previously suggested.

Chapter 2

COMPARATIVE GENEFINDING BY GIBBS SAMPLING

With the publication of the mouse [*Waterston et al.*, 2002] and rat [*Gibbs et al.*, 2004] genomes, it has become apparent that comparative-based gene finding methods such as SGP2 [*Parra et al.*, 2003], SLAM [*Alexandersson et al.*, 2003] and TWINSCAN [*Flicek et al.*, 2003] improve upon single organism gene finding methods as implemented in GENSCAN [*Burge and Karlin*, 1997] or GENIE [*Kulp et al.*, 1996]. Comparative-based gene finders are more accurate because conserved regions in genomes are more likely to be functional (and in particular coding), and therefore an alignment of a pair of homologous sequences can be used to assist in gene identification. It appears intuitive that the addition of more sequences (and their alignments) should improve the quality of gene predictions. However there are multiple serious issues in developing multiple species gene prediction algorithms. The problem of accurately aligning large genomic regions is non-trivial (especially in the case of sequence inversions and rearrangements). The problem of accurate alignment of orthologous genes is further exacerbated by the fact that conserved exons are

much shorter compared to highly divergent introns. In addition, many of the pairwise gene prediction methods become computationally intractable when generalized to more than two sequences (for example, the running time of SLAM for k sequences of length m is $O(m^k)$).

The Gibbs sampling method has been widely used for sequence analysis after it was successfully applied to the problem of identifying regulatory motif sequences upstream of genes [*Lawrence et al.*, 1993]. Since then numerous variants of the original idea have emerged: however, in all cases the application has been to finding short motifs in collections of short sequences (typically less than 100 nucleotides long). In this chapter, we introduce a Gibbs sampling approach for identifying genes in multiple large genomic sequences up to hundreds of kilobases long. This approach leverages the evolutionary relationships between the sequences to improve the gene predictions, without explicitly aligning the sequences. As we will see, the approach we propose avoids the need for a pre-processed multiple alignment of the sequences, and in fact implicitly produces a partial alignment which is robust with respect to genomic rearrangements, large insertions/deletions and other evolutionary events which complicate the multiple alignment of large genomic regions.

We begin by describing in more detail our approach to finding genes in multiple vertebrate sequences. We apply Gibbs sampling to learn the parameters of a suitable hidden Markov model, from which we can infer gene annotations. This is equivalent to the missing data formulation of *Tanner and Wong* [1987]. In section 2.2, we present results of tests of the method on multiple large genomic regions.

2.1 The collapsed Gibbs sampler for hidden Markov models

We briefly review the fundamentals of Gibbs sampling for the missing data problem, and proceed to describe its application in the context of gene finding. Our notation and presentation borrow from a number of sources including standard texts [*Durbin et al.*, 1998] and the papers by Liu [*Lawrence et al.*, 1993; *Liu et al.*, 1995]. We denote a state sequence path by z, and the i^{th} state in a state sequence path by z_i. An HMM is described by *transition probabilities*:

$$a_{kl} = P(z_i = l | z_{i-1} = k) \tag{2.1}$$

and *output probabilities*:

$$e_k(b) = P(x_i = b | z_i = k). \tag{2.2}$$

Here the letters l, k denote states and b an output symbol. The set of all parameters is denoted by θ. In hidden semi-Markov models (or generalized HMMs) we also have duration distributions from which we sample durations for each state.

We will assume that we have a set of n sequences $\mathbf{y} = y^1, \ldots, y^n$ that have been generated independently by the hidden Markov model but for which the state paths are unknown. We would like to maximize the log likelihood

$$log\ P(y^1, \ldots, y^n | \theta) \ = \ \sum_{j=1}^{n} log\ P(y^j | \theta). \tag{2.3}$$

This is of course the standard parameter estimation problem.

Suppose we have a distribution $p(x_1, \ldots, x_n)$ from which we would like to sample, but it is difficult to do so because of complex dependencies between the random variables. Gibbs sampling is an iterative (provably correct under appropriate assumptions) method for sampling from the distribution in the case where the conditional

distributions are easy to sample from. The method is simply to iteratively sample from the conditional distributions

$$p(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)(1\leq i\leq n). \tag{2.4}$$

In the case of HMM parameter estimation where the path information is missing, we would like to sample from the joint posterior distribution

$$p(\mathbf{z},\theta|\mathbf{y}). \tag{2.5}$$

The random variables $\mathbf{z} = z^1,\ldots,z^n$ (denoting state paths) and θ now play the roles of the x_is above and the method is to iteratively sample from the conditional distributions

$$p(z^i|\mathbf{z}^{[-i]}\mathbf{y},\theta)\ (1\leq i\leq n) \tag{2.6}$$

and the posterior distribution

$$p(\theta|\mathbf{y},\mathbf{z}). \tag{2.7}$$

In the first equation, $\mathbf{z}^{[-k]}$ denotes the data $\mathbf{z}$ with z^k missing. A practical implementation of the Gibbs sampler requires that we are able to efficiently sample both the paths and the parameters θ (from their posterior distributions). We provide the details of these two key steps of the Gibbs Sampler in the next two sections.

2.1.1 The path sampling step

As we have mentioned, the key ingredient of a Gibbs sampler for an HMM is an efficient method for sampling a state path $z^i = z^i_1,\ldots,z^i_L$ in the sequence $y^i = y^i_1,\ldots,y^i_m$ from

$$p(z^i_1,\ldots,z^i_L|y^i_1,\ldots,y^i_m,\mathbf{z}^{[-i]},\mathbf{y}^{[-i]}). \tag{2.8}$$

This can be done using a standard forward-backtrack type algorithm (e.g. *Durbin et al.* [1998]). For completeness, and since it is rarely explicitly outlined in hidden

Markov model texts or tutorials, we provide the detailed algorithm here, and illustrate it in a slightly more general graph theoretic setting:

Lemma 1 *Let G be a directed acyclic graph with source s and sink t. Let each edge $e = (v_i, v_j)$ of G have a weight $w(e)$ (we also use the notation $w(e) = w(v_i, v_j)$), and each node weight $w(v_i)$. Assume without loss of generality that*

$$\sum_{paths\ \ P=(v_1,\ldots,v_{k(P)})} w(v_1) \prod_{i=2}^{k(P)} w(v_{i-1}, v_i) w(v_i) = 1.$$

It is possible to pick a path P consisting of $s = v_1, v_2, \ldots, v_{k(P)} = t$ at random in time $O(n)$ so that

$$Pr(picking\ P) = w(s) \prod_{i=2}^{k(P)} w(v_{i-1}, v_i) w(v_i).$$

Proof: The proof is by induction on the maximal length of a path between s and t. The base case $n = 1$ is trivial. Suppose the theorem is true for the case when the length of the longest path between s and t is n, and consider the case where the max. path has length $n+1$. Suppose the edges out of s have weights $w(e_1), \ldots, w(e_k)$, and are adjacent to vertices $v_1, \ldots, v_k$ respectively. Suppose we have computed weights $\beta(v_i)$ for all the vertices adjacent to s, where $\beta(v_i)$ is the sum of the weight of all the paths from v_i to t (this step can be done using dynamic programming, and is the backward algorithm for HMMs). Our path picking algorithm is to pick an edge from s at random with probability $w(s)w(e_i)\beta(v_i)$, at which point the distance from v_i to t is at most n, so by induction we can choose from there a path P which has weight z, and which has been selected with probability $\frac{z}{\beta(v_i)}$. Observe that the weight of the path from s to t is $w(s)zw(e_i)$, and that it has been selected with probability $\frac{z}{\beta(v_i)} w(s)w(e_i)\beta(v_i) = w(s)zw(e_i)$.

The sampling method above can also work in reverse by backtracking from t instead of starting from s, in which case one first needs to compute the forward

variables $\alpha(v)$. This algorithm also has fast memory efficient implementations [*Cawley and Pachter*, 2003].

In a hidden Markov model we have two types of parameters: the output probabilities and the transition probabilities. These parameterize multinomial distributions and geometric distributions respectively. Thus, for prior distributions on θ we use the conjugate priors: the Dirichlet distributions for the multinomial output data and the β distribution for the geometric data. In semi-hidden Markov models, we have states whose lengths are distributed according to arbitrary distributions which may not be convenient to deal with in the Bayesian framework. Fortunately, in the case we are interested in (generalized HMMs for gene finding), the only generalized states are the exon states, which are conveniently modeled with Gamma distributions, and these have the Gamma distribution as the conjugate prior.

2.1.2 The collapsed Gibbs sampler

In the traditional Gibbs sampling setup, the parameters of the hidden Markov model are sampled together with state paths (in an alternating fashion). Liu [*Liu*, 1994] has pointed out that the sampling of parameters can be avoided by integrating out the parameters and if the sequences are sufficiently long and there are enough of them, the integration can be efficiently approximated. We now review the argument, and in the process correct some small mistakes in [*Liu et al.*, 1995], in which the proof incorrectly skips the necessary requirement of many sequences (and not just long sequences).

We begin with the observation that

$$p(\mathbf{z}|\mathbf{y}) \propto p(\mathbf{z},\mathbf{y}) = \int_\theta p(\mathbf{z},\mathbf{y}|\theta) f(\theta) \tag{2.9}$$

$$\propto \prod_i \Gamma(h(\mathbf{y_i}) + \alpha_i). \tag{2.10}$$

where $f(\theta)$ is the prior distribution for θ and consists of a product of gamma distributions each with parameters α_i. The notation $h(y)$ denotes counts obtained from the data y. Note that in the final product term above, $\mathbf{y_i}$ consists of subsets of the data which are determined by $\mathbf{z}$.

Now $p(z^k|\mathbf{z}^{[-k]}, \mathbf{y}^{[-k]}) \propto p(\mathbf{z}|\mathbf{y})$ and therefore

$$p(z^k|\mathbf{z}^{[-k]}, \mathbf{y}^{[-k]}) \propto \prod_i \frac{\Gamma(h(\mathbf{y}_i^{[-k]}) + \alpha_i + h(y_i^k))}{\Gamma(h(\mathbf{y}_i^{[-k]}) + \alpha_i)} \tag{2.11}$$

where $\mathbf{y}^{[-k]}$ denotes the data $\mathbf{y}$ with sequence k missing (and similarly for $\mathbf{z}$). The subscript i and the notation $\mathbf{y}_i$ denotes the fact that the product will range of subsets of the data, these being determined by the particular state paths.

We now use the fact that if $b << a$

$$\frac{\Gamma(h(a+b))}{\Gamma(h(a))} \approx h(a)^{h(b)} \tag{2.12}$$

where notation $h(a+b)$ denotes the "sum" of the counts, in other words the counts obtained from considering the union of the two data sets.

From this we infer that the predictive distribution $p(z^k = i|\mathbf{z}^{[-k]}, \mathbf{y}^{[-k]})$ is proportional to the probability of path i in the hidden Markov model. Although not explicitly pointed out by Liu, the validity of the approximation depends on both the number and size of the sequences, and on the amount of data available for learning each state. For example, in the block motif application there is a lot more background sequence data than position data (for the former one needs long sequences, for the latter many of them).

2.1.3 The block-motif Gibbs sampler

A special case of the above framework is the Gibbs sampler for block-motifs of which was introduced in a seminal paper in *Science* [*Lawrence et al.*, 1993] and which

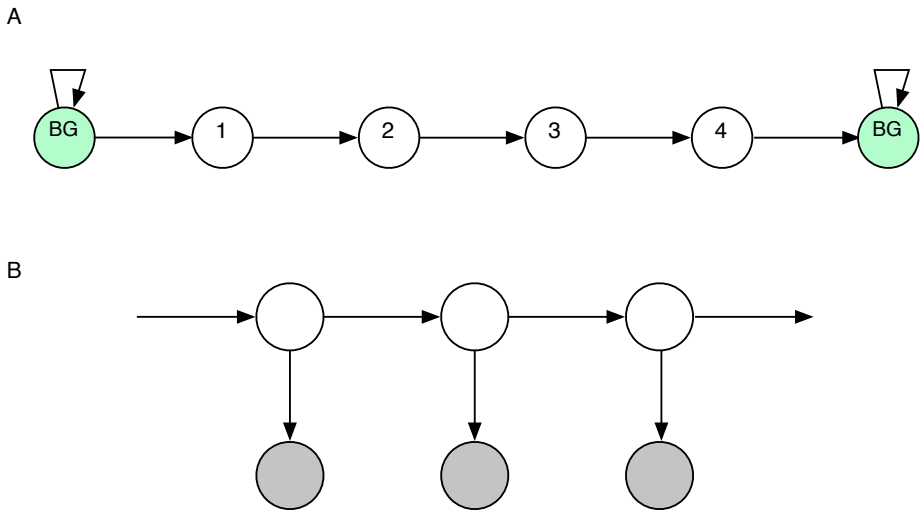


Figure 2.1. The hidden Markov model representation of the block-motif sampler for W=4. Panel A shows the the allowable transitions between states. Panel B shows the graphical model. In this representation, shaded circles correspond to observed random variables and unshaded circles to hidden random variables. The arrows represent conditional dependencies among random variables. The BG (background) states are used for sequence before and after the motif.

has found widespread application in the detection of binding site motifs in genomic sequences. Suppose that our motifs have width W. In this case, we have k sequences each of length n and generated from an HMM (shown in figure 2.1 for W=4).

The description of θ is as follows: The self transition probabilities on the end states are fixed and equal. The remaining parameters consist of output probabilities for states 1, ... W, which do not have dependence on previous sequence. It is straightforward to derive the Liu block-motif predictive update formula (equation 5 in *Liu et al.* [1995]) using the above model.

2.1.4 The gene finding Gibbs sampler

We applied the collapsed Gibbs sampler to a gene finding HMM (state space shown in Figure 2.2). We therefore consider the coding exons and introns of a gene (and

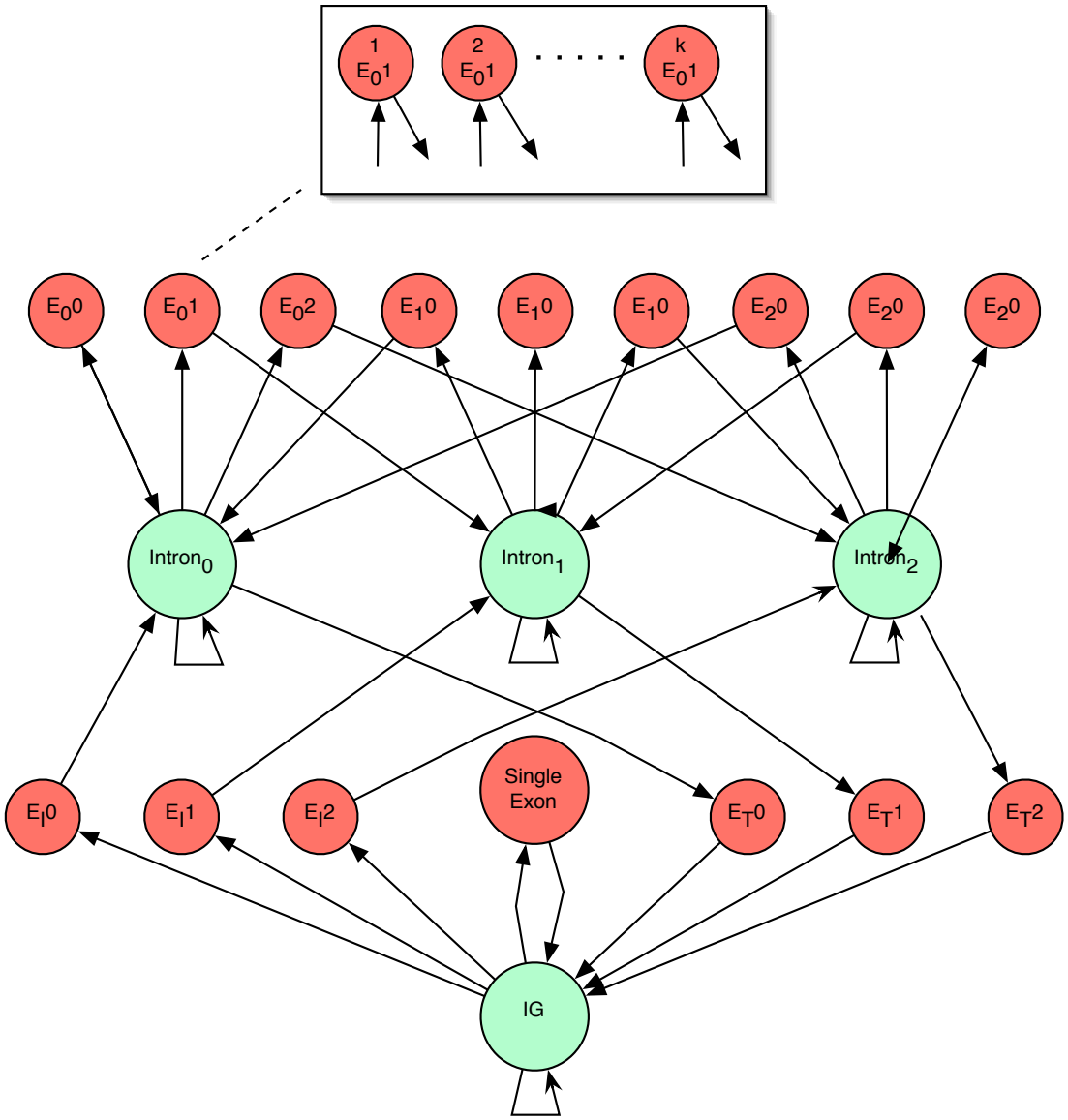


Figure 2.2. States space of the gene finder. Only the forward strand states are shown. The triplets of exon states are required to ensure a contiguous open reading frame across introns. The state BG is the intergenic state, and $Intron_0, Intron_1$ and $Intron_2$ are for the three phases of intron. The exon states are divided into initial (E_Ij), terminal (E_Tj) and internal (E_ij) types, with an additional special state for single exon genes. Further details about the basic gene finding model can be found in [3]. Our model is an extension of the basic model, in that each exon state actually consists of k states (corresponding to k different gene models). Details are shown only for the $E_0 1$ state (see box in figure). Thus, if the model is to be used to predict up to k genes, it contains 3 intron states, one intergenic state, and $16k$ exon states

possibly other features) to be encoded in the hidden state paths of an HMM, and comparative sequence data is viewed to consist of independent realizations of sequences from the model. This setup assumes the sequences are related via a "star" evolutionary tree, that is, a tree where all the leaf edges meet at one point. This assumption is realistic for the sequences we test with in this thesis (mammalian sequences roughly equidistant from each other); however in general it is of course preferable to use a model that incorporates the true phylogenetic relationships between the sequences. We defer discussion of this until the final section of the chapter. In summary, we view the gene finding problem as the solution to the parameter estimation problem for the HMM.

