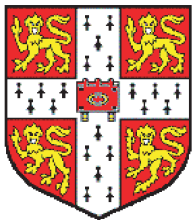
Quantifying evolution and natural selection in vertebrate noncoding sequence

Michael Milner Hoffman

Trinity College



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

EMBL–European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
England

Email: hoffman@ebi.ac.uk

3 June 2008

To my parents

This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgments.

This thesis does not exceed the specified length limit of 300 pages as defined by the Biology Degree Committee.

This thesis was typeset in 12 pt Palatino using LaTeX2ε and Sweave (Leisch 2002) according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Quantifying evolution and natural selection in vertebrate noncoding sequence

Summary

Michael Milner Hoffman
Trinity College

3 June 2008

When studying genomic evolution, biologists find it important to identify varying patterns of natural selection. Many traditional methods of classifying directional selection have relied on models that categorize mutations as function-altering or neutral, and then comparing the rates of the two categories of mutations. The most well-known methods specifically compare nonsynonymous and synonymous substitutions in protein-coding sequence. The recent availability of whole genome sequences, especially those of various mammals and other vertebrates, enables us to develop alternative methods for analyzing molecular evolution and selection that rely on noncoding sequence. Furthermore, our greater understanding of the importance of noncoding DNA demands such methods.

This thesis contains the results of the first in-depth genomic-scale analysis using intron substitutions to estimate the neutral rate of evolution. Performing this analysis across several genomes requires the development of a new model of gene evolution and related methods. I find strong correlation between estimates of the neutral rate made with intron methods and estimates made with synonymous coding nucleotides for both human–dog and mouse–rat comparisons. However, the two estimates cannot be considered directly equivalent.

This thesis also describes a novel method that estimates a rate of function-affecting evolution in promoter regions by inspecting the effect of simulated mutations on transcription factor binding. This involves the development and use of a probabilistic method that uses a hidden Markov model to predict the binding of transcription factors. I report the results of applying these new methods to the human genome for the identification of transcription factor binding sites, and for the identification of natural selection.

Preface

This thesis describes work carried out at the European Bioinformatics Institute in Cambridge, England, between October 2003 and March 2008. The EBI is an outstation of the European Molecular Biology Laboratory.

Many people have helped me to reach this point. I would like to thank the members of my thesis advisory committee, Nick Goldman, Simon Tavaré, and Lars Steinmetz, who have provided guidance and useful pointers along the way. I thank my colleague Alison Meynert, the first other user of Sunflower, who not only found bugs but also fixed a few.

I would like to thank everyone in the Ensembl project at the EBI and the Wellcome Trust Sanger Institute. In particular, Andy Jenkinson and Eugene Kulesha set up a DAS server for Sunflower. Others provided assistance and answered many questions, in addition to producing a web site, database, and code that were used extensively in this work.

I am grateful to long-patient system administrators Tim Cutts, Guy Coates, and Nicolas Rodriguez. Trinity College provided a home and a community. I wish to thank my tutor Jean Khalfa for his help.

The development communities of Python and R, and a number of other open source projects, provided tools essential for this work. This thesis began with a LaTeX template graciously provided by Anton Enright, which was passed to him through a series of other research students.

A number of friends and colleagues reviewed portions of this document. I wish to thank Alison Meynert, Elma Brenner, Richard Hoffman, Daniel Zerbino, Jing Su, Matt Hestand, Martin Hammond, Markus Fritz, Stephanie Bates, Ben Paten, Emile Chabal, and Ed Minor for useful discussions and comments.

I would like to thank the Marshall Aid Commemoration Commission for financial support. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this document are those of the author and do not necessarily reflect the views of the National Science Foundation.

Most of all, I thank my supervisor Ewan Birney. Without his support, advice, and inspiration, none of this would have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, I explore the use of some novel techniques to understand the evolution of noncoding sequence of vertebrates, primarily of mammals. My work involved creating two new measurements, one to study neutral evolution using introns, and the other to investigate the evolution of transcription factor binding sites, primarily in promoters. In establishing these measurements, I developed models and tools that have other uses, and I discuss some of these as well.

This chapter provides a review of some of the existing methods for examining molecular evolution, and covers some essential topics in bioinformatics used in the rest of the thesis. The rest of the chapters each have a brief introduction with additional background material, followed by an explanation of methods used in the chapter, results, and discussion. In chapter 2, I discuss the use of introns to estimate neutral rates of evolution. In chapter 3, I describe a probabilistic model to detect transcription factor binding sites. In chapter 4, I describe the use of this model to identify genomic regions where changes in sequence will affect transcription factor binding the most. Finally, in chapter 5, I outline a new distance measurement that summarizes the rate of phenotype-affecting promoter substitutions, and its use in studying the evolution of promoters.

1.1 Evolutionary distance measurements[1]

1.1.1 Directional selection and neutral evolution

Evolutionary biologists have long had an interest in determining which regions of the genome are under selective pressure, and if so, what kind of selective pressure affects them (Nei and Kumar 2000; Yang and Bielawski 2000; Eyre-Walker 2006). Directional selection occurs when natural selection drives one allele in a population to fixation. The direction can be positive or negative (Graur and Li 2000). Positive selection, also known as adaptive evolution, occurs when a newly-derived allele carries a phenotype sufficiently adaptive that it is preferentially driven towards fixation. Negative selection, also called purifying selection, occurs when newly-derived alleles are sufficiently maladaptive or deleterious that selection drives them towards elimination. Loci not under any kind of selection are said to evolve neutrally, and can therefore be used to estimate the rates of the mutational process driving evolution while reducing the confounding factors of selection. The heterogeneity of mutational processes across the genome (Lercher and Hurst 2002; Ellegren and co-workers 2003; Webber and Ponting 2005) necessitates the estimation of a local neutral rate, which is used as a null model when testing for directional selection (Nielsen 2001).

About 5 percent of the human genome undergoes purifying selection, including almost all protein-coding genes (Waterston and co-workers 2002; Ponting and Lunter 2006). Purifying selection indicates selective pressure against mutations, as most potential mutations are deleterious in these regions. One finds purifying selection by observing that the rate of *substitutions*, or fixed mutations, in functional positions lies below the expected substitution rate in local neutral positions. The rate of neutral substitution is equal to the rate of neutral mutation (Gillespie 2004). I discuss the mechanics of these observations further in subsection 1.1.2.

Simple positive and purifying selection are not the only classes of selection found in nature. Genes may also be subject to selective forces such as disruptive selection, balancing selection, and overdominant selection. These other selective forces are more difficult to study at the genomic level, partly because they

[1] Parts of this section were previously published (Hoffman and Birney 2007).

require information on heterozygosity or polymorphism less widely available than simple reference genomes for various species. I do not consider these other kinds of selection further in this thesis.

For protein-coding genes, many researchers determine selective pressure by comparing observed substitution frequencies in nonsynonymous and synonymous coding nucleotides, as described in subsection 1.1.2. Determining where purifying selection or positive selection influences gene evolution has proven useful in several areas of genome research (Hurst 2002). In mammals, researchers have found many genes with regions under positive selection, some of which are reviewed by Yang and Bielawski (2000) and by Sabeti and co-workers (2006). Many of these genes represent biological functions known to undergo adaptive evolution. These functions can be characterized by their role in competition and co-evolution (Nielsen and co-workers 2007). They include some of the defining biology that sets mammals apart from other taxonomic classes, as well as functions that set inframammalian clades and even individual species apart from each other. Knowing which human genes could have been influenced by positive selection helps us understand the evolution of humans into their present form, and helps us to identify genomic regions of functional and medical importance (Nielsen 2001).

Genes involved in immune function have experienced the most positive selection in humans and other well-characterized mammals (Nielsen and co-workers 2007). They provide an excellent example of the sort of evolutionary arms race that can drive positive selection, as they must counter the improving fitness of pathogens by natural selection. These immune genes include genes such as β-globin (*HBB*), defensin, immunoglobulin V_H, and genes in the major histocompatibility complex. Another category of genes subject to positive selection consists of reproductive genes such as *Sry*, the androgen-binding protein subunits (*Abpa*, *Abpb*, *Abpg*; Karn and Nachman 1999; Karn and Laukaitis 2003), and protamine 1 (*PRM1*). Olfactory genes also undergo positive selection (Emes and co-workers 2004). In humans, forkhead box protein P2 (*FOXP2*), involved in speech and language, has also evolved through positive selection (Enard and co-workers 2002; Krause and co-workers 2007). Other examples of genes under positive selection exist, but most human and mammalian coding genes undergo purifying selection (Yang and Bielawski 2000), and purifying selection affects perhaps as many as 75% of amino acid substitutions (Eyre-Walker and

co-workers 2006; Kryukov and co-workers 2007).

1.1.2 Protein-coding neutral and functional rates

The most sophisticated models of selective pressure (Yang and Nielsen 2000), use the evolutionary distances d_N (nonsynonymous changes) and d_S (synonymous changes), and call their ratio $\omega = {}^{d_N}/_{d_S}$. Some researchers (Hurst 2002) use the notations K_A, K_S, and ${}^{K_A}/_{K_S}$, and perhaps different models, but they all rely on the same underlying set of assumptions. These models usually posit that d_S, the number of synonymous substitutions per potentially synonymous site, represents neutral changes unaffected by selective pressure. In contrast, the models posit that d_N, the number of nonsynonymous substitutions per potentially nonsynonymous site, represents functional differences affected by selective pressure.

The simplest way to identify synonymous coding positions is to focus on *fourfold degenerate* (4D) sites, which can change freely without any associated change in amino acid sequence, and occur only at the third position of some codons. For example, `GCT` codes for alanine, and changes to its third position result in `GCA`, `GCC`, or `GCG`, which also code for alanine. Therefore the third position of `GCT` is a 4D site.

A distance estimate based wholly on 4D sites is sometimes called d_4. The model underlying this estimate ignores partial synonymity due to *twofold degenerate* (2D) and *threefold degenerate* (3D) sites, and does not consider synonymity in the first two codon positions. Miyata and Yasunaga (1980) developed a more elaborate method, which included partial degeneracy by considering all possible pathways between any pair of codons. Later, Nei and Gojobori (1986) devised a similar but simpler method by removing parameters shown to be unnecessary in computer simulations. A brief description of this method based on that of Nei and Kumar (2000) follows.

I use the notation f_i to refer to the proportion of synonymous changes at position $i \in \{1, 2, 3\}$ of a codon. This is the number of potential changes that result in the same amino acid residue divided by the number of potential changes that result in any amino acid residue. Changes that result in a stop codon are ignored in the denominator because they are far less likely to survive in a functional protein-coding gene. For any codon, the quantity of potentially syn-

Table 1.1: Fraction of potentially synonymous changes s for each codon in the standard genetic code.

	_T_	_C_	_A_	_G_
T__	Phe TTT 1/3	Ser TCT 1	Tyr TAT 1	Cys TGT 1/2
	Phe TTC 1/3	Ser TCC 1	Tyr TAC 1	Cys TGC 1/2
	Leu TTA 2/3	Ser TCA 1	Stop TAA —	Stop TGA —
	Leu TTG 2/3	Ser TCG 1	Stop TAG —	Trp TGG 0
C__	Leu CTT 1	Pro CCT 1	His CAT 1/3	Arg CGT 1
	Leu CTC 1	Pro CCC 1	His CAC 1/3	Arg CGC 1
	Leu CTA 4/3	Pro CCA 1	Gln CAA 1/3	Arg CGA 3/2
	Leu CTG 4/3	Pro CCG 1	Gln CAG 1/3	Arg CGG 4/3
A__	Ile ATT 2/3	Thr ACT 1	Asn AAT 1/3	Ser AGT 1/3
	Ile ATC 2/3	Thr ACC 1	Asn AAC 1/3	Ser AGC 1/3
	Ile ATA 2/3	Thr ACA 1	Lys AAA 1/3	Arg AGA 5/6
	Met ATG 0	Thr ACG 1	Lys AAG 1/3	Arg AGG 2/3
G__	Val GTT 1	Ala GCT 1	Asp GAT 1/3	Gly GGT 1
	Val GTC 1	Ala GCC 1	Asp GAC 1/3	Gly GGC 1
	Val GTA 1	Ala GCA 1	Glu GAA 1/3	Gly GGA 1
	Val GTG 1	Ala GCG 1	Glu GAG 1/3	Gly GGG 1

onymous changes $s = \sum_{i=1}^{3} f_i$, and the quantity of potentially nonsynonymous changes $n = 3 - s$. So, for a codon such as ATT, which codes for isoleucine,

$$\begin{aligned} s &= f_1 + f_2 + f_3 \\ &= 0 + 0 + \frac{2}{3}, \end{aligned}$$

because no potential changes in the first two positions result in isoleucine, but a third position change to ATA or ATC does. A 4D codon degenerate only one position, such as ACG (threonine), would have $s = 1$, $n = 2$. Table 1.1 displays values of s for every codon.

Adding up s and n for each of C codons gives the total numbers of synonymous sites S and nonsynonymous sites N where $N = 3C - S$. The length of the sequence $L = 3C = S + N$.

Having calculated quantities of synonymous and nonsynonymous sites, we must now quantify how these sites change in a pairwise comparison with

another sequence. Using the amino acid translations of the two sequences, we obtain a higher-quality alignment than would be possible with nucleic acid alignment. Then we project this alignment back to the nucleic acid sequence so that nucleotide positions relative to codon boundaries are conserved. Then we can compare the aligned sequences of codons.

We call the number of actual nonsynonymous differences in an aligned codon n_d and the number of actual synonymous differences s_d. If there is only one nucleotide difference between two aligned codons, we can simply assign n_d and s_d based on whether this results in an amino acid change or not. So for a synonymous change such as CAA $\leftrightarrow$ CAG (both glutamine), then $s_\mathrm{d} = 1$ and $n_\mathrm{d} = 0$. If the change, instead, had been CAA (Gln) $\nleftrightarrow$ AAG (Lys), then $n_\mathrm{d} = 1$ and $s_\mathrm{d} = 0$. I use the straight arrow $\leftrightarrow$ to indicate synonymous substitutions and the squiggle arrow $\nleftrightarrow$ to indicate nonsynonymous substitutions in the rest of this subsection.

If the number of overall differences l_d between the two codons is greater than 1, then we identify the pathways of length l_d that do not include a stop codon (it is unlikely that a nonsense mutation could survive in functional sequence). We add up the numbers of nonsynonymous substitutions $\check{n}_\mathrm{d}$ and synonymous substitutions $\check{s}_\mathrm{d}$ for each pathway and calculate their means, n_d and s_d. For example, CCG (Pro) $\nleftrightarrow$ CGT (Lys) gives rise to two pathways:

$$\begin{aligned}
&\texttt{CCG}\ (\text{Pro}) \nleftrightarrow \texttt{CGG}\ (\text{Arg}) \nleftrightarrow \texttt{CGT}\ (\text{Lys}) \quad [\check{s}_\mathrm{d} = 0, \check{n}_\mathrm{d} = 2] \\
&\texttt{CCG}\ (\text{Pro}) \leftrightarrow \texttt{CCT}\ (\text{Pro}) \nleftrightarrow \texttt{CGT}\ (\text{Lys}) \quad [\check{s}_\mathrm{d} = 1, \check{n}_\mathrm{d} = 1]\,.
\end{aligned}$$

Averaging the values for these two pathways together, we get $s_\mathrm{d} = 0.5$ and $n_\mathrm{d} = 1.5$. In paired codons with three substitutions, there can be up to $3! = 6$ pathways. In these cases, the total number of differences in the codon l_d is still equal to $n_\mathrm{d} + s_\mathrm{d}$, just as the total number of differences in the sequence L_d is $N_\mathrm{d} + S_\mathrm{d}$.

Using these quantities we can estimate overall proportions of synonymous and nonsynonymous nucleotides that change, $p_S = {}^{S_\mathrm{d}}/_{S}$ and $p_N = {}^{N_\mathrm{d}}/_{N}$. One may also refer to these proportions as *p distances*. For very closely related sequences, one may use a p distance as an estimate of the number of substitutions per site d, a concept explained more fully in subsection 1.1.3. For more accurate results on more distantly-related sequences, one must use the models in that subsection.

Regardless of the method used, one can use p_N and p_S to estimate the number of substitutions per nonsynonymous site d_N, and the number of substitutions per synonymous site d_S. One may divide the two to get the nonsynonymous/synonymous rate ratio ω. This yields a quantity that summarizes selective pressure after correction for the local variation in neutral evolution. One must estimate this ratio locally since the character of neutral evolution varies in different genomic regions. For any given gene (or portion of a gene), if $\omega > 1$, the gene is under positive selection, while if $\omega < 1$, the gene is under purifying (or negative) selection. It is also possible for part of the sequence to be under positive selection, even if $\omega < 1$ for the whole sequence. Since $\omega < 1$ for almost all genes, sometimes comparisons of ω-ordered lists of genes are used, where the genes with the highest ω values are said to be the most likely to have been affected by positive selection (Waterston and co-workers 2002).

1.1.3 Distance measurement models

The p distances presented above estimate the proportion of nucleotides that have changed, but it would be more useful to know the number of changes that have occurred. While we could theoretically observe up to one substitution per aligned position, in reality many of those positions may have changed an arbitrary number of times, meaning that the total number of differences is unbounded. Thankfully, models exist to estimate difference distances, denoted d, from p distances. These models account for multiple substitutions in the same position over time.

The simplest evolutionary distance model, the *Jukes-Cantor model* (Jukes and Cantor 1964; Nei and Kumar 2000), assumes that there is a uniform rate of mutation α regardless of the start and end nucleotides. One can call the rate of mutation to any of the three other nucleotides $u = 3\alpha$. If we are comparing two aligned sequences, we can call p_t the proportion of differing nucleotides, and q_t the proportion of identical nucleotides at time t. The probability that a nucleotide identical in both sequences at time t does not change in either sequence by time $t + 1$ is

$$(1 - u)^2 = 1 - 2u - u^2 \approx 1 - 2u,$$

when one selects a time unit such that u is small enough that we can ignore the u^2 term. The probability that differing nucleotides at time t become identical at time $t+1$ is double the sum of the probability that one of the nucleotides stays the same $1-u$, and the probability that the other one changes to be identical to the first α

$$2\alpha(1-u) = \frac{2u(1-u)}{3} = \frac{2u-u^2}{3} \approx \frac{2u}{3}.$$

From this, we can calculate that the proportion of identical nucleotides at time $t+1$ is the sum of the probabilities of identical nucleotides staying identical and differing nucleotides becoming identical

$$q_{t+1} = (1-2u)q_t + \frac{2u}{3}p_t.$$

We can then find the change in identical nucleotides in one unit of time $q_{t+1} - q_t$:

$$\begin{aligned} q_{t+1} &= (1-2u)q_t + \frac{2u}{3}(1-q_t) \\ &= (1-6\alpha)q_t + 2\alpha(1-q_t) \\ &= q_t - 6\alpha q_t + 2\alpha - 2\alpha q_t \\ q_{t+1} - q_t &= 2\alpha - 8\alpha q_t. \end{aligned}$$

If we switch to a continuous time model, where $\frac{dq}{dt} = q_{t+1} - q_t$, we get

$$\frac{dq}{dt} = 2\alpha - 8\alpha q.$$

If we multiply both sides by the integrating factor $e^{8\alpha t}$, which differentiates with regard to t to $8\alpha e^{8\alpha t}$, we get

$$\begin{aligned} e^{8\alpha t}\frac{dq}{dt} &= 2\alpha e^{8\alpha t} - 8\alpha e^{8\alpha t} q \\ 8\alpha e^{8\alpha t} q + e^{8\alpha t}\frac{dq}{dt} &= 2\alpha e^{8\alpha t} \\ \frac{d}{dt}(e^{8\alpha t} q) &= 2\alpha e^{8\alpha t} \\ &= \frac{d}{dt}(\frac{1}{4}e^{8\alpha t}) \\ e^{8\alpha t} q &= \frac{1}{4}e^{8\alpha t} + C. \end{aligned}$$

In order for $q = 1$ at $t = 0$, the constant of integration C must be $\frac{3}{4}$. Using this, we can solve for q:

$$\begin{aligned} e^{8\alpha t} q &= \frac{1}{4}e^{8\alpha t} + \frac{3}{4} \\ q &= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}. \end{aligned}$$

Finally, we can solve for the expected number of nucleotide substitutions $d = 2ut = 6\alpha t$:

$$\begin{aligned} 1 - p &= \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d} \\ \frac{3}{4} - p &= \frac{3}{4}e^{-\frac{4}{3}d} \\ 1 - \frac{4}{3}p &= e^{-\frac{4}{3}d} \\ \ln(1 - \frac{4}{3}p) &= -\frac{4}{3}d \\ d &= -\frac{3}{4}\ln(1 - \frac{4}{3}p). \end{aligned}$$

This equation is used to estimate the Jukes-Cantor distance d from the proportion of changed nucleotides p.

More sophisticated models with more parameters also exist. *Kimura's two-parameter model* (Kimura 1980) represents a single step up in sophistication, using separate rates for *transitions* (more frequent purine-to-purine and pyrimidine-to-

pyrimidine mutations) and *transversions* (less frequent purine-to-pyrimidine and pyrimidine-to-purine mutations). The most complex single-nucleotide mutation model is the *unrestricted general 12-parameter model*, which has a separate rate for each possible substitution, in either direction. There are several other models spanning the gamut of sophistication between these two models, which Nei and Kumar (2000) and Felsenstein (2004) both review.

