Comparative Gene Identification in Mammalian, Fly, and Fungal Genomes

by

Michael F. Lin
S.B., Computer Science and Engineering (2005)
Massachusetts Institute of Technology

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology

May 2006

Copyright © 2006 Massachusetts Institute of Technology
All rights reserved

Author..............................................

Department of Electrical Engineering and Computer Science

May 26, 2006

Certified by..............................................................

Manolis Kellis
Thesis Supervisor

Accepted by..............

Arthur C. Smith
Chairman, Department Committee on Graduate Theses
Comparative Gene Identification in Mammalian, Fly, and Fungal Genomes

by

Michael F. Lin

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology May 2006

ABSTRACT

An important step in genome interpretation is the accurate identification of protein-coding genes. One approach to gene identification is comparative analysis of the genomes of several related species, to find genes that have been conserved by natural selection over millions of years of evolution. I develop general computational methods that combine statistical analysis of genome sequence alignments with classification algorithms in order to detect the distinctive signatures of protein-coding DNA sequence evolution. I implement these methods as a software system, which I then use to identify previously unknown genes, and cast doubt on some existing gene annotations, in the genomes of the fungi Saccharomyces cerevisiae and Candida albicans, the fruit fly Drosophila melanogaster, and the human. These methods perform competitively with the best existing de novo gene identification systems, and are practically applicable to the goal of improving existing gene annotations through comparative genomics.

Thesis Supervisor: Manolis Kellis
Title: Assistant Professor of Electrical Engineering and Computer Science
1 Introduction

In 2001, the Human Genome Project published the sequence of the human genome. Encoded within these three billion DNA base pairs are the blueprints for constructing the molecular machines that make our cells work and a complex regulatory code which, in ways that we are only beginning to understand, governs the development of a single fertilized egg into a fully formed human being. Also contained somewhere within are the genetic factors that can predispose us to obesity, heart disease, diabetes, cancer, and many other diseases. Now that the sequence is available, the challenge for the next generation is to interpret and understand the information contained within, both to advance fundamental biology and to drive the development of revolutionary treatments for human disease.

Analyzing the sequence of the human genome is a daunting task. Only a small fraction of it encodes biologically functional elements, while a great majority seems to be an evolutionary artifact that serves no present-day function at all. How can we distinguish functional elements from non-functional sequence? Moreover, how can we distinguish different types of functional elements in the genome? A promising new approach to this problem is comparative genomics, the comparative analysis of the genomes of several evolutionarily related species.

Over millions of years of evolution, functional elements in the genome are subject to natural selection. Mutations in functional elements are likely, on average, to reduce the fitness of the organism, and thus tend to be selected against. In contrast, non-functional “junk” sequence undergoes essentially random mutations that have no effect on the organism, and are therefore not selected against. In comparisons of the genomes of related species, then, shared functional elements stand out as having been more strongly conserved by evolution than the surrounding nonfunctional sequence. Thus, in principle, comparative genomics has the power to precisely define the functional elements shared among a set of related genomes. Moreover, it is possible to choose a set of species to analyze such that the vast majority of functional elements are expected to be shared among their genomes, while the nonfunctional sequence has diverged sufficiently that it can be distinguished as such. Hence, comparative analysis of a target genome with the genomes of several, carefully chosen related species can be used to reveal essentially all of the functional elements within.
In this thesis, I propose a methodology for systematically identifying protein-coding genes by comparative analysis of several related genomes. Although protein-coding genes are only one category of functional genomic elements, they have distinctive and well-studied properties that make them particularly amenable to comparative analysis. I have implemented this method as a software system, and applied it to the genomes of several different species: baker’s yeast, *Saccharomyces cerevisiae*, the fungal pathogen *Candida albicans*; the fruit fly *Drosophila melanogaster*; and the human. In each case, this analysis leads to the prediction of many novel protein-coding genes strongly supported by comparative evidence. Moreover, this methodology can be used to systematically analyze and evaluate existing gene annotations, and thus to identify where they may be incorrect or misleading. Thus, the comparative evidence we examine can, in concert with other lines of evidence, facilitate the completion of an accurate gene catalog for the human and other genomes, which will ultimately be necessary to fulfill the great potential of genomics in the 21st century.
2 Background

In this chapter, I provide relevant background on genomes, genes, and comparative genomics. In the next chapter, I rely on this background to thoroughly define the problem that this thesis addresses.

2.1 DNA and Genomes

As computer scientists, we can understand a DNA sequence as a string over an alphabet of four characters, A, G, C, and T, representing four different nucleotides that can form a lengthy molecular chain: adenine, guanine, cytosine, and thymine. The genome of an organism is the complete DNA sequence, or set of sequences, that defines its genetic identity. Although every organism in a species has a unique genome insofar as there is genetic variation within the population, this variation is small enough that it is meaningful to study the genome of a species. (The genomes of any two humans, for example, are expected to be at least 99.9% identical to each other.)

Genomes, from the simplest bacteria up to the human, are a mosaic of many different types of elements, only some of which are currently understood. The most well-studied functional elements are protein-coding genes, which are instructions for synthesizing proteins, the building blocks of all cellular machines. Other genes code not for proteins, but rather molecules made of RNA, a chemical cousin of DNA that can perform catalytic and regulatory functions in the cell. Also contained in the genome are cis-regulatory elements which, in concert with much cellular machinery, control when genes are transcribed, the first step in protein synthesis. Some genome sequence is comprised of highly repetitive heterochromatin that serves as structural scaffolding for the chromosome (DNA molecule) in the cell. And especially in higher species, genomes contain large amounts of nonfunctional sequence, left over as historical artifacts of evolution. Some of these include pseudogenes, sequences that were once protein-coding genes but have since ceased to function and degraded, and transposons, self-replicating sequences that increase in number over evolutionary time but serve no useful purpose to the organism. Only about 1.5% of the human genome is protein-coding, while the vast majority, by some estimates 95% or more, is thought to be nonfunctional "junk" sequence. However, it is likely that there are genomic
mechanisms that remain unknown to us, and it cannot be ruled out that a substantially larger fraction of the human genome than currently thought might have some biological function.

2.2 Protein-Coding Genes and the Genetic Code

In this thesis, I focus on protein-coding genes. Protein-coding genes are instructions for synthesizing a protein. When protein synthesis begins, the gene, which is an otherwise undistinguished substring in the genome, is copied into RNA, a chemical cousin of DNA, in a process called transcription. The transcript is then translated into a chain of amino acids, the basic chemical building blocks of proteins. The chain of amino acids, called a polypeptide, later folds up in three dimensions to form a functional protein.

With very few exceptions, twenty different amino acids are the building blocks of proteins across all kingdoms of life. The genetic code defines the mapping between sequences in the RNA transcript, and by extension the genome, and amino acids. Every three nucleotides in the transcript specify a single amino acid in the polypeptide. A nucleotide triplet is called a codon in this context. For example, the codon TCA corresponds to the amino acid serine. During translation, the cellular machinery reads each codon in the transcript one-by-one and attaches the corresponding amino acid to the nascent polypeptide. The genetic code itself, the mapping between codons and amino acids, is nearly universal among known forms of life.

```
DNA    GAACGGCTCGACTCCGATAGCAGG
RNA    GAACGGCUUCGACUCCGAUAGCAGG
protein E R L D S D S R
```

Figure 1. Information flow from a protein-coding gene. Protein-coding DNA sequence in the genome is transcribed into RNA, a chemical cousin of DNA. The RNA nucleotides are translated in groups of three (codons) into a sequence of amino acids such as aspartic acid (D) and serine (S), which form the primary structure of a protein.

Since there are four DNA nucleotides and three nucleotides in a codon, there are $4^3 = 64$ possible codons, used to specify only twenty different amino acids. The genetic code is degenerate: more than one codon may code for a single amino acid. A few amino acids (such as
methionine) have only one corresponding codon, while several (such as serine and alanine) have as many as four. There are four special codons that control translation: one start codon, which indicates where in the transcript translation should begin, and three stop codons, which indicate where translation ends. The start codon, ATG, also codes for methionine when it occurs in the coding region of the transcript. The stop codons, TAA, TAG, and TGA, do not code for any amino acids under normal circumstances, but rather cause translation to terminate.

The transcript of a protein-coding gene typically contains an upstream untranslated region (UTR), followed by the actual protein-coding sequence, followed by a downstream UTR. The protein-coding sequence consists of the start codon, followed by a number of codons, followed by one of the three stop codons. The portion of the sequence with no stop codons is referred to as an open reading frame (ORF). (See Figure 2.)

2.3 Introns, Exons, and Splicing

In the genomes of simple organisms such as bacteria, the RNA transcript of a gene directly corresponds to the sequence in the genome. However, in higher forms of life, the relationship of the final transcript sequence to the genomic sequence is more complicated. In these genomes, including all the species studied in this thesis, the protein-coding sequence may not occur as a contiguous open reading frame in the genome, but rather may be partitioned into several pieces separated by noncoding sequence. The coding portions of these genes are called exons, and the intervening noncoding portions are called introns. When the transcript is copied from the genome, it subsequently undergoes splicing, where the introns are cut out from the transcript to form a contiguous open reading frame. Specifically, processing machinery in the cell recognizes splice sites in the transcript, which are short sequences (8-20 nucleotides in length) that flank the introns, and recruit enzymes that cut the introns out of the transcript. There are two distinct types of splice sites: acceptor sites, which occur at the beginning of an exon, and donor sites, which occur at the end of an exon. The transcript after splicing is called the messenger RNA (mRNA), while before splicing it is called the pre-mRNA.
Figure 2. Structure of a eukaryotic protein-coding gene. (a) Transcription from the genomic DNA into an RNA transcript starts and ends at certain points in the genome. Introns in the RNA transcript are spliced out to form (b) the final mRNA. The open reading frame (ORF) between the ATG start codon and the stop codon (in this case TAA) codes for the protein sequence as shown in Figure 1. The ORF is flanked by untranslated regions (UTR).

There are several biological reasons for splicing. Most notably, it allows for alternative splicing: the selective inclusion or exclusion of individual exons during transcript processing. Alternative splicing allows one gene to encode several different proteins, and is thought to be a major progenitor of structural and functional protein diversity from a comparatively limited number of protein-coding genes.