The SLAM gene finding program [*Alexandersson et al.*, 2003] was modified to work as a single organism non-homogeneous gene finder, thus being very similar to the GENSCAN program [*Burge and Karlin*, 1997] (except for the non-homogeneity and some extra states, described below). We omit details of the signal models used and refer the reader to *Burge and Karlin* [1997]. We used the parameters of SLAM in single species mode as the priors or pseudo-counts for the Gibbs Sampler. For the results described below, we only learned exon frequencies (modeled with a 5th order HMM) and lengths for the generalized exon states.

A particular sequence can have two or more genes with widely different characteristics (different exon frequencies and exon lengths) and we need to have different models for different gene types in the sequence. In our Gibbs sampler, every exon state is composed of k states, where k is the number of gene models (see box in Figure 2.2). Furthermore, the hidden Markov model is *non-homogeneous*; in particular, the transition probabilities to the different gene classes change with the sequence location. Formally, the probability of the exon e under our model is $\sum_{i=0}^{k} a_i^t P(e|M_i)$, where a_i^t is the probability of choosing gene class M_i at position t and $P(e|M_i)$ is the probability of exon e in gene class M_i. It is important to note that the probabilities

a_i^t depend on t, that is they are allowed to change over time (in our case "time" is the location of the exon).

In principle, it is necessary to learn the number of classes k, and also the non-homogeneity of the chain as part of the sampling process. In order to speed up the computations, we circumvented this problem by directly comparing predicted peptides in the learning step of the Gibbs sampler to identify the number of classes, and where they appear in each sequence, on the basis of significant hits.

More precisely, we constructed an undirected graph $G = (V, E)$, with one node v_i for each predicted gene. The predicted peptides of the genes were compared using translated BLAT [*Kent*, 2002]. Two genes were defined to be similar if there was a significant hit, and this was represented by an edge (v_i, v_j) in the graph. The connected components of the graph were used to define the gene classes M_i. The exon parameters for gene class M_i were in turn learned from the ith connected component. Exon counts were adjusted by "pseudo-counts" based on standard gene finding parameters. This can be interpreted probabilistically as a mixture model for the exon states.

In order to obtain a probability for a gene in the model, it was necessary to know (or learn) the transition probabilities a_i^t. These were set so that at position t, $a_j^t = 1$ for some j and $a_i^t = 0$ for all $i \neq j$. This condition enforced the use of only one gene model per exon. The transition probabilities were set by assigning an exon the probability: $P(e) = \max_i P(e|M_i)$.

2.2 Results

2.2.1 Data

We tested our results on mammalian sequences from the NISC Comparative Vertebrate Sequencing Project. A region was selected for testing if it satisfied a number of criteria: No alternatively spliced genes were allowed to lie in the region, and the length in each organism was required to be less than 0.3Mb (in order to reduce memory usage). This latter restriction should not be necessary in general if the approach is used with a memory efficient gene finder on a machine with large amounts of RAM. GENSCAN is often run on sequences megabases long. Finally, regions were required to contain sequence from the human (where reliable annotations could be obtained), as well as three other mammalian species roughly equidistant from each other. The final criterion was ensured by selecting regions with one sequence from cat or dog, one from cow or pig, and one from mouse or rat. Thus, each region had sequence from four organisms.

We identified ten suitable gene regions from the NISC comparative sequencing project which satisfied our criteria. One of these was the cystic fibrosis (CFTR) gene region, which is fairly large and was therefore subdivided into five smaller regions, each of size less than 0.3Mb. This gave us a total of 14 regions with 20 genes, which we call test set 1. In addition, we created two artificial test sets to evaluate the robustness of various genefinding programs with respect to evolutionary events such as rearrangements. These sets are described in the next section.

2.2.2 Testing

The Gibbs sampler was compared to four freely available and widely used programs that have been used for whole genome annotations: the GENSCAN program

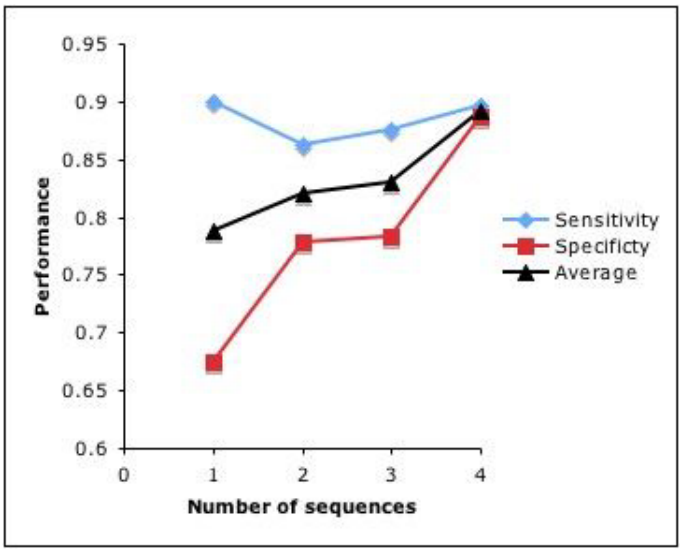


Figure 2.3. Variation of the performance of the Gibbs Sampler with the number of sequences used

Program	Nucleotide Level			Exon Level		
	Sn	Sp	Avg	Sn	Sp	Avg
GENSCAN	0.918	0.548	0.733	0.777	0.518	0.648
TWINSCAN	0.692	0.856	0.774	0.440	0.513	0.477
SGP2	0.943	0.586	0.764	0.755	0.530	0.642
SLAM	0.791	0.881	0.836	0.632	0.527	0.580
Gibbs sampler	**0.897**	**0.886**	**0.891**	**0.714**	**0.628**	**0.671**

Table 2.1. Performance of the gene finders on test set 1.

	Test set 2 (before rearrangement)				Test set 3 (after rearrangement)			
Program	Nucleotide Level		Exon Level		Nucleotide Level		Exon Level	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
GENSCAN	0.911	0.680	0.771	0.612	0.866	0.652	0.748	0.594
TWINSCAN	0.694	0.895	0.465	0.604	0.665	0.853	0.465	0.598
SGP2	0.957	0.723	0.771	0.620	0.914	0.704	0.763	0.621
SLAM	0.927	0.911	0.718	0.566	0.438	0.936	0.250	0.646
Gibbs sampler	**0.939**	**0.950**	**0.763**	**0.735**	**0.885**	**0.910**	**0.740**	**0.703**

Table 2.2. Effect of rearrangements. Performance of the gene finders before and after artificially induced rearrangements

used for single organism gene prediction, and the SGP2, SLAM and TWINSCAN programs used for pairwise comparative-based prediction. SLAM simultaneously aligns

and annotates a pair of sequences, and the alignment is required to be global. TWINSCAN and SGP2 predict genes in one organism, but modify the exon probabilities based on TBLASTX local alignments to the second sequence. SLAM, TWINSCAN and SGP2 were tested with human and mouse/rat. Note that comparison based gene finding approaches give similar results with either mouse or rat as they are equidistant to human. The results of the comparisons are shown in Table 2.1.

A second test was performed to measure the performance of the Gibbs sampler on sequences with a rearrangement. A randomly picked subset of 8 regions from test set 1 was selected (we call this test set 2). Pairs were then concatenated, and then one of the sequences in human was reverse complemented (i.e. we performed an artificial inversion in human). We call this test set 3. The results of the gene finders on test sets 2 and 3 are shown in Table 2.2.

A final test was performed to assess the effect of the number of species on the performance of the Gibbs Sampler. In this experiment, the number of sequences input to the Gibbs Sampler were varied from one to four species. The results of this experiment are summarized in Figure 2.3.

There are two standard tests for measuring the accuracy of a gene finder. At the nucleotide level, sensitivity and specificity calculations are performed to find the fraction of coding bases covered by predictions and conversely the fraction of predicted bases covering true exons. Another standard test is the exon level test, which measures the accuracy of gene predictions at the exon level; i.e. exons are required to be predicted with exact matching boundaries in order to be counted. For more details about these tests, the reader is referred to *Burset and Guigo* [1996].

2.2.3 Conclusions

We found that our Gibbs Sampler outperformed the other comparison based gene finders in both nucleotide and exon level results. It is well known that GENSCAN suffers from an over-prediction problem and this is reflected in very poor specificity in the tested regions, especially in nucleotide level results. There is scope for improvement in exon level results for our Gibbs sampler, though the results are already better than for the other comparative-based programs. Future development will include the learning of splice sites, which holds the promise of improving these results.

The rearrangement test confirms that the results for single organism gene finders such as GENSCAN don't change with regard to rearrangements. The sensitivity of SLAM drops dramatically, because SLAM only predicts genes which have a consistent open reading frame in both organisms, and this is impossible with a rearrangement in the human. The results of TWINSCAN and SGP2 are fairly stable because they use local TBLASTX alignments (although there did appear to be some drop in exact exon sensitivity, perhaps an artifact from a bad local alignment). The Gibbs sampler maintains its high level of specificity and sensitivity in spite of rearrangements. It is interesting to observe that the gene predictions made by the sampler can therefore be used to locate rearrangements. Applications include the "seeding" of global multiple alignment algorithms.

We also performed an experiment to assess how the number of species affected the performance of the Gibbs Sampler. For a single sequence, the Gibbs Sampler is equivalent to a single species genefinder and has high sensitivity/low specificity. For two species, there is a small drop in sensitivity but a significant increase in specificity. There are small increases in both sensitivity and specificity as we go from two to three species. The increase in specificity and sensitivity is much larger when we go from three to four species. We therefore see that increasing the number of sequences

from two to four leads to a significant improvement in the accuracy of the genefinder. It will be interesting to see whether there is room for further improvement as more multiple species data become available.

2.3 Discussion

This chapter describes one of the first genefinding programs for homologous sequences from multiple (> 2) species. The running time for the program is $O(kNL)$ where k is number of sampling iterations, N is the number of sequences and L is the maximum length of a sequence. We have found that for our application the sampler converges after six iterations (with the exception of minor changes to some boundary predictions of exons). The memory used is proportional to the memory requirements for a single species gene finder (linear in the size of the sequences). Thus, the Gibbs sampling strategy for gene finding is extremely efficient, especially in comparison to existing comparative-based gene finding methods which require either a local or global alignment of the sequences (quadratic time in the lengths of the sequences for a pair of sequences).

A key feature of the Gibbs sampling approach to gene finding is that the method is robust with respect to rearrangements. Gene rearrangements (where the order of genes is not preserved between organisms) have been the Achilles heel of comparative-based gene prediction programs, because sorting out the rearrangements during an alignment phase is usually non-trivial. Because the Gibbs sampler only predicts in one sequence at a time, rearrangements present no problem. In fact, it should be possible to apply our strategy to infer the locations of rearrangements between sequences. This should be very useful for annotating multiple *Drosophila* genomes, where high transposon activity has resulted in frequent rearrangements. Furthermore, the approach is robust with respect to gene duplications in any of the sequences.

Another improvement over other comparative-based based species gene finders is that the sequences can be of draft quality (some sequence missing, or contigs not fully assembled). Draft sequence can be annotated irrespective of its order, so contigs can just be glued together for analysis. This feature should be extremely useful in the coming years as draft sequence emerges for multiple sequences (primarily from comparative BAC mapping and sequencing projects).

In the current implementation, we have not taken into account the evolutionary tree of the species involved. By selecting species which are mutually distant from each other we have circumvented this problem by effectively treating the sequences as independent realizations from a GHMM. This is analogous to many of the models used for motif finding. We have experimented with different weighting schemes based on phylogeny, and with the possibility of integrating probabilistic phylogenetic methods into the Gibbs sampling framework, but initial tests indicate that a rather sophisticated solution will be required. Further analysis along these lines is beyond the scope of this thesis, but is an obvious direction for future research. Other improvements consist of learning more parameters. We are currently exploring the use of new, flexible gene finders for which we can easily tune the parameters, and with which we can learn splice site probabilities, transition probabilities to exons and other gene features. States will also be added for repetitive sequence and conserved noncoding sequence. As with phylogenetic sampling, the exhaustive search of parameter space and careful analysis of the best parameters to use for prediction is beyond the scope of this thesis.

In summary, the results we have obtained are very encouraging and suggest that the Gibbs sampling approach to gene finding is accurate, scalable, and well suited for comparative gene finding with multiple organisms.

Chapter 3

GENEMAPPER : REFERENCE BASED ANNOTATION

With large scale sequencing of vertebrate, fly and worm genomes now underway, it is imperative to develop methods that produce high quality annotations of these newly sequenced genomes. Lack of genome-wide full length cDNA sequences for these species will make it virtually impossible to completely annotate these genomes using cDNA based methods such as Aceview[*Thierry-Mieg et al.*, unpublished]. An alternative approach is to transfer reference annotation from a well-annotated genome (such as human and D. melanogaster) to other (possibly draft) genomes. We call this *reference based annotation*. In fact, annotation systems such as ENSEMBL[*Birney et al.*, 2004b] already incorporate reference based annotation as part of their gene prediction pipelines.

The rationale behind the reference based approach is that a lot of resources have been invested in annotating genomes of model organisms, and it is unreasonable to expect similar efforts to be expended for the myriad of genomes that are now being sequenced. The status of current annotation projects for various insect and chordate

genomes is shown in Table 3.1. In the case of vertebrate genomes, the human genome provides an excellent source of reference annotations suitable for transfer. In addition to having extensive numbers of cDNA sequences and a fairly complete RefSeq gene annotation, the human genome annotation also consists of a manual annotation component. By contrast, the other vertebrate genomes have insufficient cDNA sequence. In fact, many genome projects lack sufficient resources to run some of the existing ab-initio gene prediction programs. The reference based annotation tool we have developed, called GeneMapper, can be used in such cases to transfer human annotations. GeneMapper provides a comprehensive annotation that, as we show, is surprisingly accurate. A similar argument can be made for other clades. For example, Drosophila melanogaster is an extensively studied model organism and there is a well curated FlyBase database[*Drysdale et al.*, 2005] of supporting annotations. GeneMapper has been used to provide high quality annotations of the newly sequenced fruitfly genomes by transferring the FlyBase annotations.

GeneMapper has been influenced by and is in the same category of gene-finding methods as Projector[*Meyer and Durbin*, 2004]. Projector uses gene annotations from a reference species as evidence to predict the gene structure in a target sequence. In analogy to cDNA based methods, Projector aligns mRNA from a reference gene to a target sequence, however it exploits additional information about splice sites. This is accomplished by using a pair hidden Markov model to transfer annotations from the reference species to the target sequence.