Simulations generated using the *Tamura-Nei model* (which has rate parameters for purine transitions, pyrimidine transitions, and transversions, plus a frequency parameter for each of the four nucleotides; Tamura and Nei 1993) reveal that for close evolutionary distances ($d \leq 0.25$), estimates of d produced by any of the models of evolutionary distance show only negligible differences from each other. The various models only start to show substantial differences at $d \geq 0.6$. Even at $d = 1.5$, the simpler Kimura and Jukes-Cantor models give estimates of within 0.1 of each other, so there is little advantage to be had in only a slight increase in complexity (Nei and Kumar 2000).

The simpler distance measurements have the advantages of fewer parameters to discover, and smaller variance. They are also considerably easier to implement and are computationally inexpensive. For these reasons, I use the Jukes-Cantor model throughout this thesis.

1.1.4 Noncoding neutral rates

Researchers have historically focused on identifying selection in protein-coding sequence using d_N and d_S. Both coding sequence in general, and the 4D sites used to estimate a neutral rate of evolution, are easier to identify than any noncoding functional or neutral sequence (Hurst 2002; Ponting and Lunter 2006). Still, noncoding sequence plays a critical role in many areas of biology, and there are almost certainly roles and classes of noncoding sequence that are not yet understood (ENCODE Project Consortium 2007). Therefore, it is important to understand the patterns of evolution and natural selection in noncoding sequence, despite the difficulties.

Most noncoding genomic sequence is believed to evolve neutrally (Kimura 1968; Hellmann and co-workers 2003), but comparative genomics reveals that a substantial portion is under constraint (International Chicken Genome Sequencing Consortium 2004; Woolfe and co-workers 2005). Lunter and co-workers

(2006) analyzed the human genome looking for sequence that was constrained against insertions and deletions (*indels*). Most of the indel-conserved sequence is noncoding, so they conclude that a majority of human functional sequence is noncoding (Ponting and Lunter 2006).

There are several classes of DNA widely considered to evolve neutrally, because they are the parts of the genome we are most certain are nonfunctional. Pseudogenes and ancient repetitive elements are readily identifiable, yet serve no apparent function. Also, 4D sites are usually considered functionless, excepting the influence of some confounding factors I discuss in section 2.1. In general, these classes appear to evolve with the least evolutionary constraint (Ponting and Lunter 2006). Also identified as neutrally evolving are interior intron sites (Castresana 2002; Ogurtsov and co-workers 2004; Hoffman and Birney 2007), discussed in detail in chapter 2.

While substitution rates in neutral sequence are constant in regions of less than 100 kbp, there is much variation in rates both between and within chromosomes (Ellegren and co-workers 2003). Ponting and Lunter (2006) suggest several causes of this, including CpG dinucleotide hypermutability, recombination, germ-line sequence repair, and base composition disequilibrium.

I discuss my approach to estimating the neutral rate of evolution in noncoding sequence using introns in chapter 2. That chapter also includes a comparison to some other methods for estimating local neutral rates in subsection 2.4.1.

1.1.5 Noncoding functional rates

In the past few years, we have begun to see the exciting prospect of being able to identify positive selection widely in noncoding sequence. Sabeti and co-workers (2006) discuss several genes where significant positive selection for noncoding mutations has been found, including a mutation in the promoter of the Duffy blood group chemokine receptor (*DARC*) gene that protects against malaria (Hamblin and Rienzo 2000), and a mutation in regulatory regions near the lactase (*LCT*) gene (Bersaglieri and co-workers 2004). These discoveries have provided crucial insights into human biology, and improved our understanding of human evolution.

There are many methods based on population genetics that one may use to detect selection in genomic sequence that do not require the regions of interest

to code for proteins. These include searching for reductions in genetic diversity, high-frequency derived alleles, allele frequency differences between populations, and long haplotypes (reviewed in Sabeti and co-workers 2006).

These methods have some limitations—they can only detect selection that occurred less than $\sim 4N_e$ generations ago, where N_e is the *effective population size* (Nielsen and co-workers 2007). Additionally, they make demographic assumptions, which may account for apparent deviations from a neutral model, without providing evidence against selective neutrality (Nielsen 2001). While d_N/d_S methods can detect selection in protein-coding genes before $4N_e$ generations ago and lack demographic assumptions, there has been less work in detecting ancient selection in noncoding sequence. I discuss a novel method to do this in chapter 4, which relies on predictions of transcription factor binding and the effects of mutations on the same.

1.2 Transcription factor binding

1.2.1 Biological overview

Transcription factors (TFs) are proteins that bind to genomic DNA near a gene, and promote transcription of a protein-coding gene by *RNA polymerase II* (Pol II) to *messenger RNA* (mRNA), or of a noncoding gene by other RNA polymerases to other forms of RNA. Some transcription factors form a part of the transcription initiation complex essential for RNA polymerase activity. Others encourage processivity during transcriptional elongation. Some transcription factors are not strictly necessary for transcription but instead accelerate or repress further transcription under certain cellular conditions (Brown 2006). Transcription factors are only one type of many factors that influence gene expression. Other factors include those involved in mRNA degradation, RNA binding and translation initiation, and protein degradation. When gene expression is discussed in this thesis, however, it is primarily with a focus on the effects of TFs, especially in the *promoter* regions adjacent to *transcription start sites* (TSSs).

Most transcription factors recognize specific sequences of DNA by physically penetrating into the major or minor grooves of a double helix. The positions where TFs bind are known as *transcription factor binding sites* (TFBSs), and the sequence is called the transcription factor's *motif*. The motif can vary somewhat

across the genome, so it is sometimes crudely represented as a consensus sequence of unambiguous and ambiguous DNA letter codes. For example, the consensus sequence `TATAWAAR` describes the TATA box, the sequence bound by the TATA binding protein (TBP) transcription factor (Brown 2006), an essential part of the pre-initiation complex of some genes (Sandelin and co-workers 2007).

A genetic code for transcription factors is the holy grail for bioinformaticians studying transcriptional regulation. Unfortunately, such a code has proven far more difficult to establish than the genetic code of protein synthesis. The translational genetic code is quite simple, taking 64 different codon inputs on nonoverlapping boundaries and usually producing the same 21 discrete outputs. A transcriptional code, however, would be complex and highly degenerate, producing continuous output with the input of not only discrete sequence, but also a complicated probabilistic and energetic dance of various factors, cofactors, and epigenetic effects. Nevertheless, this has not stopped researchers from trying to come to a more mechanistic understanding of the molecular interplay between proteins and nucleic acids (Hoffman and co-workers 2004), or the relationship between sequence and transcription levels (Benos and co-workers 2002; Frith and co-workers 2008).

Earlier laboratory studies have focused on high-confidence TFBSs using techniques such as DNAse I footprinting, dimethyl sulfate modification protection assays, gel retardation analysis, and reporter gene assays (Brown 2006; Sandelin and co-workers 2007). These techniques may ignore influences beyond the sequence and transcription factor. This leads to some ascertainment bias, as the techniques might miss finding weak binding sites. The results are also limited in their power to identify other TFBSs outside those areas of sequence being tested and significantly similar sequence. More recently, researchers have begun more systematic investigations of TF binding. ChIP-chip (Bulyk 2006) and ChIP-seq (Robertson and co-workers 2007) techniques reveal binding across the whole genome, and SELEX (Ellington and Szostak 1990; Tuerk and Gold 1990) reveals affinity for sequences not extant *in vivo*. While SELEX carries the problems of an *in vitro* technique, it is relatively unbiased and therefore more powerful, allowing the construction of models of affinity to any sequence (see subsection 1.2.2).

1.2.2 Discovery methods

Most methods for locating transcription factor binding sites rely on *position weight matrices* (PWMs), also called *position-specific scoring matrices* (PSSMs) (Berg and von Hippel 1987; Stormo 1990; Wasserman and Sandelin 2004). PWMs provide a way to characterize TF binding that is more sophisticated than the use of consensus sequences. To define a PWM for a TFBS, we use the alphabet of unambiguous DNA, $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, and therefore $|\mathcal{A}| = 4$. PWMs are $m \times |\mathcal{A}|$ matrices representing the probability that at position $k \in [1, m]$ in a TFBS, the TF binds to each of the nucleotides with the probabilities in column k of the matrix. PWMs can be visualized through a graphic representation known as a sequence logo (Schneider and Stephens 1990), which shows the relative importance and flexibility of various positions within the TFBS motif. For example, Figure 1.1 depicts the sequence logo for TBP from the JASPAR database (Vlieghe and co-workers 2006). This representation contains vastly more information than the consensus sequence `TATAWAAR`, showing, among other things, that the second thymine is more important than the first thymine. It even includes some low information positions at the beginning and end that would be inappropriate in a consensus sequence, because the sequence at these positions affects binding significantly less than at the core positions.

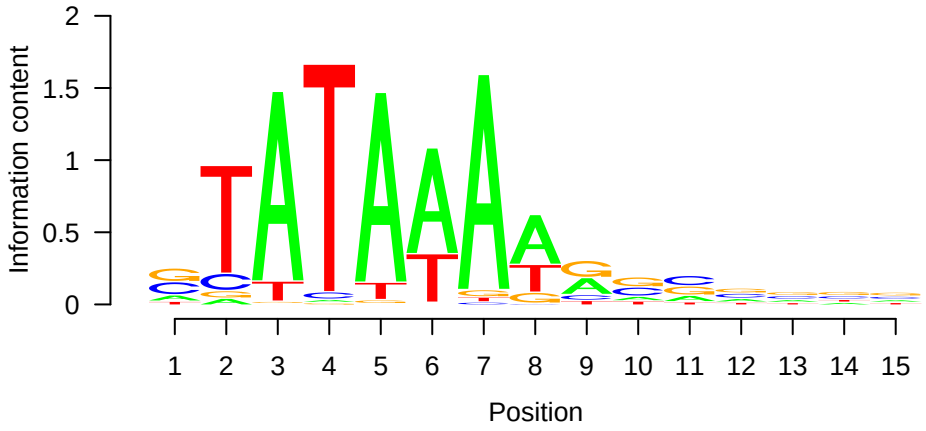


Figure 1.1: Sequence logo for TBP, generated by the seqLogo package of Bioconductor (Gentleman and co-workers 2004).

Position frequency matrices (PFMs) generated by summing observations of binding sequences can be converted to PWMs by transforming frequencies into probabilities. This is done by dividing the individual frequencies by the sum of all frequencies in their column, potentially with the addition of pseudocounts.

A number of bioinformatics tools use PWMs to identify putative TFBSs by examining how well a sequence matches the matrix (Rajewsky and co-workers 2002; Loots and co-workers 2002; Dermitzakis and co-workers 2003; Kel and co-workers 2003; Lenhard and co-workers 2003). Sometimes these are combined with genomic sequence conservation information in a method known as phylogenetic footprinting (Moses and co-workers 2004; Das and Dai 2007). A number of methods also exist to identify *cis-regulatory modules* (CRMs), made up of a number of TF binding sites, which may occur several times across the genome. I focus on individual TFBSs and do not discuss CRMs further.

In chapter 3, I discuss a probabilistic method I developed to locate TFBSs using a hidden Markov model (introduced in section 1.3), and data from known PWMs.

1.2.3 CpG dinucleotides and islands

In neutrally evolving sequence, one frequently assumes that the neutral mutation rates of single nucleotides are independent of each other. Most of the widely-used models of nucleotide evolution share this simplistic assumption. The following example shows that this is not always accurate.

In mammals, nuclear methyltransferases commonly methylate CpG into m^5CpG. Then DNA repair machinery frequently deaminates the 5-methylcytosine to thymine, resulting in TpG, or CpA if the methylation and transition occurred on the minus strand. The hypermutable m^5CpG dinucleotides mutate far in excess of the average mutation rates for single nucleotides. As a consequence, CpG is depleted in the genome with respect to what one would expect by squaring the G+C content fraction. Jabbari and Bernardi (2004) found that this observed/expected ratio was 0.23 in humans, and that the mean ratio of the occurrence of CpG to the occurrence of GpC in a sample of mammalian genomes was 0.27 ± 0.04.

The depletion of CpG, however, is heterogeneous across the genome. In certain regions, known as *CpG islands*, there is an excess of CpG dinucleotides over

the low levels usually seen. These are often associated with promoters, which I call *CpG island promoters*. All other promoters I call *CpG desert promoters*. CpG islands exist where transcription factors bind to a genomic region in the germline. Transcription factor activity and CpG methylation are inversely correlated, so active areas of transcription initiation do not suffer an increased propensity for deamination. Only when the transcription factors are active in the germline is the local increase in CpG frequency propagated to further generations. Because of this, one can use CpG islands to identify constitutively expressed genes. *Housekeeping genes* (genes which are expressed in most cells) frequently have unmethylated CpG islands, while tissue-specific genes are either not associated with CpG islands or may be associated with CpG islands methylated only in the tissues where the relevant gene is expressed (Alberts and co-workers 2002; Brown 2006).

1.3 Hidden Markov models

1.3.1 The generative model

Hidden Markov models (HMMs) provide a useful framework for modeling probabilistic processes. At their core, HMMs are directed graphs consisting of nodes, called *states*, that emit *symbols*, as a path traverses the graph. The models are called *hidden* because the path through the graph is not obvious from the sequence of symbols emitted. It is possible for a symbol to be emitted by more than one state, and for a state to emit more than one symbol (Durbin and co-workers 1998).

An HMM consists of the graph of states and associated parameters. The set of all states $\mathcal{K}$ is the union of a set of one or more emitting states $\mathcal{E}$ and a set of zero or more silent states $\mathcal{S}$. There can be no cycles made entirely of silent states, but other cycles are acceptable. The parameters consist of transition probabilities that determine how likely it is that a path continues from one state to another (or back to the same state), and emission probabilities that determine how likely it is that a particular state emits a particular symbol from an alphabet $\mathcal{A}$.

A simple representation of an entire HMM with $m = |\mathcal{K}|$ states consists of a transition probability matrix $\mathbf{A} = (a_{k \to l})_{m \times m}$ and an emission probability matrix $\mathbf{E} = (e_{k,x})_{m \times |\mathcal{A}|}$, which together contain all the parameters. These ma-

trices also define the topology of the graph since two unconnected states have transition probabilities of 0. Silent states have all of their emission probabilities set to 0.

Generating symbols is a simple matter. The GENERATE algorithm yields an infinite sequence of symbols:

```
GENERATE(A = (a_{k→l})_{m×m}, E = (e_{k,x})_{m×|A|})
1  k ← 0
2  repeat forever
3        if Σ_{a∈A} e_{k,a} ≠ 0
4           then yield WEIGHTED-RANDOM-CHOICE(e_k)
5        k ← WEIGHTED-RANDOM-CHOICE(a_k)
```

This algorithm can be stopped after generating an arbitrary length of sequence or when a special end state is reached.

1.3.2 Dynamic programming algorithms

One may make inferences about the path through an HMM that generated a sequence of symbols $\boldsymbol{X} = (x_1 \dots x_n)$ using various algorithms. Researchers frequently want to find the most likely single path, which is called the Viterbi path. Another possibility is to derive the posterior probability that each symbol in the sequence was generated by each state, leading to an $m \times n$ posterior probability matrix $\mathbf{P}$. The FORWARD and BACKWARD algorithms can do this over all possible paths simultaneously (Durbin and co-workers 1998).

The FORWARD algorithm returns the posterior probability $P(\boldsymbol{X})$ that a sequence $\boldsymbol{X}$ was generated by the model defined by $\mathbf{A}$ and $\mathbf{E}$. In order to produce that result, the algorithm calculates the forward probability $f_{l,i}$ that at each position i in the sequence, the model has generated the sequence $(x_1 \dots x_i)$ and the path π is in state l. This can be expressed as

$$f_{l,i} = P(x_1 \dots x_i, \pi_i = l).$$

We can obtain this by multiplying the emission probability e_{l,x_i} of the symbol x_i at the current position i by the summation of the forward probabilities of each

state k at the previous position $i-1$, multiplied by the transition probability $a_{k\to l}$ from k to l. That is,

$$f_{l,i} \sum_{k\in\mathcal{K}} f_{k,i-1} a_{k\to l}.$$

Since the forward probability for a state at one position incorporates information from all of the forward probabilities at the previous position, by induction we can see that the forward probabilities at the last position incorporate information from every possible path of states through the model. The complete FORWARD algorithm, with some complexity added to deal with potential runs of silent states between symbols, is:

FORWARD$(\mathbf{A} = (a_{k\to l})_{m\times m}, \mathbf{E} = (e_{k,x})_{m\times|\mathcal{A}|}, \boldsymbol{X} = (x_1 \ldots x_n))$

1 $f_{0,0} \leftarrow 1$
2 $f_{k,0} \leftarrow 0$ for all $k > 0$
3 **for** $i \leftarrow 1$ **to** n
4 **do for each** l **in** $\mathcal{E}$
5 **do** $f_{l,i} \leftarrow e_{l,x_i} \sum_{k\in\mathcal{K}} f_{k,i-1} a_{k\to l}$
6 **for each** l **in** $\mathcal{S}$
7 **do** $\mathcal{L} \leftarrow \mathcal{E} \cup \{k : k \in \mathcal{S} \text{ and } k < l\}$
8 $f_{l,i} \leftarrow \sum_{k\in\mathcal{L}} a_{k\to l} f_{k,i}$
9 $P(\boldsymbol{X}) \leftarrow \sum_{k\in\mathcal{K}} f_{k,n} a_{k\to 0}$
10 **return** $(P(\boldsymbol{X}), \mathbf{F} = (f_{k,x})_{m\times n})$

The BACKWARD algorithm performs a similar task to that of the FORWARD algorithm, but it relies on the information from the symbols beyond a position rather than the symbols preceding a position. It calculates the backward probability $b_{k,i}$ that at each position i in the sequence, the model will next generate the sequence $(x_{i+1} \ldots x_n)$ and the path π is in state k. This can be expressed as

$$b_{k,i} = P(x_{i+1} \ldots x_n, \pi_i = k),$$

and calculated by

$$b_{k,i} = \sum_{l\in\mathcal{K}} a_{k\to l} e_{l,x_{i+1}} b_{l,i+1}.$$

The complete BACKWARD algorithm is similar in construction to the FORWARD algorithm:

BACKWARD($\mathbf{A} = (a_{k\to l})_{m\times m}, \mathbf{E} = (e_{k,x})_{m\times|\mathcal{A}|}, \boldsymbol{X} = (x_1 \dots x_n)$)
1 $b_{k,n} \leftarrow a_{k\to 0}$ for all k
2 **for** $i \leftarrow (n-1)$ **downto** 1
3 **do for each** k **in** $\mathcal{E}$
4 **do** $b_{k,i} \leftarrow \sum_{l\in\mathcal{K}} a_{k\to l} e_{l,x_{i+1}} b_{l,i+1}$
5 **for each** k **in** $\mathcal{S}$
6 **do** $\mathcal{L} \leftarrow \mathcal{E} \cup \{l : l \in \mathcal{S} \text{ and } k < l\}$
7 $b_{k,i} \leftarrow \sum_{l\in\mathcal{L}} a_{k\to l} b_{l,i}$
8 $P(\boldsymbol{X}) \leftarrow \sum_{l\in\mathcal{K}} a_{0\to l} e_{l,x_1} b_{l,1}$
9 **return** $(P(\boldsymbol{X}), \mathbf{B} = (b_{k,x})_{m\times n})$

The overall posterior probability $P(\boldsymbol{X})$ generated from both algorithms is theoretically the same. In practice, it may be slightly different due to accumulated rounding error from floating point calculations.

The matrices of forward and backward probabilities produced by the FORWARD and BACKWARD algorithms give the marginal probability that a path through the model is in a particular state at a position only using the information from symbols to one side of that position. But using the matrices produced by these algorithms, one can calculate the posterior probabilities of each state at each position incorporating all of the known symbol information using the formula

$$P(\pi_i = k \mid \boldsymbol{X}) = \frac{f_{k,i} b_{k,i}}{P(\boldsymbol{X})}.$$

This process is called *posterior decoding*. Combining the FORWARD and BACKWARD algorithms with this posterior decoding step yields the FORWARD-BACKWARD algorithm:

FORWARD-BACKWARD($\mathbf{A}, \mathbf{E}, \boldsymbol{X}$)
1 $(P(\boldsymbol{X}), \mathbf{F} = (f_{k,x})_{m\times n}) \leftarrow$ FORWARD($\mathbf{A}, \mathbf{E}, \boldsymbol{X}$)
2 $(P(\boldsymbol{X}), \mathbf{B} = (b_{k,x})_{m\times n}) \leftarrow$ BACKWARD($\mathbf{A}, \mathbf{E}, \boldsymbol{X}$)
3 **for** $i \leftarrow 1$ **to** n

4 **do for each** k **in** $\mathcal{K}$

5 **do** $P(\pi_i = k \mid \boldsymbol{X}) = \dfrac{f_{k,i} b_{k,i}}{P(\boldsymbol{X})}$

6 **return** $\mathbf{P} = P(\pi_i = k \mid \boldsymbol{X})_{k=0,\ldots,m-1;\ i=1,\ldots,n}$

The returned posterior probability matrix **P** can be used for further results. I will determine the posterior probability that nucleotides in a sequence were emitted by a particular TF in chapter 3, and examine the impact of mutations on this probability in chapter 4.