### 2.4 Comparative Genomics

The genomes of modern species are a product of their evolution over millions of years. Thus, comparative analysis of the genomes of related species can yield insights into their evolution, and conversely, our knowledge of evolutionary theory can guide the comparative analysis of genomes. The field of **comparative genomics** is one of the most active areas in computational biology research, as the technologies developed by the Human Genome Project are being used to sequence the genomes of dozens of species in all kingdoms of life.

The major motivation behind comparative genomics for the identification of functional elements is the observation that functional elements evolve subject to the constraints of natural selection, whereas nonfunctional sequence undergoes essentially random mutation. Comparative analysis of the genomes of related species, which have evolved independently since they diverged, can reveal selection on functional elements. Most apparently, functional elements generally show higher sequence identity among related genomes than nonfunctional sequence,
since any individual mutation in a functional element is more likely to reduce the fitness of the organism than a mutation in a nonfunctional element. More specifically than indicating that a sequence is functional, however, comparative analysis can reveal distinctive signatures of selection that are clues to how the sequence is functioning. This thesis will investigate in detail how one class of these signatures can reveal the presence of protein-coding sequence. However, before it is possible to compare genomes at such a fine granularity, there several major challenges that must be addressed.

**Sequence alignment.** A classic problem in computational biology is sequence alignment. Given two or more related biological sequences (for example, DNA sequences of related genes in two different species), a sequence alignment algorithm computes the optimal pattern of insertions, deletions, and nucleotide substitutions in each sequence in order to match up the parts of the sequences that are similar to one another. Sequence alignment among biological sequences that are known to be related is a reasonably well-understood algorithmic problem, and there are a variety of tools available for this purpose.

![Figure 3. Example of a sequence alignment. A sequence alignment algorithm computes the optimal placement of gaps (indicated by dashes) in order to line up several biological sequences, highlighting their similarities and differences. The stars indicate perfect matches down the corresponding column of the alignment. Gaps are also referred to as “indels” referring to the inference that, during evolution, some sequence was inserted or deleted at that point in the sequence.](image)

**Genome alignment.** Unfortunately, sequence alignment is not by itself sufficient for comparative genomics. To understand why, consider that while humans have 24 distinct types of chromosomes, mice have 21, and dogs have 39. While the DNA sequence for each of these chromosomes is known, no simple mapping exists between the chromosomes of the different species. If a certain chromosome in human contains a certain set of genes, those genes might be dispersed across several different chromosomes in other species, and vice versa. Moreover, individual genes, and even whole chromosomes, can be duplicated or lost during evolution. Thus, in addition to a tool for nucleotide-level sequence alignment, it is necessary to have a methodology for genome alignment, determining at a large scale which parts of related genomes
correspond to each other, in order to comparatively analyze those genomes. The effectiveness of any fine-grained comparative genomic analysis is strongly dependent on the completeness and quality of the genome alignments for the species under analysis. Genome alignment, however, remains a research area with many unresolved problems, especially on large genomes such as those of the mammals. I will rely on tools still very much under development in order to proceed.

**Choosing informant species.** Finally, it is crucial to carefully choose the species to compare, so that their genomes are far enough diverged that the conservation of functional elements, such as genes, is measurable against the background, but not so far that the desired elements are not well-conserved across those species. For example, the genome of the chimpanzee is so similar to that of a human that a comparative analysis of the two is virtually uninformative for finding conserved genes, while human and fly are so distant that probably only a modest subset of their genes are even shared.

In the case of the human genome, it is thought that eutherian mammals outside of primates, such as dog and mouse, are appropriate candidates for comparative genomic analysis\(^3\). The vast majority of human genes should be conserved in these species, while the nonfunctional sequence has diverged substantially more. The common ancestor of these species is thought to have lived about 125 million years ago\(^4\).

Another important concern is the number of informant genomes to use. Because each genome contributes some unique information about how random mutation and natural selection acted over an independent evolutionary lineage, more genomes in principle provide more power to distinguish functional elements under selection\(^5\). However, a number of practical issues arise as more genomes are added to the analysis. Each step in genome alignment and sequence alignment is slightly error-prone, and using more genomes provides more opportunities for alignment problems at any given point. Likewise, the sequences of the informant genomes themselves are often incomplete or error-prone due to the limitations of genome sequencing technology. Finally, mammalian genomes take several gigabytes to store, and managing many times this amount of data can be challenging from a bioinformatics perspective. The most effective choice for how many informant genomes to use strongly depends on the type of analysis being pursued and the properties of the genomes, and I will explore in this thesis how various combinations of informant genomes affect my analysis.
3 The Gene Identification Problem

Protein-coding genes are otherwise undistinguished DNA sequences within the genome, with no obvious, specific markers in the sequence indicating where they begin and end. If the genome is represented as a string, then the gene identification problem is to identify which substrings comprise protein-coding genes. In genes with introns, only the exons must be predicted as protein-coding sequence. It may also be desirable to identify gene structures: to predict how individual exons splice together to form a complete open reading frame. A gene identification system as defined here can predict that a certain sequence is protein-coding, but does not attempt to predict the structure or function of the resulting protein. The latter is extremely difficult based on the coding sequence alone, and is generally considered a separate problem.

3.1 Traditional Methods for Gene Identification

Gene identification has been a fundamental problem in genomics since its inception. The most direct way to locate individual protein-coding genes is to search the genome for a sequence that is similar to the sequence of a protein or transcript known to be produced in the cell. The sequences of proteins and transcripts are directly related to the sequence of the gene that coded for them, and it is a relatively straightforward computational problem to efficiently search the genome for such sequences\(^6\)\(^7\). A limitation of such an experimental evidence-based technique is that it requires costly large-scale sequencing of protein and transcripts in addition to the genome. Also, while every cell in an organism contains a full copy of the genome, different cells produce different subsets of proteins, in different amounts and at different times. Thus, while protein and transcript sequences provide the strongest evidence of a protein-coding gene in the genome, they may not be readily available for many genes. Hence, it is also desirable to identify genes solely by examination of the genomic sequence, without relying on experimentally-derived evidence. Such \textit{de novo} or \textit{ab initio} strategies for gene identification are the focus of this thesis.

\textit{De novo} gene prediction strategies rely on identifying sequence properties that can broadly be categorized as \textit{signals}, specific sequences or sequence patterns that suggest the presence of a gene, and \textit{content}, distinctive properties of protein-coding sequence itself. For example, protein-coding sequences are flanked by signals such as start codons, stop codons, and splice sites, and
usually exhibit a variety of content properties, such as periodicities in nucleotide composition at each of the three codon positions, and biases in the usage of the 61 sense codons\(^8\). All known sequence signals (such as start codons and splice sites) are highly nonspecific, occurring frequently in locations not associated with genes. Content properties are subtle and difficult to detect in short sequences, such as the typical length of a human exon. Hence, any of these properties are individually insufficient to reliably identify genes. But they can be simultaneously combined to identify protein-coding sequence with better accuracy. Different measurements are often combined using a hidden Markov model (HMM), a type of probabilistic model that is naturally applicable to the problem of parsing a biological sequence into separate subsequences, such as exons and introns. HMM-based gene finders such as Glimmer\(^9\) are highly effective in prokaryotic and viral genomes, which have long, contiguous open reading frames and a low proportion of noncoding sequence.

The gene identification problem is more difficult in the genomes of higher organisms such as the human, with its short exons and many, long introns. In addition to signal and content measurements, HMM-based gene finders for higher eukaryotic genomes use the constraints imposed by a model of gene structure. In particular, the first exon in a predicted gene must begin with a start codon; the last exon must end with a stop codon; each exon in between must be flanked by splice sites; and the reading frame of translation must carry over from one exon to the next. These constraints can be incorporated into the hidden Markov model in a natural way. However, these systems, and all \textit{ab initio} gene finders, have had limited success in the human genome. This can be attributed to the human genome's very low signal-to-noise ratio: the median human exon length is just 122 nucleotides, and they make up only about 1.5% of the entire genome. The signals and content biases that \textit{ab initio} gene finders measure have limited discriminatory power with such rare and short exons. The most notable HMM-based gene finder designed for the human genome is GENSCAN\(^{10}\), but its predictions have generally proven unreliable for practical use\(^{11}\).

### 3.2 Problem Statement: Comparative Gene Identification

Extrinsic protein and transcript evidence may not be obtainable for all genes, and known statistical properties of protein-coding regions are insufficient to identify human genes with high accuracy. In order to have more success at accurately identifying genes in complex genomes, it is
necessary to increase the signal-to-noise ratio. Comparative genomics is a particularly promising avenue towards this goal, since the signatures of random mutation and natural selection can be used to distinguish functional from nonfunctional sequence.

In this thesis, I address the problem of using multiple sequence alignments of several related genomes in order to identify evolutionarily conserved protein-coding sequence. The key questions I address are:

1. What are the signatures of mutation and natural selection that distinguish evolutionarily conserved protein-coding sequence from nonfunctional and conserved non-coding sequence?
2. How can we identify and measure these signatures in multiple sequence alignments of several related genomes?
3. How can we use these measurements to generate de novo predictions of individual protein-coding sequences (exons) in the target genome?
4. How well does this prediction strategy perform across a variety of species, and how sensitive is its performance to the choice and number of informant genomes and genome alignments?

Additionally, in the process of evaluating the performance of this prediction strategy against imperfect existing gene annotations, I identify many new genes and call into question some existing gene annotations in each of the species I analyze.

I do not address in this thesis the important problem of constructing gene structure models from individual exon predictions. This remains a direction for future research, as I will discuss.

### 3.3 Related Work

Comparative genomics approaches for gene identification have been under development for several years. Any such system is dependent upon the availability of genome sequence for appropriate informant species, as well as that of the species primarily under investigation. Initially, human and mouse were the only mammals for which high-quality genome sequence
was available. The earliest comparative gene finders for mammals were thus designed to consider pairwise conservation between two species. Meyer & Durbin\textsuperscript{12} and Korf et. al.\textsuperscript{13} developed DOUBLESCAN and TWINSCAN, respectively, tools similar to GENSCAN that were able to predict genes in human DNA sequence with increased accuracy by simultaneously analyzing related mouse sequence. Several other tools\textsuperscript{14,15,16} are available that also measure pairwise sequence conservation to predict genes, typically between human and mouse. While these systems are substantially more accurate than purely \textit{ab initio} systems like GENSCAN, they are essentially incremental improvements.