GeneMapper uses a bottom up approach to predict the gene structure. First, each reference exon is aligned to a target genome and these alignments are then joined to build a gene structure. As exons are much shorter than introns, this approach makes use of dynamic programming with a fairly sophisticated codon evolution model to provide detailed alignment of exons. GeneMapper also uses a novel mapping process that exploits the phylogeny of the reference and target species to obtain more

Organism	**ESTs**	**mRNAs**	**RefSeqs**	**Manual**	**Ab-initio**
Homo sapiens	6134812	207905	24293	22421	5
Pan troglodytes	4983	947	None	None	3
Macaca mulatta	52754	1766	None	None	None
Canis familiaris	349306	1666	None	45	2
Bos taurus	702434	8046	None	None	2
Mus musculus	4686082	241865	18757	5501	3
Rattus norvegicus	701072	23,017	9012	None	5
Oryctolagus cuniculus	28046	2669	None	None	None
Dasypus novemcinctus	None	None	None	None	None
Loxodonta africana	None	4	None	None	None
Monodelphis domestica	50	363	None	None	1
Gallus gallus	578445	29743	3848	None	4
Xenopus tropicalis	1038272	10712	None	None	1
Dan rerio	673076	25094	10689	3546	None
Tetraodon nigroviridis	99	107945	None	None	2
Takifugu rubripes	25850	978	None	None	1
Drosophila melanogaster	383407	19931	19697	None	4
D. simulans	5013	80	None	None	2
D. yakuba	11015	808	None	None	2
D. erecta	None	6	None	None	1
D. ananassae	None	11	None	None	1
D. pseudoobscura	35042	40	None	None	4
D. virilis	663	41	None	None	1
D. mojavensis	361	2	None	None	1
D. grimshawi	None	None	None	None	1

Table 3.1. The table summarizes the annotation status of vertebrate and fly genomes as of October 2005. The number of EST sequences were obtained from the NCBI dbEST database [*Boguski et al.*, 1993]. The number of manually annotated genes was obtained from the VEGA annotation project site[*Ashurst et al.*, 2005]. The number of genebank mRNAs, RefSeq genes and ab-initio tracks were obtained from the UCSC genome browser database[*Karolchik et al.*, 2003].

precise annotations. If a gene is to be mapped from a reference species to multiple target species, GeneMapper makes use of characteristic properties extracted from all the available orthologous genes in the family. In other words, the program works with profiles of orthologous genes, which are not unlike protein profiles. The gene profile is built up progressively as the gene is mapped into successive target species. Therefore, the profile becomes more complete as the gene is mapped into additional

target species. The profile is especially useful in mapping genes to evolutionarily distant species that may have diverged a lot from the reference species. The rationale behind the profile based approach is that information from all orthologous sequences results in a more comprehensive representation of the gene than is possible with a single sequence.

GeneMapper was tested on a set of orthologous human and mouse genes. Results were compared with GeneWise and Projector annotations. We show that GeneMapper outperforms both GeneWise and Projector, and also establish that the addition of multiple sequences from chimpanzee, rat, and chicken further improves performance through the use of gene profiles.

3.1 The GeneMapper Algorithm

3.1.1 ExonAligner

GeneMapper is a bottom up algorithm that first predicts the ortholog of each reference exon in the target sequence and then combines the exon predictions to determine the gene structure. Therefore, the most critical step of the algorithm is to predict the ortholog of each reference exon by aligning it with the target sequence. A module called ExonAligner was developed to carry out this step in GeneMapper. ExonAligner takes as input two sequences, the annotated exon from the reference species and a target sequence *containing* its ortholog. A fairly intricate dynamic programming model is then used to align the reference exon with the target sequence. The bottom panel shows the dynamic programming matrix used by ExonAligner. Only the edges into top right node are shown. The solid edges represent matches/mismatches and gaps in codon space. The dotted edges represent translation frame disrupting events such as frameshifts.

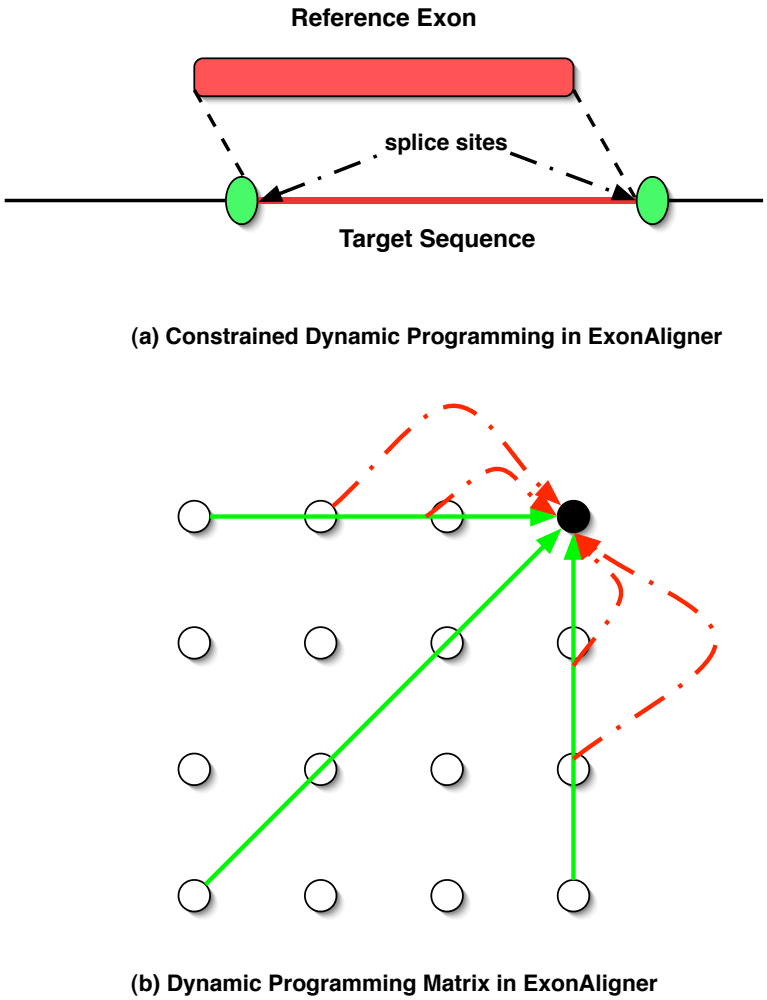


Figure 3.1. The ExonAligner Algorithm

ExonAligner uses a version of the Smith Waterman algorithm to find the best alignment of the reference exon with a *subsequence* of the target sequence. Figure 3.1(a) is a representation of constrained dynamic programming used by ExonAligner. It aligns the reference exon with a subsequence of the target sequence. In this version of the standard dynamic programming algorithm, overhanging ends are penalized in the reference exon but not in the target sequence. This subsequence is additionally constrained to have splice sites at its ends, which are represented by blobs in the cartoon. The splice sites are scored using StrataSplice

(http://www.sanger.ac.uk/Software/analysis/stratasplice/) to improve splice site detection.

ExonAligner uses a special dynamic programming matrix to model the evolution of codons and to allow for sequencing errors and frame shifts. The dynamic programming matrix is shown in Figure 3.1(b). There are two types of edges in the matrix, solid edges representing transitions in codon space and dotted edges representing events that cause disruptions in the translation frame. The solid edges model insertions, deletions and pairing of codons and cover three nucleotides in the X and(or) Y coordinates. On the other hand, the dotted edges cover one nucleotide in the X or Y directions. They model events such as sequencing errors and frameshifts which cause disruptions in the translation frame. As these events are very rare, a big penalty is charged for traversing these edges.

ExonAligner models the evolution of codons by using 64×64 matrices, which we call COD matrices. COD matrices define distances between codons and are very similar to PAM and BLOSUM matrices [*Dayhoff et al.*, 1978; *Henikoff and Henikoff*, 1992] that define distances between amino acids. The COD matrices are learned from whole genome alignments. In the case of vertebrates, the COD matrices are extrapolated from human and chimpanzee whole genome alignments. The whole genome alignment of the human and chimpanzee genomes was obtained from the UCSC genome browser database[*Karolchik et al.*, 2003]. The alignments of human genes with the chimpanzee genome were extracted from this data. The gene alignments were then used to learn parameters for evolution of codons between human and chimpanzee genomes. The human/chimpanzee parameters were extrapolated to obtain parameters for other species.

The ExonAligner algorithm predicts the reference exon's putative ortholog in the target species. The putative ortholog is used as a prediction by GeneMapper only

if its alignment with the reference exon passes a test of statistical significance. The testing of statistical significance of alignments is a well studied problem. The reader is referred to the book by Durbin [*Durbin et al.*, 1998] for an overview. ExonAligner uses the Bayesian likelihood ratio test as its core test. In this test, the calculated score is the ratio of the likelihood of the alignment in the match model to its likelihood in the random model. As the score is dependent upon length, short exons may fail to pass the ratio test. Therefore, ExonAligner also allows highly conserved short exons to pass the test of statistical significance.

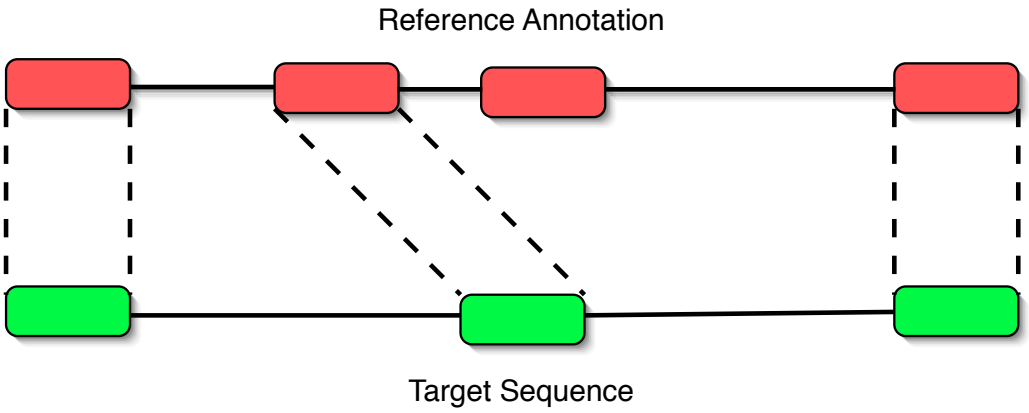
3.1.2 The Pairwise GeneMapper Algorithm

In this section, we describe the pairwise version of GeneMapper that maps gene annotations from a reference species to a single target species. The GeneMapper pipeline consists of three stages and is depicted in Figure 3.2. In the first stage, only the most conserved exons are mapped to the target sequence. At the end of this stage, an approximate outline of the gene in target sequence is obtained, as depicted in Figure 3.2(a). In the second stage, this outline is used to predict the orthologs of exons that are unmapped in the first stage. The exons mapped in the first stage narrow down the possible locations of neighboring unmapped exons and thus help in mapping them with more confidence. For example, in Figure 3.2(b), the search for the third exon in the target sequence can be narrowed down between the second and fourth exons (which were mapped in the first stage of the algorithm). In the first two stages, it is assumed that there are equal numbers of exons in orthologous genes of the reference and target species. But studies [*Waterston et al.*, 2002] have shown that this is not entirely true. In case of human and mouse, for instance, about 15% of orthologous genes do not have the same number of exons. Therefore, GeneMapper

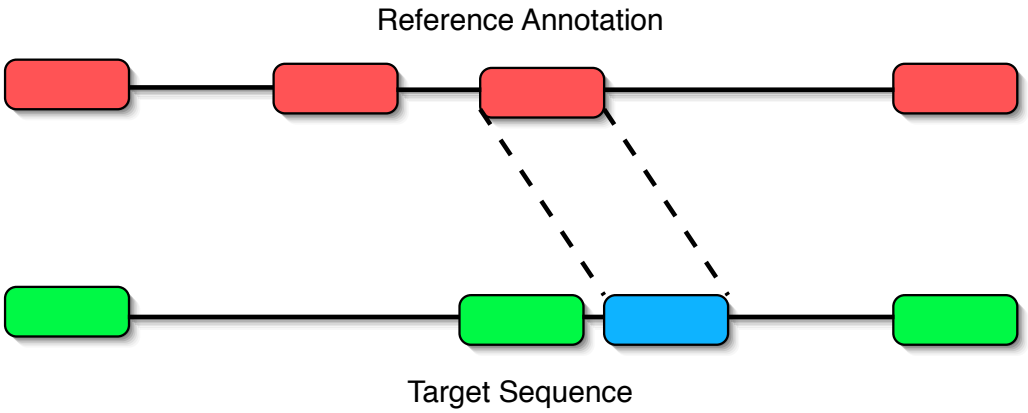
searches for exon splitting and exon fusion events in the third stage. We now describe in detail each stage of the pipeline.

In the first stage of the GeneMapper algorithm, only the highly conserved exons are mapped. GeneMapper initially searches for the approximate locations of the ortholog of each exon in the target sequence by using translated BLAST. If any significant hits are found for an exon, the best hit is extended to get an approximate location of the exon's ortholog in the target sequence. The ExonAligner algorithm is then used to predict the exact ortholog of the exon. The alignment of the predicted ortholog with the reference exon is checked for statistical significance using a combination of tests described in the previous section. These tests are made quite stringent so that only the most conserved exons pass them. This choice is made by design as we are able to obtain an outline of the gene structure in the target sequence that can be utilized to map less conserved exons more confidently in the next stage of the algorithm.

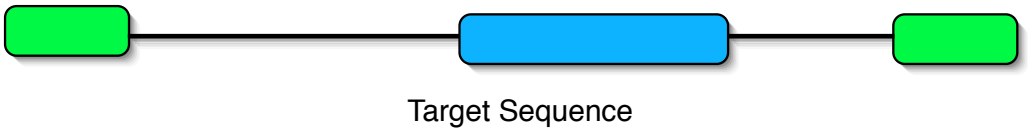
In the second stage of GeneMapper, linearity of transcription is used to map exons that are missed in the first stage of the algorithm i.e. already mapped exons are used to find out the approximate locations of unmapped exons. The details of the use of extrapolation to pinpoint the location of unmapped exons are shown in Figure 3.3. We assume that the gene is in the same strand in both species. If an unmapped exon has mapped exons both upstream as well as downstream, the unmapped exon should be mapped between the orthologs of its nearest mapped upstream and downstream exons. This is depicted in Panel 3.3(a). If only the exons upstream of an unmapped exon are mapped, then the unmapped exon should be mapped downstream of the ortholog of its closest mapped exon. This is depicted in Panel 3.3(b). If only the exons downstream of an unmapped exon are mapped, then the unmapped exon should be mapped upstream of the ortholog of its closest mapped exon. This is depicted in Panel 3.3(c). Once the possible location of an unmapped exon has been narrowed



(a) Step 1: Map the highly conserved exons



(b) Step 2: Use extrapolation to map less conserved exons



(c) Step 3: Find cases of exon splitting and exon fusion

Figure 3.2. The three stages of the GeneMapper pipeline. Panel a shows the first stage, where only the most conserved exons are mapped. Panel b depicts the second stage, where the algorithm uses exons mapped in the first stage as signposts to map already mapped exons. In this example, the possible locations of the second and third exons is narrowed down as they must be between the first and fourth exons. Panel c shows the last stage, in which the algorithm searches for cases of exon splitting and exon fusion.

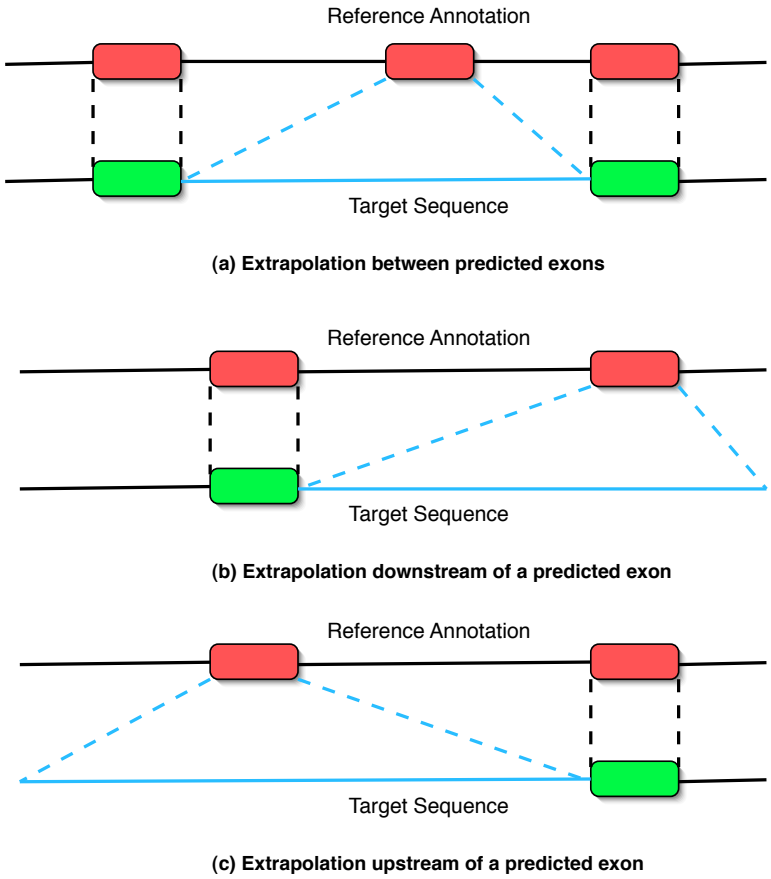


Figure 3.3. Extrapolation in GeneMapper. The blue sequence shows the possible location of the unmapped exon in the target sequence.

down, translated BLAST and ExonAligner are used to map the exon in the target sequence by a procedure that is similar to the first stage of the algorithm. However, the statistical significance tests are made less stringent in the second stage. This is because the position of the exon was narrowed down using already predicted exons and this makes us more confident about the accuracy of the prediction.