Chapter 2

Intron measurements of neutral evolution[1]

2.1 Introduction

Using d_S as a measure of the neutral rate of evolution presents several problems. First, the underlying assumption that synonymous coding nucleotides can change freely is not correct in all cases. For example, synonymous sites may overlap with exonic splicing enhancers (Cartegni and co-workers 2002) or silencers (Chamary and co-workers 2006). Synonymous sites may also be under selective pressure due to their effects on mRNA stability, either through G+C content or the appearance or avoidance of certain sequence motifs (Chamary and co-workers 2006). In some animals, there is definitely selection in synonymous coding sites in the form of codon bias related to tRNA abundance (Chamary and co-workers 2006), but whether it affects synonymous coding sites in mammals is unclear (Bernardi 1995, 2000; Graur and Li 2000; Iida and Akashi 2000). There is also indirect evidence that selection affects synonymous sites, revealed by comparing base composition and evolutionary rates in synonymous sites with other potentially neutral types of sequence (Chamary and co-workers 2006). Second, the usually unstated assumption that synonymous coding sites undergo the same mutational processes as other sites in the genome ignores the fact that they occur next to a limited number of flanking nucleotides, and therefore are unequally affected by effects such as CpG hypervariability (Hardison and co-workers 2003). Finally, there are so few synonymous coding nucleotides for a

[1] Parts of this chapter were previously published (Hoffman and Birney 2007).

given gene that it can be difficult to estimate precisely the number of changes per nucleotide. This is especially true when comparing species such as human and chimpanzee, which are so closely related that there has not been enough time for many of these sites to change.

Primarily to overcome the issue of low numbers of synonymous coding sites, some researchers have estimated the neutral rate using either ancestral repeats near genes (Chimpanzee Sequencing and Analysis Consortium 2005) or substitutions in introns (Castresana 2002). Introns in homologous genes provide natural and well-defined boundaries for alignment regions, and the recent release of several high-quality vertebrate genome sequences allows for a robust comparison across several species. One cannot assume that intron changes are always neutral, but one cannot assume that synonymous coding nucleotide changes are always neutral either. Intron substitutions still provide a useful alternative for dealing with the small-sample error inherent in the synonymous coding nucleotide model.

While intron nucleotide substitutions have been used previously to estimate neutral rates, this work contains the first thorough comparison with a neutral rate estimate based on synonymous coding nucleotides. It also contains an investigation of the evolutionary distances at which using introns is sensible. This chapter examines properties of a neutral rate estimate based on intron substitutions compared with one based on synonymous coding sites by highlighting their similarities and differences, with particular regard to their use in estimates of selective pressure.

I call the estimated neutral rate from nucleotide substitution in introns d_I, defined as the number of intron substitutions per intron nucleotide. I wish to estimate d_I for a large portion of the genome, but the mechanics involved are not trivial, as one must first identify truly orthologous intronic sequences to compare. When studying tens of thousands of gene pairs, it is not practical to use hand-crafted alignments. Naive assumptions about the congruence of gene structures between species, however, lead to errors due to differences in gene annotations or changes in the biological transcript structure. One cannot, for example, assume that the first intron in a mouse gene is orthologous to the first intron in its orthologous rat gene. In addition, the existence of alternative splice forms in eukaryotic genes increases the challenge in determining which intron regions are homologous. I found the common approach of using the longest

translation of a gene as a stand-in for the whole gene inadequate as it ignores the selective pressure on introns that contain non-constitutive exons. To address this need, I created a model of gene evolution, which I call *Metascript*, which incorporates alternative splice forms and considers varying selective pressures on introns and coding sequence.

I chose to analyze data from mouse–rat and human–dog comparisons. Evolutionary biologists believe that the mouse and rat lineages diverged 16 Myr to 35 Myr ago (Adkins and co-workers 2001; Springer and co-workers 2003), and that the human and dog lineages diverged 90 Myr to 95 Myr ago (Springer and co-workers 2003). Although the human–murid divergence occurred later (85 Myr to 88 Myr ago; Springer and co-workers 2003), I use comparisons between human and dog because the human–dog evolutionary distance is shorter than the human–murid evolutionary distance. This is probably partly due to shorter generation time in the murid lineage (Li and co-workers 1996).

I used a relatively simple Jukes-Cantor model in a consistent manner for both d_S and d_I. More sophisticated models require more parameters to be estimated, such as transition/transversion bias (Nei and Kumar 2000). An advantage of the Nei-Gojobori method I used over more sophisticated methods such as that of Yang and Nielsen (2000) is consistency in choosing parameters. The Yang and Nielsen method includes parameters to account for codon bias, for which there is no comparable analog in introns.

2.2 Methods

2.2.1 Ortholog identification, genomic sequence, and transcript predictions

I used Ensembl Compara 28 (Birney and co-workers 2006; http://feb2005.archive.ensembl.org/) to identify ortholog pairs predicted by the unique best reciprocal hit method for both human–dog (builds NCBI 35 and BROADD1) and mouse–rat (builds NCBI m33 and RGSC 3.4). For the genes in these ortholog pairs, I then obtained genomic sequence, transcript predictions, and peptide prediction from Ensembl 28. I excluded from further analysis genes with introns longer than 1,000,000 bp, and genes with introns that overlap with predicted exons from other genes. I used RepeatMasker (Smit and co-workers 1996) to

mask repetitive elements in the nucleotide sequences.

2.2.2 The metascript model of alternative splicing

I constructed a metatranscript or "metascript" model for each gene that incorporates all of its alternative splicing forms. Figure 2.1 depicts an artificial example. Genomic regions expressed as protein in some or all of the predicted transcripts are called metaexons, and regions that never express as protein are called metaintrons. Each metaintron has a phase code, determined by the phases of the coding sequence upon interruption by the metaintron's constituent introns. A metaintron's phase code can indicate that its constituent introns all start in phases 0, 1, or 2, or in untranslated region (UTR). The phase code can also indicate one of $2^4 - 4 - 1 = 11$ specific mixtures of the four simple possibilities. Frameshift errors and introns shorter than 50 bp are each assigned special phase codes. I represent the metascript model as a string of nucleotides and phase codes.

2.2.3 The Introndeuce algorithm for pairing orthologous introns

I developed a new algorithm (Introndeuce) for robustly assigning orthologous introns in the presence of alternative splicing, without requiring genomic alignments. For each gene, I project all exons into genomic coordinates, and produce a novel sequence-like model called a metascript (Figure 2.1). The metascript is the concatenation of the nucleotide sequence of all annotated exonic regions with phase codes to indicate the phase of the intervening introns. The Introndeuce algorithm then aligns the exonic sequence and intronic phase codes of the resulting metascripts with an extension of standard dynamic programming methods. This results in the pairing of introns if and only if the surrounding exonic sequence is truly orthologous.

I perform the Smith-Waterman alignment (Smith and Waterman 1981) of orthologous metascripts using PSW from the Wise2 package (Birney 2002). I created a distance matrix based on HOXD70 (Chiaromonte and co-workers 2002), but that includes the 17 possible phase codes, and makes it extremely unlikely that out-of-phase metaintrons align with each other. This has the effect of penalizing intron sliding, which, if it happens at all, is thought to be exceedingly rare

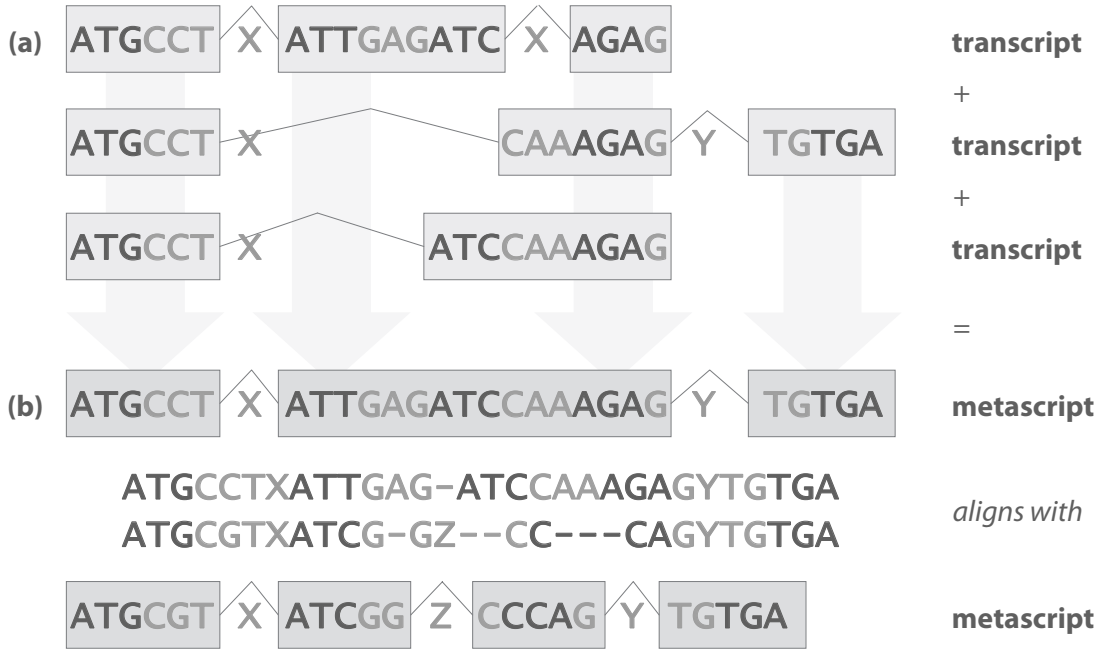


Figure 2.1: Using the metascript model and the Introndeuce algorithm to align alternatively spliced genes. This example shows (a) how the alternative splice forms of an artificial gene are combined to create a single metascript, and (b) how that metascript is aligned with another hypothetical metascript.

(a) Three conventional transcript models for this artificial gene are shown, featuring exons and their constituent nucleotides (letters alternating between dark gray and medium gray by codon). Introns are replaced by phase codes that indicate how many nucleotides are between the previous codon boundary and the start of the introns (X=0, Y=1, Z=2). The conventional transcript models are overlaid to form a single metascript, consisting of metaexons, regions of the gene that are sometimes or always expressed in transcribed mRNAs, and metaintrons, regions of the gene that are constitutively spliced out.

(b) The metascript generated in part (a) and another metascript are aligned to identify homologous introns. The nucleotide sequences of the metaexons and the phase codes of the metaintrons are concatenated to make a single sequence, which is then locally aligned using the Smith-Waterman algorithm, using a scoring matrix that makes it extremely unfavorable for a particular phase code to align with a nucleotide or a different phase code.

except in distantly related species, and rare even then (Stoltzfus and co-workers 1997; Rogozin and co-workers 2000).

2.2.4 Software availability

I have made available a Python (van Rossum 2006) package that constructs metascripts from Ensembl gene models and implements the Introndeuce algorithm (http://www.ebi.ac.uk/~hoffman/software/metascript/).

2.2.5 Estimation of d_N and d_S

I estimated the proportion of different nonsynonymous coding nucleotides p_N and the proportion of different synonymous coding nucleotides p_S for each gene pair using the Nei-Gojobori method (Nei and Gojobori 1986), based on PSW alignments of the longest translations of the genes. I then estimated the number of nonsynonymous substitutions d_N and the number of synonymous substitutions d_S based on p_N and p_S respectively using the Jukes-Cantor model (see subsection 1.1.3).

2.2.6 Estimation of d_I

I took the nucleotide sequence of each orthologous metaintron identified by Introndeuce, and masked out the first 10 bp and the last 30 bp of each metaintron to exclude conserved intron splicing signals. I then used BLASTZ (Schwartz and co-workers 2003) with a reduced stringency (reducing the maximal scoring pair score threshold to 2200) to align the sequences. Before making further calculations, I masked five nucleotides from both edges of each aligned block. This reduces edge wander effects (discussed further in subsection 2.3.1) and decreases uncertainty about the correctness of accepted alignment columns. If we let I_d be the number of mismatches in the remaining aligned sequence, and I the number of matches plus mismatches, we divide the two quantities to estimate a proportion of differing intron nucleotides

$$p_I = \frac{I_\mathrm{d}}{I}.$$

I then estimated d_I using the Jukes-Cantor model as above. I also estimated

an alternative version $d_{I,\text{maskedges}=0}$ without masking the edges of aligned blocks. I calculated the relative error between the two versions

$$\eta = \frac{2|d_{I,\text{maskedges}=0} - d_{I,\text{maskedges}=5}|}{d_{I,\text{maskedges}=0} + d_{I,\text{maskedges}=5}}.$$

2.2.7 Phylogenetic tree construction

I used MartShell (Kasprzyk and co-workers 2004) to identify all the human members of the MAGE family, ENSF00000000336, in Ensembl Compara 30 (http://apr2005.archive.ensembl.org/), and estimated d_S and d_I as above for each possible gene pairing. I converted the pairwise d values into distance matrices, using a distance of 1 where the distance could not be estimated because the sequences were too divergent. I then used the FITCH program of PHYLIP 3.64 (Felsenstein 2005) through the Pylip interface (http://www.ebi.ac.uk/~hoffman/software/pylip/) to generate a tree that I visualized with TreeView 1.6.6 (Page 1996).

2.2.8 Miscellaneous statistics

Many of the statistics and calculations reported in this chapter were performed using the R statistical programming environment (R Development Core Team 2007). I measured G+C content in the longest translation of the human or mouse gene of an orthologous gene pair, as G+C is highly correlated across the species pairs I used (human–dog: Spearman's rank-order correlation coefficient $r_S = 0.95$; mouse–rat: $r_S = 0.97$). I found overrepresented Gene Ontology (Gene Ontology Consortium 2006) terms using GO-TermFinder (Boyle and co-workers 2004), with a cutoff of $p = 0.01$, and annotations by GOA (Camon and co-workers 2004).

2.3 Results

Using the Introndeuce algorithm, most introns pair with consistent phases between the species considered, as expected. The 15,176 orthologs between human and dog, and the 16,183 orthologs between mouse and rat produced

96,476 human–dog paired metaintrons in 10,443 gene pairs and 105,560 mouse–rat paired metaintrons in 12,566 gene pairs (Table 2.1). In the metascripts that could be produced, the median number of metaintrons is 7 in humans and dogs, and 6 in the murids. I then aligned these metaintrons using BLASTZ individually.

I estimated d_S and d_I values for the ortholog pairs with at least one aligned metaintron (Figure 2.2; Table 2.1) using the Jukes-Cantor model (see subsection 1.1.3) in an identical manner for each statistic (see subsection 2.2.5). The median d_S value is 0.370 for human–dog and 0.212 for mouse–rat, whereas the corresponding median d_I values are lower at 0.305 and 0.158 respectively. This may be due to a lower substitution rate or the effects of more selective constraint in introns than in synonymous coding sites (Chen and Li 2001). As expected, the variance of the d_I values is smaller than that of d_S—the median absolute deviations (MADs; Huber 1981) of the human–dog and mouse–rat d_S values (0.113 and 0.062 respectively) are two to three times the MADs of the corresponding d_I values (0.053 and 0.022). Even considering extreme outliers, the d_I range (human–dog: [0.0285, 0.700]; mouse–rat: [0, 0.535]) is still smaller than the d_S range (human–dog: [0.0259, 4.813]; mouse–rat: [0, 2.328]).

Figure 2.2 shows a scatterplot of d_S versus d_I values for human–dog and mouse–rat gene pairs. As expected, these two variables are strongly correlated when analyzing the data from both species pairs ($r_S = 0.75$). One can clearly see, however, that a different model for each species pair would better fit the data from that species pair, despite lower correlation on a per–species-pair basis (human–dog $r_S = 0.57$; mouse–rat $r_S = 0.46$) than when examining all the gene pairs together. This means that one cannot universally predict a d_I value from a d_S value without reference to a particular species pair. One can see a clear separation between the range of d_I values for the two species pairs analyzed, while the d_S ranges overlap quite a bit. Additionally, within the same species pairs, the variance of d_I values is smaller. This suggests that d_I might provide a more distinctive characterization of genome-wide neutral evolutionary distance for species pairs in this range, which would allow better phylogenetic tree construction. When looking at 8095 genes that have a 1:1:1:1 ortholog relationship in human–dog–mouse–rat, I find that $d_{I,\text{human}-\text{dog}} > d_{I,\text{mouse}-\text{rat}}$ as expected in 8078 (99.8%) cases, but $d_{S,\text{human}-\text{dog}} > d_{S,\text{mouse}-\text{rat}}$ in only 7420 (92%) cases.

If one assumes that neutral mutation occurs at the same rate in exonic and

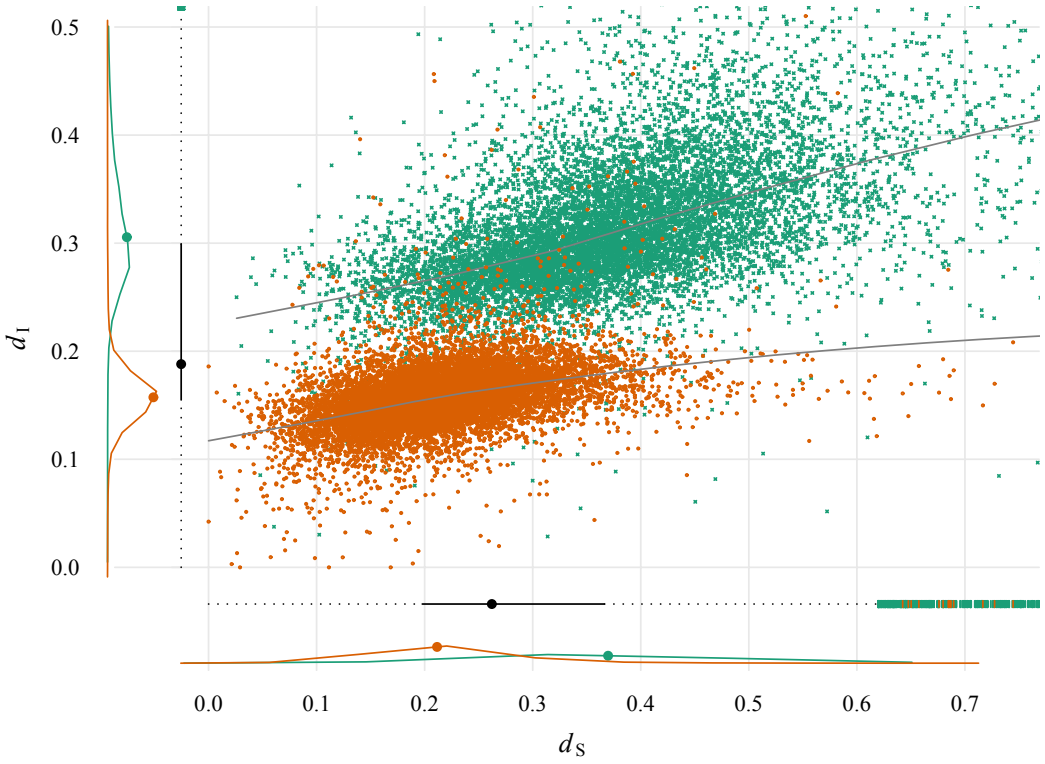


Figure 2.2: Dot-dash-density plot comparing the intron measurement d_I with the synonymous coding nucleotide measurement d_S. Values are shown for 12,329 mouse–rat ortholog pairs (dark orange circles) and 10,241 human–dog ortholog pairs (cyan crosses), representing only those gene pairs where d_S and d_I were below the pooled 99th percentile. The actual maximum d_S value is 4.81, and the maximum d_I value is 0.70. Local linear regression lines are shown for each species pair.

Decorations at the bottom and left sides of the graph are kernel density plots for each variable, grouped by species pair. The median of each variable grouped by species pair is marked on the density plot with a colored dot. Between the density plots and the scatterplot are condensed box plots with rugs (Tufte 2001) for the pooled values of each variable. Created with the xyddplot package (Hoffman 2006) for R.

Table 2.1: Summary statistics for genomes and genome pairs.

Species	Genes[a]	… with orthologs[b,c]	… with metascript models[d]	… with aligned metaintrons[b]	Metaintrons	… that align[b]	$\langle d_S \rangle^e$	$\langle d_I \rangle^e$	$\langle d_N \rangle^e$
Human	22,216	15,176 (83%)	11,882 (78%)	10,443 (88%)	131,532	96,476 (73%)	0.370 ± 0.113	0.305 ± 0.053	0.044 ± 0.046
Dog	18,201		13,469 (89%)		140,184				
Mouse	25,397	16,183 (73%)	13,693 (85%)	12,566 (92%)	145,665	105,560 (82%)	0.212 ± 0.062	0.158 ± 0.022	0.022 ± 0.023
Rat	22,159		13,827 (85%)		128,816				

[a] In Ensembl 28, excluding pseudogenes.

[b] Percentage value is the quantity in this column as a fraction of the smaller quantity in the previous column.

[c] Satisfy criteria in subsection 2.2.1 and have at least one metaintron.

[d] Percentage value is the quantity in this column as a fraction of the quantity in the previous column.