An important development for comparative gene finding is the simultaneous analysis of more than two related genomes to extract the strongest conservation signals. Several different groups\textsuperscript{17,18,19} have recently used comparative analysis of the genomes of several species of yeast, an important model organism in biology, to find new protein-coding genes and other conserved functional elements in the yeast genome. In these cases, by analyzing not just two, but four or more related genomes, the comparative approach was so effective that these studies led to sweeping revisions of the yeast gene catalog, which is now considered substantially more reliable as a result.

High-quality sequences for a number of mammalian genomes have become available in the past few years. It is, therefore, now possible to undertake similar efforts to identify mammalian genes with unprecedented accuracy. Siepel and Haussler\textsuperscript{20} proposed the "phylo-HMM", a complex probabilistic model that measures conservation across many different species, including mammals, to identify individual exons. Similar approaches include EHMMs\textsuperscript{21} and GHMPs\textsuperscript{22}. Most recently, Gross and Brent\textsuperscript{23} have extended TWINSCAN to additionally consider evolutionary conservation across multiple related genomes to obtain N-SCAN, the current state-of-the-art in human gene prediction.
4 Signatures of Protein-Coding Sequence Evolution

In this chapter, I explain how unique properties of the genetic code lead protein-coding sequence to evolve in distinctive ways, and propose quantitative measurements of these signatures. I apply these measurements to test sets of genes from a variety of different species to demonstrate that they indeed distinguish protein-coding sequence from noncoding sequence. One application of these measurements is to evaluate existing gene annotations to determine whether the putative genes appear to have been conserved by evolution. In the next chapter, I will develop a methodology to use these measurements in order to generate de novo predictions of protein-coding regions in a genome.

As we have already observed, functional sequence is more strongly conserved across related genomes than nonfunctional sequence because natural selection tends to weed out mutations in the functional sequence, while it is neutral towards nonfunctional sequence. Typically, however, natural selection does not operate directly on the genomic sequence; when a mutation occurs in a gene, it is a consequence of that mutation, such as a reduction in the efficiency of an enzyme, that selection may act against. It is possible for a mutation to the gene sequence to affect the function of the resulting protein to a lesser or greater degree, or not at all. Hence, some mutations in the genomic protein-coding sequence are more easily tolerated than others by natural selection. Over evolutionary time, this leads to certain distinct signatures of evolution in protein-coding sequence, as certain types of mutations are preserved more than others.

4.1 Reading Frame Conservation

As the translation machinery reads codons in the transcript one-by-one, it is crucial for it to process each nucleotide triplet in the correct reading frame: it must read the nucleotides in positions 0, 3, 6, 9, ... of the coding sequence as the first position in each codon. A frameshift error, when the transcript is read in the incorrect reading frame, leads to a completely different protein translation. The cellular machinery has elaborate mechanisms to abort protein synthesis when a frameshift errors occur during transcription or translation, but if a frameshift occurs in the genomic sequence, then there is no way for the correct protein to be produced.
Figure 4. Frameshift mutations. (a) an example protein-coding sequence and its corresponding translation. (b) the insertion or deletion of a single nucleotide, or any number of nucleotides not divisible by three, mangles the translation of the remainder of the protein because the remaining nucleotides are read in the incorrect reading frame. (c) in contrast, the insertion or deletion of a number of nucleotides divisible by three only affects the protein in the immediate vicinity of the event, in this case deleting a single amino acid.

During evolution, sequence insertions and deletions in the genome occur at random as a result of errors during DNA replication. (Since it is usually difficult to distinguish a sequence insertion in one species from a deletion in another during evolution, these events are often referred to as indels.) If a frameshift indel occurs somewhere inside a protein-coding gene or exon, then it is likely to severely mangle the remainder of the protein, as shown in Figure 4. Frame-shifting indels are therefore strongly selected against over evolutionary time. However, an indel that is a multiple of three in length will affect the protein sequence only in the immediate vicinity of the indel, typically through the insertion or deletion of individual amino acids. Such an event is much more likely to be tolerated by natural selection. Hence, in alignments of protein-coding sequence, indel events are much rarer than in noncoding sequence, and when they do occur, they are strongly biased to be a multiple of three in length. This bias in indel patterns is a powerful discriminator between coding and non-coding sequence when the genomic region under examination is lengthy enough that several indels have occurred within during evolution.
4.1.1 Measuring Reading Frame Conservation

In order to quantitatively assess whether there has been an evolutionary pressure to conserve the codon reading frame in a sequence alignment, I adopt the reading frame conservation (RFC) test proposed by Kellis\textsuperscript{8}. Briefly, the RFC test inspects the gap pattern in a pairwise alignment between a target and informant sequences to determine where frameshifting indels occur, and evaluates how much of the informant sequence appears in a consistent reading frame with respect to the target sequence. Specifically, I compute the percentage of the nucleotides in the target sequence that have an informant nucleotide aligned to it in the same reading frame. In the case of lengthy intervals ($\geq 200$nt) to be measured, I compute this percentage in 100nt sliding windows overlapping by 50nt, and take the average of the windows to obtain the RFC score for that interval.

In a multiple alignment, we have one target and several informant sequences. Each informant casts a "vote", +1, 0, or -1. If the informant sequence aligns to fewer than 50\% of the bases in the target sequence, it votes 0. Otherwise, I use a species-specific cutoff on the RFC score to decide whether it votes +1 or -1. This cutoff is chosen by observing the distribution of RFC scores in alignments of known genes and intergenic regions between the target and each informant sequence. This distribution is usually clearly bimodal, and I choose the maximum likelihood estimator as the cutoff, assuming the distribution is a mixture of real genes and noncoding intervals. Across a wide range of species, this cutoff is usually between 70\% and 80\%. If the RFC score for a species is above this cutoff, it votes +1; otherwise, it votes -1. I then sum the votes to obtain an overall RFC score for an interval in a multiple alignment, which may range between $-n$ and $n$, where $n$ is the number of informant sequences.

4.1.2 Reading Frame Conservation in Known Genes

In this section, I compute RFC scores for a test set of known exons and noncoding regions in the genome of the fruit fly, \textit{Drosophila melanogaster}, in order to demonstrate that protein-coding sequence indeed exhibits reading frame conservation, whereas noncoding regions usually do not. I choose to present this data in fly for two reasons: first, because a large number (11) of informant genomes are available, allowing us to try different combinations to understand how this effects our analysis, and second, because the distribution of fly exon lengths is intermediate.
among the genomes I study (human exons tend to be shorter, and yeast exons tend to be longer). Details about the twelve fly genomes, the genome alignments, and the existing gene annotations can be found in chapter 6.

Figure 5. RFC scores of coding and noncoding regions in alignments of fly genomes. RFC scores were computed as described in section 4.1.1 over intervals covering approximately 5% of the fruit fly (Drosophila melanogaster) genome. The bars represent the total number of distinct intervals with the corresponding RFC score on the x-axis. The yellow (light) bars represent known protein-coding exons from a gene annotation database (total 2,464) while the blue (dark) bars represent other open reading frames (total 10,178). The scores were computed using one, two, four, and eleven informant genomes (see chapter 6 for details about these fly genomes). The fraction and length distribution of the putatively noncoding intervals accurately reflects the composition of ORFs the genome. As more informant genomes are used in the analysis, the power of the RFC score to distinguish coding from non-coding intervals increases.
I chose a test set of sequence alignments covering approximately 5% of the fly genome, and used an existing gene annotation database to identify known exons within these alignments. I then computed the RFC scores over these known exons as well as a sample of other, presumably noncoding intervals, chosen with a length distribution that accurately reflects the composition of the genome. As shown above, coding sequence indeed tends to exhibit excellent RFC. 55% (1,360 out of 2,464) of known exons demonstrate RFC in all eleven informant genomes, while only 0.9% (93 of 10,178) of other intervals meet the same criterion. 96% (2,371) of known exons and only 25% (2,504) of other intervals have a positive RFC score using all eleven informants. It should be noted that the existing annotation database is imperfect, so these figures may somewhat underestimate the true accuracy of the RFC score as a discriminator between protein-coding and noncoding sequence. Still, it is evident from these figures that while the RFC score is a powerfully informative signal, it is not alone sufficient to build a highly accurate gene finder, at least in the fly genome. Additional signals are needed to specifically identify protein-coding sequence.

4.2 Codon Substitution Biases

Individual substitutions in the nucleotide sequence of a gene may affect the resulting protein by substituting a different amino acid. For example, substitution of the codon TCC with GCC would cause the protein to contain the amino acid alanine instead of serine at the corresponding position in the amino acid sequence. Such an event is known as a missense mutation. However, because the genetic code is degenerate, not all codon substitutions lead to amino acid substitutions. A silent mutation occurs when a codon changes to another codon that codes for the same amino acid. Silent mutations are tolerated more easily than missense mutations by natural selection because they usually do not affect the function of the resulting protein.