In the third and final stage of GeneMapper, the algorithm searches for exon fusion and exon splitting events. For detecting exon fusion, we use the fact that introns must be of at least a minimum length to maintain the intron splicing reaction. Thus, if two adjacent exon predictions in the target sequence are closer than the minimum intron length, they must have fused during evolution. This rule is very effective in

detecting all cases of exon fusion in the Projector data set. On the other hand, the rule for detecting exon splitting is comparatively crude and is dependent on having an accurate alignment of the reference exon with the predicted ortholog. The alignment is searched for gaps of length greater than the minimum intron length and having splice sites at their ends. Such gaps are best explained by exon splitting events. The rules for detecting exon splitting are preliminary and improvements are planned in future versions of GeneMapper.

3.1.3 Multi species GeneMapper

Several studies [*Dubchak et al.*, 2000; *Dewey et al.*, 2004; *Chatterji and Pachter*, 2004; *Gross and Brent*, 2005] have shown that increasing the number of species help in improving the performance of comparative ab-initio gene-finding programs. It therefore appears intuitive that increasing the number of species (and thus increasing the amount of available data) should enhance the accuracy of evidence based gene-finding methods. The multiple species version of the GeneMapper algorithm makes use of two key ideas to improve upon the pairwise algorithm. First, a profile of the gene is built and updated each time we map the gene into a new target species. The gene profiles are very similar to protein profiles which are used extensively in protein informatics. The profiles help us to map genes more accurately into species that are evolutionarily distant from the reference species. Second, there is a specific order in which a gene is mapped from the reference species into the multiple target species and this order is designed to take full advantage of the profile.

Gene profiles are alignments of one or more orthologous genes that are used to search for new orthologs. As shown in Figure 3.4, gene profiles work in codon space and each column in the profile contains orthologous codons. As with standard profiles, a gene profile can include gaps of length 3 that cover a codon. For example, the fifth

Human : **AGT TTG** *GGA* GAA TCG *TCC* **TTT GGG** *AGC* **CAT** CTG **CCT GAC**

Chimp : **AGT TTG** *GGA* GAA TCG *TCC* **TTT GGG** *AGT* **CAT** CTG **CCT GAC**

Mouse : **AGT TTG** *GGT* GAC ____ *TCT* **TTT GGG** *AGC* **CAT** CCA **CCT GAC**

Rat : **AGT TTG** *GGA* GAC ____ *TCT* **TTT GGG** *AGC* **CAT** CCA **CCT GAC**

Figure 3.4. A gene profile. A portion of the gene profile of the *Neurod4* gene orthologs in human, chimpanzee, mouse and rat. Each column in the profile contains orthologous codons and is used to obtain residue scoring matrix for dynamic programming. Columns with conserved codons are shown in bold, whereas columns with synonymous substitutions are italicized.

column in the figure has codon gaps in the mouse and rat sequences. In addition, a gene profile can contain non-codon gaps that cover one nucleotide. These gaps account for rare translation disrupting events such as frameshifts and sequencing errors and are not shown in the figure.

ExonAligner is modified to align gene profiles with sequences. As with pairwise ExonAligner, COD matrices are used to model the evolution of codons. To evaluate the residue scoring matrix for the profile, ExonAligner calculates the COD matrices defining the distances between the codons in the target species and each species in the profile. The COD matrices are then used to get the pairwise residue scoring matrix for each species. The residue scoring matrix for the whole profile is the sum of the pairwise scores. We illustrate the procedure by calculating the residue scoring matrix for species s at the third column in Figure 3.4. We first calculate the pairwise COD matrices between s and human, chimpanzee, mouse and rat, and call them COD_{sh}, COD_{sc}, COD_{sm} and COD_{sr} respectively. The score for codon c is the sum of the pairwise scores:

$COD_{sh}(c, GGA) + COD_{sc}(c, GGA) + COD_{sm}(c, GGT) + COD_{sr}(c, GGA)$

ExonAligner uses two evolutionary models to take into account the variations

in mutability of codons. The first model represents codons that are under negative selection and have low mutation rate. The second model represents codons that are not under any selection pressure and therefore have a high rate of mutability. A simple heuristic is employed to determine the model for a particular site. The first model is used if all the mutations in the site are synonymous, otherwise the second model is used. In addition, the program uses position sensitive gap scores whereby sites represented by the second model have a lower gap penalty.

The mapping of the gene into each target species takes place in three stages, in exactly the same manner as described for pairwise GeneMapper. The sequence in which the target species are mapped is ordered by the evolutionary distance from the reference species i.e. the gene is first mapped to the target species closest to the reference species, then to the next closest species and so on. This particular order is used because it is comparatively easy to map genes to a species that is evolutionarily close to the reference species than to a species that is more distant. Each time an orthologous gene is predicted in a target species, it is added to the profile. The updated profile is a more complete representation of the statistical properties of the gene family and therefore helps us in getting a more accurate prediction of the ortholog in the next species.

3.2 Results

GeneMapper was implemented in C and tested on a standard Linux machine. The running time of GeneMapper on a single gene is $O(\sum_{i=1}^{N_e}(l_i)^2)$, where N_e is the number of exons in the gene and l_i is the length of the i^{th} exon. A loose upper bound on this running time is $O(L^2)$, where L is the length of coding sequence in the gene. However the running time is expected to be appreciably smaller than quadratic for

multiple exon genes. GeneMapper can be downloaded from the GeneMapper website (http://bio.math.berkeley.edu/genemapper/).

Two tests were carried out to evaluate the performance of GeneMapper. In the first test, GeneMapper was compared with GeneWise and Projector, two commonly used reference based programs. For the second test, a data set of orthologous genes from the human, chimpanzee, mouse, rat and chicken genomes was created. This data set was then used to test the hypothesis that adding more species improves the performance of GeneMapper. The tests are described in detail in the next two sections.

3.2.1 Performance

GeneMapper was compared with Projector and GeneWise on the Projector data set[*Meyer and Durbin*, 2004]. This data set consists of 491 orthologous genes that are reciprocal best matches between mRNA supported human and mouse ENSEMBL genes. The set can be divided into two subsets. The first subset contains 465 genes where the number of exons is the same in the human and mouse orthologs. The second subset has 26 genes where the human and mouse orthologs have different number of exons, in some cases due to exon fusion and splitting events. Some of the genes in this subset were not true orthologs and the data set was refined manually to remove any such errors.

To compare the performance of the programs, the human annotations were used to predict the gene structure in the orthologous mouse sequences. GeneWise and Projector predictions were taken from the Projector paper[*Meyer and Durbin*, 2004]. The eval package [*Keibler and Brent*, 2003] was then used to calculate the nucleotide, exon and gene level sensitivities and specificities of the programs. For more details about these metrics, the reader is referred to [*Burset and Guigo*, 1996]. The perfor-

Program	Nucleotide Level		Exon Level		Gene Level	
	Sn	Sp	Sn	Sp	Sn[1]	Sp[1]
GeneWise	99.86	99.91	92.8	93.4	61.3	60.8
Projector	99.78	99.70	94.2	90.5	59.9	59.5
GeneMapper	99.88	99.94	97.2	97.8	81.7	81.7

Table 3.2. The table summarizes the performance of GeneWise, Projector and GeneMapper on the Projector data set consisting of 491 orthologous human and mouse genes. The human annotations was used to predict the gene structure in the mouse sequence. Performance is reported in terms of nucleotide, exon and gene level sensitivities and specificities.

mance of the three programs are compared in Table 3.2. The exon level sensitivity and specificity of GeneMapper is 97.15% and 98.19% respectively and the error rate is less than half of that of the other programs. The gene level sensitivity and specificity is improved by more than 20% compared to GeneWise and Projector. We believe that the primary reason for GeneMapper's accuracy is the use of a proper exon model for the alignment and mapping of exons. The results clearly indicate that GeneMapper is a significant improvement over existing programs and will be a useful tool for accurately transferring annotations from reference genomes to the newly sequenced genomes.

Program	Nucleotide Level		Exon Level		Gene Level	
	Sn	Sp	Sn	Sp	Sn[1]	Sp[1]
Pairwise GeneMapper	99.95	99.93	91.3	95.1	52.2	52.2
Multiple Species GeneMapper	99.95	99.93	91.5	95.2	52.6	52.6

Table 3.3. The table summarizes the effect of additional species on the performance of GeneMapper. To test pairwise GeneMapper, only the human annotations was used to predict the gene structure in the chicken sequence. For testing the profile based approach, additional orthologous sequences from the chimpanzee, mouse and rat genomes were used to create a profile for each gene. The profiles were then employed to predict genes in the chicken sequences. The table compares the accuracy in predicting the gene structure in the chicken sequences.

[1]GeneMapper predicts exactly one gene per reference annotation and the number of predicted genes is equal to the number of genes in true or gold standard annotation. Consequently, gene sensitivity is equal to gene specificity for GeneMapper.

3.2.2 Using additional species to improve performance

The second test used a data set of orthologous human, chimpanzee, mouse, rat and chicken genes to measure the improvement in accuracy of GeneMapper with the addition of multiple species. RefSeq annotations of human, mouse and chicken genomes were downloaded from the UCSC genome browser database [*Karolchik et al.*, 2003]. The gene set was refined to remove annotations with common errors such as the absence of start or stop codons. BLAT [*Kent*, 2002] was then used to find mutually best hits among the proteomes. The pairwise hits were further joined together to obtain orthologous triplets of human, mouse and chicken genes. The human and mouse orthologs were then mapped into the chimpanzee and rat genomes respectively resulting in a set of orthologs from all five species. The data set obtained by this process consisted of 895 potential orthologous segments from the five vertebrate genomes. We should note here that this standard method of obtaining orthologs by reciprocal best hits cannot distinguish between paralogs. However the accuracy of reference based programs such as GeneMapper is not affected as long as the potential orthologs are sufficiently conserved.

To assess the performance of pairwise GeneMapper, human annotations were used to predict the gene structure in the orthologous chicken sequences. For multiple species GeneMapper, additional orthologous sequences from chimpanzee, mouse and rat were utilized. The profiles were initialized with the human genes, and were then used to predict gene structures incrementally in the chimpanzee, mouse and rat genomes. As gene structures were predicted in each new species, they were added to the profiles. Finally, the profiles were used to predict the gene structures in the chicken sequence. The performance of the pairwise and multiple species versions of GeneMapper on the chicken genome is summarized in Table 3.3. The table demonstrates that multiple species GeneMapper is an improvement upon pairwise GeneMapper. We

point out later that most of the errors in the predictions are caused by factors that cannot be corrected computationally. Consequently, it is quite significant that multiple species GeneMapper is able to correct 18 wrong exon predictions of pairwise GeneMapper with just three additional species. We thus believe that with the addition of more species, multiple species GeneMapper will comes close to the limit of computational reference based methods.

3.3 Discussion

We have shown that GeneMapper is able to transfer reference annotations with remarkably high accuracy and is a substantial improvement over existing programs. This suggests that reference based gene finding is a feasible approach for accurately annotating the large number of genomes that are now being sequenced.

It is important to note that the idea of transferring annotations is not a new concept and methods such as DPS, Procrustes, GeneWise, Genomescan and Projector have been designed to perform exactly the same task. GeneWise and Procrustes align proteins with genomic sequences from target species. The principal disadvantage of the protein alignment approach is that it does not utilize information about exon/intron boundaries and therefore does not perform very well on less conserved genes. On the other hand, methods such as Projector and GeneMapper utilize the exon/intron structure of the gene and thus are more accurate in identifying splice sites. However, it should be noted that GeneMapper and Projector are not suitable for mapping genes from very distant species where the exon/intron structure of the gene might not remain conserved. For example, if one wants to find the homolog of a novel fruitfly gene in the human genome, it is probably best to use methods such as Procrustes and GeneWise.

Both GeneMapper and Projector use the exon/intron structure of the gene to predict the ortholog of a reference gene in a related species, but they have different approaches to the prediction problem. Projector uses the Viterbi algorithm for a pair hidden Markov model to predict the gene structure. Since the running time of the Viterbi algorithms for pair hidden Markov models is quadratic, Projector uses a heuristic to decrease the search space. On the other hand, GeneMapper uses a bottom up algorithm that first maps each exon and then joins the exon predictions together to obtain the gene structure. As exons are much shorter than introns, a more sophisticated model can be used for the exon alignment. The optimal alignment is still obtained using dynamic programming, albeit a more complex one. We believe that the use of our exon alignment model makes GeneMapper more accurate compared to Projector. Furthermore, unlike Projector, GeneMapper models sequencing errors and frameshifts and we believe that this makes GeneMapper more suitable for draft genomes.

When a gene has to be mapped into multiple species, GeneMapper uses profiles to obtain a more complete characterization of the gene and thus make more precise predictions. This is because a profile of orthologous genes can help us in obtaining much more information about the gene family than a single reference gene. We show that the use of additional species and the application of the profile based approach outperforms the pairwise approach. The use of profiles is particularly appropriate for annotating the newly sequenced vertebrate, insect and worm genomes as the profile can exploit information from all related genomes while making gene predictions.

Even though GeneMapper is remarkably accurate and has an error rate of less than 3% in transferring exons from human genes to orthologous mouse sequences, we investigated the sources of these errors to gain more insight into the GeneMapper algorithm. Most errors can be classified into the following categories:

1. *Highly divergent exons :* Exons that have diverged a lot between the reference and the target genes are not able to pass the statistical significance tests of ExonAligner. This is because a choice was made of reporting only highly reliable predictions at the cost of missing a few true exons.

2. *Exon Splitting :* As described in Section 3.1.2, GeneMapper's procedure for detecting exon splitting is comparatively crude and depends on the accurately aligning the reference exon with the orthologous target sequence (which contains an inserted intron). The presence of the inserted intron makes it difficult to accurately align these regions, especially if it is a long intron. Such wrongly aligned exons are partially predicted and this problem can probably be solved by having a more sophisticated alignment model that allows inserted introns.

3. *Assembly and sequencing errors :* The GeneMapper algorithm is unable to account for certain assembly and sequencing errors. For example, we found many cases of duplicated chicken exons, most probably due to errors in the assembly. In such cases, there is no way to distinguish between the duplicate exons and the prediction is made randomly among the duplicates. GeneMapper also constrains the predicted exons to have splice sites at their ends. Therefore, we are unable to deal with sequencing errors at splice sites.

4. *Differential splicing :* Differential splicing in the reference and target species can also cause errors in GeneMapper predictions. For example, if an exon is transcribed in the reference species but its ortholog is not transcribed in the target species, GeneMapper predicts a wrong exon in the target species. However, it is not clear whether this is a wrong prediction considering that this exon might be part of an alternate transcript in the target species. In fact, it is an open question whether alternative spliced forms are conserved among related species such as human and mouse, and we believe that GeneMapper predictions

could be an appropriate starting point for any experiment that seeks to answer this question.

An analysis of these errors will facilitate future improvements in GeneMapper. For example, we intend to work on statistical significance tests that are able to do a better job in discriminating between true and false exon predictions. Future enhancements of GeneMapper will also include improved handling of exon splitting. GeneMapper only transfers the *coding sequence* of a reference gene to a target sequence. We intend to modify GeneMapper to map 5' and 3' untranslated regions(UTRs). This will also help in mapping short initial/terminal coding exons, which are more divergent compared to internal exons.

Although, as we have pointed out, there is still room for improvement, we believe that multiple species GeneMapper comes close to the limit of gene prediction accuracy possible with computational reference based gene finding.