[e] Median and median absolute deviation.

intronic sequence, there are two plausible reasons for the systematically different estimates of neutral substitution in synonymous coding nucleotides and introns. First, the two different kinds of sequence are subject to different kinds of selective pressures. Some of these selective pressures, such as those discussed in the introduction, differently affect the fixation of point substitutions at certain sites in introns and synonymous coding sequence. Additionally, indels are much more likely to be selected against in coding sequence. The other possibility is that alignment effects cause the systematic difference. Even when corrected for by masking the edges of aligned blocks, edge wander effects may still lead to an observed similarity within the same species pair that underestimates the divergence of the two species.

2.3.1 Variability in d_I and d_S measures

I decided to look at the variability of d_S and d_I in a number of ways. First, I examined the variance for each data point as calculated by an analytical formula. The variance of a single Jukes-Cantor distance (representing the error in that particular estimation, rather than the dispersion of the whole population as discussed earlier) varies inversely with the number of nucleotides examined (Nei and Kumar 2000). Because of this, and the fact that I is generally much greater than S, the error for a single estimate of d_I (which is the square root of the variance, and called s_{d_I}) is usually lower than the error for a single estimate of d_S (called s_{d_S}). This is the case for 22,541 gene pairs, or 98% of those examined.

To consider the influence of generally larger d_S values on the error, I compared the coefficient of d_I variation, which is $V_I = {}^{s_{d_I}}/_{d_I}$, with the coefficient of d_S variation V_S, calculated in a similar way, and plotted the comparison in Figure 2.3. This reveals that V_S is greater for 22,817 gene pairs, or 96% of those examined. For 1476 gene pairs (6%), V_S is more than 10 times greater than V_I. One must consider, however, that the uncertainty of the intron nucleotide alignment is greater than that of the information-rich and selectively constrained amino acid alignment. Therefore, the intron method works best when one has reliable alignments, which happens at small evolutionary distances.

I examined the small number of points where V_I or V_S was greater than 100%. For the nine genes where $V_I > 100\%$, this is mainly due to genes in conserved regions of the genome (such as *Hoxd9* and *Hoxc4*) with small d_I values (all

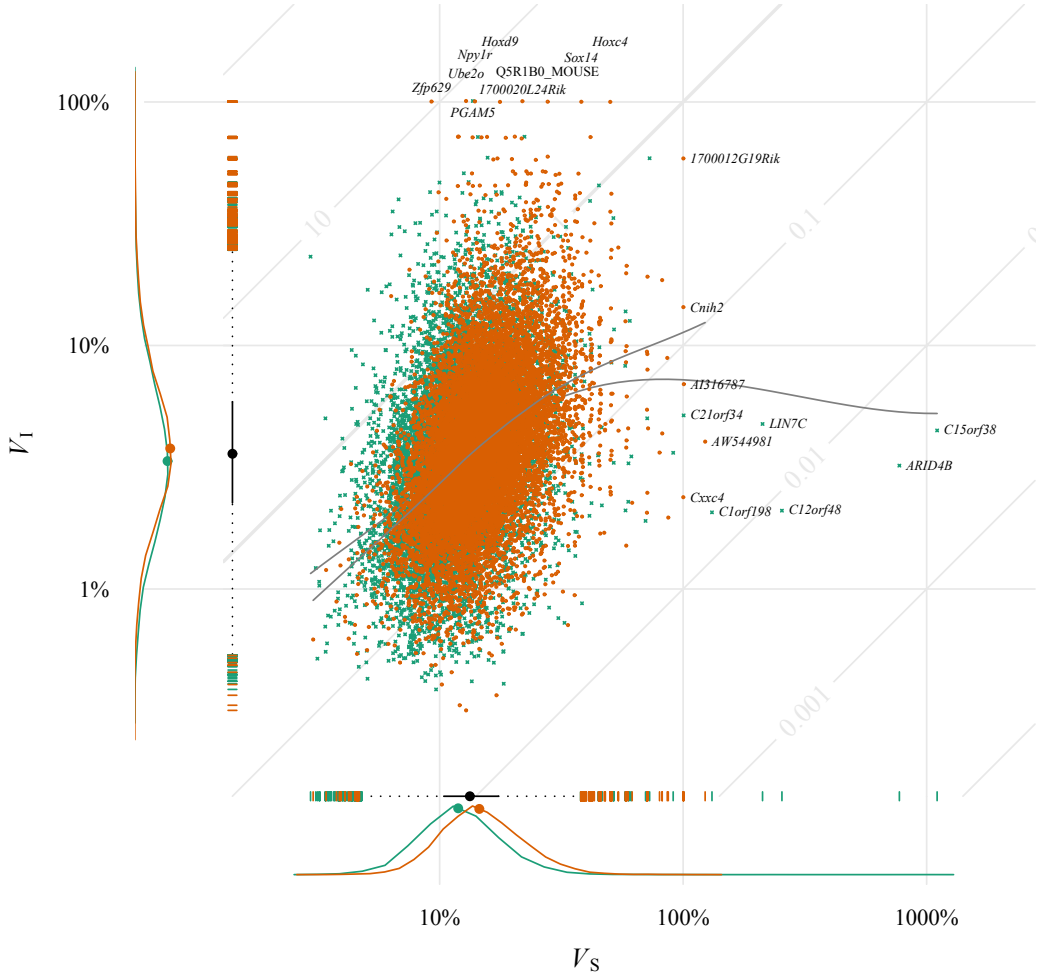


Figure 2.3: Dot-dash-density plot comparing V_I (the coefficient of variation in d_I), and V_S (the coefficient of variation in d_S). Values were calculated for 12,561 mouse–rat ortholog pairs (red circles) and 10,443 human–dog ortholog pairs (green crosses), and shown on a logarithmic scale. Two mouse–rat points where $d_S = 0$ and three mouse–rat points where $d_I = 3$ are not shown. Local linear regression lines are shown for each species pair. Gray labels on the diagonal grid lines indicate the ratio V_I/V_S for points along that line. Gene pairs where either V_I or V_S is greater than 100% are labeled with their HUGO Gene Nomenclature Committee or Mouse Genomic Nomenclature Committee symbols in an italic typeface, or with a UniProt entry name in a roman typeface. Decorations at the bottom and left sides of the graph are as in Figure 2.2.

< 0.06), which means that the coefficient of variation is large even if the error is still small in absolute terms. I found a similar pattern in the five outliers where V_S is only slightly greater than 100%, which all had $d_S < 0.04$. In the other six V_S outliers, which have the largest d_S, both d_S and s_{d_S} are quite large (> 1). These are examples of cases where ω_S underestimates positive selection, and merit further study, since $\omega_I > 1$ for each of the six.

I then investigated the effects of other factors, such as G+C content, on the differences between d_I and d_S, as well as their variability. To visualize the difference between d_I and d_S and the effect of G+C content on this difference, I created Tukey mean-difference plots (Cleveland 1993) (Figure 2.4) split at the quartiles of G+C content. These plots allow a visual assessment of shift between d_I and d_S, by analyzing deviation from the zero line in both distance and slope. As the neutral evolutionary distance between two orthologs increases, the difference between d_I and d_S becomes more pronounced (d_S increases proportionally more than d_I). The median $d_I - d_S$ is below 0 for both species pairs for all ranges of G+C content, indicating that median d_S is larger than median d_I. The slope indicates the difference in variance between d_S and d_I.

One of the key reasons for estimating an evolutionary distance from introns is that the number of intron sites I is an order of magnitude greater than the number of synonymous coding sites S or the number of nonsynonymous coding sites N, where fractional values in S or N indicate partial degeneracy. The distribution of the number of each kind of site in a gene pair is positively skewed. In human–dog, the quartiles of S are (185.6, 297.5, 462.5), the quartiles of N are (590.5, 952.6, 1497.1), and the quartiles of I are (1393, 4046, 10944). In mouse–rat, the distribution of the number of different kinds of sites shows a similar relationship: S quartiles = (174.9, 285.6, 440.0), N quartiles = (562.9, 913.5, 1404.3), I quartiles = (2067, 5224, 12,497).

A striking difference between the human–dog comparison and the mouse–rat comparison is the effect of G+C content. As G+C content ($q_{\mathtt{G+C}}$) measured in human increases, human–dog d_I increases ($r_S = 0.41$; significant at $p < 0.001$). Human–dog d_S increases with G+C content as well, but much less of the proportion of variation in d_S is attributable to a change in G+C content, as the correlation is weaker ($r_S = 0.27$; $p < 0.001$). Mouse–rat d_S and d_I decrease as mouse G+C content increases, but the correlation between G+C content and d_S is much weaker ($r_S = -0.15$; $p < 0.001$), as is the correlation with d_I ($r_S = -0.11$;

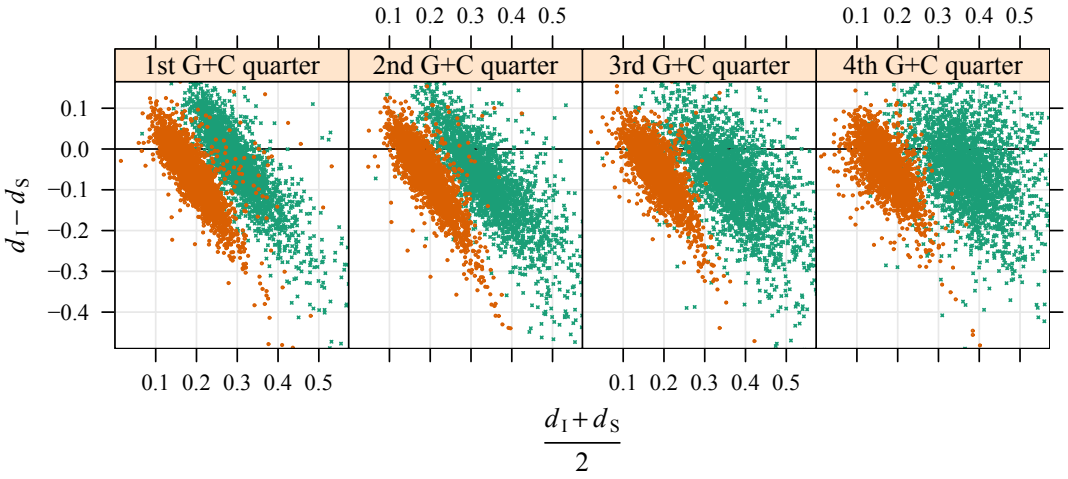


Figure 2.4: Tukey mean-difference plots comparing d_S and d_I at different levels of G+C content. The mean of the quantities d_S and d_I for 12,381 mouse–rat (dark orange circles) and 10,276 human–dog (cyan crosses) ortholog pairs is plotted against the difference of the quantities. The domain of the plot includes only values below the 99th percentile of mean values, and the range includes only data between the 0.005 and 0.995 quantiles of difference values. The actual maximum mean is 4.93 and the actual minimum and maximum differences are -4.59 and 0.32, respectively. The data points are placed in quarters depending on the G+C content of the longest translation of the relevant human or mouse gene. The minimum, quartiles, and maximum $q_{\mathtt{G+C}}$ are at (31.1, 45.6, 52.8, 59.5, 79.7)% for human and (31.9, 48.2, 53.0, 57.1, 73.6)% for mouse.

$p < 0.001$). This effect is best seen in the comparison of the first quarter of G+C content to the fourth quarter in Figure 2.4 showing dramatically increased variance in the human–dog d_I value at high G+C. This effect is unchanged if the G+C content is measured with dog or rat respectively. This change in responsiveness to G+C content between the species pairs suggests that d_I is a more labile measure over evolutionary time (see section 2.4).

One benefit of the d_S measure is that the information-rich amino acid alignment provides a robust scaffold for identifying orthologous synonymous nucleotides. This contrasts with nucleotide alignments, where the placement of substitutions and indels is determined by an alignment program that must penalize gaps and substitutions. This process is therefore more error prone in

particular in the placement of substitutions near insertions where the highest-scoring alignment is less likely to reflect the true evolutionary relationships between bases, an effect known as edge wander (Holmes and Durbin 1998). I generally estimated d_I by counting only substitutions in aligned blocks at least 5 nucleotides away from the nearest indel ($d_{I,\text{maskedges}=5}$) to remove these errors. To examine how well this strategy removed errors, I also estimated it by counting all substitutions in the intron nucleotide alignment ($d_{I,\text{maskedges}=0}$). For mouse–rat values, both versions of d_I are very close (Figure 2.5) and the median relative error η between the two measurements is 1.7%. The two methods of estimating d_I for human–dog values produce values that are close with median $\eta = 3.7\%$, but this is twice the median for mouse–rat. Absolute error increases with the mean of the two d_I methods for human–dog ($r_S = 0.67$; $p < 0.001$), but actually decreases very slightly for mouse–rat ($r_S = -0.04$; $p < 0.001$). The increase in variance indicates that the human–dog distance might be at the edge of where such alignment artifacts dominate (see section 2.4).

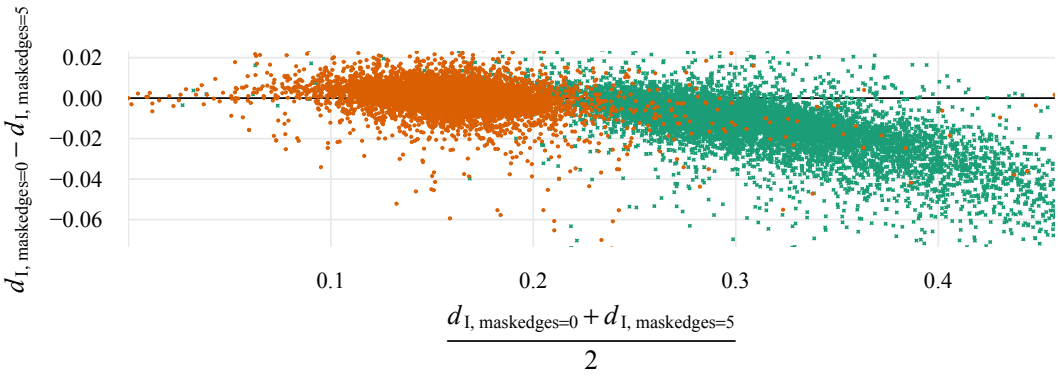


Figure 2.5: Tukey mean-difference plot comparing d_I computed with edge masking against d_I computed without edge masking. The mean of d_I computed when masking five nucleotides on both sides of alignment gaps, and d_I computed without masking for 12,336 mouse–rat (dark orange circles) and 10,255 human–dog (cyan crosses) ortholog pairs, is plotted against the difference of d_I computed with each of these methods. The domain of the plot includes only values below the 99th percentile of mean values, and the range includes only data between the 0.005 and 0.995 quantiles of difference values. The actual maximum mean is 0.67, and the actual minimum and maximum differences are -0.17 and 0.11, respectively.

To examine the effects of selective pressures on supposedly synonymous sites, I examined d_S and d_I for genes with known functional elements in synonymous sites. I inspected a set of 34 human genes where synonymous coding mutations can alter splicing (Cartegni and co-workers 2002; Chamary and co-workers 2006). In this set, none of the human–dog values for d_S, d_I, or $d_I - d_S$ significantly changed from the broader population of human–dog gene pairs (Mann-Whitney tests; $p > 0.2$), probably because the splice-altering elements affected only a small proportion of the synonymous sites. Nonetheless, it would be sensible to mask the beginnings and ends of exons when estimating d_S in the future, since exonic splicing enhancers and silencers more frequently occur in these regions. I also looked at human–dog *GRIA2*, a glutamate receptor with a known conserved intronic sequence that induces RNA editing (Seeburg 2002). Again, d_S, d_I, and $d_I - d_S$ fit well within the general population of human–dog gene pairs.

2.3.2 Effect on the estimation of selection

One of the major uses of d_S is to calculate a ratio $\omega = {}^{d_N}/_{d_S}$ to determine what sort of selective pressure a region is under. Here, I refer to this established ratio as ω_S, and define an equivalent that uses intron substitutions to estimate the non-neutral rate of evolution $\omega_I = {}^{d_N}/_{d_I}$. Looking at the human and dog genes with the largest ω_I (Table 2.2), there are a variety of kinds of genes, including transcription factors, RNA-binding proteins, reproductive-related genes, and genes of unknown function. The human and dog genes with the largest ω_S (Table 2.3) include immune-related genes and similar kinds of genes to those with the highest RNA- and DNA-binding genes, and the genes with the highest ω_I. Surprisingly, only two genes are shared amongst the top 10 ω_I and top 10 ω_S groups: *C1orf198*, and *CENPA*, suggesting that many of the previously noted outliers in ω_S are potentially due to variance in neutral rate estimation, and not to changes in nonsynonymous rate. No genes are shared amongst the top 10 ω_I and top 10 ω_S groups for mouse and rat (Table 2.4 and Table 2.5).

When broader groups are considered, such as the ω_I and ω_S values above the 95th percentile, more genes (human–dog: ${}^{341}/_{523} = 65\%$; mouse–rat: ${}^{397}/_{629} = 63\%$) are shared in both lists, but a substantial number are still unique. In contrast, when considering the ω_I and ω_S values below the 5th percentile, many more

Table 2.2: Human–dog gene ortholog pairs with the 10 largest values of ω_I.

Human gene accession ID[a]	Human gene symbol	Human gene name	Dog gene accession ID[a]	ω_S	ω_I
ENSG00000140832	*MARVELD3*	MARVEL domain containing 3	ENSCAFG00000020193	0.58	2.28
ENSG00000184650	*ODF4*	Outer dense fiber of sperm tails 4	ENSCAFG00000017061	0.29	2.28
ENSG00000115163	***CENPA***	**Centromere protein A, 17kDa**	**ENSCAFG00000004472**	**0.89**	**2.39**
ENSG00000164808	*KIAA0146*	KIAA0146 protein	ENSCAFG00000009134	0.56	2.44
ENSG00000185480	*C12orf48*	Chromosome 12 open reading frame 48	ENSCAFG00000007291	0.22	2.45
ENSG00000172346	*CSDC2*	Cold shock domain containing C2, RNA binding	ENSCAFG00000001051	2.07	2.59
ENSG00000107736	*CDH23*	Cadherin-like 23	ENSCAFG00000014229	0.66	2.65
ENSG00000054267	*RBM34*	RNA binding motif protein 34	ENSCAFG00000011439	0.15	3.25
ENSG00000163867	*ZMYM6*	Zinc finger, MYM-type 6	ENSCAFG00000003599	0.68	3.67
ENSG00000119280	***C1orf198***	**Chromosome 1 open reading frame 198**	**ENSCAFG00000011960**	**0.86**	**6.24**

Rows in boldface denote ortholog pairs that also occur in Table 2.3.

[a] Ensembl genes.

Table 2.3: Human–dog gene ortholog pairs with the 10 largest values of ω_S.

Human gene accession ID[a]	Human gene symbol	Human gene name	Dog gene accession ID[a]	ω_S	ω_I
ENSG00000089505	*CKLF*	Chemokine-like factor	ENSCAFG00000020419	0.84	0.80
ENSG00000119280	***C1orf198***	**Chromosome 1 open reading frame 198**	**ENSCAFG00000011960**	**0.86**	**6.24**
ENSG00000115841	*FAM82A*	Family with sequence similarity 82, member A	ENSCAFG00000006160	0.86	0.73
ENSG00000164047	*CAMP*	Cathelicidin antimicrobial peptide	ENSCAFG00000012896	0.87	0.80
ENSG00000115163	***CENPA***	**Centromere protein A, 17kDa**	**ENSCAFG00000004472**	**0.89**	**2.39**
ENSG00000150076	*CCDC7*	Coiled-coil domain containing 7	ENSCAFG00000003847	0.90	0.74
ENSG00000102021	*LUZP4*	Leucine zipper protein 4	ENSCAFG00000018242	1.00	1.31
ENSG00000121644	*C1orf121*	Chromosome 1 open reading frame 121	ENSCAFG00000015825	1.08	1.53
ENSG00000100721	*TCL1A*	T-cell leukemia/lymphoma 1A	ENSCAFG00000017719	1.08	0.74
ENSG00000172346	*CSDC2*	Cold shock domain containing C2, RNA binding	ENSCAFG00000001051	2.07	2.59

Rows in boldface denote ortholog pairs that also occur in Table 2.2.

[a] Ensembl genes.

Table 2.4: Mouse–rat gene ortholog pairs with the 10 largest values of ω_I.

Mouse gene accession ID[a]	Mouse gene symbol[b]	Mouse gene name[b]	Rat gene accession ID[a]	ω_S	ω_I
ENSMUSG00000042101[c]	*Ppp1r12b*	Protein phosphatase 1, regulatory (inhibitor) subunit 12B	ENSRNOG00000005011	0.48	3.15
ENSMUSG00000025527	*Satl1*	Spermidine/spermine N1-acetyl transferase-like 1	ENSRNOG00000022476	0.68	3.47
ENSMUSG00000057763	Q5R1B0_MOUSE	TRAV17 (Fragment)	ENSRNOG00000008906	0.27	3.47
ENSMUSG00000022514	*Il1rap*	Interleukin 1 receptor accessory protein	ENSRNOG00000001928	0.49	3.68
ENSMUSG00000054863	*AW049604*	Expressed sequence AW049604	ENSRNOG00000004460	0.62	4.92
ENSMUSG00000020661	*Dnmt3a*	DNA methyltransferase 3A	ENSRNOG00000012555	0.71	5.15
ENSMUSG00000064373	*Sepp1*	Selenoprotein P, plasma, 1	ENSRNOG00000015849	1.02	6.28
ENSMUSG00000026638	*A130010J15Rik*	Interferon regulatory factor 6	ENSRNOG00000005082	0.40	6.63
ENSMUSG00000027338	*Prnd*	Prion protein dublet	ENSRNOG00000021259	1.19	6.89
ENSMUSG00000043342	*Hoxd9*	Homeo box D9	ENSRNOG00000001580	0.52	27.80

[a] Ensembl genes.