Moreover, some missense mutations are tolerated more easily than others. The amino acids have different chemical properties, such as electrical charge or hydrophobicity. The substitution of a hydrophobic residue with another hydrophobic residue is more likely to be tolerated than the substitution of a hydrophobic for a hydrophilic residue. Finally, a nonsense mutation occurs when a sense (coding) codon mutates into a stop codon. Nonsense mutations in
protein-coding sequence are selected against particularly strongly, because they lead to truncation of the protein, which is very likely to disrupt its function.

\[
\begin{align*}
\text{DNA} & \quad \text{GAA} & \quad \text{CGG} & \quad \text{CTC} & \quad \text{GAC} & \quad \text{TCC} & \quad \text{CGA} & \quad \text{AGC} & \quad \text{AGG} \\
\text{protein} & \quad \text{E} & \quad \text{R} & \quad \text{L} & \quad \text{D} & \quad \text{S} & \quad \text{R} & \quad \text{S} & \quad \text{R}
\end{align*}
\]

aspartic acid  

serine  

\[
\begin{align*}
\text{DNA} & \quad \text{GAA} & \quad \text{CGG} & \quad \text{CTG} & \quad \text{GAC} & \quad \text{GCC} & \quad \text{TGA} & \quad \text{TGC} & \quad \text{AGG} \\
\text{protein} & \quad \text{E} & \quad \text{R} & \quad \text{L} & \quad \text{D} & \quad \text{A}
\end{align*}
\]

alanine  

\text{(a)}  

\text{(b)}  

\text{(c)}  

Figure 6. Possible effects of point nucleotide mutations in protein-coding sequence. (a) \textit{synonymous mutations} lead to no change in the protein sequence because the new codon corresponds to the same amino acid as the old codon. (b) \textit{missense mutations} cause a different amino acid to be incorporated at the corresponding position. Some amino acid substitutions are more favorable than others due to biochemical similarities and differences between the different amino acids. (c) \textit{nonsense mutations} change a sense codon into a stop codon, truncating the protein sequence.

4.2.1 Measuring Codon Substitution Biases  

The biases in codon substitutions observed in alignments of protein-coding regions can be formalized in the following way. Consider a certain codon \( i \) in the target genome, where \( i \) ranges over the 64 possible codons. Given codon \( i \) in the target, there is some probability of observing any codon \( j \) aligned to \( i \) in an informant sequence. In particular, in conserved protein-coding sequence, it is most likely that \( i = j \). But in case \( i \neq j \), codons \( j \) that correspond to synonymous substitutions and favorable amino acid substitutions for \( i \) will be more likely to appear than other missense or nonsense substitutions. Now, we can place these probabilities for each possible pair of target and informant codons in a 64x64 matrix \( M \), where \( M_{ij} \) is the probability of a substitution of codon \( i \) with codon \( j \) over the evolutionary time separating the two sequences, \textit{conditional} on \( i \) and \( i \neq j \). I refer to \( M \) as a \textit{codon substitution matrix} (CSM).
Under a random model of nucleotide sequence evolution, the probabilities of substitution within the CSM would distribute in a uniform manner, structured only by edit distance (e.g. it takes two nucleotide substitutions to go from CGC to CAG, but only one to go from CGC to CGA; so under a random model of nucleotide evolution, the former will take place less frequently than the latter). In protein-coding sequence, however, the CSM is structured at multiple levels: first, preferring synonymous codon substitutions over missense substitutions; second, preferring biochemically favorable amino acid substitutions; and strongly avoiding nonsense substitutions.

Because the actual probabilities of a codon substitution being preserved by evolution depend on varying selective pressures, rates of mutation, amino acid biochemistry, and other subtle effects, it would be difficult to derive an accurate CSM from a mathematical model. Instead, we can estimate the values in the CSM by observing the frequency of codon substitutions in actual alignments of protein-coding sequence, based on trusted annotations of well-known genes. For example, in order to estimate the CSM for human and mouse genes, I examined sequence alignments of thousands of human genes with their mouse orthologs (equivalents). For each human codon \( i \) aligned to a different mouse codon \( j \), I incremented the corresponding entry in a matrix of counts. Finally, I normalized each row of this matrix to obtain an estimate of the CSM. A CSM can be computed in this way for any desired pair of genomes.

I also used the same approach to compute CSMs for non-coding sequence, for the purposes of comparison. Figure 7 shows visualizations of coding and non-coding CSMs measured from sequence alignments of human and mouse genomes, compared to a random model of nucleotide evolution.
Figure 7. Codon substitution biases versus a random model. Each colored square represents the probability of observing the indicated mouse codon substituted in place of the indicated human codon in a sequence alignment of the two species, as compared to a random model. Blue (dark) squares indicate that the corresponding codon substitution occurs more frequently than expected under the random model, while yellow (light) squares indicate that codon substitution occurs less frequently than expected. (Left) In coding regions, note the bias for synonymous substitutions and substitutions preserving amino acid biochemistry, and the strong bias against nonsense substitutions. (Right) In contrast, these probabilities measured over noncoding regions do not show the same intricate structure.

Since the biases captured by the CSM reflect specific selective pressures on an amino acid sequence, any sequence alignment that exhibits them systematically is likely to be protein-coding. I next propose a specific scoring scheme with which to measure the strength of these biases in any given alignment.

Measuring codon substitution biases in pairwise alignments. I first consider the simple case of evaluating codon substitution biases in an alignment of orthologous sequence from two species. Specifically, I seek to assign a score to any interval in the alignment, commensurate with how strongly it exhibits evolutionary signatures of codon substitution consistent with the CSM for the two species under analysis.
Let $M_{ij}$ be the probability of observing codon $j$ in the informant sequence aligned to target codon $i$ in protein-coding regions, conditioned on $i$ and $i \neq j$, as previously defined. Let also $N_{ij}$ be the same probability measured within non-coding regions. We can assign to any observed codon substitution the likelihood ratio $M_{ij} / N_{ij}$. This quantity measures how much more likely that codon substitution is to occur in a protein-coding region than in a non-coding region. For example, synonymous substitutions are more likely to occur in protein-coding regions than in noncoding regions; conversely, nonsense substitutions are much less likely to occur in coding regions.

For any given interval in the alignment, we can assign a score to each observed codon substitution within, and multiply them together to obtain a composite likelihood ratio as a score for the interval (in practice, it is necessary to add the log-likelihood ratios rather than multiply the likelihood ratios, due to numerical precision issues).

$$M_{ij} = P(\text{informant codon } j \mid \text{target codon } i, i \neq j)$$
$$N_{ij} = \text{same, computed in non-coding regions}$$

**Figure 8. Computation of the Codon Substitution Matrix (CSM) score in a pairwise alignment.** Each observed codon substitution is assigned a likelihood ratio measuring how much more likely it is to occur in a protein-coding region versus a non-coding region, based on the probabilities in the corresponding CSMs. The likelihood ratios are multiplied for each observed codon substitution in the desired interval of the alignment to obtain a composite score for that interval (in practice, the log-likelihood ratios are added).

An important detail in computing this score is selecting the correct reading frame in which to examine the codons; in particular, whether to consider the nucleotides at positions 0, 3, 6, 9, ... as the first nucleotide in each codon, or 1, 4, 7, 10, ..., or 2, 5, 8, 11, .... In practice, I simply compute the score over all three possible reading frames, and take the maximum of these three scores as the score for that interval. In real protein-coding sequence, the correct reading
frame almost always has a much higher score than the other two, due to the fact that most possible synonymous substitutions occur at the third codon position, and thus the expected pattern of nucleotide substitutions has a clear, frame-specific periodicity.

The CSM and CSM score as defined here notably ignore codons that are perfectly conserved between the two sequences, that is, where no substitution has occurred. On average, codons are certainly perfectly conserved more often in coding regions than in non-coding regions. However, examining only the substitutions provides much more specific information about the evolutionary pressures that have operated to preserve the amino acid sequence. For example, consider some biologically functional sequence that is evolutionary conserved, but not protein-coding. This sequence would tend to exhibit higher-than-normal sequence identity, and thus would appear to contain more perfectly conserved codons. But, where sequence divergence has occurred, it is unlikely that it would exhibit the same patterns of substitution that we expect in protein-coding sequence, because these patterns reflect evolutionary pressure to preserve an amino acid sequence. Hence, in examining only codon substitutions, we seek to identify signatures of natural selection that specifically apply to protein-coding sequence, and would not be exhibited by conserved noncoding sequence. I consider this point in more detail in section 4.3.

**Measuring codon substitution biases in multiple alignments.** I now turn to extend the CSM score to obtain a single score for an interval in a multiple sequence alignment, that is, an alignment of more than two sequences. In this case there is one target sequence and several informant sequences. A simple approach would be to compute the score separately for each pair of target and informant sequences, as described in the last section, and then combine these several scores together by multiplying the likelihood ratios (or adding the log-likelihood ratios). However, this effectively assumes that the observed codon substitutions in each species are independent of the other species, which is often not the case. For example, consider an alignment of human, mouse, and rat sequences. Because mouse and rat are so closely related, shared codon substitutions in these two genomes relative to the human are quite likely to have arisen from a single evolutionary event before mouse and rat diverged, as depicted in Figure 9. In this event, computing the CSM scores separately between each pair of species would "double count" the evidence.
Several observed codon substitutions may be due to a single evolutionary event. The diagram shows the phylogenetic tree of human (H), mouse (M), and rat (R). In this scenario, an ancestral species had the codon CCC. A single substitution to CTC occurs on the rodent lineage before mouse and rat diverge. Using the human sequence as the target, we observe codon substitutions in both mouse and rat informants, but these actually arise from a single evolutionary event. Given only the observed codons, it cannot be ruled out that two separate substitution events occurred in both mouse and rat since they diverged, but this is much less likely than the depicted single event.