Chapter 4

EVOLUTION OF GENE STRUCTURE

The ultimate goal of this thesis is to systematically study the evolution of gene structure in eukaryotic genomes by using the multiple mammalian and fruit-fly genomes that are currently available to us. To perform any such study, we need accurate gene annotations in all the genomes being studied. As discussed in Chapter 3, scarce experimental evidence in many of these newly sequenced genomes necessitates the development of computational annotation methods. In Chapters 2 and 3, we have developed algorithms for accurate gene prediction. We now develop a pipeline that utilizes these methods to annotate protein coding genes in the newly sequenced genomes. In Sections 4.2 and 4.3, these annotations are used for studying gene structure evolution in eukaryotic genomes. We use our annotations to study variations in the rates of intron gain and loss in various clades. We then test various previously proposed mechanisms of intron gain and loss. We also try to find the relationship of gene structure changes to gene duplications and selection pressure.

4.1 Annotation Pipeline

We have developed an annotation pipeline for predicting protein coding genes in target genomes by transferring annotations from a well curated reference genome. The steps of this reference based pipeline are depicted in Figure 4.1. Initially, homology maps identifying evolutionary relationships between reference and target genomes are created. These homology maps are used to determine the approximate location of the ortholog of each reference gene in the target genomes. GeneMapper then uses the orthology information and reference annotations to annotate the target genomes. In the process, multiple alignments of each reference gene and its orthologs in the other genomes are also created. We now discuss these steps in greater detail.

4.1.1 Generation of Homology Maps

Mercator (http://bio.math.berkeley.edu/mercator/) was used to create an initial orthology map relating the reference and target genomes. The large scale evolutionary relationships detected by Mercator may be incomplete because of various factors. For instance, Mercator uses genomic landmarks or *anchors* to identify evolutionary relationships between various genomes. As a result, incomplete anchor coverage might lead to an incomplete homology map. In addition, homologous segments with low sequence identity might also be missed. Lastly, Mercator mainly identifies large scale evolutionary relationships between the target and reference genomes. Consequently, the program might not be able to detect small scale rearrangements and inversions.

As some evolutionary relationships might be missed by Mercator, we extend the Mercator orthology map by using extrapolation. The extrapolation algorithm is depicted in Figure 4.2 and is similar to the extrapolation algorithm in the second stage of the GeneMapper algorithm (cf. Section 3.1.2). For example, if an unmapped re-

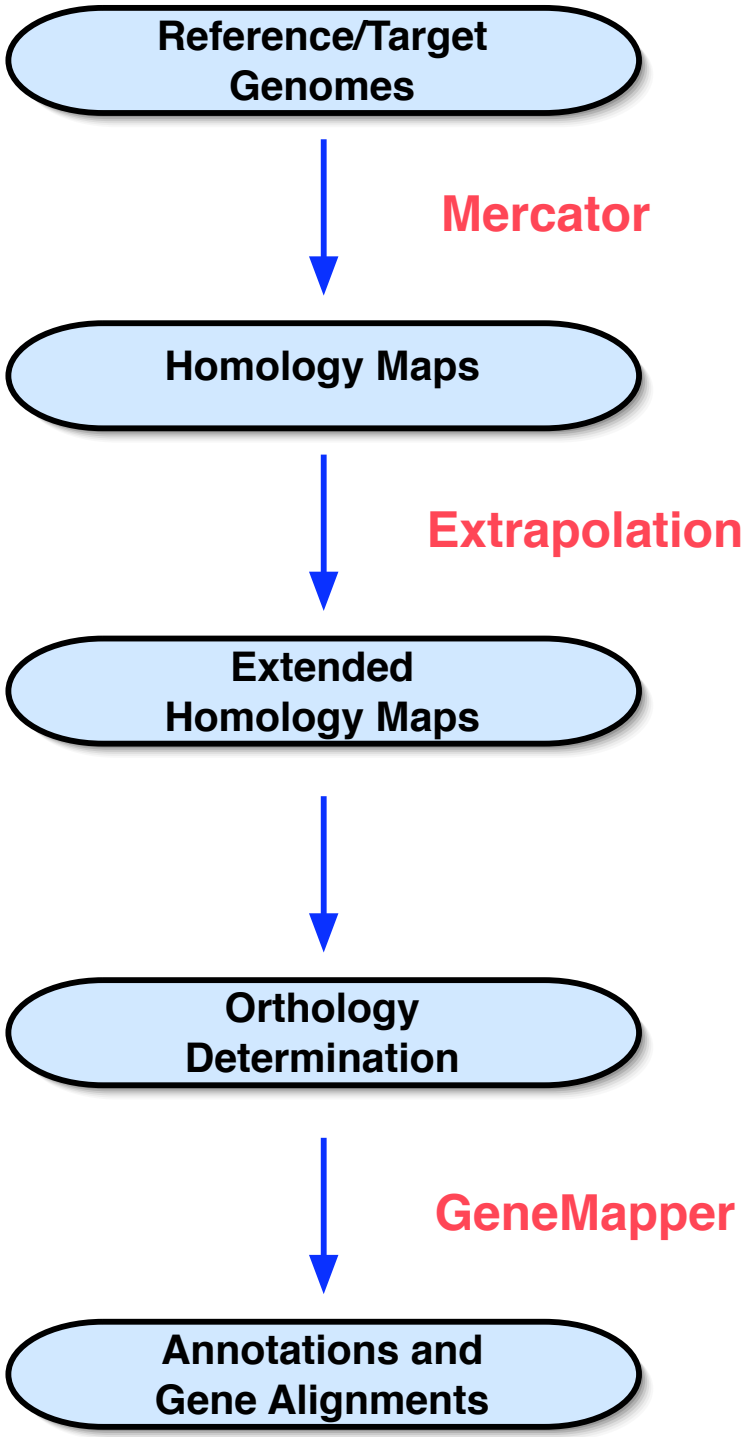


Figure 4.1. The gene prediction pipeline.

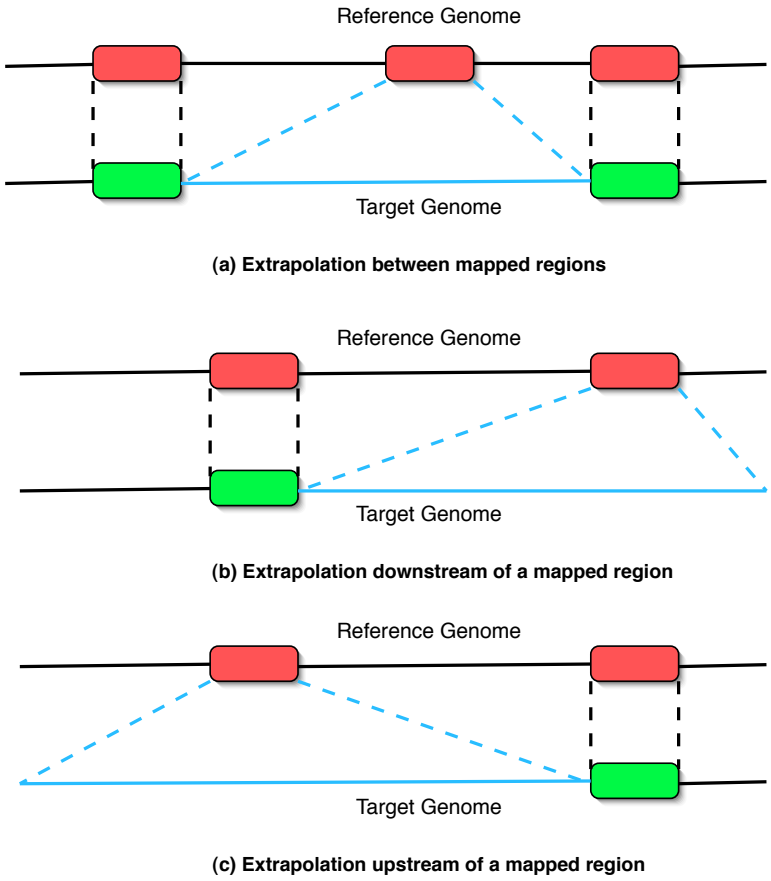


Figure 4.2. Extrapolation in the annotation pipeline.

gion had mapped regions both upstream and downstream, we looked for the orthologs of the unmapped region between the orthologs of its nearest mapped upstream and downstream region (Figure 4.2 (a)). However, there is a subtle difference between this algorithm and the extrapolation used in the GeneMapper algorithm. While extending the Mercator homology maps, we search for homologous regions in both strands and this helps us detect inversions that might have been missed by Mercator. Note that this is not required in GeneMapper as all the exons of a gene are in the same strand.

4.1.2 Annotations and Gene Alignments

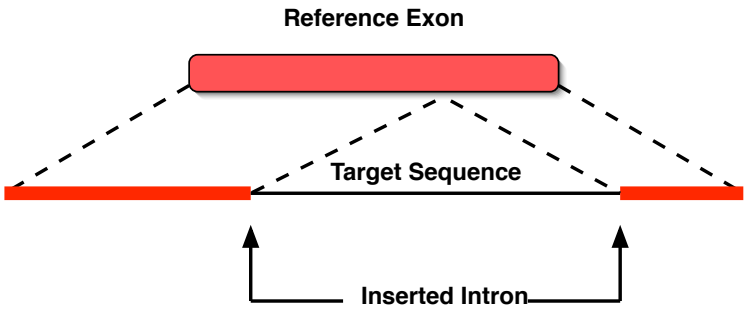
The extended orthology map created in the previous step was used to determine the approximate location of the ortholog of each reference gene in the target genomes. GeneMapper was then used to annotate every target species by transferring the reference annotation to the target genome. Separate clade specific parameters were used for fly and mammalian genomes. In the case of mammals, the COD matrices were calculated from human-chimp whole genome alignments, whereas fruitfly specific parameters were calculated from *Drosophila melanogaster-Drosophila yakuba* whole genome alignments. A description of COD matrices and their derivation from whole genome alignments is provided in Section 3.1.

As discussed in Section 3.1.3, GeneMapper iteratively creates a gene profile of orthologous genes while transferring genes from the reference species to multiple target species. The use of the profile helps us map genes accurately to evolutionarily distant species. In addition, the profiles are used to create gene alignments for each reference gene. The gene profile that is created by GeneMapper while transferring annotations is essentially an alignment of the reference gene and its orthologs. Therefore, the gene profile can be used to guide a gene alignment to study gene evolution. Unlike global alignment programs that are not conscious of the patterns of gene evolution, GeneMapper carefully models the evolution of genes, taking into account the fact that they have a codon structure and splice sites (Figure 3.1). The evolution of codons is modeled using 64×64 COD matrices. Furthermore, GeneMapper uses exact dynamic programming while adding each ortholog to the gene profile. Consequently, GeneMapper gene alignments are much more accurate compared to gene alignments obtained from global alignment programs. We provide these gene alignments as a resource for researchers studying the evolution of genes and they are available on the GeneMapper website (http://bio.math.berkeley.edu/genemapper/).

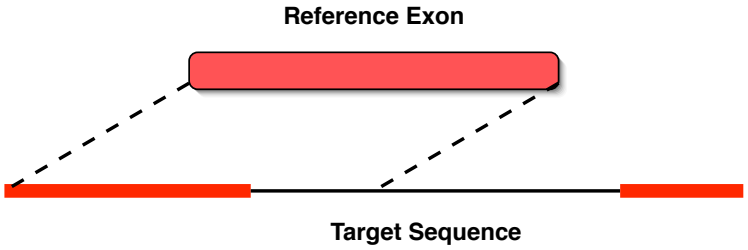
4.1.3 Determination of Gene Structure Changes

The main application of the annotations generated by the pipeline is to investigate gene structure evolution in eukaryotes. As a result, it is imperative for the annotation pipeline to reliably detect changes in gene structure. However, as discussed in Section 3.1.2, the stand alone version of GeneMapper has a comparatively crude algorithm for detecting inserted introns. The algorithm assumes that we have an accurate alignment of the reference exon and the target sequence (containing an inserted intron). However, the dynamic programming alignment algorithm used by GeneMapper can misalign such exons, especially if the inserted intron is long. This problem is illustrated in Figure 4.3. Panel (a) shows the true alignment when there is an inserted intron in the target sequence. The true alignment should align the coding sequences with a gap for the inserted intron. However, a dynamic programming algorithm that doesn't allow for inserted introns would produce a misalignment, as is illustrated in Panel (b). Consequently, we have modified the GeneMapper algorithm to detect such misalignments and thus accurately detect exon splitting events and the modified algorithm is described below.

The reference exon and the target sequence are first aligned using the dynamic programming algorithm described in Section 3.1.1. This algorithm doesn't take into account the possibility of inserted introns in the target sequence and can therefore cause misalignments (cf. Figure 4.3(b)). We detect such misalignments by making the following observation: if the alignment algorithm aligns the exon only partially (due to the presence of an inserted intron), either the $5'$ or $3'$ end of the reference exon will be aligned to the target intron. Consequently, the alignment should have low sequence identity at one of the boundaries. For example, in Figure 4.3(b), the $3'$ end of the reference exon is aligned to the intron and the alignment should have low sequence identity at the $3'$ end. Therefore, we can look at the boundaries of the



(a) An example of inserted intron and the true alignment.



(b) An example of mis-alignment due to an inserted intron

Figure 4.3. An example demonstrating the problem of accurately aligning orthologous coding sequences with inserted introns. Panel (a) shows the "true" alignment. The target sequence orthologous to the reference exon contains two exons (colored red) and an inserted intron (colored black). Panel (b) illustrating the misalignment that can be caused if a naive alignment algorithm (that doesn't allow inserted introns) is used to align the reference exon and the target sequence. This algorithm aligns the reference exon with one of the target exons and the contiguous intronic sequence.

alignments to find misalignments (due to inserted introns). The misaligned exons are realigned by employing a pair-HMM based algorithm described below.

The states of the pair-HMM and the transitions between the states are depicted in Figure 4.4. The evolution of coding sequences is modeled in a manner similar to ExonAligner (Section 3.1.1). In addition, we have an inserted intron state (**IN**) to model the inserted intron. The Viterbi algorithm is then used to find the best alignment between the reference and target exon. The running time of the Viterbi algorithm on this pair-HMM is cubic in the length of the sequences. As a result, we use this algorithm only to realign misaligned exons detected by the algorithm

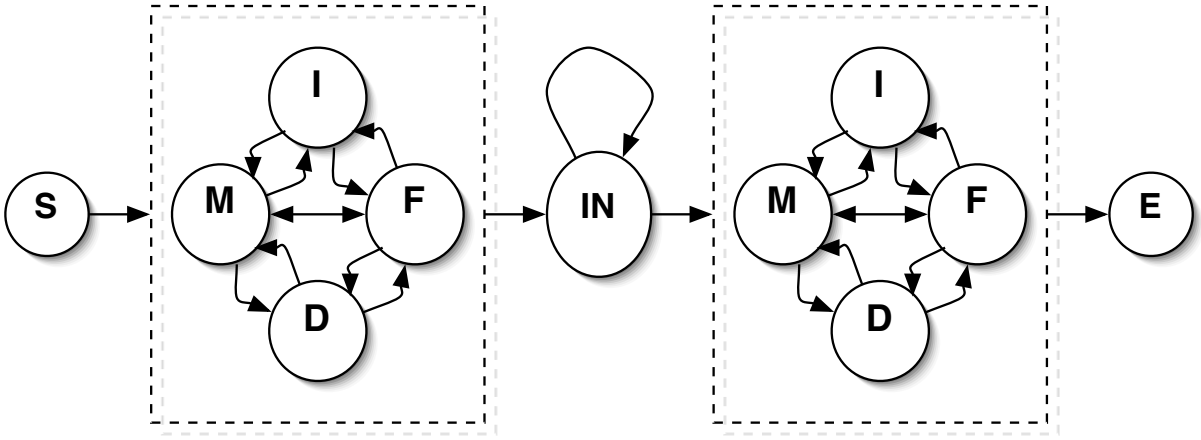


Figure 4.4. The pair-HMM used to align reference exons and target sequence with an inserted intron. **S** and **E** are the standard start and end states of pair-HMMs. To keep the figure simple, we have collapsed the states used to model evolution of coding sequences into dotted squares. Each inbound edge into a dotted square implies that there are corresponding inbound edges into every state inside the square. Similarly, every outbound edge from each dotted square represents corresponding outbound edges from every state in the dotted square. The states in the each dotted square and the transitions between them are equivalent to the dynamic programming matrix in Figure 3.1(b). Each square has the standard match(**M**), insert(**I**) and delete(**D**) states of standard pair-HMMs. In addition, the **F** state is used to model frame shifts. The intron state(**IN**) is used to model the inserted intron in the target sequence.

described above. It is also important to notice that the algorithm cannot detect multiple inserted introns. Therefore as discussed earlier, the GeneMapper algorithm is not suitable when the gene structure has changed drastically.