[b] If a Mouse Genomic Nomenclature Committee (Blake and co-workers 2000) gene symbol is not available, the UniProt (Wu and co-workers 2006) entry name is used as the gene symbol (in a roman typeface, rather than italics) and the UniProt description is used as the gene name.

[c] Not valid in Ensembl 37.

Table 2.5: Mouse–rat gene ortholog pairs with the 10 largest values of ω_S.

Mouse gene accession ID[a]	Mouse gene symbol[b]	Mouse gene name[b]	Rat gene accession ID[a]	ω_S	ω_I
ENSMUSG00000027404	*Snrpb*	Small nuclear ribonucleoprotein B	ENSRNOG00000006961	1.24	0.38
ENSMUSG00000056260	*4933421E11Rik*	RIKEN cDNA 4933421E11 gene	ENSRNOG00000017784	1.24	0.58
ENSMUSG00000032758	*Kap*	Kidney androgen regulated protein	ENSRNOG00000005858	1.27	2.54
ENSMUSG00000061460[c]	*LOC229550*	RIKEN cDNA 9130204L05 gene	ENSRNOG00000027525	1.31	1.37
ENSMUSG00000036925	*Spt1*	Salivary protein 1	ENSRNOG00000026202	1.33	1.63
ENSMUSG00000024379	*Tslp*	Thymic stromal lymphopoietin	ENSRNOG00000027355	1.39	1.45
ENSMUSG00000030858	*1700007K09Rik*	RIKEN cDNA 1700007K09 gene	ENSRNOG00000026419	1.39	2.72
ENSMUSG00000018914	*Il3*	Interleukin 3	ENSRNOG00000026786	1.43	1.49
ENSMUSG00000020702	*Ccl1*	Chemokine (C-C motif) ligand 1	ENSRNOG00000021851	1.79	1.42
ENSMUSG00000045391	*Herc4*	Hect domain and RLD 4	ENSRNOG00000000382	2.28	1.53

[a] Ensembl genes.

[b] If a Mouse Genomic Nomenclature Committee (Blake and co-workers 2000) gene symbol is not available, the UniProt (Wu and co-workers 2006) entry name is used as the gene symbol (in a roman typeface, rather than italics) and the UniProt description is used as the gene name.

[c] Not valid in Ensembl 37.

genes are shared in both lists (human–dog: $^{475}/_{523} = 91\%$; mouse–rat: $^{606}/_{629} = 96\%$). One can explain this asymmetry by noting that as d_N, and therefore ω, increases, the difference between d_S and d_I also increases. This means that it is particularly important to use both the coding site and intron methods to determine neutral rate, especially when considering genes under positive selection,

Human–dog genes with ω_I values above the 95th percentile (but with ω_S values below the 95th percentile) significantly overrepresent only the Gene Ontology (Gene Ontology Consortium 2006) biological process terms involved in immunity and defense (immune response; defense response; response to pest, pathogen or parasite). So do human–dog genes above the 95th percentile exclusively for ω_S (response to biotic stimulus; defense response; immune response). The mouse–rat genes above the 95th percentile exclusively for ω_S overrepresent only similar terms, but no terms are significantly overrepresented in the genes above 95th percentile exclusively for ω_I.

2.3.3 Use of d_I for the investigation of paralog relationships

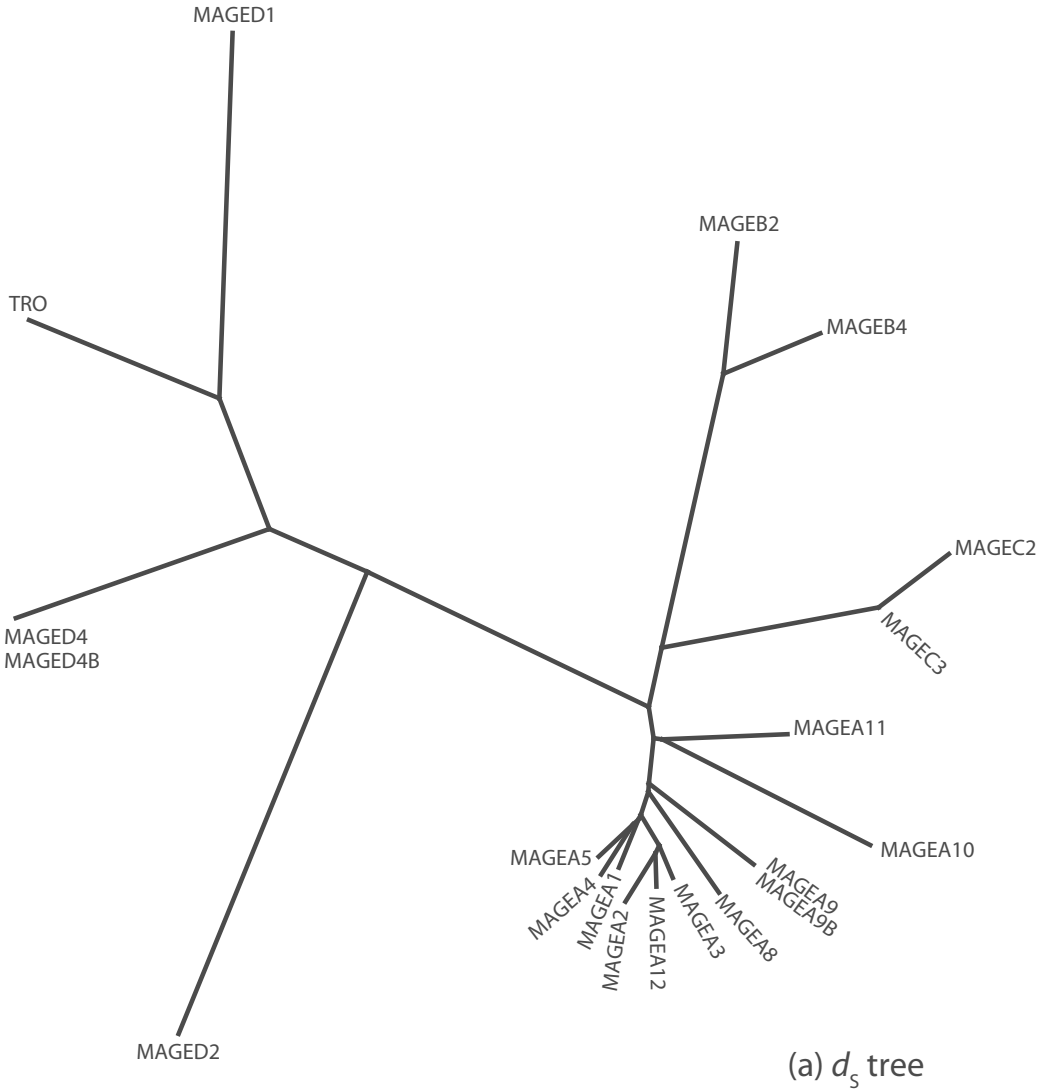
Both d_S and d_I have applications beyond orthologs in examining the relationship between paralogous genes. For recently evolved gene families, d_S is likely to be small (often 0) between two paralogs due to the short evolutionary time separating the two genes. In these cases, using a d_I measure becomes crucial to understanding gene relationships, because d_I is usually greater than d_S (Figure 2.4) at small evolutionary distances. Figure 2.6 shows the MAGE family of paralogs (Chomez and co-workers 2001), where different trees are found when using d_S or d_I measures. As expected, the d_I tree resolves some of the recent duplication events, including the cases where the coding sequence is identical at the DNA level. For more diverged sequences, however, the d_I measurement appears to be saturated due to edge wander. This, along with the arbitrary assignment of distances that cannot be estimated, produces artificially small branch lengths.

2.4 Discussion

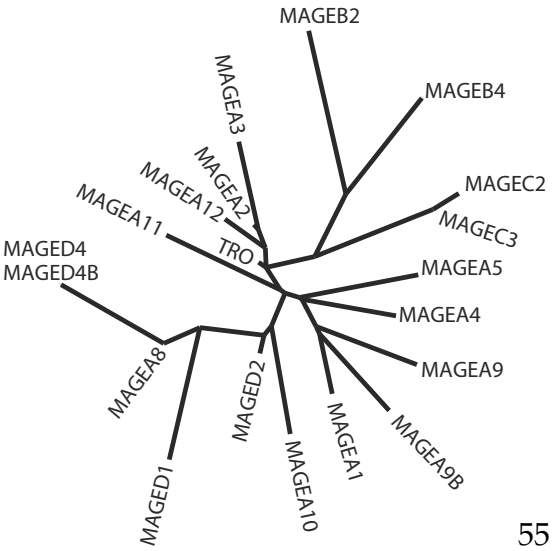
In order to sensibly compare d_S to d_I, one must have a large number of correctly paired orthologous introns. Previously, researchers have either (a) annotated orthologous introns manually, (b) assumed the colinearity of introns in orthologous genes with identical numbers of introns (Castresana 2002), or (c) used genomic nucleotide alignments and relied on the gene structure in only one organism (Chimpanzee Sequencing and Analysis Consortium 2005). Manual annotation obviously does not scale well to genome-scale analyses. Assuming colinearity does not account for the apparent loss and gain of introns when comparing annotated genes, which is either due to an actual biological change in the number of introns, or a result of dissimilar methods and data for annotation. Genomic alignments have some useful properties for identifying orthologous introns—notably that one only needs gene annotation for one species. As with assuming colinearity, however, using genomic alignments also assumes that the

---

Figure 2.6 *(following page)*: Comparison of phylogenetic trees constructed with the two methods. Phylogenetic trees for the human MAGE family constructed with distances from (a) d_S and (b) d_I. Gene symbols are as assigned by the HUGO Gene Nomenclature Committee (Bruford and co-workers 2008).



(a) d_S tree



0.1

(b) d_I tree

intron and exon assignments are effectively unchanged, and that the genomic alignment method has enough sensitivity to correctly align bases. Genome-wide alignments also have great difficulty in correctly handling lineage-specific duplications. Finally, genome-wide alignments are inherently computationally expensive and require complex engineering.

I have introduced a new method (Introndeuce) that automatically pairs orthologous introns in the absence of detectable intronic alignments with modest computational requirements. This method may have applications beyond those discussed in detail here. For example, metascript alignment is an elegant way to consider gene-level selection when alternative splicing is present. Previous approaches ignored the selective pressure on intron regions due to nonconstitutive exons in these regions.

The large amount of data has allowed me to investigate the properties of d_I compared to d_S. It is clear that the two variables are correlated, but that they are measuring different properties of the genome—one cannot consider d_I directly equivalent to d_S because d_I would systematically underestimate d_S as it increases. My opinion is that both measures have their flaws, both from a conceptual perspective of potential non-neutral bases in each case, and from a pragmatic issue of alignability and observed mutations.

One surprise has been the more marked species difference in d_I compared to d_S. The d_I measure shows far greater change in median between human–dog versus mouse–rat, when compared to d_S. Although the greater variance of d_S when compared to d_I in both species pairs must explain this partially, Figure 2.2 clearly shows that the relationship between d_S and d_I is not equivalent in the two species pairs. I have investigated both alignment artifact effects (Figure 2.5) and G+C content effect (Figure 2.4) to explain this difference. It seems clear that d_I has more specific variation due to G+C effects in human–dog but not in mouse–rat. Humans and dogs have more closely related G+C content distributions, which are more likely to reflect the boreoeutherian ancestor, and in these species evolutionary distance rates are correlated with G+C content along chromosomal positions (Lindblad-Toh and co-workers 2005). The variation in G+C content in genomes is a complex phenomenon that probably interacts with the d_I measure in multiple ways, such as their shared correlation with local recombination rate (Eyre-Walker 1992; Fullerton and co-workers 2001; Lander and co-workers 2001; Lercher and Hurst 2002; Waterston and co-workers 2002;

Hardison and co-workers 2003; Hellmann and co-workers 2003), hypermutability of CpG dinucleotides, which more frequently occur in elevated G+C areas (Fryxell and Moon 2005), biased gene conversion to GC (Marais 2003) causing G+C content elevation in regions of high recombination, increased SINE insertion in elevated G+C regions (Jurka 1997), and other effects. In particular, Webber and Ponting (2005) have observed elevated G+C content and human–dog d_S values in dog genes in subtelomeric and pericentromeric regions, or in syntenic blocks of < 4 Mb, as well as human genes in subtelomeric regions, while noting that these phenomena are much less striking in the mouse and rat genomes. Due to the correlation between d_S and d_I, similar phenomena may affect d_I, although the CpG hypermutability affects intron sites differently from synonymous coding sites, as noted earlier. Genome-wide changes in neutral mutation rates and genomic landscape are more likely to affect d_I than d_S, as runs of intron nucleotides have fewer constraints than runs of coding site nucleotides, which include nonsynonymous sites. Consequently d_I measures are less useful over broader evolutionary time.

Preliminary research into estimating d_I values for human–mouse, human–rat, and *Tetraodon nigroviridis–Takifugu rubripes* species pairs indicated that the edge wander problem was too severe to trust the d_I measure (data not shown). I surmised that there had been too much neutral evolutionary change at these kinds of distances for the nucleotide substitution d_I method to be useful. Judging by the increase in variance as d_I rises in the alignment artifact investigation (Figure 2.5), the human–dog evolutionary distance (76% median identity in intronic local alignments) is close to the limit of where d_I measures are useful.

For evolutionary distances shorter than human–mouse, d_I may be a more useful measurement than d_S. The small number of synonymous coding sites becomes more crucial when there is less time over which one can observe changes. In particular, I have shown that it can be used to make the evolutionary relationship between young paralogs less ambiguous when compared to a phylogeny generated with the use of d_S. While d_I can provide a less ambiguous tree with greater distances for closely related paralogs than a d_S tree, for more distantly related paralogs the distances are far shorter than those estimated with d_S. The d_I measurement is probably saturated for distances this far out, but d_S is useful here because the measurement still retains some dynamic range. Additionally, the difference between d_I and d_S varies for different kinds of genes.

Genes that have RNA editing sites, distant intron branch points (Gooding and co-workers 2006), or other conserved intronic elements have a particularly low d_I, whereas genes with conserved DNA and RNA regulatory elements in their exons have a low d_S.

I used BLASTZ local alignments to identify the orthologous nucleotides within introns. Since I was only inspecting aligned blocks, I needed only the high-scoring areas of alignment that a local alignment algorithm could produce. Preliminary investigations with a global alignment algorithm, LAGAN (Brudno and co-workers 2003), convinced me not to pursue a global alignment approach, as it forces even more edge wander. Replacing BLASTZ with LAGAN in this analysis led to a bimodal distribution of d_I values for gene pairs, with modes for those pairs where alignable intron sequence was either sufficient or insufficient (data not shown).

Using the Jukes-Cantor model means that a lack of parameters to estimate simplifies the analysis. Nei-Gojobori is similar to maximum likelihood if transition/transversion and codon-usage bias are ignored (Yang and Bielawski 2000). Ignoring transition/transversion bias leads to overestimation of d_S, because transitions at the third position of a codon are more likely to be synonymous (Ina 1995; Nei and Kumar 2000). This may partially explain the tendency of d_I to underestimate d_S.

The difference in d_S and d_I suggests that one should use caution when investigating outliers of ω_S or ω_I, because many of these outliers are not consistent between the two measures of neutral rate. I suggest that wherever appropriate, such as within reasonably close species, researchers should quote both values, as the most robust outliers have support from both the ω_S and ω_I measures.

2.4.1 Comparison with previous research

Researchers have suggested still more methods to estimate the neutral rate; I summarize some of these in Table 2.6. The most similar method to the intron method in this chapter is that of Castresana (2002), who also suggested a method to estimate an evolutionary distance using intron substitutions in aligned blocks, and used it on a manually selected set of 63 human–mouse gene pairs with 504 introns. It differed from the method described in this chapter in several notable ways. First, he did not use an algorithm like Introndeuce to identify orthologous

Table 2.6: Neutral rate estimation methods and selected uses on particular species pairs.

Neutral nucleotides	Mutation process	Species pairs
Synonymous coding	Substitution	Many
Intron	Substitution	Human–mouse (Castresana 2002), human–dog and mouse–rat (Hoffman and Birney 2007)
Intron	Indel	Mouse–rat and human–Old World monkey (Ogurtsov and co-workers 2004)
Intron	Length change	Human–mouse (Ogurtsov and co-workers 2004)
Ancestral repeat	Substitution	Human–mouse (Waterston and co-workers 2002), human–chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005), human–dog (Lindblad-Toh and co-workers 2005), human–chicken (Webster and co-workers 2006)

introns, instead assigning intron orthology colinearly on orthologous genes, and discarding genes that had a different number of introns in either species. This does not work in cases of intron gain or loss, or differences in annotated gene structure, although it might have worked for the small hand-selected set Castresana uses. His method also does not take alternative splicing into account, which would affect the results. Castresana uses a Needleman-Wunsch global alignment (Needleman and Wunsch 1970) followed by Gblocks (Castresana 2000) to identify aligned blocks based both on distance from gaps and anchoring by highly conserved positions. This differs from my approach of BLASTZ local alignment followed by identifying aligned blocks solely on the basis of distance from gaps. Castresana uses the HKY model of evolution (Hasegawa and co-workers 1985), which is more sophisticated than the Jukes-Cantor model I use. However, it would be possible to use the alignments produced by both methods as input to any model of DNA evolution.

While I tested my method on a genome-wide 23,009 gene pairs, Castresana's input data were winnowed from a larger hand-selected set of 77 gene pairs (Jareborg and co-workers 1999). This means that my set has true outliers, allowing me to analyze corner cases such as the ω values in the top 95th percentile for both

methods. Furthermore, the correlation between d_S and d_I is much better in my dataset ($r_S = 0.75$ overall and $r_S = 0.46$ for mouse–rat, as opposed to $r_S = 0.34$ for the human–mouse set). Castresana's data also lead him to claim that genes with fast-evolving exons have fast-evolving introns. I used my method at closer evolutionary distances than human–mouse, and have concluded that it is not actually useful at this distance due to edge wander problems. This is borne out by the lower correlation at this distance.

Ogurtsov and co-workers (2004) produced a maximum likelihood estimate of a neutral evolutionary distance between two genes based on the rate of insertions and deletions in introns. Two different approaches were taken: measuring indels with alignments on very short evolutionary distances (human–Old World monkey and mouse–rat), and measuring change in intron length for longer distances (human–mouse). The former was used to calibrate the latter. The estimation based on change in intron length may be confounded by possible selective effects on intron length, which have been suggested in organisms such as *Drosophila* (Carvalho and Clark 1999; Yandell and co-workers 2006). Ogurtsov and co-workers did not publish a comparison of their measurement with d_S on a gene-by-gene basis so it might fluctuate depending on gene type. Other researchers have reported (Gibbs and co-workers 2004) a low correlation between indel and substitution rate. They report that their measurement is not saturated for human–mouse and might work for distances up to human–*Takifugu*. When possible, however, they say that for short evolutionary distances (such as those examined in this chapter) a substitution-based measurement is preferable. I agree that it would be preferable since it measures the same character of mutation as d_S—within well-aligned blocks, the character of selectively neutral mutation should be the same. While indel-based evolutionary distances might correlate with d_S when both are measured for the whole genome, fluctuations might affect individual gene pairs or classes of gene pairs. There is no evidence to suggest that the molecular clock for indels covaries with the molecular clock for substitutions over long evolutionary timescales. They might vary independently.

Another recently-introduced method for estimating the neutral rate of nucleotide substitution is by counting substitutions in *ancestral repeat* or AR sites—aligned nucleotides in transposable elements that predate the common ancestor of the genomes being compared (Waterston and co-workers 2002; Hardison and co-workers 2003). This relies on the assumption that the character of a particular

AR site can change synonymously without affecting the fitness of the organism. Kamal and co-workers (2006) report, however, that at least some ancient repeats are under strong selection. Substitutions in ARs, just like those used to estimate d_S and d_I, show a G+C content-dependent pattern (Waterston and co-workers 2002; Hardison and co-workers 2003). Hardison and co-workers argue that a measure of evolutionary distance based on AR sites is "roughly similar" to one based on 4D sites. However, they report better correlation between substitutions in intron sites and substitutions in 4D sites than between substitutions in AR sites and substitutions in 4D sites. Hardison and co-workers also state that AR sites may provide a better model of neutral evolution than 4D sites, since they are not affected as much by CpG hypermutability bias. However, more sophisticated synonymous coding nucleotide models such as Nei-Gojobori allow more variation in bases flanking a synonymous site, which mitigates this effect. The reported AR methods do not have any provision to manage edge wander effects or problems arising from spurious AR orthology assignment after a gene conversion event involving these repeats. The need to start with an anchor of known orthologous sequence when estimating evolutionary distance with ARs means that they would not be as useful for analyzing close paralogs.