There are formal probabilistic approaches\textsuperscript{20,23,24} to accounting for all such possibilities given the extant residues (nucleotides, codons, or amino acids observed at the leaves of the phylogenetic tree). Here, however, I choose to adopt an \textit{ad hoc} approach to combining the evidence from multiple species. Consider a multiple alignment with a target sequence and $n$ informant sequences. As shown in Figure 10, I compute a log-likelihood ratio for each individual codon substitution between the target and any informant species, as in the pairwise case. Each codon “column” of the alignment has between zero and $n$ scores, depending on how many informant sequences display a codon substitution there. If there is more than one score in a column, I combine them using a certain function, $f$, to obtain a composite score for that column. I then add together the score of each column to obtain a score for the entire interval. As before, I compute this score over all three possible reading frames and then take the maximal score.
<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>CTGTTTTTCTTTTGTAGGAAGTCAC</td>
</tr>
<tr>
<td>Dog</td>
<td>CTGTTTTTCTTTTGTAGTAAGTCAC</td>
</tr>
<tr>
<td>Mouse</td>
<td>CTGTTATTTCTTTTGTAGAAAGTCAC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human/dog</td>
<td>0 0 0 e2 0 0 e4 0 0</td>
</tr>
<tr>
<td>Human/mouse</td>
<td>0 e1 0 e3 0 0 e5 0 0</td>
</tr>
</tbody>
</table>

\[ e1 + f(e2, e3) + f(e4, e5) \]

Figure 10. Computation of the CSM score in a multiple alignment. Each observed codon substitution between the target species (human) and an informant species is assigned a certain score or “evidence”, exactly the likelihood ratio shown in Figure 8 based on the appropriate CSMs. When multiple codon substitutions are observed in a certain column of the alignment, several pieces of evidence are available in that column. These scores in each column are combined using a function \( f \) and the results added over each column in the interval with at least one piece of evidence. I use the median as \( f \) but there are other reasonable choices, such as the mean, minimum, or something more complex. This approach has several evident limitations. For example, it often under-counts the available evidence: if several different codon substitutions are observed in a certain column, then only one of them will actually contribute to the final score. Also, the final score loses its meaningful interpretation as a likelihood ratio. However, this approach also has several practical advantages. First, it is much simpler than rigorous probabilistic phylogenetic models, which require intricate training and inference algorithms. Second, it scales to an arbitrary number of informant sequences with very little cost: to add any additional informant species, we need only train the pairwise coding and non-coding CSMs between the target and new informant. The training algorithms for probabilistic phylogenetic models may have difficulty scaling to large numbers of informant sequences. Third, it has reduced sensitivity to certain classes of alignment errors, where essentially spurious sequence is aligned in one or more informant species. A probabilistic model that examines all available evidence may excessively penalize an interval with such a spurious alignment. The ad hoc approach proposed here is relatively insensitive to such outliers, which may occur frequently with current methods for genome alignment.
4.2.2 Codon Substitution Biases in Known Genes

In this section, I compute CSM scores for test sets of known exons and noncoding regions in a several different species, in order to demonstrate that this score can indeed distinguish protein-coding sequence from noncoding sequence.

Figure 11. CSM scores in a region of the human genome. CSM scores were computed over a well-studied region of human chromosome 7 surrounding the cystic fibrosis transmembrane conductance regulator (CFTR) gene, containing 120 known exons. (a) CSM scores computed in a pairwise fashion using mouse and dog as informant sequences. The red points represent the known exons and the blue points represent other, presumably noncoding intervals. (b) The pairwise human/mouse CSM score vs. interval length (in nucleotides) over the same intervals. Longer intervals tend to provide more observed codon substitutions, leading to higher scores.

Figure 11 shows CSM scores computed in a pairwise fashion over a small region of the human genome, using mouse and dog sequences as informants. It is evident that, even when computed in a pairwise fashion against individual informant sequences, the CSM score is a powerful discriminator between coding and noncoding regions. It is also clear, however, that the scores computed pairwise against different species strongly correlate with each other. The use of the two scores together does not seem to provide much more information than either individually – in particular, observe that the human/dog score seems to provide very little additional discriminatory power over the human/mouse score. I next consider the question of whether the multiple-species CSM score provides increased discriminatory power by simultaneously
considering evidence from several informant sequences, in a larger dataset from the fly genome. As more informants are added to the analysis, the multiple-species CSM score observes more codon substitutions, and thus accumulates more evidence that should increase its discriminatory power.

Figure 12. Distribution of CSM scores with various species combinations. Multiple-species CSM scores were computed over a test set comprising approximately 5% of the fruit fly genome. (a) Each plot shows the distribution of the CSM score computed using a different combination of informant sequences, for known exons (yellow/light, $n = 2,464$) and other, presumably noncoding open reading frames (blue/dark, $n = 10,178$). (See chapter 6 for details about these fly genomes.) The length distribution and fraction of noncoding intervals accurately reflect the composition of ORFs in the genome. (b) Cumulative distributions of CSM scores with two (dashed) and twelve (solid) flies. As additional genomes are added to the analysis, the discriminative accuracy of the CSM score at any given cutoff increases.

Figure 12 shows CSM scores computed over the same test set of fly exons and noncoding intervals used to benchmark the RFC score, using one, two, and eleven informant genomes. To understand the importance of these distributions, consider classifying the intervals simply by whether they have a positive or negative CSM score in the cumulative distribution in Figure 12b. Almost all noncoding intervals have a negative CSM score, while the vast majority of known
exons have a positive CSM score. As additional informant genomes are added, the fraction of noncoding intervals with a positive CSM score decreases significantly. With one informant (*D. pseudoobscura*, dashed lines), 9.0% (916 of 10,178) of noncoding intervals and 91.9% (2,264 out of 2,464) of coding intervals have positive CSM scores. With eleven informants (solid lines), 3.7% (377 out of 10,178) of noncoding intervals and 91.4% (2,252 out of 2,464) of coding intervals have positive CSM scores. Thus, adding ten additional informant genomes decreased the false positive rate of this simple classification by nearly 60%, while decreasing the sensitivity (perhaps due to alignment errors) by only 0.5%. And again, because the existing annotations are imperfect, these figures may somewhat underestimate the true discriminatory power of the CSM score.

### 4.3 Summary

In this chapter, I presented two different and largely independent signatures of protein-encoding sequence evolution, and gave evidence, based on a test set of coding and non-coding intervals spanning approximately 5% of the fly genome, that they are powerful discriminators of protein-encoding sequence from noncoding sequence, and that they are able to leverage the evidence provided by several informant genomes.

The basic motivation for using comparative genomics to identify functional elements in a genome is that functional sequence tends to be conserved by evolution. An important theme of the measurements I presented in this chapter, however, is that they are primarily concerned with sequence divergence, not conservation *per se*. The RFC test inspects whether nucleotide insertions and deletions have occurred consistent with the constraints of the codon reading frame. Similarly, the CSM score inspects whether codon substitutions have occurred consistent with the constraints of preserving the properties of the amino acid sequence. In perfectly conserved sequence, these scores have essentially no discriminatory power. I propose that this is consistent with both our own intuition in identifying conserved protein-encoding sequence (see Figure 13), and the underlying biology. The selective pressures on protein-encoding sequence generally do not require absolute conservation of the genomic coding sequence. The observation of perfect conservation over lengths comparable to protein-encoding sequence has in fact been considered a surprising indicator of some other, unknown mechanism. It also appears that there is more well-conserved noncoding sequence than coding sequence in mammalian genomes, and
distinguishing the two is a significant challenge for comparative gene finders. By observing signatures of sequence divergence rather than conservation per se, I seek to identify evidence that is specifically indicative of protein-coding sequence evolution.

Figure 13. Sequence conservation vs. signatures of divergence. Consider a sequence that is extremely well-conserved across four mammalian genomes, as shown in (a). There would be little doubt, based on its conservation, that this sequence is somehow biologically important; but there is no specific evidence to indicate that it is protein-coding. Given the apparent abundance of conserved noncoding sequence in mammalian genomes, we cannot draw any specific conclusions about what purpose this sequence serves. In contrast, the sequence alignment shown in (b) demonstrates signatures of sequence divergence that are characteristic of protein-coding sequence: indels of length multiple of three, reflective of the evolutionary pressure to conserve the codon reading frame, and codon substitution patterns heavily biased towards synonymous substitutions (most of which occur at the third codon position, resulting in the observable periodicity in nucleotide substitutions). In this case, we can conclude by inspection, with high confidence, that the sequence shown in (b) is protein-coding. Hence, the signatures of divergence in (b) are, in an important sense, much more informative than the conservation observed in (a).
5 A Methodology for Comparative Gene Identification

In this chapter, I develop an integrated approach for de novo prediction of protein-coding sequence (exons) in alignments of several related genomes, based primarily on the measurements of protein-coding sequence signatures proposed in the last chapter. This system can be used both to systematically evaluate existing gene annotations and to predict novel protein-coding regions.

5.1 A Classification Approach

In the previous chapter, I proposed several quantitative measurements that proved to be powerful discriminators of protein-coding sequence from noncoding sequence. The distributions of the RFC and CSM scores shown in the last chapter suggest that these measurements can be used to classify any given interval in a sequence alignment as coding or non-coding. For example, consider the cumulative distributions of the CSM scores computed over twelve fly genomes shown in Figure 12b. By choosing a simple cutoff on the CSM score, we could classify the intervals in this fly test set with quite good accuracy: a cutoff of zero, for example, would lead to a sensitivity of 91% (2,252 out of 2,464 of positive examples classified correctly) and a specificity of 96% (9,751 out of 10,178 negative examples classified correctly).

However, this initially promising observation ignores a number of important underlying considerations. The examples in the test set contain disjoint genomic intervals whose boundaries were selected based on known annotations; to evaluate a genome without the benefit of some pre-existing segmentation, we must have a strategy for choosing which intervals to evaluate and classify. In the fly test set, noncoding intervals outnumber coding intervals by a factor of four, and the imbalance will be even greater in the human genome, only 1.5% of which is protein-coding. The length of the intervals we choose to examine will also influence the performance of our approach, since the score distributions depend on the interval length (see Figure 5 and Figure 11b).

Finally, we should be able to make better decisions about whether an interval is coding based on several different measurements simultaneously, not just a simple cutoff on one score. The RFC score and the CSM score are based on largely independent lines of evidence, and together they should provide increased discriminatory power than either individually. We might
also wish to incorporate measurements of nucleotide periodicities and other sequence properties that have proven useful in single-genome gene finders. Hence, we are led to undertake a full classification approach, within which we can leverage well-known algorithms such as support vector machines to decide whether an interval is coding based on many measurements or features.

5.1.1 Identifying Candidate Intervals

The goal of our system is to predict protein-coding exons given multiple sequence alignments of related genomes, including exons that are full open reading frames (beginning with the start codon and ending with the stop codon), initial exons (beginning with the start codon and ending with a donor site), terminal exons (beginning with an acceptor site and ending with the stop codon), or internal exons (beginning and ending with splice sites). The first step is to identify candidate intervals in the target genome to evaluate. A simple approach would be to divide up the target genome into windows at regular intervals such as 100nt, and evaluate each window. However, there are several problems with this approach. If a true coding sequence happens to be smaller than the window size, then it might be undetectable when measurements are taken across the entire window. The human genome in particular has many very short exons, and diluting their signal would make them even harder to discover. It is also generally desirable to predict the exact boundaries of each exon, since this information is necessary to construct full gene structure models.