Tandem Repeats : Tandem repeats are two or more adjacent and approximate copies of a sequence of nucleotides. There is a widespread tandem repeat polymorphism in human protein coding genes [*O'Dushlaine et al.*, 2005]. We also found a lot of inter-species variation in tandem repeat length and our algorithm for finding inserted introns predicts some tandem repeats (in the target sequence) as inserted introns. Consequently, we use the program Tandem Repeat Finder [*Benson*, 1999] to remove any tandem repeats that have been incorrectly predicted as inserted introns.

4.1.4 Reconstructing the Evolutionary History of Introns

To understand the biology of intron gain/loss, it is necessary to reconstruct the evolutionary history of introns in genes undergoing structure changes. The evolutionary history of an intron can be reconstructed by comparing the presence/absence of the intron in the phylogenetic tree relating the orthologous genes. There are two approaches to reconstructing this evolutionary history from the phylogenetic tree. The maximum parsimony approach [*Rogozin et al.*, 2003; *Nielsen et al.*, 2004] infers the evolutionary history that can explain the phylogenetic tree most parsimoniously in terms of intron gain and loss events. The parsimony approach assumes that introns gain and loss are comparatively rare. However, if species being studied are phylogenetically sparse, the parsimony approach may give incorrect or ambiguous answers because of parallel intron gain and loss. On the other hand, the maximum likelihood approach [*Roy and Gilbert*, 2005a; *Qiu et al.*, 2004; *Nguyen et al.*, 2005] infers the evolutionary history with the highest probability according to a particular model of intron evolution. The results of the likelihood approach depend on the assumptions in the underlying model and different likelihood models infer vastly disparate results for phylogenetically diverse data sets such as the one in *Rogozin et al.* [2003].

In this thesis, we work with phylogenetically dense data sets where we can make inferences about the evolutionary history of an intron with high confidence by using parsimony. Indeed, gene structure changes among related species are so rare that the location of intron gain/loss could be identified by manual inspection of the phylogenetic tree. We use GeneMapper annotations to identify genes that underwent gene structure changes. For each instance of intron gain and loss, the presence/absence of the intron in each species was used to label the leaves of the phylogenetic tree relating the species. A parsimony analysis similar to *Rogozin et al.* [2003] was then used to locate intron gain and loss in the tree.

4.2 Gene Structure Evolution in Mammals

The ENCODE Project [*Feingold et al.*, 2004] aims to study functional elements by rigorously analyzing a portion (about 1%) of the human genome. 44 regions across the human genome were chosen for investigation. The ENCODE project, although focused on the identification of functional elements in the human genome, offers an unprecedented opportunity to study the evolution of functional elements. A key part of the project has been the sequencing of multiple species orthologous to the human sequence. The September 2005 release of ENCODE contains 546 Mb of genomic sequence from 44 vertebrates. This includes about 500 Mb of sequences from 38 mammalian genomes. In addition, the human ENCODE sequences have been rigorously annotated as part of the GENCODE project (http://genome.imim.es/gencode/). Thus the dense phylogenetic sampling of genomes in the ENCODE regions offers an unprecedented opportunity to study the evolution of gene structure in mammalian genomes.

The ENCODE sequences have a well curated set of human annotations. However, ENCODE sequences in non-human species have little experimental evidence to support gene annotation. Therefore, human GENCODE annotations were used as a reference to annotate the non-human sequences. We have also generated high quality alignments of the GENCODE genes which should be a useful resource for other studies of gene function and structure. All these resources are publicly available at the supplementary website (http://bio.math.berkeley.edu/genemapper/encode).

GeneMapper annotations were used to search for changes in gene structure in the mammalian sequences. Because this is a comparatively small data set, all cases of putative gene structure change were manually verified for any discrepancies. A phylogenetic analysis of gene structure changes in the mammalian lineages was used to identify 11 genes with instances of intron loss (Table 4.1). No intron gains were

Gene	ENCODE Region	Species	Introns Lost
RP11-126K1.1	ENr231	Rat	1
AC009404.1	ENr121	Mouse, Rat	1
AC116366.3	ENm002	Rat	1
RP11-505P4.2	ENr223	Shrew	2
XX-FW83128A1.1	ENm006	Shrew	1
XX-FW83563B9.3	ENm006	Bat	1
AF277315.16	ENm006	Shrew	1
AP001187.10	ENr332	Mouse, Rat	1
AC011330.7	ENr233	Rat	1
AC018512.8	ENr233	Mouse, Rat	3
AP000313.5	ENm005	Shrew	2

Table 4.1. Intron loss events in the ENCODE regions.

observed. Some genes were found to have lost more than one intron, resulting in 15 distinct cases of intron loss. A single instance of intron loss was detected in the bat lineage whereas the rest of the instances of intron loss were in the rodent (mouse/rat) and shrew lineages. A particularly interesting example is the gene AC018512.8 (a microfibrillar-associated protein), where the second and third introns were lost in the mouse lineage and the fourth intron was lost in the rat lineage. In fact, introns are lost in this gene in fugu and zebrafish also. This example suggests the presence of hot-spots for structural changes.

The dense phylogenetic sampling in the ENCODE regions allows us to infer the evolutionary history of an intron with high confidence. The only previous gene structure evolution study in mammals was done in human, mouse and rat, with fugu as the out-group [*Roy et al.*, 2003]. To illustrate the limitations inherent in using such a phylogenetically sparse species set, it is instructive to analyze the fourth intron of the gene AC018512.8, where introns are lost in both fugu and rat. Without more species, it is impossible to decide with confidence whether these events are due to parallel intron gains in human/mouse or intron losses in fugu and rat. However, a

phylogenetic analysis of the gene structure in all the ENCODE species makes it clear that the scenario of two intron losses is the most parsimonious explanation.

Rates of Intron Gain/Loss : It is apparent from the results above that intron loss occurs at a much higher rate compared to intron gain in mammalian lineages. In fact, to the best of our knowledge, no instance of recent intron gain has been detected in mammalian lineages. It also appears that some lineages (such as rodents and shrew) have a much higher rate of gene structure change compared to other lineages (such as primates). The difference in rates might be related to differences in generation times. These observations are consistent with results in a previous study comparing the structure of human and rodent genes [*Roy et al.*, 2003].

Mechanisms of Intron Loss : The classical theory of intron loss states that introns are lost by recombination of reverse transcribed mRNA transcript with the genome [*Bernstein et al.*, 1983]. As reverse transcriptase operates from the 3' to 5' end and may terminate prematurely, this theory predicts that more introns should be lost from the 3' end compared to the 5' end. Because of the involvement of reverse transcriptase, this theory also predicts that many introns should be lost in tandem. While we did not find that the lost introns show bias towards the 3 end of genes, all the cases of multiple intron loss did occur in tandem. An alternative theory of intron loss hypothesizes that introns are lost by genomic deletion [*Kent and Zahler*, 2000; *Cho et al.*, 2004]. This theory predicts that intron lost is inexact in which a small number of codons are added or lost from the flanking coding sequence. However, all the intron losses in our data set are exact.

Gene Expression : For a gene structure change to be passed on to subsequent generations, it has to occur in the germline. Indeed, it has been previously observed [*Coghlan and Wolfe*, 2004] that genes expressed in the germline are more susceptible to gene structure change. Gene expression levels in 79 human and 61 mouse tissues

were obtained from the GNF Gene Expression Atlas 2 [*Su et al.*, 2002]. For each gene with gene structure change, the maximum expression level across all germline tissues was compared to the median value across all tissues. It was found that all the genes had moderate to high expression levels in at least one germline tissue (more than 0.9 above the median on the log scale). Of these genes, four were highly expressed (more than 2 above the median on the log scale). It should be pointed out that it is possible that the genes with moderate expression levels might be expressed at higher levels in other germline tissues. This is because not all the genes were covered by the mouse experiments and the coverage of some other genes was incomplete. It is also possible that these genes are expressed in tissues that were not sampled in the experiment. Furthermore, some of the gene structure changes occurred in the rat, shrew and bat lineages and expression levels might have changed in these species. In any event, the evidence seems to indicate that genes that have undergone structural change are expressed in at least moderate levels in germline cells.

Selection : GeneMapper was used to create multiple alignments of all GENCODE genes and their orthologs. We used these alignments to measure ω, the ratio of synonymous and non-synonymous substitution rates for the genes undergoing gene structure evolution. The value of ω is a measurement of the nature of selection undergone by a gene. If $\omega << 1$, a gene is likely to be under purifying selection. On the other hand, a value of $\omega >> 1$ suggests that a gene is under positive selection. As the biological functions of most genes are expected to be conserved during evolution, genes are expected to be under purifying selection. All the 11 genes with intron loss were under strong to moderate purifying selection ($\omega < 0.20$). In addition, 6 genes were under very strong purifying selection ($\omega < 0.05$). Therefore, it appears that changes in gene structure evolution are not related to any drastic changes in coding sequence.

Gene Duplication : It has been suggested that intron gain/loss is accelerated

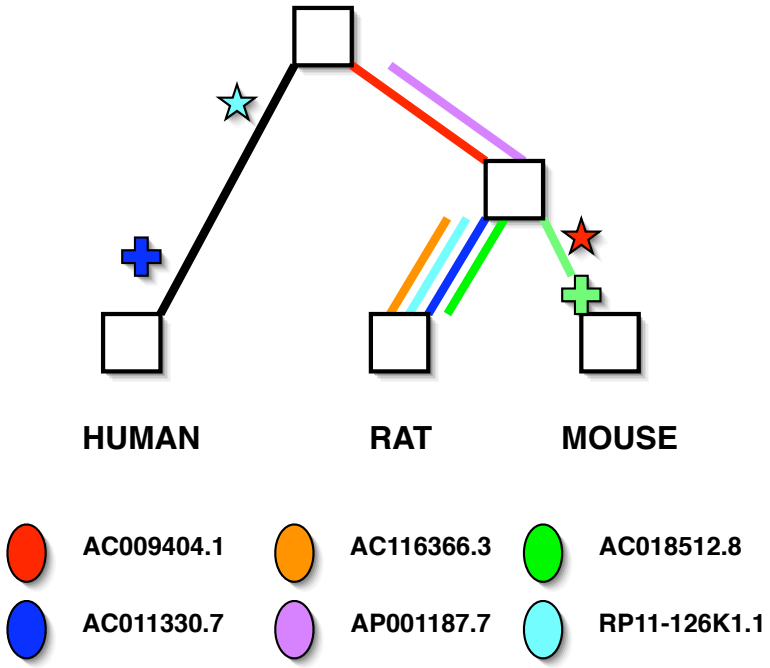


Figure 4.5. The relationship between duplication events and intron losses. Each gene is assigned a separate color. Colored edges on the tree show when intron losses occurred. The stars and plus signs show when retro-transposition events and local duplication events occurred. The locations of the symbols and edges indicate the relative order of the associated events. The gene AC018502.8 (green) is interesting because intron loss occurred twice in separate introns (one in the mouse and the other in the rat). In the mouse lineage, both the loss of the intron and local duplication occurred after separation from the mouse/rat ancestor. Moreover, we were able to infer that the duplication event occurred after the intron loss

in genes with duplications as a result of a reduction in selective pressure [*Castillo-Davis et al.*, 2004; *Lin et al.*, 2006]. If this hypothesis is true in mammalian lineages, most cases of intron loss should follow gene duplication. We tested this hypothesis by studying duplication events in the six genes with intron loss in mouse and rat using the complete genome sequences available for those species. We searched for homologs of each gene in the human, mouse and rat genomes using BLAT [*Kent*, 2002]. Four genes had multiple copies in at least one of the three genomes. For each gene with a homolog, the homolog with the highest sequence identity was identified as the one formed by the most recent duplication event. The location of the duplication

event as well as intron loss was then identified on the phylogenetic tree relating the three species. This association of gene duplication with intron loss is depicted in Figure 4.5. It is interesting to note that in two genes (AC009404.1 and AC018512.8), the most recent duplication event occurred after the intron loss. In two other genes (AC011330.7 and RP11-126K1.1), the most recent duplication occurred in the human lineage (which had no structure change). It is also interesting to note that all the genes are under strong or moderate purifying selection. Therefore, all available evidence indicates that intron loss does not occur due to relaxation of selection pressure (caused by duplication). But it appears that genes undergoing intron loss are also susceptible to duplication and that indeed, many of the duplication events may occur after intron loss.

Summary : Our study of the mammalian lineages provides evidence that gene structure changes may not be caused by reduction of selection pressure due to duplication. In fact, we show that many duplication events occur after gene structure change. Our conclusions are also supported by the fact that all the genes with gene structure change are under purifying selection ($\omega < 0.20$). In addition, it appears that genes with intron loss are susceptible to duplication. This provides evidence for a common underlying cause for intron loss and gene duplication. We speculate that changes are induced by a mechanism mediated by reverse transcriptase. The fact that the genes we identified are moderate to highly expressed in germline cells is also consistent with a reverse splicing mechanism.

The data set used in this study was comparatively small and a follow-up study based on larger amounts of data expected from the forthcoming phase of the ENCODE project will be necessary (and, we believe sufficient) for reaching definitive conclusions about gene structure evolution in mammals.

4.3 Gene Structure Evolution in Diptera

The sequencing of twelve fruitfly genomes (http://rana.lbl.gov/drosophila/) offers another opportunity to study gene structure evolution. A comparison with mammals also allows us to study variation in the mode of gene structure evolution. Among these twelve genomes, *Drosophila melanogaster* is well annotated by FlyBase [*Drysdale et al.*, 2005], whereas other species have comparatively sparse experimental evidence for gene annotation. Consequently, we have used the *D. melanogaster* FlyBase gene annotations as a reference to annotate the other fruitfly genomes. Gene alignments for each FlyBase protein coding gene and its orthologs were also generated at this step.

To study gene structure evolution in diptera, we have used closely related species in the Melanogaster subgroup (Figure 4.6). The phylogenetic relationship between *D. melanogaster*, *D. yakuba* and *D. erecta* is unsettled [Daniel A. Pollard, personal communication]. This is because many gene trees are not consistent with the branching in the consensus species tree. However, *D. annanassae* is unequivocally an out-group for these three species. Therefore, we have used pairwise comparisons between these three species and employed *D. annanassae* as an out-group to detect evolutionary history of introns. Specifically, we have compared gene structures of orthologous *D. melanogaster* and *D. yakuba* genes and also the structures of orthologous *D. melanogaster* and *D. erecta* genes to find recent cases of intron gain and loss. Since these three species are separated by less than 6 million years, we can assume that cases of parallel intron gain and loss are non-existent. Consequently, a parsimony based analysis should accurately retrieve the evolutionary history of introns.

A pairwise comparison of gene structure in the three species in the Melanogaster subgroup (*D. melanogaster*, *D. yakuba* and *D. erecta*) found 87 cases of intron loss and 161 cases of intron gain. This suggests that the dipteran genomes are gaining

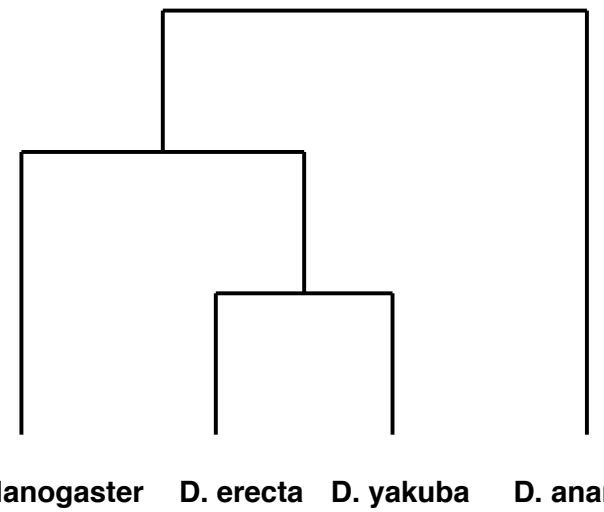
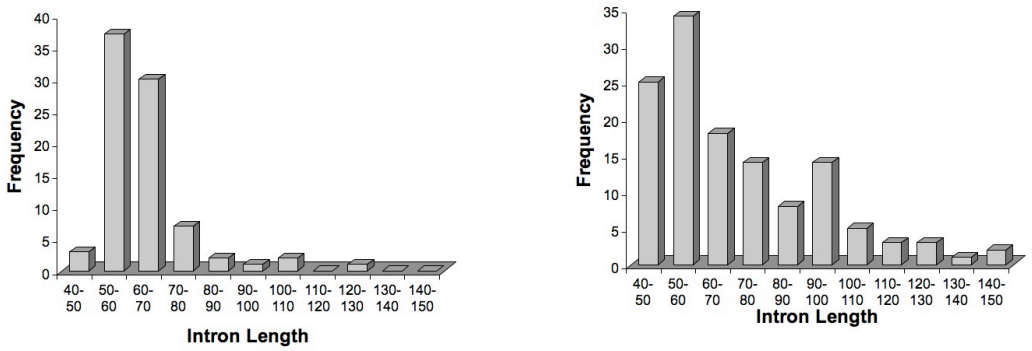


Figure 4.6. The consensus tree relating the species used to study gene structure evolution in diptera. The phylogenetic relationship between D. melanogaster, D. yakuba and D. erecta is unsettled because the consensus species tree is incongruous with many gene trees.

introns at a faster rate than they are losing introns. In contrast, our study in mammalian genomes (Section 4.2) found no cases of intron gain. In fact, to the best of our knowledge, no case of recent intron gain has ever been reported in mammalian genomes. Our study indicates that there is great diversity in the mode and tempo of gene structure evolution between eukaryotic clades. Our data set is also one of the largest data sets of recently gained/lost introns in eukaryotic genomes. We have used this comparatively large data set to study the statistical properties of the recently gained and lost introns, and infer any stochastic process underlying intron gain and loss.