Lunter and co-workers (2006) introduce an interesting method for modeling the neutrality of any kind of genomic sequence using indels, which they apply to find functional elements under purifying selection in human–mouse–dog. However, since their method does not estimate a separate quantity for neutral evolution as it varies over chromosomal position, it is not really comparable to the other methods discussed in this chapter.

Chapter 3

Sunflower: a probabilistic model of transcription factor binding

3.1 Introduction

In the remaining chapters of this thesis, I present a new method to determine the nature of selection in promoter nucleotide positions with respect to the binding of transcription factors. In this chapter I outline the design and some key features of a model I call *Sunflower*. It is designed to model the simultaneous binding of a full set of transcription factors on a given region of DNA, and changes in this model due to potential mutations. I discuss the basic model in this chapter, and the mutational effects in chapter 4.

Sunflower is somewhat similar to biophysical models which posit that gene expression can be estimated using the equilibrium probability of transcription factor binding (Bintu and co-workers 2005). While those models estimate these probabilities using thermodynamic calculations involving the Boltzmann factor and partition functions, I simplify matters by eliminating energy terms and relying solely on probabilistic modeling.

Sunflower is implemented as a hidden Markov model (see section 1.3) of the ensemble of transcription factors bound to the whole promoter. I use the forward-backward algorithm on this model to estimate the posterior probability of each of the transcription factors being bound at each position. I call the totality of probabilities together the promoter's *binding profile*.

I have focused my investigation on the 1000 bp on either side of a TSS, because

cis-acting elements in a subset of these regions are usually sufficient to promote transcription (Trinklein and co-workers 2003). I focus further on ± 100 bp to include only bases where I have higher confidence that they affect transcription. Tabach and co-workers (2007) found that location-specific functional binding sites are most likely in the range $[-200, 100]$. Recent research, including data from the ENCODE Project Consortium (2007), emphasizes the important role of sequence downstream of the TSS in mediating transcription. Sequence that is downstream of the TSS, however, is still more difficult to analyze for evolutionary signals of transcriptional regulation because it is confounded by signals of post-transcriptional regulation and protein-coding sequence.

Carninci and co-workers (2006) classified putative promoters into "shape classes" based on the distribution of cap analysis gene expression (CAGE; Shiraki and co-workers 2003) tag clusters. Some of the promoters were assigned to a single dominant peak (SP) class with most of the transcriptional activity beginning at a single position TSS as in traditional gene models. However, the plurality of the promoters are in the broad (BR) shape class, which instead leads to the initiation of transcripts anywhere within a range of positions typically spanning50–100 bp (Sandelin and co-workers 2007). Carninci and co-workers (2006) observed that the BR promoters had a strong association with CpG islands, and that 90% of TATA-independent transcription initiation was associated with a CpG island, while TATA boxes were strongly associated with SP promoters. This means that CpG island association can be used to discriminate between different classes of promoters. This is not a perfect assignment, since even in the experiments of Carninci and co-workers, there are still TATA boxes in BR tag clusters and CpG islands associated with SP tag clusters.

CpG island promoters are believed to rely less on strong affinity binding to particular transcription factors and primarily to express housekeeping genes (Larsen and co-workers 1992). Tissue-specific genes, by contrast, are much more likely to be associated with CpG deserts (Larsen and co-workers 1992) and TATA boxes (Schug and co-workers 2005). In tissue-specific gene expression, transcriptional switches rely on the concentration of a single transcription factor or a small cohort of factors to produce consistent transcription in both time and space (Sandelin and co-workers 2007). This results in a more highly conserved promoter (Lee and co-workers 2005). Additionally, TATA box promoters evolve more slowly than CpG island promoters (Taylor and co-workers 2006).

3.2 Methods

3.2.1 The model

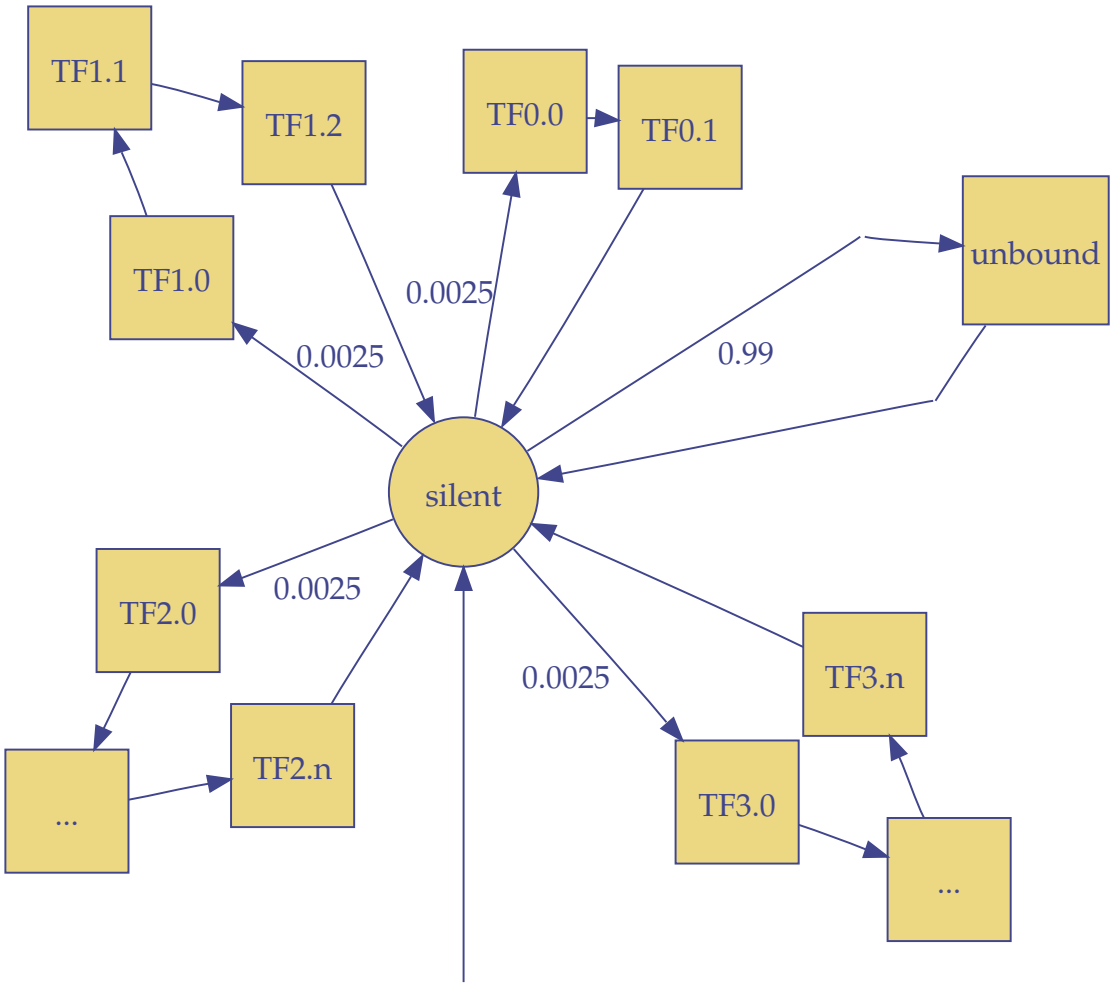


Figure 3.1: Toy example schematic of a Sunflower model for transcription factors. Circle: silent state; squares: emitting states. Arcs indicate transitions between states with nonzero probability. The transition probability is either designated by a label, or is 1 in the case of unlabeled areas. Squares labeled by ellipses represent an arbitrary number of sequential states. The arc from empty space indicates the initial state of the model.

Sunflower is implemented as a probabilistic model using a hidden Markov model approach (see section 1.3), similar to the suggestions of Rajewsky and co-workers (2002). A schematic of a very simple Sunflower model is in Figure 3.1. The model consists of a single silent state and a number of states that emit a

single nucleotide. These are arranged into a number of *petals* around the silent state.

One of these petals consists of a single emitting *unbound* state. The unbound state is trained with emission probabilities from the background distribution of unambiguous nucleotides in the sequenced genome of the target species. The other petals each represent the characteristic motif of a transcription factor. There is a one-to-one correspondence between columns of a TF position weight matrix and states in a petal. Each frequency in the PWM is transformed into an emission probability. The addition of a pseudocount prevents the possibility of zeros eliminating a particular path through the model for a given sequence. Because there are no zero emission probabilities, Sunflower takes every possible path between states into account during calculations.

In the model, every nucleotide in a sequence is either unbound or bound to one of several transcription factors. A nucleotide is bound only if it is part of a continuous subsequence of nucleotides that are all bound to the same transcription factor. This subsequence corresponds to a transcription factor binding site, and the transcription factor is more likely to bind a subsequence congruent with its motif. Nucleotides cannot bind to more than one transcription factor simultaneously, modeling the biophysical constraint of steric hindrance. The model, however, does allow for the simultaneous binding of multiple transcription factors to a DNA sequence. At the end of a particular TFBS subsequence, the model can transition to any of the other strings of bound states with equal probability, or to the unbound state with a much higher probability.

Most of the transition probabilities are set to either 0 or 1 by the model's intrinsic architecture. The only exception is the only place where a choice of path is possible—when exiting the unbound state. I have set the prior transition probability $a_{\text{silent}\to\text{unbound}}$ to 0.9, 0.99, and 0.999, for comparison in three separate computational runs. Use of these values assumes that most positions in the genome are unbound, although the stringency of this assumption is greater for higher silent-to-unbound transition probabilities. Essentially, higher $a_{\text{silent}\to\text{unbound}}$ values represent lower concentrations of each of the modeled transcription factors. While I have initially set the transition probabilities from the silent state to each TF petal to be equal, I discuss the possibility of adjusting them to reflect the relative concentrations of each transcription factor in subsection 3.4.5.

Every path through the model begins in the silent state. Before emission of the first nucleotide, the model transitions to either (a) a state representing the nucleotide unbound (with high probability), or (b) one of several states representing the nucleotide being the first of a subsequence bound to a particular transcription factor (with low probability). After a nucleotide emission in the unbound state, the model returns to the silent state. After an emission in a particular transcription-factor–bound state, the model continues through a fixed number of emitting states for the transcription factor before returning to the silent state. The example model in Figure 3.1 includes only three transcription factors: one with a motif 2 nucleotides long, another with a motif 3 nucleotides long, and two with a motif n nucleotides long.

It was necessary to limit the Sunflower model in a few ways. As stated above, the model does not allow multiple TFs to be bound to the same position in the same path. This is probably rare. Regardless, all of the TFs bind to every position in at least one path. If this is improbable, it makes a correspondingly small contribution to the final output of a posterior decoding process. The model also assumes that transcription factors only exhibit steric hindrance in a binary fashion and only within the recognition sequence of the TF. This was deemed acceptable, as to incorporate nonuniform steric hindrance one would need a markedly different model, and the data needed to parametrize it would be sparse to nonexistent today.

The model, as described here, does not incorporate information about cooperativity between transcription factors, which may play an important role in binding and and transcriptional regulation (examples in Uemura and co-workers 1997; Yuh and co-workers 2001). By avoiding the consideration of cooperativity effects in the base version of the model, one does not need to parametrize these connections, which is a difficult problem outside the scope of this work. Also, the problems become easier computationally with fewer connections between states. One can still add cooperativity effects to the base model, however, using Sunflower's extensibility, as seen in subsection 3.4.4.

3.2.2 Transcription factor affinity data

A useful property of the Sunflower model is that the emission probabilities of a string of bound states are equivalent to the position weight matrix of a

transcription factor. There are several databases of transcription factor PWMs, but I chose to parametrize the model with JASPAR CORE 2006 (Vlieghe and co-workers 2006), a high-quality, curated set of transcription factor binding profiles determined either from SELEX experiments or experimentally-observed binding to genomic binding sites.

The diagrams in the previous subsection are just toy examples. The graph of the actual model used in the rest of this thesis is produced by taking all of the matrices for vertebrate TFs in JASPAR CORE. It contains 89 transcription factors (listed in Table 3.1), and 2000 states. These states consist of the silent state, the unbound state, and a plus strand state and a minus strand state for each of the 999 PWM columns. While most of the analysis here is shown on human sequence, the model includes affinities determined by experiments on mouse homologs of human transcription factors. This assumes that TFs in one species recognize similar binding sites in related species.

Table 3.1: JASPAR CORE vertebrate transcription factors.

Name	Class	Taxon	Method
Ar	nuclear receptor	*Rattus rattus*	SELEX
Arnt	bHLH	*Mus musculus*	SELEX
Arnt-Ahr	bHLH	*Mus musculus*	SELEX
Bapx1	homeo	*Mus musculus*	SELEX
c-ETS	ETS	*Gallus gallus*	SELEX
cEBP	bZIP	——	compiled
Chop-cEBP	bZIP	*Rattus norvegicus*	SELEX
CREB1	bZIP	*Homo sapiens*	SELEX
deltaEF1	Zn-finger, C2H2	*Gallus gallus*	SELEX
E2F1	——	*Homo sapiens*	compiled
ELK1	ETS	*Homo sapiens*	SELEX
ELK4	ETS	*Homo sapiens*	SELEX
En1	homeo	*Mus musculus*	SELEX
ESR1	nuclear	——	compiled
Evi1	Zn-finger, C2H2	*Mus musculus*	SELEX
Fos	bZIP	*Mus musculus*	SELEX
Foxa2	forkhead	*Rattus norvegicus*	compiled
FOXC1	forkhead	*Homo sapiens*	SELEX
FOXD1	forkhead	*Homo sapiens*	SELEX
Foxd3	forkhead	*Rattus norvegicus*	SELEX
FOXF2	forkhead	*Homo sapiens*	SELEX
FOXI1	forkhead	*Homo sapiens*	SELEX

Table 3.1: (continued)

Name	Class	Taxon	Method
FOXL1	forkhead	*Homo sapiens*	SELEX
Foxq1	forkhead	*Rattus norvegicus*	SELEX
GABPA	ETS	*Homo sapiens*	compiled
Gata1	Zn-finger, GATA	*Mus musculus*	SELEX
GATA2	Zn-finger, GATA	*Homo sapiens*	SELEX
GATA3	Zn-finger, GATA	*Homo sapiens*	SELEX
Gfi	Zn-finger, C2H2	*Rattus norvegicus*	SELEX
HAND1-TCF3	bHLH	*Homo sapiens*	SELEX
HLF	bZIP	*Homo sapiens*	SELEX
HNF4	nuclear	——	compiled
IRF1	TRP-cluster	*Homo sapiens*	SELEX
IRF2	TRP-cluster	*Homo sapiens*	SELEX
Klf4	Zn-finger, C2H2	*Mus musculus*	SELEX
MafB	bZIP, MAF	*Rattus norvegicus*	SELEX
MAX	bHLH-ZIP	*Homo sapiens*	SELEX
MEF2A	MADS	*Homo sapiens*	SELEX
Myb	TRP-cluster	*Mus musculus*	SELEX
MYC-MAX	bHLH-ZIP	*Homo sapiens*	SELEX
Mycn	bHLH-ZIP	*Mus musculus*	SELEX
Myf	bHLH	*Homo sapiens*	compiled
NF-kappaB	REL	Vertebrata	compiled
NFIL3	bZIP	*Homo sapiens*	SELEX
NFKB1	REL	*Homo sapiens*	SELEX
NHLH1	bHLH	*Homo sapiens*	SELEX
Nkx2-5	homeo	*Mus musculus*	SELEX
NR1H2-RXR	nuclear receptor	*Homo sapiens*	SELEX
NR2F1	nuclear receptor	*Homo sapiens*	compiled
NR3C1	nuclear	——	compiled
Pax2	paired	*Mus musculus*	SELEX
Pax4	paired-homeo	*Mus musculus*	SELEX
Pax5	paired	*Mus musculus*	compiled
Pax6	paired	*Homo sapiens*	SELEX
Pbx	homeo	*Homo sapiens*	SELEX
PPARG	nuclear receptor	*Homo sapiens*	SELEX
PPARG-RXRA	nuclear receptor	*Homo sapiens*	SELEX
Prrx2	homeo	*Mus musculus*	SELEX
REL	REL	*Homo sapiens*	SELEX
RELA	REL	*Homo sapiens*	SELEX
Roaz	Zn-finger, C2H2	*Rattus norvegicus*	SELEX
RORA	nuclear receptor	*Homo sapiens*	SELEX

Table 3.1: (continued)

Name	Class	Taxon	Method
RORA1	nuclear receptor	*Homo sapiens*	SELEX
RREB1	Zn-finger, C2H2	*Homo sapiens*	SELEX
RUNX1	RUNT	*Homo sapiens*	SELEX
RUSH1-alfa	Zn-finger, GATA	*Oryctolagus cuniculus*	SELEX
RXR-VDR	nuclear receptor	*Homo sapiens*	SELEX
Sox17	HMG	*Mus musculus*	SELEX
Sox5	HMG	*Mus musculus*	SELEX
SOX9	HMG	*Homo sapiens*	SELEX
SP1	Zn-finger, C2H2	*Homo sapiens*	SELEX
SPI1	ETS	*Homo sapiens*	SELEX
SPIB	ETS	*Homo sapiens*	SELEX
Spz1	bHLH-ZIP	*Mus musculus*	SELEX
SRF	MADS	*Homo sapiens*	SELEX
SRY	HMG	*Homo sapiens*	SELEX
Staf	Zn-finger, C2H2	*Xenopus laevis*	compiled
T	T-box	*Mus musculus*	SELEX
TAL1-TCF3	bHLH	*Homo sapiens*	SELEX
TBP	TATA-box	——	——
TCF1	homeo	Vertebrata	compiled
TCF11-MafG	bZIP	*Gallus gallus*	SELEX
TEAD	TEA	*Homo sapiens*	compiled
TFAP2A	AP2	*Homo sapiens*	SELEX
TP53	p53	*Homo sapiens*	SELEX
USF1	bHLH-ZIP	*Homo sapiens*	SELEX
YY1	Zn-finger, C2H2	*Homo sapiens*	compiled
ZNF42_1-4	Zn-finger, C2H2	*Homo sapiens*	SELEX
ZNF42_5-13	Zn-finger, C2H2	*Homo sapiens*	SELEX

3.2.3 Data structures

A program called pwm2sfl generates the matrices in the "model" and "optimization" categories of Table 3.3 from position weight matrix input. A brief description of this process follows.

Let there be an input set of μ position weight matrices on an alphabet $\mathcal{A}$. For each PWM, generate another PWM for the reverse complement of the motif, leading to $\mu_{\text{revcom}} = 2\mu$ matrices with an aggregate total of m_0 columns. Then, the number of states m is $m_0 + 2$ after the addition of silent and unbound states. The emission probability matrix $\mathbf{E}$ has dimensions $m \times |\mathcal{A}|$. The transition

Table 3.2: Descriptions of dimensions of matrices used in Sunflower.

dimension	Value[a]	Description
$\lvert\mathcal{A}\rvert$	4	number of letters in alphabet $\mathcal{A}$
μ	89	number of real transcription factors used to produce model
μ_{revcom}	178	$= 2\mu$, number of transcription factors in model, including generated reverse complement TFs
m	2000	number of states
n	2000	length of nucleotide sequence $\boldsymbol{X}$

[a] This is the value used throughout this thesis, except where otherwise specified.

Table 3.3: Dimensions of vectors, sets, and matrices used in Sunflower.

Use	Name	Dimensions[a]
model	$\mathcal{A}$	$\lvert\mathcal{A}\rvert$ nucleotides
	$\mathcal{K}$	k states
	$\mathbf{A}$	$m \times m$ floats
	$\mathbf{E}$	$\lvert\mathcal{A}\rvert \times m$ floats
optimization	$\mathbf{c}_\text{f}$	m sets
	$\mathbf{c}_\text{b}$	m sets
input	$\boldsymbol{X}$	n nucleotides
intermediate	$\mathbf{F}$	$m \times n$ floats
	$\mathbf{B}$	$m \times n$ floats
output	$\mathbf{P}$	$m \times n$ floats

[a] Quantities more fully described in Table 3.2.

probabilities are stored in an $m \times m$ matrix, $\mathbf{A}$.

Two vectors of m connection sets each, the forward connection sets $\mathbf{c}_\text{f}$ and the backward connection sets $\mathbf{c}_\text{b}$, are calculated from $\mathbf{A}$. These sets indicate which arcs connect states, and are used by an optimization in the algorithm. If and only if there is a connection between two states k and l such that the transition probability $a_{k\to l}$ is nonzero, then k is in the connection set $c_{\text{f},l}$ and l is in the connection set $c_{\text{b},k}$.

The transition probability data structure is inefficient for the sort of sparse data it holds, as it takes $O(n^2)$ space. It has the advantage of extensibility, however, as described in subsection 3.4.4. The simplicity has been a boon in implementation while the inefficiency has not presented a roadblock so far.

Instead of storing and acting on actual floating point probabilities, their natural logarithms are used. This allows me to obtain the result of multiplying these small values by each other many times by adding their logarithms without fear of an underflow error (Durbin and co-workers 1998).