![Diagram of gene structure with acceptor and donor sites](image)

Figure 14. Identifying candidate coding intervals by enumeration of potential exon boundaries. First, I identify all start codons, stop codons, acceptor sites, and donor sites. Then, a candidate exon is any interval beginning with a start codon or acceptor site and ending with a stop codon or donor site. An interval may extend past a candidate donor site but not an in-frame stop codon. As shown above, this can lead to many overlapping intervals.
To identify candidate exons, I take the approach of exhaustively enumerating all intervals in the target genome that are flanked by the canonical exon boundaries, namely start codons, stop codons, and splice sites. The start codons and stop codons are easy to identify since they are simply nucleotide triplets. Splice site sequences are more complex, but they have certain known sequence patterns. I use an advanced local sequence model\textsuperscript{29} that can assign a score to every location in the genomic sequence indicating how strongly the local sequence appears to be a splice site. I use a configurable cutoff on this score to control which locations are considered splice sites for the purpose of enumerating candidate exons. Having identified all start codons, stop codons, and splice sites, it is then simple to exhaustively enumerate all possible candidate exons in the target genome. In principle, the precisely correct exons should appear somewhere in this list.

In noncoding sequence, start codons, stop codons, and sequences that resemble splice sites occur frequently by chance, so this enumeration captures many noncoding intervals in addition to the desired coding regions. As shown in Figure 14, this definition can also lead to many overlapping candidate exons, depending on where different exon boundaries appear in the genomic sequence. Also, the enumeration must be performed on both strands, and in different reading frames, since the location of start and stop codons depends on the reading frame. The total number of candidate exons is kept under control for two reasons. First, under normal circumstances, protein translation cannot continue past a stop codon, so no candidate exons can span an in-frame stop codon. (Because my method for detecting candidate donor splice sites is by design overinclusive, I allow enumeration to continue past putative donor sites, up to the next in-frame stop codon.) Second, I apply the simple heuristic that each candidate exon must be \textit{aligned} in at least one informant sequence, defined as having at least 50\% of the target nucleotides aligned to informant nucleotides, regardless of sequence identity. These two constraints prevent the total number of candidate exons from combinatorially exploding. In practice, I typically find several thousand candidate exons per million nucleotides of target sequence, and this number grows approximately linearly with the length of the target sequence.

5.1.2 Evaluating Candidate Exons

The algorithm described in the last section identifies all candidate exons in the target genome that are aligned to at least one informant sequence. I then use genome sequence
alignments to compute RFC and CSM scores for each candidate exon, as described in chapter 4. I additionally compute several other heuristic features that I found to empirically improve prediction accuracy. These are described in detail below, and graphically summarized in Figure 15.

![Figure 15. Measurements taken on each candidate exon.](image)

1. RFC, CSM
2. If the exon has a splice site boundary, the splice site, enhancer, and silencer motifs are scored according to a single-genome evaluator model (see text; distances not to scale).
3. The conservation of the exon boundaries (splice sites or start or stop codons) is evaluated.
4. Small windows inside and outside of the candidate exon are compared for differential sequence conservation (not to scale).

**Splice site, enhancer, and silencer motifs.** The RFC and CSM scores are effective at identifying stretches of coding sequence in the target genome. Many candidate exons may, however, overlap such a stretch, depending on the distribution of candidate splice sites and terminal codons nearby. In order to both improve candidate exon evaluation and to later resolve overlapping predictions, I developed an advanced splice site evaluator similar to that used in ExonScan\(^{30}\) that augments a sophisticated local sequence model\(^{29}\) with information about distant sequence motifs, including intronic splicing enhancers, exonic splicing enhancers, and exonic splicing silencers. While the precise nature and mechanism of these motifs is not yet well understood, this additional information has been shown to substantially improve splice prediction accuracy\(^{30}\). I use this model to obtain a score for each putative splice site in a candidate exon.

**Splice site conservation.** Because the exon boundaries (start codons, stop codons, and splice sites) are important for correct splicing and translation, they tend to be very well conserved in the genomic sequence. I compute simple measures of the conservation of exon
boundaries. For splice sites, I report the number of informant species in which the boundary aligns and scores above a certain threshold. (A substitution in the crucial “AG” of a 3’ splice site, for example, would cause the score to fall below this threshold). For start and stop codons, I report the number of informant species with start or stop codons aligned at the same position.

Conservation boundaries. I observed that conserved exons very often appear as “islands” of conservation in a “sea” of degraded noncoding sequence. I compute a measure of the relative strength of sequence conservation within a candidate exon compared to small windows a short distance upstream and downstream of the candidate exon, computed as the ratio of the percentage nucleotide identity within the candidate exon to the percentage nucleotide identity within the upstream/downstream windows. This metric may help distinguish real coding exons from candidate exons that arise when terminal codons or possible splice sites happen to occur within lengthy stretches of conserved non-coding sequence.

Additional variables. Finally, I included simple variables such as the length of each candidate exon and its overall nucleotide conservation rate.

5.1.3 Classification with a Support Vector Machine

Having taken a number of measurements to evaluate each candidate interval, we must somehow combine them to decide whether each candidate interval is coding or not. In principle, we could use any of the many available classification algorithms for this task, such as neural nets, random forests, “k nearest neighbors”, or our own custom cutoffs. A relevant characteristic of the classification problem is that the input is highly unbalanced: especially in the human genome, we identify, evaluate, and must correctly classify far more spurious candidate exons than real, coding exons. I chose to use a support vector machine (SVM) for classification, because their effectiveness in unbalanced classification tasks has been demonstrated in text document retrieval problems31.

The SVM takes as input each candidate exon and its associated measurements as features, and outputs for each a numerical score, positive to indicate that the candidate exon is real and negative to indicate it is not. The SVM makes this decision based on a previously trained model. In order to have good predictive accuracy, the model be derived from good training data. For each target species and particular set of genome alignments I analyze, I selected a set of reliably annotated, well-conserved genes. I computed the measurements described in the last section over
all the exons in these genes, as well as a set of spurious candidate exons chosen to accurately exemplify the distribution of candidate exon lengths in each genome. I then used these examples as training data for a different SVM model for each species. I found that the SVM typically requires training examples from only a few dozen genes, constituting several hundred exons and several thousand noncoding regions, in order to derive effective models.

5.1.4 Postprocessing

The SVM classifier often emits several overlapping exon predictions for each actual coding exon, when various combinations of potential splice sites or terminal codons led to several candidate exons containing well-conserved coding sequence according to the RFC and CSM scores. The SVM produces a single composite score for each candidate exon (positive for a prediction, negative for a rejection). In case of overlapping predictions, I simply take the highest scoring of the group. A future extension of this system might preserve different overlapping possibilities and later try to apply gene structure constraints in order to resolve the ambiguity. I will explore this possibility in discussion.

5.2 Implementation

I implemented the approach outlined above as a collection of libraries, programs, and scripts written in Objective Caml. The system contains a total of approximately 10,000 lines of code. For the support vector machine, I used SVM-Light\textsuperscript{32}, an off-the-shelf package that includes programs for training SVM models and classifying new examples.

From start to finish, it takes about 30 minutes to analyze the yeast genome on a desktop PC and 3 hours to analyze the fly genome. However, like many computational biology applications\textsuperscript{33}, this analysis is “embarrassingly parallelizable” in that it can be divided up and distributed to run in parallel, with little or no coordination between jobs required. To analyze the human genome, which is approximately 20 times larger than the fly genome, I divide up the genome alignments into many individual segments, and distribute the evaluation and classification stages on each part to a large cluster of machines. This analysis completes in well under one hour.
5.3 Summary

In this chapter, I proposed an integrated approach for using multiple sequence alignments of several related genomes in order to identify conserved protein-coding exons. I framed this as a classification problem, and used a well-studied algorithm, the support vector machine, to classify candidate exons as real or not based on measurements of evolutionary conservation and other features. A key advantage of the classification approach is that it has the flexibility to incorporate many different types of discriminatory signals (features), including those that may be very difficult to encode in the framework of an HMM or phylo-HMM.
6 Results: Comparative Gene Identification in Several Species

In this chapter, I apply my methodology for comparative gene identification to the genomes of a variety of different species in order to benchmark its performance against existing gene annotations. Because the existing annotation sets are generally far from perfect, my methodology leads to the prediction of many new protein-coding regions that are likely to be real, and suggests many corrections to existing gene annotations. For each species, I will give a brief introduction, describe the basic properties of its gene and genome structure, identify the informant species used for comparative analysis, and finally present the results of novel gene prediction and annotation evaluation.

6.1 Baker’s Yeast, *Saccharomyces cerevisiae*

Baker’s yeast, *Saccharomyces cerevisiae*, is a unicellular fungal species that has been used since ancient times to ferment alcoholic beverages and make bread rise. It is also an important model organism in eukaryotic biology, and extensive experimental techniques have been developed to manipulate it in the laboratory. The yeast genome was sequenced between 1989 and 1996, and has been extensively studied since then. It is the most thoroughly understood of all eukaryotic genomes. Hence, it is a valuable platform on which to validate my strategy.

**Genomic landscape.** The yeast genome is sixteen chromosomes containing approximately 12 million nucleotides and around 5,700 genes. A small fraction (~5%) of yeast genes have introns, but most are a contiguous open reading frame in the genome. For this reason, I configured my system to ignore splice sites during candidate exon enumeration. The median gene length is 1,169nt (sd = 1,118). Approximately 70% of the yeast genome is protein-coding.

**Informant species.** I use alignments with three other *Saccharomyces* species of yeast, *S. paradoxus*, *S. mikita*, and *S. bayanus*, for comparative analysis with *S. cerevisiae*. These three genomes were sequenced several years ago for the purpose of comparative analysis, which has
already led to major revisions in the yeast gene catalog\textsuperscript{18}. Their most recent common ancestor with \textit{S. cerevisiae} diverged about 20 million years ago.