Intron Lengths : The lengths of recently lost introns are shown in Figure 4.7 (a). The distribution is similar to the distribution of lengths of all introns in the *D. melanogaster* genome [*Deutsch and Long*, 1999], with a peak between 50 and 70 bp. However the length distribution is not smooth as the number of data points in the sample is comparatively small. The number of inserted introns is much higher and



(a) **(b)**

Figure 4.7. The lengths of recently gained and lost introns in the *Drosophila* subgroup. Panel (a) shows the distribution of lengths of 87 introns that have been lost in the *Melanogaster* subgroup whereas Panel (b) shows the distribution of lengths of 161 recently gained introns.

the length distribution is much smoother with a peak between 50 and 70 bp. This distribution is also similar to the distribution of length of all introns in the genome. Therefore it appears that there is no bias in the length of introns that are being lost and gained in *Drosophila* genomes.

Intron Phases : The phases of inserted and lost introns have profound implications on the theories of the origin of introns [e.g. *Long et al.*, 1995]. About 42% of *Drosophila* introns are phase zero (compared to 33.3% that would be expected if there was no phase bias). This phase bias is also observed in other eukaryotic genomes. The abundance of phase zero introns has been used to support the introns early or the exon theory of genes. The excess of phase zero introns is conjectured to be the legacy of exon shuffling events. It is argued that in the event that genes were assembled through exon shuffling, an exon shuffling product would be viable only if the all the introns were phase zero, otherwise the resultant gene will have frame-shifts and will code for an in-viable protein. However, there can be alternative explanations for the excess of phase zero introns in eukaryotic genomes. For example, intron insertion can be phase biased [e.g. *Coghlan and Wolfe*, 2004]. It has also been suggested that the

excess of phase zero introns is due to their selective advantage [*Lynch*, 2002]. We now use our data set to test the merits of these theories.

The phase bias in the introns that have been lost and gained recently in the Melanogaster subgroup is shown in Figure 4.8. We found that that phase zero introns were preferentially gained and lost. We found that about 50% of recently inserted introns were phase zero (compared to 42% that would be expected by chance). There was an even greater likelihood for phase zero introns to be lost and about 62% of the recently lost introns were phase zero. Therefore, we did not find any evidence for selective advantages of phase zero introns. But we did not have enough evidence to either substantiate or refute the other two theories. It is clear that a phase zero bias in the inserted introns. However, there is an even greater phase zero bias in the introns that are being deleted. Consequently, it is not clear whether the excess of phase zero introns are due to phase biased insertion or due to the legacy of exon-shuffling events. We should also point out that the phase zero bias might be related to some diptera specific mechanism of intron gain and loss. A more extensive study examining these biases in multifarious eukaryotic clades might help us answer these questions more precisely.

Intron Positions : The preferential loss of 3' introns has been observed in some eukaryotic genomes and is used to support the reverse transcriptase mediated theory of intron loss [*Roy and Gilbert*, 2005c]. It has also been observed that there is an excess of 5' introns in eukaryotic genomes [*Sakurai et al.*, 2002; *Lin and Zhang*, 2005]. We now try to explain the molecular mechanisms behind patterns by looking at the positions of recently gained and lost introns in the coding sequence. For each inserted and lost intron, we compiled its position in the coding sequence. As in *Sakurai et al.* [2002], the position of each intron within its host gene is mapped into a (0,1) interval relative to its coding sequence length. The fraction of the introns in each quarter is displayed in Figure 4.9. As has been previously observed in *Lin and Zhang* [2005],

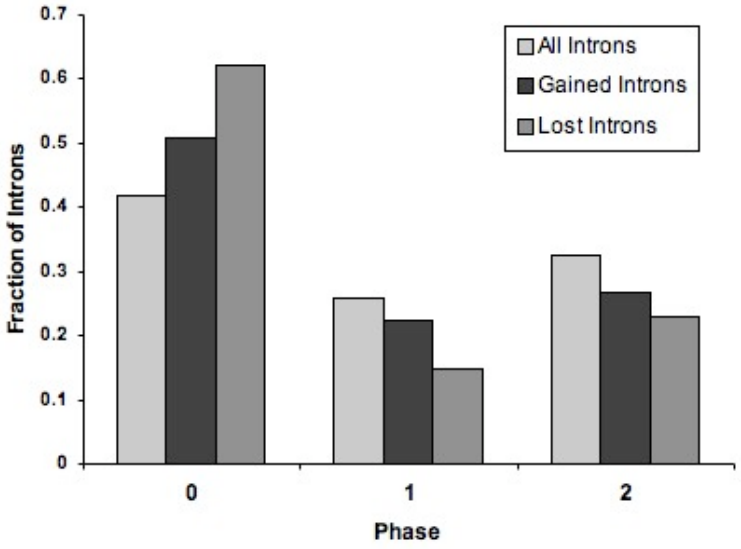


Figure 4.8. The phases of recently gained and lost introns in *Diptera*. The chart shows the fraction of all *Drosophila melanogaster* introns, recently gained introns and recently lost introns in each of the three phases.

we can see that there is an excess of 5' introns in the genome with around 31% of the introns in the 5' quarter of the gene. In contrast, only 16% of lost introns are in the $5'$ quarter and 30% of these introns are in the $3'$ quarter. Therefore, $3'$ introns seem to be preferentially lost in the *Drosophila* lineage. We are not able to detect any significant bias in the positions of inserted introns. However, it appears that introns are more likely to be inserted in second and third quarters (56% of the introns) and less likely to be lost in the first and fourth quarters (44% of the introns). Our findings suggest that the $5'$ bias in intron positions is due to preferential loss of $3'$ introns and not due to any significant bias in the position of inserted introns. We believe that this result is significant support for the reverse transcriptase mediated theory of intron loss.

Mechanisms of Intron Gain and Loss : In Section 1.2.3, we discussed various theories that have been proposed to explain the gain and loss of introns. Each of these theories makes certain testable predictions about the pattern of intron loss. We now

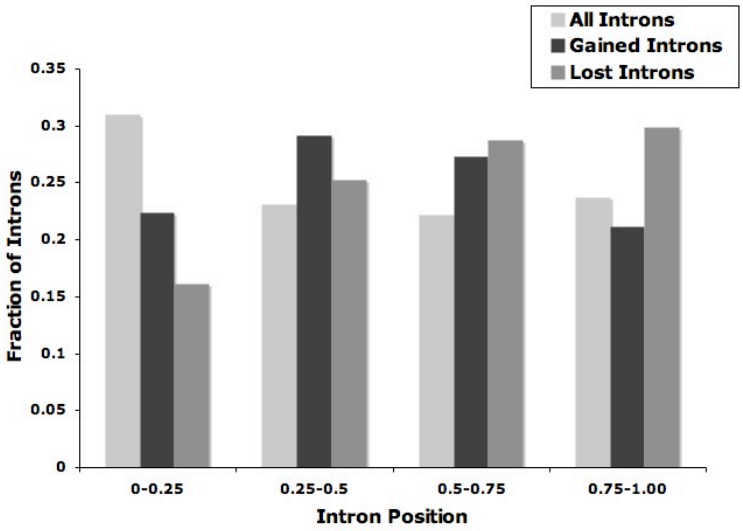


Figure 4.9. The positions of recently gained and lost introns in each quarter of the coding sequence. The chart shows the fraction of all *Drosophila melanogaster* introns, recently gained introns and recently lost introns in the each quarter of the coding sequence.

test these predictions on our data set and thus make inferences about the mechanisms of intron gain and loss in dipteran lineages.

The transposon theory of intron gain [*Crick*, 1979] postulates that novel introns arise by insertion of transposons. To test this theory we searched for transposable elements in the recently gained introns by using RepeatMasker (http://www.repeatmasker.org). 3 of the 161 newly inserted introns were found to be transposons. Therefore, it seems that even though some novel dipteran introns are formed by insertion of transposable elements into genes, this mechanism is used very rarely.

The alternative duplication theory of intron gain [*Tarrio et al.*, 1998] proposes that new introns are formed by duplication of existing introns. To test this theory we used BLAT [*Kent*, 2002] to search our data set of inserted introns for matches with existing introns. No duplications were found. Therefore, we are unable to obtain any support for the duplication theory in dipteran genomes. It is interesting to note that

this theory was proposed by looking at the gene structure of a *single* gene (Xanthine Dehydrogenase) in dipteran and related genomes. It was found that the novel intron in *Drosophila willistoni* was formed by a duplication of an existing intron. However, our systematic study clearly proves that this mechanism isn't widespread in dipteran genomes.

As discussed earlier in this section, we have found support for the reverse transcriptase mediated theory of intron loss [*Bernstein et al.*, 1983]. Reverse transcriptase works from the 3' to 5' end of a gene and therefore predicts a 3' bias in intron loss. Earlier in this section, we detected considerable preferential loss of 3' introns (Figure 4.9). In addition, we find that there is no bias in the positions of inserted introns. Consequently, the $5'$ bias in intron positions in eukaryotic genomes can be explained by this theory. Our study clearly provides substantial evidence for the reverse transcriptase mediated theory of intron loss.

In contrast, we couldn't find any evidence confirming the deletion theory of intron loss [*Kent and Zahler*, 2000; *Cho et al.*, 2004]. This theory predicts that intron loss is inexact and a small number of codons are added or lost from the flanking coding sequence during intron deletion. To test this theory, we looked at gene alignments around the lost introns. No introns with inexact intron loss were found and this evidence suggests the absence of the deletion mechanism of intron loss in dipteran lineages.

Summary : Most previous studies about gene structure evolution compared gene structures in phylogenetically diverse species. In fact, previous whole genome gene structure evolution studies compared species that are at least 100 million years apart [*Nielsen et al.*, 2004; *Coghlan and Wolfe*, 2004; *Castillo-Davis et al.*, 2004; *Roy and Hartl*, 2006]. In contrast, our study compares species that are around 6 million years apart. Therefore, we have been able to detect introns that have been gained and

lost very recently. Consequently, we expect that there wouldn't be any significant changes in the intronic sequences since they were gained/lost in one of the lineages. In addition, our study is not affected by the possibility of parallel intron gain and loss. Consequently, we believe that this is one of the most error-free studies of gene structure evolution.

We show that the rate of intron gain is much higher than the rate of intron loss in the recent evolutionary history of Diptera. This is in contrast to the rates in mammalian lineages where we couldn't find any occurrences of intron gain. Thus we are able to confirm previous studies such as *Roy and Gilbert* [2005b] and *Rogozin et al.* [2003] which found widely varying rates of gene structure evolution in different eukaryotic lineages. Our findings also provide an explanation for the 5' bias in the position of introns in eukaryotic genomes. We find that there is no discernible bias in the positions of inserted introns, but we were able to find a 3' bias in the positions of lost introns. If similar mechanisms exist in other eukaryotic clades, these dynamics of intron gain and loss will explain the 5' bias in the positions of eukaryotic introns.

We have also tested previously proposed theories of intron gain and loss. We were able to find strong evidence favoring the reverse transcriptase theory of intron loss. In addition, we were able to prove that a small fraction of new introns were formed by insertion of transposable elements. However, it is clear that not all new introns were formed by this mechanism. Furthermore, we were able to find evidence that is contradictory to the predictions of the duplication theory of intron formation [*Tarrio et al.*, 1998] and the deletion theory of intron loss [*Kent and Zahler*, 2000; *Cho et al.*, 2004], proving that these mechanisms are absent in dipteran lineages.

4.4 Concluding Remarks

Among other impacts, the sequencing of several closely related eukaryotic genomes will revolutionize evolutionary biology. In this thesis, we try to develop computational tools that will accelerate the pace of this revolution. One of the first steps in understanding these genomes would be the accurate annotation of protein coding genes in these genomes. Due to the lack of sufficient experimental evidence, we have developed computational tools for rapid and accurate annotation of these newly sequenced genomes. We then use these methods to study the evolution of gene structure in mammalian and dipteran genomes.

An interesting aspect of our study was the variations in the mode of gene structure evolution among various clades. There were particularly stark contrasts in the rate of gene structure evolution. Although no inserted introns were found in mammalian genes, the rate of intron gain was much higher than the rate of intron loss in dipteran clades. In contrast to previous studies in nematodes [*Cho et al.*, 2004], we were unable to find any evidence supporting the deletion theory of intron loss in mammalian and dipteran lineages. Similarly, we were able to find evidence validating the reverse transcriptase mediated mechanism of intron loss unlike conflicting studies in fungi and plasmodium [*Nielsen et al.*, 2004; *Roy and Hartl*, 2006]. As more genomes become available, we will be able to understand these variations and apparent contradictions much more precisely. In particular, we would be able to know if the results from the previous studies were muddled due to parallel intron gain and loss. This is because the previous studies compared genomes which were evolutionarily distant and the presence of hotspots for intron gain and loss would lead to parallel intron gain and loss. Furthermore, such studies would help us answer deeper questions about genome evolution. For example, if a mechanism is common among many eukaryotic clades, it was probably present in the eukaryotic ancestor. An understanding of the molecular

mechanisms in the eukaryotic ancestor would help us resolve the debate between the introns early and introns late theory of the origin of introns.

A population-genetic explanation for the evolution of gene structure and their diversity among eukaryotic clades has recently been proposed [*Lynch*, 2006]. We believe that this is an interesting perspective on gene structure evolution, but it is beyond the scope of the thesis. However, any future hypotheses about gene structure evolution should keep this viewpoint in mind.

Our study used only computational methods to obtain the gene annotations in the non-reference organisms. To the best of our knowledge, most previous studies also used annotations from databases that use computational methods to obtain annotations. We believe that our study is much more accurate compared to previous studies. This is because we compare gene structures of closely related organisms and transferring annotations from a reference genome to an evolutionary close species is much more accurate compared to transferring annotations from a reference genome to an evolutionarily distant species. Furthermore, we have demonstrated that GeneMapper is much more accurate compared to existing gene prediction programs. Therefore, we are very confident about the accuracy of our annotations. However, we should point out that a combination of experimental and computational methods (in which experiments are used to verify the computational gene predictions) would be a much more appropriate, albeit considerably costlier, strategy for future studies.

Finally, we should point out that we have studied only one of the modes of gene structure evolution i.e. intron gain and loss. Gene structure evolution can take place through other mechanisms such as gain/loss of coding sequence or development of alternative isoforms. It is necessary to conduct experiments to study these alternate mechanisms of gene structure evolution and was therefore beyond the scope of this thesis.

Bibliography

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, and B. Olde, Complementary DNA sequencing: expressed sequence tags and human genome project., *Science*, *252*(5013), 1651–6, 1991.

Alexandersson, M., S. Cawley, and L. Pachter, SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model, *Genome Res*, *13*(3), 496–502, 2003.

Allen, J. E., and S. L. Salzberg, JIGSAW: integration of multiple sources of evidence for gene prediction., *Bioinformatics*, *21*(18), 3596–603, 2005.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool., *J Mol Biol*, *215*(3), 403–10, 1990.

Ashurst, J. L., C.-K. Chen, J. G. R. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming, and T. Hubbard, The Vertebrate Genome Annotation (Vega) database, *Nucleic Acids Res*, *33*(Database Issue), D459–65, 2005.

Batzoglou, S., L. Pachter, J. Mesirov, B. Berger, and E. S. Lander, Human and mouse gene structure: comparative analysis and application to exon prediction, in *Proceedings of the fourth annual international conference on Computational molecular biology, April 8-11, 2000; Tokyo, Japan.*, pp. 46–53, ACM Press, New York, NY, USA, 2000.