3.2.4 Algorithm

Sunflower does posterior decoding using an algorithm I call SUNFLOWER-REFERENCE($\mathbf{A}, \mathbf{E}, \boldsymbol{X}, \mathbf{c}_\mathrm{f}, \mathbf{c}_\mathrm{b}, \mathbf{F}, \mathbf{B}, i$). The algorithm calculates the posterior probability $P_{k,i} = P(x_i \mid k)$ that a particular nucleotide x_i was emitted by a given state k in the Sunflower model. The results are the same as the FORWARD-BACKWARD($\mathbf{A}, \mathbf{E}, \boldsymbol{X}$) algorithm from subsection 1.3.2 when the silent state is the start state ($k_\mathrm{silent} = 0$).

Posterior decoding is equivalent to tracing all of the pathways through this model that can emit a single sequence, and estimating the posterior probability that the model is in each of the states at each position of that sequence. Underlying the model is the physical mechanism that transcription factors are continuously binding and leaving chromosomal sequences, at a rate related to their affinity for the sequence. The statistical mechanics of the biophysical model are approximated by the probabilities that a transcription factor is bound in the sequence model. Indeed, PWMs, which are frequently thought of as purely probabilistic concepts, were originally proposed as part of a statistical mechanics model (Berg and von Hippel 1987).

The new parameters in SUNFLOWER-REFERENCE allow two optimizations. The first is the use of the connection set vectors $\mathbf{c}_\mathrm{f}$ and $\mathbf{c}_\mathrm{b}$ to relieve the algorithm from the necessity in each round of doing calculations involving transition probabilities of zero. The other optimization is that one can specify i to indicate that the intermediate matrices $\mathbf{F}$ and $\mathbf{B}$ have already been partially calculated, such that recalculation is only necessary in the forward direction for values $> i$, and in the reverse direction for values $< i$. This partial calculation facility is used in chapter 4.

I wrote Sunflower in the Python language (van Rossum 2006) and inner loops in the C language (Kernighan and Ritchie 1988) for speed (see subsection 3.3.4).

3.2.5 Promoter sequence

For each human protein-coding gene in Ensembl 47 (Flicek and co-workers 2008; http://oct2007.archive.ensembl.org/Homo_sapiens/; build NCBI36), I downloaded genomic sequence for ± 1000 bp from the TSS of transcripts which (a) had a 5′ UTR of at least 40 bp in length, and (b) contained no ambiguous sequence in the downloaded region. When multiple transcripts from a gene met these criteria, I selected the one with the longest 5′ UTR. I stored the sequences for all the genes in a single Fasta-formatted file.

The aim of the selection procedure was to include only transcript models where I had some expectation that the TSS prediction might be based on experimental data and therefore accurate. Certainly the genes with shorter or nonexistent UTRs are more likely to be based on prematurely truncated cDNAs rather than full-length transcripts.

3.2.6 Output storage

Direct analysis

Sunflower produces substantial output. A double precision floating point number for each posterior probability multiplied by 2000 states, 2000 positions, and 17,600 genes quickly adds up. I used HDF5 (HDF Group 2007) and PyTables (Altet and co-workers 2007) for efficient and portable output storage.

The output files can either be directly loaded as arrays into Python, into R (R Development Core Team 2007) using the `hdf5` package, or into any other environment that supports HDF5. Sunflower also includes a Sunreport program that produces tab-delimited text output files from the HDF5 files. It has a variety of features, including being able to produce mean and standard deviation statistics or histograms grouped either by transcript or by position ranges. It can also separate positions by whether they are conserved in a provided sequence alignment.

Sunreport and R were used to produce the rest of the figures in this chapter and many in subsequent chapters.

Whole genome

The previous data structure is convenient for direct analysis, but produces data files too large for whole-genome runs. The expected uses for whole-genome analysis, however, are somewhat different than the expected uses for direct analysis of a single promoter or even all promoters. In the whole-genome case, end users may be interested primarily in visual exploration of the predicted probabilities through the medium of a genome browser. Therefore if one can assume that the user never wants to view a transcription factor posterior probability track with a height of more than 256 pixels[1], one can safely downsample 64-bit floating point numbers to unsigned 8-bit integers P_{uint8} such that the original posterior probability $P = P_{\text{uint8}}/255$.

When downsampling its output, Sunflower frequently produces runs of particular values, especially long runs of zeros. In genome-scale mode, Sunflower uses a form of run-length encoding to store these more efficiently. It stores these in a MySQL database that can be easily served by a Distributed Annotation System (DAS; Prlić and co-workers 2007) package such as ProServer (Finn and co-workers 2007).

3.2.7 CpG island identification

Human promoters can be divided into two groups by the prevalence of CpG dinucleotides in their promoters (Saxonov and co-workers 2006), which correlate well with the BR shape class, as discussed in subsection 3.2.1.

To discover CpG islands, I use the default parameters from Newcpgreport in the EMBOSS package (Rice and co-workers 2000), which I explain briefly. First, let us define $q_{\boldsymbol{Z}}$ on any sequence to be the number of subsequences that begin with the sequence $\boldsymbol{Z}$, allowing overlapping. I then scan an input sequence $\boldsymbol{X}$ for a CpG island subsequence $\boldsymbol{X}_{\text{CpG}} = (x_i \ldots x_{i+l-1})$ where

(a) $l \geq 200$; and

(b) over an average of 10 sliding subsequences $\{(x_j \ldots x_{j+99}), \ldots, (x_{j+9} \ldots x_{j+108})\}$ defined for every nucleotide $x_j \in \boldsymbol{X}_{\text{CpG}}$,

[1] This is 256, and not 255, because a single pixel indicates a zero value. All pixels off indicates missing data.

(1) $\frac{q_{G+C}}{l} \geq 0.5$; and

(2) $\frac{q_{CG}}{q_C q_G} \geq 0.6$.

Transcripts with at least one CpG island within ± 1000 bp of the TSS are categorized as "CpG island" transcripts, and those that do not are called "CpG desert" transcripts.

3.3 Results

3.3.1 Single transcripts[2]

To examine Sunflower's output, it seems sensible to start with results on a characteristic specimen transcript. For this purpose, I have chosen ENST00000344265, an Ensembl transcript for *POU1F1*, POU class 1 homeobox 1, a gene on human chromosome 3p11.2. The product of this gene is a POU protein, a class of DNA-binding proteins with two helix-turn-helix domains: a homeodomain and a POU domain. These two domains probably work together to bind two motifs in a segment of duplex DNA (Brown 2006), and influence pituitary development and neural fate (Gilbert 2000). In this section I conservatively consider only the ± 700 bp around the TSS in order to avoid any possibility of edge effects.

First, consider the posterior probability of the unbound state $P(\pi_i = \text{unbound} \mid \boldsymbol{X})$ for three different values of $a_{\text{silent}\to\text{unbound}}$, as shown in Figure 3.2. When $a_{\text{silent}\to\text{unbound}}$ is 0.999, this posterior probability varies little with a range of 0.954 to 0.999. When $a_{\text{silent}\to\text{unbound}}$ is only 0.9, variation is extreme, ranging from 0.089 to 0.935. When $a_{\text{silent}\to\text{unbound}} = 0.99$, a middle ground appears. While most positions show a relatively high posterior probability in the unbound state (median 0.95), there are still significant troughs (minimum 0.67). While the $a_{\text{silent}\to\text{unbound}} = 0.99$ model results in some clipping for the positions where the probability of being unbound is high, the troughs of the greatest magnitude are preserved. This means that detection of significant transcription factor binding is preserved, although the reported magnitude is less than that reported for $a_{\text{silent}\to\text{unbound}} = 0.9$. Detection of less significant and less likely transcription factor binding, however, is decreased. Noise effects in estimation of unbound state posterior probabilities decrease greatly—the median absolute deviation

[2] Results in this subsection were produced jointly with Alison Meynert.

(MAD; Huber 1981) in $P(\pi_i = \text{unbound} \mid \boldsymbol{X})$ where $a_{\text{silent}\to\text{unbound}} = 0.99$ is 0.03, while for $a_{\text{silent}\to\text{unbound}} = 0.9$, it is 0.2. As discussed further in subsection 3.3.2, these results also hold for most transcripts. For these reasons, I have generally used $a_{\text{silent}\to\text{unbound}} = 0.99$ in the rest of this thesis.

Figure 3.3 displays the posterior probability that TATA binding protein (TBP) has bound to various positions on ENST00000344265. The posterior probability displayed is actually the sum of the first state of the PWM, TBP.f0, and the first state of its reverse complement, TBP.r0. Since the TBP motif is 15 bp wide, this means a peak indicates the probability of TBP binding there and the 14 downstream positions, on either strand. Because the posterior probability of TF binding state k at position i is roughly equal to the posterior probability of binding state $k+1$ at position $i+1$, one does not need to examine more than one state of a transcription factor. I use the most upstream state as a proxy for the rest of the transcription factor states. A peak in this state indicates a high likelihood that the next few positions are bound.

These peaks can be considered areas of likely binding by the transcription factor. The probability of binding is considered to be related to the affinity of the examined transcription factor for the sequence.

The probability that TBP has bound in most positions is zero. In fact, there are only 5 positions where the probability is greater than 1 percent that the model is entering the TBP petal. Still, the probability doesn't reach above its highest value at 0.11. It is important to consider that transcription factors do not bind certain positions all of the time. This is why Sunflower uses a continuous approach rather than the discrete threshold approach of some previously existing packages (Hughes and co-workers 2000; Lenhard and Wasserman 2002; Markstein and co-workers 2002). In positions where TFs bind with high probability, just how highly probable is an important consideration, as the probability or affinity of binding may have a proportional effect on expression (Bintu and co-workers 2005). Having a continuous probabilistic value for describing TF binding rather than a discrete value becomes even more important when the information is used as input to other probabilistic calculations, such as trying to find *cis*-regulatory modules, or the mutation-analyzing method in chapter 4.

Figure 3.4 displays the posterior probabilities that the model is entering the petal for every analyzed TF. The TBP figure of Figure 3.3 is scaled down and shown as one of 89 panels. The individual posterior decoding plots of

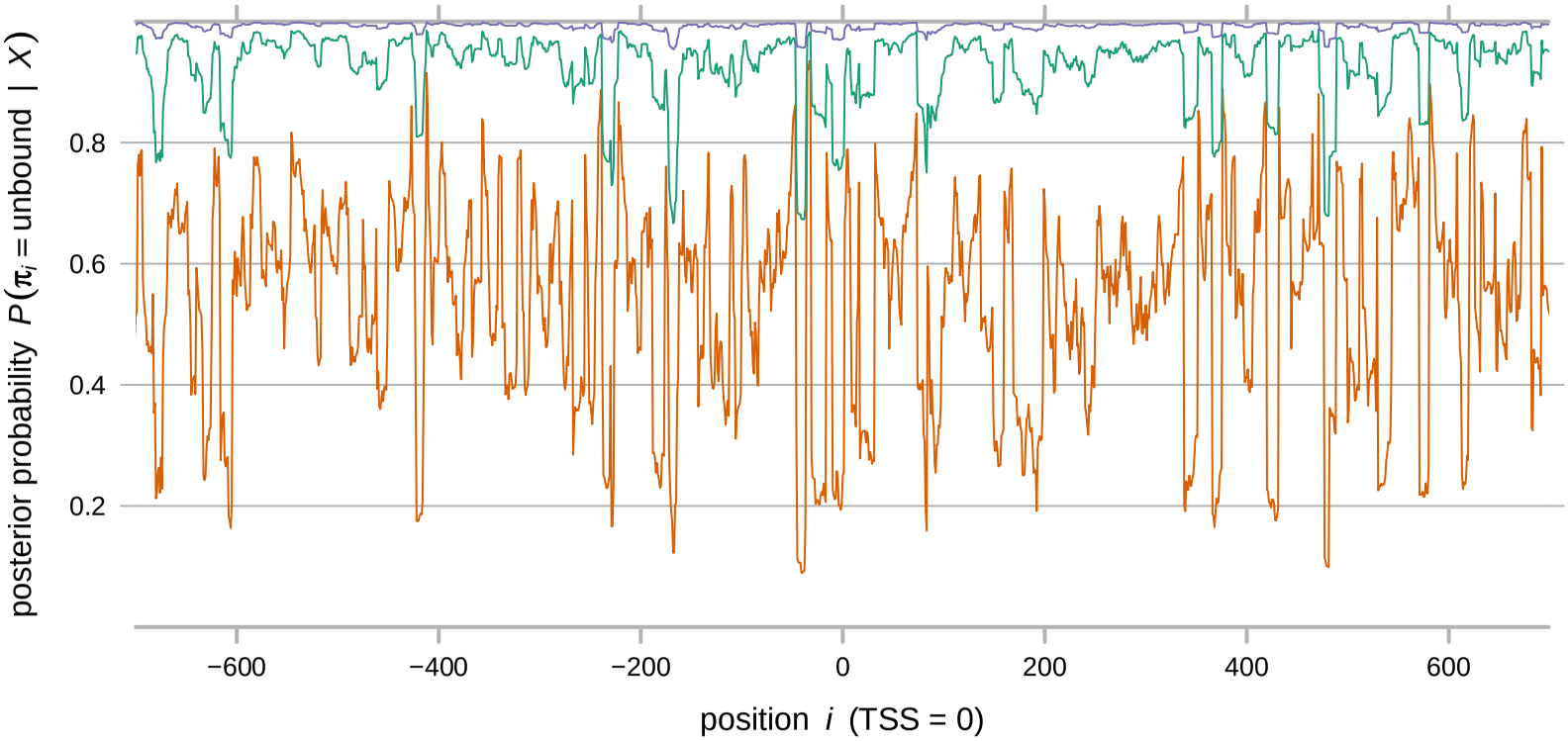


Figure 3.2: Posterior probabilities that the Sunflower model is in the unbound state at 1400 positions flanking the transcription start site of ENST00000344265, for three different models where $a_{\text{silent}\to\text{unbound}}$ is 0.9 (orange), 0.99 (teal), or 0.999 (purple).

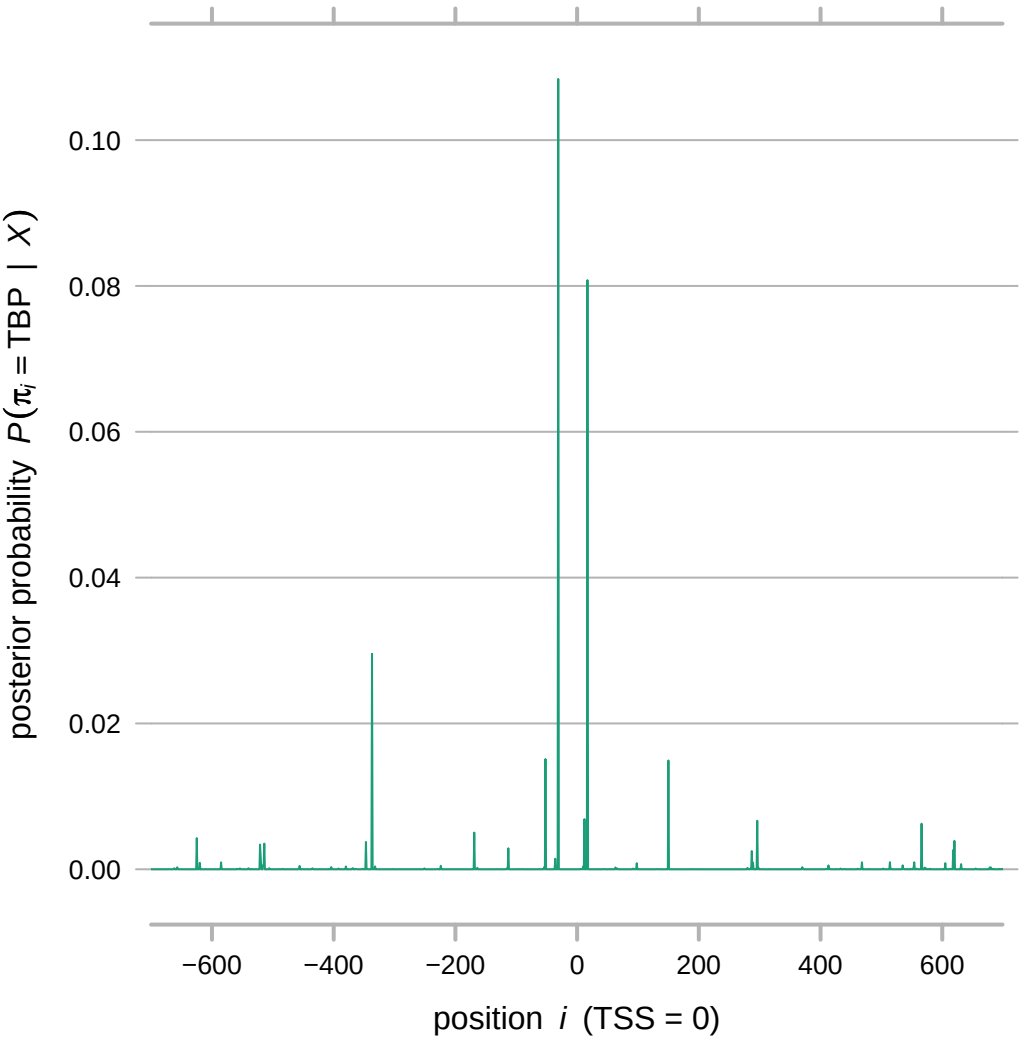


Figure 3.3: Posterior probabilities that the Sunflower model is entering a string of states representing the TBP motif at 1400 positions flanking the transcription start site of ENST00000344265. The probabilities shown here are the sums of the probabilities for the first forward and first reverse state of TBP. This means that a peak indicates a prediction of TBP bound at the peak's position and 14 bp downstream.

transcription factor states consist mainly of a small number of sharp peaks, or even no peaks, This is a result of the model architecture and the uniform low transition probabilities from the silent state into an initial transcription factor state ($^{1-0.99}/_{2\cdot 89} = 0.00006$). The TF transition probability $a_{\text{silent}\to k}$ where k is the first state of any TF petal can stand in for the concentration of the TF, or the $a_{\text{silent}\to\text{unbound}}$ value can stand in for the amount of competition a TF faces—higher values mean that there is more nonspecific competition, and therefore less TF binding, especially at weak binding sites.

The results for ENST00000344265 are typical amongst other transcripts in that they include a small number of well-defined peaks for a limited number of transcripts. Despite the continuous nature of the output data, it would still be obvious to researchers which peaks to focus on for further investigation.

3.3.2 Aggregation of transcripts[3]

Repeating the above analysis for all 17,600 transcripts produces the data displayed in Figure 3.5, which shows the mean posterior probabilities for the beginning of each petal averaged over each transcript. While most of the transcription factors have rather flat aggregate profiles, with low variance, indicating that they either have low incidence of high-affinity binding, (for example, GATA3, Nkx2-5, FOXL1, Prrx2, NFIL3), or uniform binding to the TSS-flanking region with little dependence on position (En1, FOXC1, RUSH1-alfa, Gata1, GATA2). Others show higher variances, with characteristic patterns of elevated binding around the TSS (TFAP2A, Pax5, RREB1, ESR1, Myf) with sometimes quite sharp changes in binding probability from one side of the TSS to the other. The division of the group into low- and high-variance aggregate binding profiles can be confirmed by examining the density of the distribution of MADs for each TF (Figure 3.6).

The variance of the aggregated by-position distributions is rather high. This is not unexpected as it is unlikely that all genes would share the same binding pattern exactly. The estimate of the mean itself, however, is robust due to the large number of data points. Error bars derived from the standard error of the mean would be too small to see given the large number of data points.

In many of these aggregate panels one can see characteristic patterns. The most striking is that of TATA binding protein (TBP). In genes that have TBP-

[3] Results in this subsection were produced jointly with Alison Meynert.