**Genome alignments.** I generated nucleotide-level alignments of the four genomes using a genome-wide synteny map generated by Matthew Rasmussen and the alignment program MLAGAN\textsuperscript{35}. These four yeasts have excellent conserved synteny, and the genome alignments cover 97\% of yeast genes.

**Prediction performance vs. existing annotations.** I benchmarked the gene prediction performance of my system against a version of the “official” annotations from the \textit{Saccharomyces} Genome Database\textsuperscript{36} recently refined by a comparative analysis\textsuperscript{18}. This annotation set is the cumulative result of nearly a decade of yeast genome research, already improved and refined by comparative genomics methods. Hence, it is likely to be highly reliable, and certainly the most accurate annotations available for any eukaryotic genome.

<table>
<thead>
<tr>
<th>Gene sensitivity</th>
<th>94.9% (5,408/5,701)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide sensitivity</td>
<td>94.9% (8,002,566/8,434,197)</td>
</tr>
<tr>
<td>Nucleotide precision</td>
<td>98.9% (8,002,566/8,088,648)</td>
</tr>
<tr>
<td>Exon sensitivity</td>
<td>91.4% (5,355/5,860)</td>
</tr>
<tr>
<td>Exon precision</td>
<td>98.5% (5,362/5,443)</td>
</tr>
</tbody>
</table>


Table 1 shows that my classification approach demonstrates excellent performance in the yeast genome, detecting 95\% of annotated yeast genes (versus 97\% covered by the alignments). The classifier predicted 81 exons that are not in the annotations. Upon manual inspection, it was found that 38 are known to be real and very recently added to the database, 16 appear to be real, novel genes, and 11 appear to extend existing genes (M. Kellis, personal communication). Hence, of these 81 predictions outside of the annotations, most in fact appear to be real protein-coding sequence, and only 16 appear to be spurious false positives.
6.2 A Fungal Pathogen, *Candida albicans*

*Candida albicans* is a cousin of baker’s yeast that is present in all human bodies, and can be pathogenic when it overgrows. Candidiasis, more commonly known as yeast infection, affects a majority of humans at some point in their life. Although candidiasis is usually a mere nuisance to healthy people, it can lead to serious complications in immunodeficient individuals such as those suffering from AIDS or undergoing chemotherapy. Because some close relatives of *C. albicans* are not pathogenic, it is hoped that comparative analysis of these genomes will eventually lead to an understanding of the genetic basis of pathogenicity, and perhaps lead to novel treatment vectors for candidiasis.

**Genomic landscape.** Because *C. albicans* is related to *S. cerevisiae*, the structure of its genes and genome are quite similar. The genome contains a total of about 14 million nucleotides, and the gene annotations contain 6,089 genes with a median gene length of 1,232nt (sd = 1,074). Like yeast, *Candida* has very few introns, and over 70% of the genome is protein-coding. However, the gene annotations for *C. albicans* have not been subjected to nearly as much scrutiny as yeast’s, so they are probably considerably less reliable.

**Informant species.** I use three additional *Candida* species, *C. dubliniensis*, *C. tropicalis*, and *C. guilliermondii*, as informants for *C. albicans*. *C. dubliniensis* has been sequenced by the Wellcome Trust Sanger Institute, and *C. tropicalis* and *C. guilliermondii* have been sequenced by the Broad Institute.

**Genome alignments.** As with *S. cerevisiae*, I used a synteny map provided by Matthew Rasmussen and MLAGAN to generate sequence alignments. However, these *Candida* genomes are considerably more difficult to align than the *Saccharomyces* genomes, owing in part to lower-quality genome assemblies. As a result, the alignment coverage is considerably lower: whereas the most distant yeast informant (*S. bayanus*) aligns to 94% of the *S. cerevisiae* genome, the closest *Candida* informant (*C. dubliniensis*) aligns to only 82% of the *C. albicans* genome (M. Rasmussen, personal communication).
Prediction performance vs. existing annotations.

<table>
<thead>
<tr>
<th>Gene sensitivity</th>
<th>75.1% (4,575/6,089)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide sensitivity</td>
<td>71.9% (6,437,228/8,955,466)</td>
</tr>
<tr>
<td>Nucleotide precision</td>
<td>99.5% (6,437,228/6,471,480)</td>
</tr>
<tr>
<td>Exon sensitivity</td>
<td>71.7% (4,653/6,490)</td>
</tr>
<tr>
<td>Exon precision</td>
<td>98.2% (4,737/4,819)</td>
</tr>
</tbody>
</table>


Table 2 shows that the classifier detects a considerably smaller fraction of the *C. albicans* genes than the yeast genes, while precision remains excellent. This is partially explained by the lower alignment coverage; genes must be aligned to at least one informant sequence in order for the classifier to evaluate them. Additionally, however, some of the existing gene annotations may, in fact, be spurious.

**Novel gene predictions.** The classifier predicts 82 novel protein-coding regions in the *C. albicans* genome. Again, upon manual inspection of the alignments, most of these appear to be real genes missing from the annotations.

**Dubious genes.** By applying the RFC and CSM measurements to all of the *C. albicans* gene annotations, I identified a set of 873 genes that are in fact aligned to orthologous sequence, but do not appear to have been conserved as protein-coding sequence by either metric. I designate these as “dubious” genes, since this evidence suggests that these genes may not be real. It cannot be ruled out that some may be fast-evolving, species-specific, or incorrectly aligned; therefore, this set will require detailed scrutiny of the other evidence supporting each annotation. Nonetheless, it is likely that, as with a previous analysis in yeast\(^\text{18}\), this comparative genomics evidence will prove to be a powerful indicator that most of these annotations should be removed.
6.3 The Fruit Fly, *Drosophila melanogaster*

Like baker's yeast, the fruit fly *Drosophila melanogaster* is a well-studied model organism in experimental biology, and especially genetics. As a multicellular organism with many distinct tissues and bodily systems, it is much more complex than yeast, and has more to say about the biology of higher organisms such as the human. Likewise, its genome contains mostly multi-exon genes and many other characteristics shared with vertebrate genomes, such as our own. The *D. melanogaster* genome has been available since 2000\(^38\).

**Genomic landscape.** The fly genome is about 150 million nucleotides. Most fly genes have several introns, although they are not as long or numerous as in the human. The existing fly gene annotations, FlyBase v4.2, contains slightly fewer than 14,000 genes. The median number of exons in each gene is 3 (sd = 3.5), and the median exon length is 277 nucleotides (sd = 617). Approximately 25% of the fly genome is protein-coding.

**Informant species.** I use eleven other fly genomes that have recently been sequenced for comparative analysis with *D. melanogaster*: *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*. The most distant of these, *D. grimshawi*, diverged from *D. melanogaster* approximately 40 million years ago.

**Genome alignments.** I use sequence alignments generated by the Berkeley Comparative Genomics group, using the programs Mercator and MAVID\(^39\).
Prediction performance vs. existing annotations.

<table>
<thead>
<tr>
<th>Gene sensitivity</th>
<th>87.5% (12,238/13,985)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide sensitivity</td>
<td>78.9% (17,589,254/22,299,766)</td>
</tr>
<tr>
<td>Nucleotide precision</td>
<td>96.5% (17,589,254/18,226,082)</td>
</tr>
<tr>
<td>Exon sensitivity</td>
<td>77.9% (42,630/54,707)</td>
</tr>
<tr>
<td>Exon precision</td>
<td>95.1% (44,952/47,257)</td>
</tr>
</tbody>
</table>


Table 3 shows that the classifier continues to exhibit good sensitivity and excellent precision in the fly genome. Unfortunately, although the alignments I used are publicly available, information about their quality and coverage has not been reported in detail, so it is difficult to estimate what fraction of the missed genes may be due to alignment problems.

**Novel gene prediction.** The classifier predicts 2,305 exons that are not present in existing gene annotations. As in yeast and *Candida*, I suspect that most of these are in fact protein-coding exons missing from the annotations. I partitioned these predictions into 737 that occur within known genes, and are candidate novel alternatively spliced exons. The remaining 1,568 predictions may comprise entirely novel genes. In collaboration with FlyBase, which maintains the *D. melanogaster* gene annotations, these datasets are being carefully inspected by professional annotators, and many have already been incorporated into upcoming revisions of the official annotations. We are also pursuing collaborations to obtain experimental evidence to support these predictions.

**Dubious genes.** I collected a set of 579 genes, each of which is aligned in several informant sequences, but *none* of its exons appear to exhibit signatures of protein-coding sequence evolution. Many of these are single-exon annotations that could in fact be processed pseudogenes. As with *Candida*, however, they will require close scrutiny on a case-by-case basis, since it cannot be ruled out, based on the comparative evidence alone, that some of these
are real genes that are fast-evolving or incorrectly aligned. Additionally, although I have only focused thusfar on entire genes that appear spurious, my analysis of the evolutionary signatures could be used to identify individual dubious exons in otherwise well-conserved genes.

6.4 Human

The human genome is by far the most complex of those studied in this thesis. Notably, however, the total number of human genes (~22,000) is not much greater than the fruit fly (14,000) and many other simpler organisms. There is still, however, considerable uncertainty as to the total human gene count; the latest authoritative estimate is 20,000-25,000\textsuperscript{40}, but an ongoing comparative analysis suggests the actual count may be slightly under 20,000\textsuperscript{27}.

**Genomic landscape.** The human genome is about 3 billion bp. Most human genes have numerous introns, which are frequently much lengthier than the exons. The current human genome annotation database (RefSeq) contains 22,218 genes. The median number of exons in each gene is 7 (sd = 9.5), and the median exon length is 122 nucleotides (sd = 233). Only about 1.5% of the human genome is protein-coding.

**Informant species.** I use the mouse, rat, and dog genomes as informants for the human. These mammalian genomes have been sequenced in the last several years, in part for the purpose of comparative analysis with the human\textsuperscript{26,27,41}.