Benson, G., Tandem repeats finder: a program to analyze DNA sequences., *Nucleic Acids Res*, *27*(2), 573–80, 1999.

Bernstein, L. B., S. M. Mount, and A. M. Weiner, Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites., *Cell*, *32*(2), 461–72, 1983.

Betts, M. J., R. Guigo, P. Agarwal, and R. B. Russell, Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution?, *EMBO J*, *20*(19), 5354–60, 2001.

Birney, E., M. Clamp, and R. Durbin, GeneWise and Genomewise, *Genome Res*, *14*(5), 988–95, 2004a.

Birney, E., et al., An overview of Ensembl, *Genome Res*, *14*(5), 925–8, 2004b.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. Lewis, I. Ovcharenko, L. Pachter, and E. Rubin, Phylogenetic shadowing of primate sequences to find functional regions of the human genome, *Science*, *299*(5611), 1391–4, 2003.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev, dbEST–database for "expressed sequence tags", *Nat Genet*, *4*(4), 332–3, 1993.

Brejova, B., D. G. Brown, M. Li, and T. Vinar, ExonHunter: a comprehensive approach to gene finding., *Bioinformatics*, *21*(Suppl 1), i57–i65, 2005.

Burge, C., Identification of genes in human genomic DNA, Ph.D. thesis, Stanford University, 1997.

Burge, C., and S. Karlin, Prediction of complete gene structures in human genomic DNA, *J Mol Biol*, *268*(01), 78–94, 1997.

Burset, M., and R. Guigo, Evaluation of gene structure prediction programs, *Genomics*, *34*(3), 353–67, 1996.

Castellano, S., N. Morozova, M. Morey, M. J. Berry, F. Serras, M. Corominas, and R. Guigo, In silico identification of novel selenoproteins in the Drosophila melanogaster genome, *EMBO Rep*, *2*(8), 697–702, 2001.

Castillo-Davis, C. I., T. B. C. Bedford, and D. L. Hartl, Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites., *Mol Biol Evol*, *21*(7), 1422–7, 2004.

Cavalier-Smith, T., Intron phylogeny: a new hypothesis., *Trends Genet*, *7*(5), 145–8, 1991.

Cawley, S. L., and L. Pachter, HMM sampling and applications to gene finding and alternative splicing., *Bioinformatics*, *19 Suppl 2*, II36–II41, 2003.

Chatterji, S., and L. Pachter, Multiple organism gene finding by collapsed gibbs sampling, in *Proceedings of the eighth annual international conference on Computational molecular biology, March 27-31, 2004; San Diego, California, USA.*, pp. 187–193, ACM Press, New York, NY, USA, 2004.

Chatterji, S., and L. Pachter, Large multiple organism gene finding by collapsed Gibbs sampling, *J Comput Biol*, *12*(6), 599–608, 2005.

Cho, S., S.-W. Jin, A. Cohen, and R. E. Ellis, A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution., *Genome Res*, *14*(7), 1207–20, 2004.

Coghlan, A., and K. H. Wolfe, Origins of recently gained introns in Caenorhabditis., *Proc Natl Acad Sci U S A*, *101*(31), 11,362–7, 2004.

Collins, L., and D. Penny, Complex spliceosomal organization ancestral to extant eukaryotes., *Mol Biol Evol*, *22*(4), 1053–66, 2005.

Comeron, J. M., and M. Kreitman, The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces., *Genetics*, *156*(0016-6731 (Print)), 1175–90, 2000.

Crick, F., Split genes and RNA splicing., *Science*, *204*(4390), 264–71, 1979.

Dayhoff, M., R. Schwartz, and B. Orcutt, *A model of evolutionary change in proteins*, vol. 5, pp. 345–352, National Biomedical Research Foundation, Washington, DC, 1978.

Deutsch, M., and M. Long, Intron-exon structures of eukaryotic model organisms., *Nucleic Acids Res*, *27*(15), 3219–28, 1999.

Dewey, C., J. Q. Wu, S. Cawley, M. Alexandersson, R. Gibbs, and L. Pachter, Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat., *Genome Res*, *14*(4), 661–4, 2004.

Drysdale, R., et al., FlyBase: genes and gene models, *Nucleic Acids Res*, *33*(Database issue), D390–5, 2005.

Dubchak, I., M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer, Active conservation of noncoding sequences revealed by three-way species comparisons., *Genome Res*, *10*(9), 1304–6, 2000.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

Feingold, E. A., P. J. Good, M. S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, and F. S. Collins, The ENCODE (ENCyclopedia Of DNA Elements) Project., *Science*, *306*(5696), 636–40, 2004.

Flicek, P., E. Keibler, P. Hu, I. Korf, and M. Brent, Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map, *Genome Res*, *13*(01), 46–54, 2003.

Gelfand, M., A. Mironov, and P. Pevzner, Gene recognition via spliced sequence alignment, *Proc Natl Acad Sci U S A*, *93*(17), 9061–6, 1996.

Gibbs, R. A., et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution., *Nature*, *428*(6982), 493–521, 2004.

Gilbert, W., Why genes in pieces?, *Nature*, *271*(5645), 501, 1978.

Gilbert, W., S. J. D. Souza, and M. Long, Origin of Genes., *Proc Natl Acad Sci U S A*, *94*, 7698–7703, 1997.

Giroux, M. J., M. Clancy, J. Baier, L. Ingham, D. McCarty, and L. C. Hannah, De novo synthesis of an intron by the maize transposable element Dissociation., *Proc Natl Acad Sci U S A*, *91*(25), 12,150–4, 1994.

Glusman, G., S. Qin, M. R. El-Gewely, A. F. Siegel, J. Roach, L. Hood, and A. F. A. Smit, A Third Approach to Gene Prediction Suggests Thousands of Additional Human Transcribed Regions, *PLoS Comput Biol*, *In Press*, 2006.

Gross, S. S., and M. R. Brent, Using Multiple Alignments To Improve Gene Prediction, in *Proceedings of the ninth annual international conference on Computational molecular biology, May 14-18, 2005; Cambridge, MA, USA.*, pp. 374–388, ACM Press, New York, NY, USA, 2005.

Guigo, R., E. T. Dermitzakis, P. Agarwal, C. P. Ponting, G. Parra, A. Reymond, J. F. Abril, E. Keibler, R. Lyle, C. Ucla, S. E. Antonarakis, and M. R. Brent, Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes, *Proc Natl Acad Sci U S A*, *100*(3), 1140–5, 2003.

Henikoff, S., and J. G. Henikoff, Amino acid substitution matrices from protein blocks., *Proc Natl Acad Sci U S A*, *89*(22), 10,915–9, 1992.

Huang, X., Fast comparison of a DNA sequence with a protein sequence database., *Microb Comp Genomics*, *1*(4), 281–91, 1996.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, The UCSC Genome Browser Database, *Nucleic Acids Res*, *31*(3), 51–4, 2003.

Keibler, E., and M. R. Brent, Eval: a software package for analysis of genome annotations, *BMC Bioinformatics*, *4*, 50, 2003.

Kent, W., BLAT-the BLAST-like alignment tool., *Genome Res*, *12*(4), 656–64, 2002.

Kent, W. J., and A. M. Zahler, Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment., *Genome Res*, *10*(8), 1115–25, 2000.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, The human genome browser at UCSC, *Genome Res*, *12*(6), 996–1006, 2002.

Kim, N., S. Shin, and S. Lee, ECgene: genome-based EST clustering and gene modeling for alternative splicing, *Genome Res*, *15*(4), 566–76, 2005.

Korf, I., Gene finding in novel genomes, *BMC Bioinformatics*, *5*(1471-2105 (Electronic)), 59, 2004.

Kulp, D., D. Haussler, M. Reese, and F. Eeckman, A generalized hidden Markov model for the recognition of human genes in DNA, *Proc Int Conf Intell Syst Mol Biol*, *4*, 134–42, 1996.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment., *Science*, *262*(5131), 208–14, 1993.

Lin, H., W. Zhu, J. Silva, X. Gu, and C. Buell, Intron gain and loss in segmentally duplicated genes in rice., *Genome Biol*, *7*(5), R41, 2006.

Lin, K., and D.-Y. Zhang, The excess of 5' introns in eukaryotic genomes., *Nucleic Acids Res*, *33*(20), 6522–7, 2005.

Liu, J. S., The collapsed Gibbs sampler with applications to a gene regulation problem, *Journal of the American Statistical Association*, *89*, 958–966, 1994.

Liu, J. S., A. F. Neuwald, and C. E. Lawrence, Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies, *Journal of the American Statistical Association*, *90*, 1156–1170, 1995.

Long, M., C. Rosenberg, and W. Gilbert, Intron phase correlations and the evolution of the intron/exon structure of genes., *Proc Natl Acad Sci U S A*, *92*(26), 12,495–9, 1995.

Lorenc, A., and W. Makalowski, Transposable elements and vertebrate protein diversity., *Genetica*, *118*(2-3), 183–91, 2003.

Lynch, M., Intron evolution as a population-genetic process., *Proc Natl Acad Sci U S A*, *99*(9), 6118–23, 2002.

Lynch, M., The origins of eukaryotic gene structure., *Mol Biol Evol*, *23*(2), 450–68, 2006.

Lynch, M., and J. S. Conery, The origins of genome complexity., *Science*, *302*(5649), 1401–4, 2003.

Malko, D. B., V. J. Makeev, A. A. Mironov, and M. S. Gelfand, Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes., *Genome Res*, *16*(1088-9051 (Print)), 505–9, 2006.

Martin, W., and E. V. Koonin, Introns and the origin of nucleus-cytosol compartmentalization., *Nature*, *440*(7080), 41–5, 2006.

McAuliffe, J., L. Pachter, and M. Jordan, Multiple-sequence functional annotation and the generalized hidden Markov phylogeny., *Bioinformatics*, *20*(12), 1850–60, 2004.

McNaughton, J. C., G. Hughes, W. A. Jones, P. A. Stockwell, H. J. Klamut, and G. B. Petersen, The evolution of an intron: analysis of a long, deletion-prone intron in the human dystrophin gene., *Genomics*, *40*(2), 294–304, 1997.

Meyer, I., and R. Durbin, Gene structure conservation aids similarity based gene prediction., *Nucleic Acids Res*, *32*(2), 776–83, 2004.

Meyer, I. M., and R. Durbin, Comparative ab initio prediction of gene structures using pair HMMs., *Bioinformatics*, *18*(10), 1309–18, 2002.

Morgenstern, B., DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment., *Bioinformatics*, *15*(3), 211–8, 1999.

Mott, R., EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA., *Comput Appl Biosci*, *13*(4), 477–8, 1997.

Nekrutenko, A., and W. H. Li, Transposable elements are found in a large number of human protein-coding genes., *Trends Genet*, *17*(11), 619–21, 2001.

Nguyen, H. D., M. Yoshihama, and N. Kenmochi, New maximum likelihood estimators for eukaryotic intron evolution., *PLoS Comput Biol*, *1*(7), e79, 2005.

Nielsen, C. B., B. Friedman, B. Birren, C. B. Burge, and J. E. Galagan, Patterns of intron gain and loss in fungi., *PLoS Biol*, *2*(12), e422, 2004.

O'Dushlaine, C. T., R. J. Edwards, S. D. Park, and D. C. Shields, Tandem repeat copy-number variation in protein-coding regions of human genes., *Genome Biol*, *6*(8), R69, 2005.

Ogurtsov, A. Y., S. Sunyaev, and A. S. Kondrashov, Indel-based evolutionary distance and mouse-human divergence., *Genome Res*, *14*(8), 1610–6, 2004.

Palmer, J. D., and J. M. J. Logsdon, The recent origins of introns., *Curr Opin Genet Dev*, *1*(4), 470–7, 1991.

Parra, G., E. Blanco, and R. Guigó, GeneID in Drosophila, *Genome Res*, *10*(4), 511–5, 2000.

Parra, G., P. Agarwal, J. Abril, T. Wiehe, J. Fickett, and R. Guigó, Comparative gene prediction in human and mouse, *Genome Res*, *13*(01), 108–17, 2003.

Patthy, L., Genome evolution and the evolution of exon-shuffling–a review., *Gene*, *238*(1), 103–14, 1999.

Pennisi, E., Gene counters struggle to get the right answer, *Science*, *301*(5636), 1040–1, 2003.

Qiu, W.-G., N. Schisler, and A. Stoltzfus, The evolutionary gain of spliceosomal introns: sequence and phase preferences., *Mol Biol Evol*, *21*(7), 1252–63, 2004.

Quackenbush, J., J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parvizi, G. Pertea, R. Sultana, and J. White, The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species., *Nucleic Acids Res*, *29*(1), 159–64, 2001.

Rinner, O., and B. Morgenstern, AGenDA: gene prediction by comparative sequence analysis., *In Silico Biol*, *2*(3), 195–205, 2002.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin, Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution., *Curr Biol*, *13*(17), 1512–7, 2003.

Roy, S. W., Recent evidence for the exon theory of genes., *Genetica*, *118*(2-3), 251–66, 2003.

Roy, S. W., and W. Gilbert, Complex early genes., *Proc Natl Acad Sci U S A*, *102*(6), 1986–91, 2005a.

Roy, S. W., and W. Gilbert, Rates of intron loss and gain: implications for early eukaryotic evolution., *Proc Natl Acad Sci U S A*, *102*(16), 5773–8, 2005b.

Roy, S. W., and W. Gilbert, The pattern of intron loss., *Proc Natl Acad Sci U S A*, *102*(3), 713–8, 2005c.

Roy, S. W., and W. Gilbert, The evolution of spliceosomal introns: patterns, puzzles and progress., *Nat Rev Genet*, *7*(3), 211–21, 2006.

Roy, S. W., and D. L. Hartl, Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number., *Genome Res*, *16*(6), 750–6, 2006.

Roy, S. W., A. Fedorov, and W. Gilbert, Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain., *Proc Natl Acad Sci U S A*, *100*(12), 7158–62, 2003.

Sakurai, A., S. Fujimori, H. Kochiwa, S. Kitamura-Abe, T. Washio, R. Saito, P. Carninci, Y. Hayashizaki, and M. Tomita, On biased distribution of introns in various eukaryotes., *Gene*, *300*(1-2), 89–95, 2002.

Schuler, G. D., et al., A gene map of the human genome., *Science*, *274*(5287), 540–6, 1996.

Shah, S. P., G. P. McVicker, A. K. Mackworth, S. Rogic, and B. F. F. Ouellette, GeneComber: combining outputs of gene prediction programs for improved results., *Bioinformatics*, *19*(10), 1296–7, 2003.

Siepel, A., and D. Haussler, Computational identification of evolutionarily conserved exons, in *Proceedings of the eighth annual international conference on Computational molecular biology, March 27-31, 2004; San Diego, California, USA*, pp. 177–186, ACM Press, New York, NY, USA, 2004.

Snel, B., P. Bork, and M. Huynen, Genome evolution. Gene fusion versus gene fission., *Trends Genet*, *16*(1), 9–11, 2000.

Stanke, M., and S. Waack, Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, *19*(Suppl 2), II215–II225, 2003.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes., *Proc Natl Acad Sci U S A*, *99*(7), 4465–70, 2002.

Tanner, M. A., and W. H. Wong, The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, *82*(398), 528–540, 1987.

Tarrio, R., F. Rodriguez-Trelles, and F. J. Ayala, New Drosophila introns originate by duplication., *Proc Natl Acad Sci U S A*, *95*(4), 1658–62, 1998.

Thierry-Mieg, J., M. Potdevin, and M. Sienkiewicz, Identification and functional annotation of cDNA-supported genes in higher organisms using AceView, unpublished.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, and P. An, Initial sequencing and comparative analysis of the mouse genome, *Nature*, *420*(6915), 520–62, 2002.

Wu, J. Q., D. Shteynberg, M. Arumugam, R. A. Gibbs, and M. R. Brent, Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing, *Genome Res*, *14*(4), 665–71, 2004.

Wu, T. D., and C. K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, *21*(9), 1859–75, 2005.

Yandell, M., C. J. Mungall, C. Smith, S. Prochnik, J. Kaminker, G. Hartzell, S. Lewis, and G. M. Rubin, Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes., *PLoS Comput Biol*, *2*(3), e15, 2006.

Yeh, R. F., L. P. Lim, and C. B. Burge, Computational inference of homologous gene structures in the human genome., *Genome Res*, *11*(5), 803–16, 2001.

Yoshihama, M., T. Uechi, S. Asakawa, K. Kawasaki, S. Kato, S. Higa, N. Maeda, S. Minoshima, T. Tanaka, N. Shimizu, and N. Kenmochi, The human ribosomal protein genes: sequencing and comparative analysis of 73 genes., *Genome Res*, *12*(3), 379–90, 2002.