**Genome alignments.** I use mammalian sequence alignments generated by the UCSC Genome Bioinformatics group using the MULTIZ program\textsuperscript{42}. These alignments are generated using a substantially different strategy than the sets of alignments I used for the other species. Whereas the alignments I generated for the fungi, and the MAVID/Mercator alignments I used for flies, are based on genomewide gene synteny maps, MULTIZ alignments are generated based on identifying, for each region of the target genome, the most similar region from anywhere in the informant genome. It is beyond the scope of this thesis to thoroughly investigate how these two genome alignment strategies compare, but anecdotally, the MULTIZ approach seems to lead to more artifacts and spurious sequence alignments, and this may affect the performance of any comparative analysis.
**ENCODE regions.** For this analysis, I will present two sets of benchmarks: one computed across the whole genome, and one computed within the ENCODE regions, a selection of intervals totaling about 1% of the human genome that have been selected by an international consortium for thorough analysis of functional elements\(^4^3\). The existing gene annotations in some of these regions have been subjected to close manual scrutiny, and are likely to be more reliable overall than in the remainder of the genome.

**Prediction performance vs. existing annotations.**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene sensitivity</strong></td>
<td>87.5% (329/376)</td>
</tr>
<tr>
<td><strong>Nucleotide sensitivity</strong></td>
<td>76.0% (510,087/670,840)</td>
</tr>
<tr>
<td><strong>Nucleotide precision</strong></td>
<td>90.1% (510,087/566,091)</td>
</tr>
<tr>
<td><strong>Exon sensitivity</strong></td>
<td>72.9% (2,833/3,886)</td>
</tr>
<tr>
<td><strong>Exon precision</strong></td>
<td>91.9% (2,845/3,095)</td>
</tr>
</tbody>
</table>


Table 4 shows that the classifier performs well in the ENCODE regions, but clearly not as well as in fly or fungi. This partially reflects the harder fundamental nature of the problem, in identifying very short exons within a very large fraction of noncoding sequence, and perhaps also lower alignment quality. In fact, these performance statistics are comparable to other comparative human gene prediction systems.

Table 5 shows that the performance of the classifier appears to be significantly worse across the whole genome than within the ENCODE regions. This analysis was performed using exactly the same classifier model, and the alignments I used of the ENCODE regions are simply subsets of the whole-genome alignments. Hence, it seems plausible that the apparent worse performance in the whole genome, compared to the ENCODE regions, could be largely attributed to lower quality gene annotations, on average, across the whole genome.

**Novel gene prediction.** My analysis predicted 19,403 exons outside of existing gene annotations. From these, I constructed a set of 445 “genes” by clustering nearby predictions in intergenic regions. 90 of these were subjected to experimental verification by Michael Brent at Washington University. These experiments consist of performing RT-PCR to determine whether predicted exons are transcribed in the cell. Of the 90 predictions tested, 50 were confirmed. Hence, in a small sample of my novel predictions, a majority appear to be real genes.

**Dubious genes.** I identified a set of 1,065 human genes that align to orthologous sequence but do not show the characteristic signatures of protein-coding sequence evolution in any exon. As before, these annotations merit careful scrutiny by annotators to determine which are plausibly fast-evolving or specific-specific genes, such as olfactory receptors and genes related to immune response, and which may in fact be spurious annotations. The total number of questionable human annotations I have identified is comparable to the number found by another comparative analysis of the mammalian genomes (M. Clamp, personal communication).
6.5 Summary

I demonstrated that my classification approach to comparative gene identification is effective across a wide range of species, from the unicellular yeast up to the human. This system can contribute to improving the gene annotations for all of these genomes, by predicting unknown protein-coding sequence with high precision, and by identifying existing annotations that require close scrutiny due to negative evidence from related genomes. In fact, these contributions have already begun in fly and human, through our collaborations with annotators and experimentalists.

A major lesson learned is how strongly the performance of this comparative approach depends on the quality of the genome alignments. For example, while the general structure of the yeast and Candida genomes are very similar, lower alignment coverage in Candida partially led to much lower gene prediction sensitivity. As state-of-the-art methods for genome alignment improve, the effectiveness of any comparative analysis will immediately benefit. Another lesson learned is the difficulty of meaningfully benchmarking a gene predictor against imperfect existing annotations. The sensitivity and precision figures presented in this chapter are indicative of general trends, but they can hardly be interpreted absolutely: in each species, a majority of the “false positive” predictions in fact appear to be real protein-coding sequence missing from the annotations. Likewise, many of the “missed genes” may in fact be spurious annotations. Of course, if the gene annotations were perfect, then this project would be in some sense pointless; it is through the careful inspection of these cases that this analysis will lead to improved annotations.
7 Conclusion

In this thesis, I identified distinctive signatures of protein-coding sequence evolution, and proposed quantitative measurements to detect these signatures. A key feature of these metrics is that they specifically seek characteristic signatures of sequence divergence, not conservation per se, which distinguish protein-coding sequence from conserved noncoding sequence. I gave evidence that these measurements indeed distinguish coding from non-coding sequence in multiple sequence alignments of related genomes, and that their discriminatory power increases when more informant genomes are used. Finally, I used these measurements as a basis for building a de novo exon predictor based on a classification approach, and applied this system to the genomes of four different species: baker’s yeast, Saccharomyces cerevisiae; the fungal pathogen Candida albicans; the fruit fly Drosophila melanogaster; and the human. A major advantage of a classification approach is that it can also be used to systematically evaluate existing gene annotations, and thus to identify where they may require correction. The predictions of novel protein-coding regions that this system makes, and the sets of genes that it calls into doubt, can be used to improve the gene annotations of these species. We have already begun to do so for the fly and human in collaboration with other researchers, and we will continue to expand these efforts towards a comprehensive revisiting of these gene sets based on the evidence from comparative genomics.

There are many future directions for this work. A virtue of the classification approach is that it can naturally accommodate a variety of different measurements and features. My system as it stands does not make any use of the variety of single-genome sequence models, developed over several decades of research in computational gene identification, that distinguish protein-coding sequence. Additionally, ongoing research into the sequence motifs that govern transcription, splicing, and post-transcriptional gene regulation may lead to an entirely new class of metrics that can be used to inform gene prediction. The incorporation of all these additional metrics could lead to improved performance, although I expect that the signatures of evolution will remain the most informative signals at the level of individual exons.

A significant limitation of my approach, particularly in the human genome, is that it predicts only individual exons. Most human genes have multiple exons, and the identification of gene structures is important both for improving de novo prediction accuracy, since gene structures
impose certain constraints on adjacent exons that can resolve ambiguities, and for driving experimental validations of novel gene predictions, since these experimental designs usually rely on gene structures. There are plausible natural extensions of my system for recovering gene structures; for example, one could envision a dynamic programming algorithm that computes an optimal chaining of exon predictions into gene structures, making use of the various alternative possibilities for the exact boundaries of each exon. It would also be possible to abandon the classification approach entirely, and use the measurements I developed within a flexible probabilistic framework such as a conditional random field\cite{conditional_random_field}.

In general, however, no \textit{de novo} method has thusfar achieved striking success in correctly predicting complete gene structures in the human genome. Even state-of-the-art techniques produce correct results in only a minority of cases\cite{state_of_the_art}, and do not satisfactorily capture important and common phenomena such as alternative splicing and genes appearing within introns of other genes. Hence, I am wary of undertaking any such approach now in lieu of some major advance in the field -- which could be forthcoming. In particular, perhaps much more so than individual exon prediction, gene structure prediction stands to benefit profoundly from ongoing research to uncover the sequence motifs governing transcription, splicing, and post-transcriptional gene regulation, for these are the signals by which the cell itself recognizes its genes.

The ideal solution for gene identification would take a genome sequence or sequences and produce a complete, trustworthy set of gene structures, including likely alternative splice forms. In the human genome, all existing \textit{de novo} methods fall far short of this ideal. Nonetheless, the methods I developed in this thesis are very well suited the practical goal of using the evolutionary evidence from comparative genomics to improve existing gene annotations. In concert with high-throughput experimental data such as large-scale transcript sequencing\cite{large_scale_transcript_seq} and genome tiling microarrays\cite{genome_tiling_microarrays}, it seems plausible within a few years, we will know all the protein-coding genes in the human genome. Ultimately, the identification of all the biological "parts" in the genomes of the human and important model organisms, of which protein-coding genes are only one category, promises to lead to profound advances in fundamental biology, and even to drive the development of revolutionary new treatments for human disease.
8 Appendix

This section contains supplemental information that may be useful for the discriminating reader’s reference.

\[ S. \textit{cerevisiae} \]
\[ \text{median} = 1169 \text{ sd} = 1118 \]

\[ C. \textit{albicans} \]
\[ \text{median} = 1232 \text{ sd} = 1074 \]

\[ D. \textit{melanogaster} \]
\[ \text{median} = 277 \text{ sd} = 617 \]

\[ \text{Human} \]
\[ \text{median} = 122 \text{ sd} = 233 \]

Figure 16. Exon length distributions in the species studied in this thesis, based on existing annotations (SGD, CGD, FlyBase, and Ensembl). Note differing axis scales.
Figure 17. Dependence of the RFC score on interval length. RFC scores were computed as described in section 4.1.1 over intervals covering approximately 5% of the fruit fly (Drosophila melanogaster) genome. The bars represent the total number of distinct intervals with the corresponding RFC score on the x-axis. The yellow (light) bars represent known protein-coding exons from a gene annotation database (total 2,464) while the blue (dark) bars represent other open reading frames (total 10,178). The scores were computed using eleven informant genomes (see chapter 6 for details about these fly genomes) and segregated according to the length of the interval. The fraction and length distribution of the putatively noncoding intervals accurately reflects the composition of ORFs in the genome. Note differing y-axis scales. The RFC score has less discriminatory power in very short intervals because, by chance, no indels may occur in short noncoding intervals.
Figure 18. Dependence of the CSM score on interval length. Multiple-species CSM scores were computed over a test set comprising approximately 5% of the fruit fly genome, using eleven informant genomes. Each plot shows the distribution of the CSM score for known exons (yellow/light) and other, presumably noncoding open reading frames (blue/dark), for intervals 0-90nt in length (283 exons/4192 noncoding intervals), 90-150nt (468/2019), 150-300nt (809/2873), and 300+nt (904/1094). The length distribution and fraction of noncoding intervals accurately reflect the composition of ORFs in the genome. The CSM score has more discriminatory power with longer intervals because more codon substitutions tend to be observed.
9 References