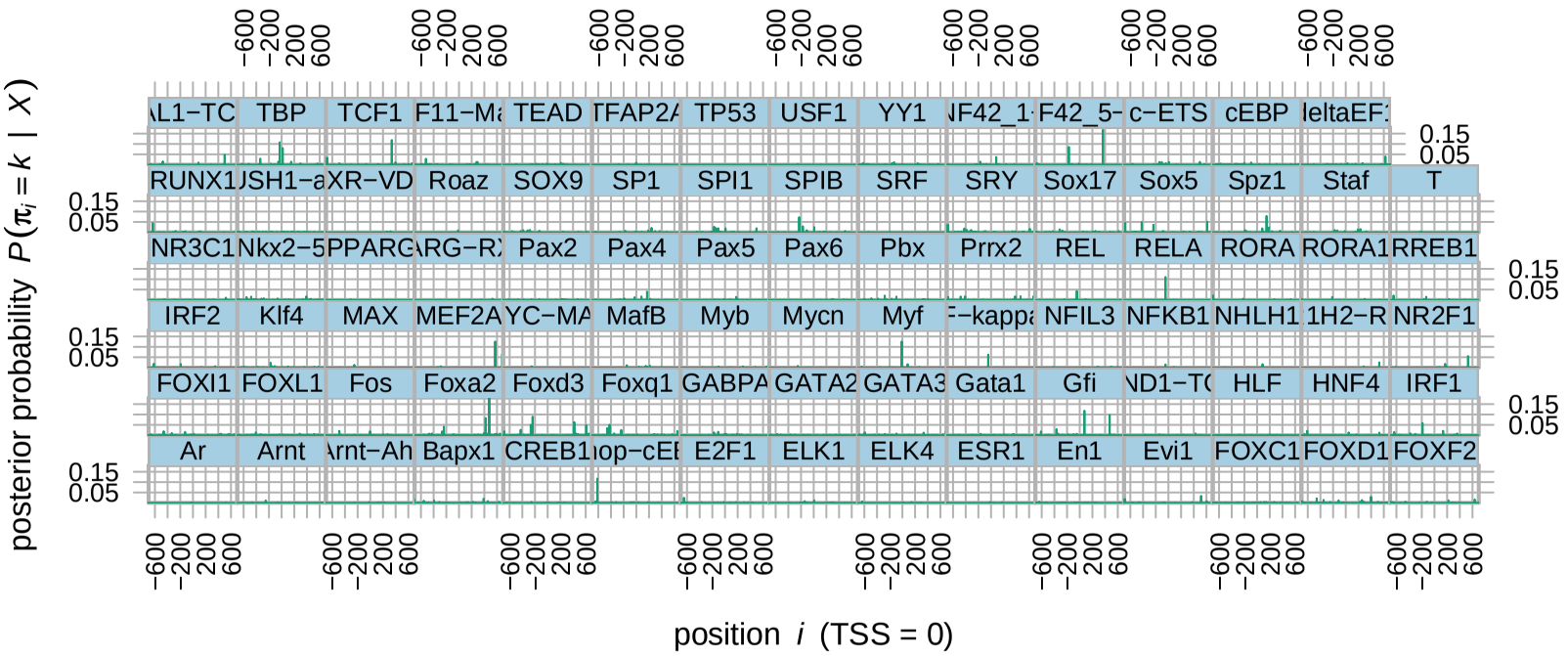


Figure 3.4: Posterior probabilities that the Sunflower model is entering a string of states representing one of 89 different motif at 1400 positions flanking the transcription start site of ENST00000344265. This figure shows a trellis of line plots corresponding to the unbound state and the first states of the 89 vertebrate transcription factors with binding profiles in JASPAR CORE. Each plot shows posterior probabilities as in Figure 3.3, and the TBP panel shows the same data that was shown in that figure.

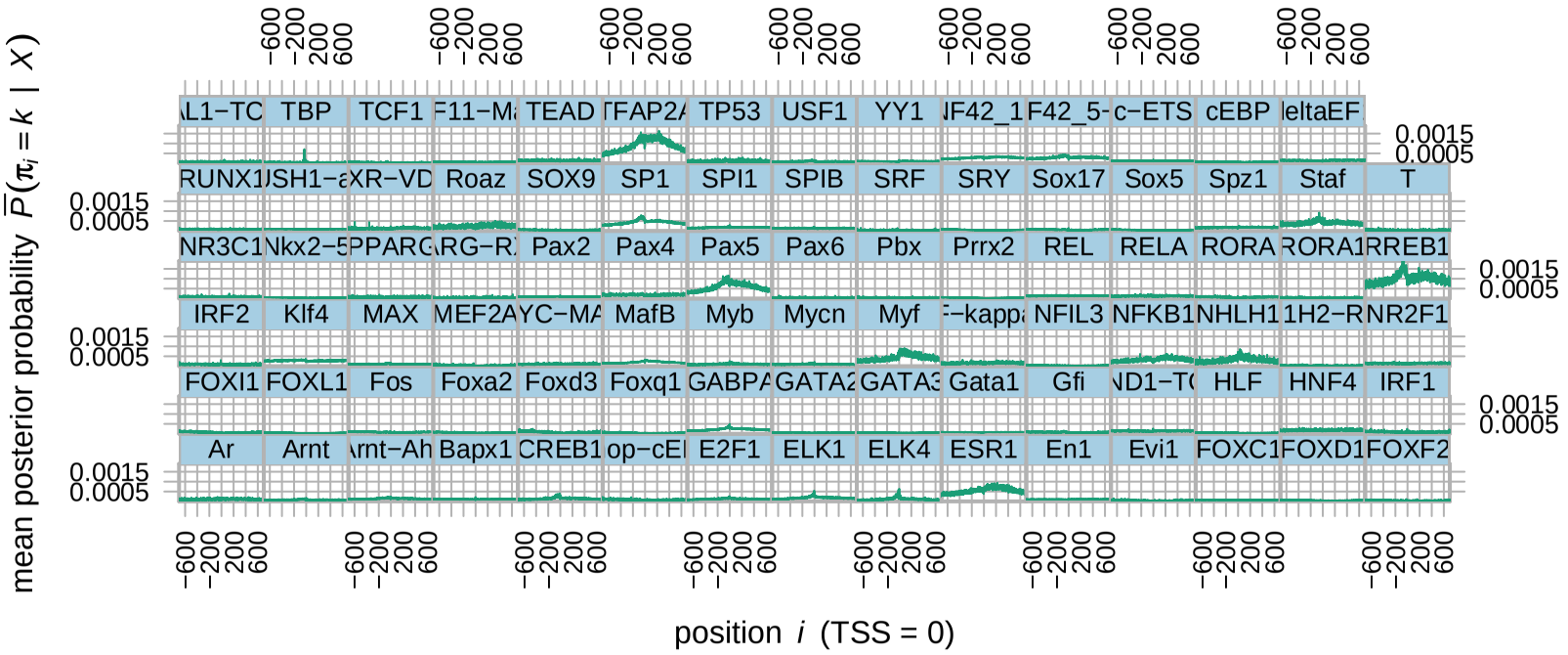


Figure 3.5: Mean posterior probabilities that the Sunflower model is entering a string of states representing one of 89 different motif at 1400 positions flanking the transcription start sites of 17,600 human transcripts. Probabilities are as in Figure 3.3, except that they are averaged by position over the 17,600 TSS-flanking regions.

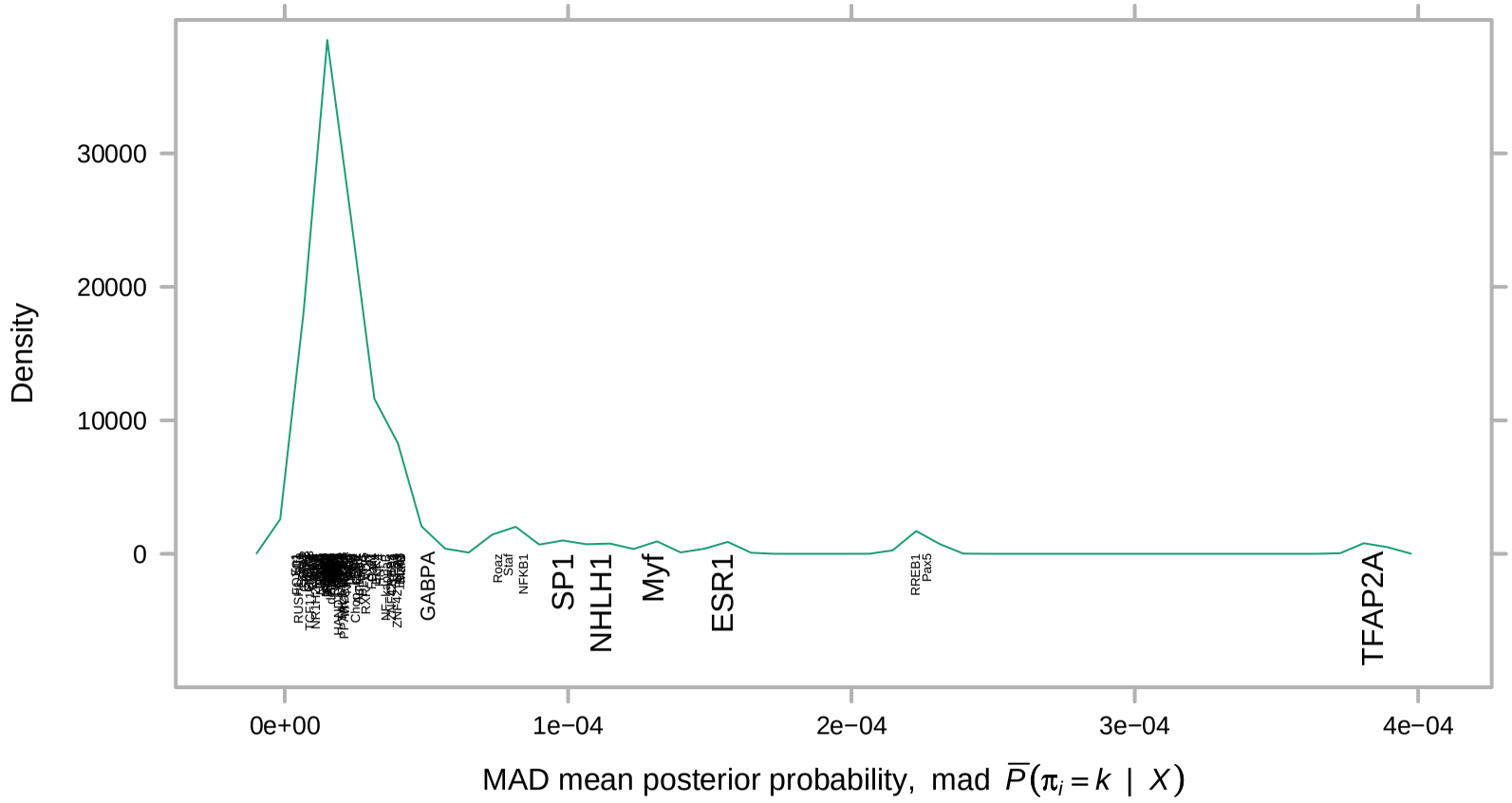


Figure 3.6: Kernel density plot of the MAD of each TF's distribution of mean posterior probability values at $\pm$ 700 bp. Some labels were made smaller to reduce overlapping.

dependent transcription initiation there is a sharp peak at approximately -45 bp relative to the TSS. These genes tend to be in CpG deserts, as transcripts associated with methylated cytosine are indicative of CpG islands.

Repeating the analysis with different start and end points gives the same results (data not shown). The peak is at -45 relative to the TSS, not relative to midway through the aligned sequence. This is good evidence that Sunflower is showing something biological rather than an artifact of the particular algorithms being used.

By segregating the transcripts on CpG classification, and then taking the mean posterior probabilities by TF and position within each group, we get Figure 3.7 for ± 700 bp around the TSS. Figure 3.8 displays the same data for the smaller region ± 100 bp around the TSS. The number of transcripts in each category is:

11822	CpG island
5778	CpG desert

The factor that reaches the highest binding probabilities in aggregate is TFAP2A, transcription factor AP-2α, an important factor in vertebrate embryonic development, which plays a role in cell-type–specific growth and inhibition of terminal differentiation (Eckert and co-workers 2005). Figures 3.7 and 3.8 show that this effect is accentuated in transcripts associated with CpG islands. The reason is obvious if one examines the sequence logo, shown in Figure 3.9. The PWM was determined by McPherson and Weigel (1999), who assigned it the consensus sequence `GCCCBVGGS`. Unusually, the first three columns all contain two bits of information and allow only one nucleotide. This is because they have counts of 146 for one nucleotide, and zero for the other nucleotides. Since CpG islands necessarily have elevated G+C content (see subsection 3.2.7), they also have a much higher likelihood of containing the sequence `GCC`, which all but ensures that the Sunflower model transitions into the TFAP2A petal. This is even more true in the area immediately surrounding the TSS.

This PWM has an overabundance of zero counts because the researchers in the original study presupposed these results by selecting for a biased sequence $\texttt{GCCN}_6$. This eliminated the noise in the first three positions that one would expect even in a TF with very high affinity for a sequence. One could solve this problem by omitting the troublesome matrix. If it were deemed important to accurately

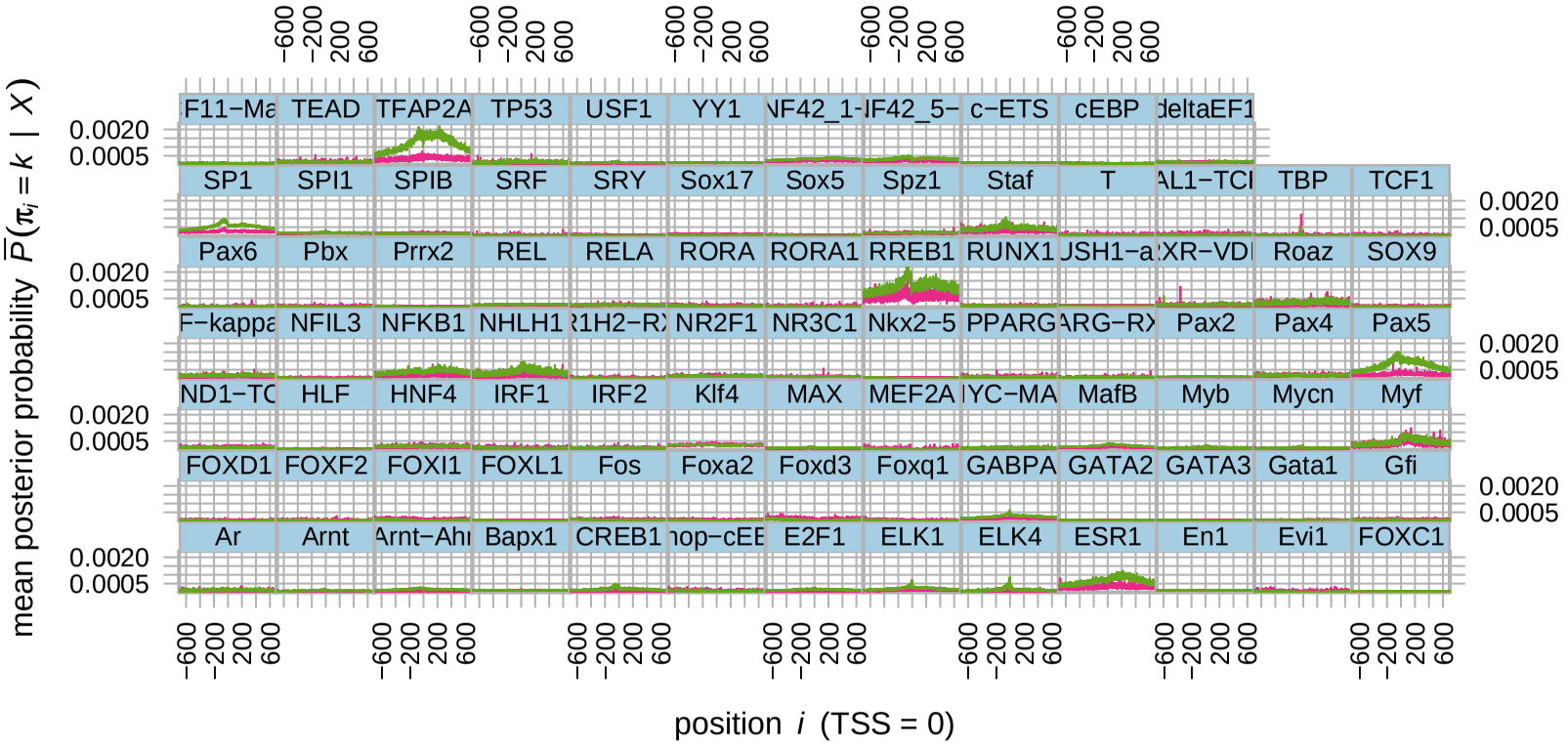


Figure 3.7: Mean posterior probabilities that the Sunflower model is entering a string of states representing one of 89 different motif at 1400 positions flanking the transcription start sites of 17,600 human transcripts, factored on CpG association. Probabilities are as in Figure 3.5. There are two lines in each panel that represent the average of only promoters containing CpG islands (lime) or in CpG deserts (magenta).

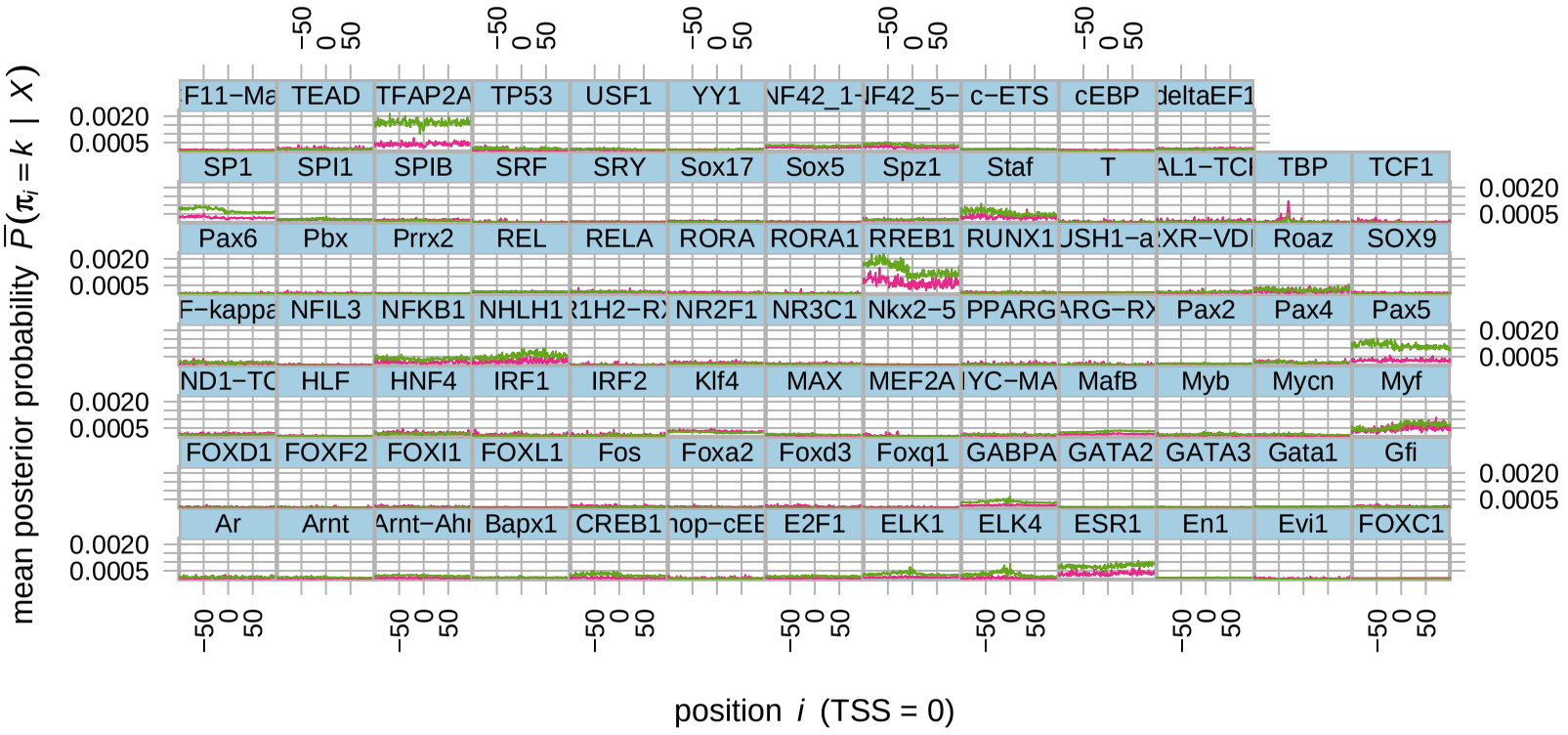


Figure 3.8: Mean posterior probabilities that the Sunflower model is entering a string of states representing one of 89 different motif at 200 positions flanking the transcription start sites of 17,600 human transcripts, factored on CpG association. Probabilities are as in Figure 3.7.

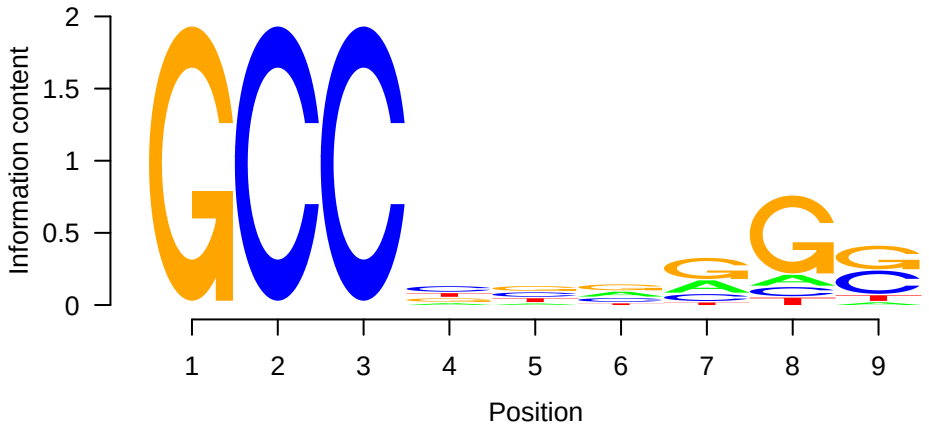


Figure 3.9: Sequence logo for TFAP2A, generated by the seqLogo package of Bioconductor (Gentleman and co-workers 2004).

detect TFAP2A binding sites, however, one could add pseudocounts to lower the information content of the first three columns of the PWM. The eighth column was determined experimentally with counts of (A=12, C=9, G=9, T=116), so the columns with counts of 146 in one cell and 0 in the others could be reduced to a similar level of information content by reducing the strong cell to a count of 116 and setting the others to 10.

SP1, specificity protein 1, is a ubiquitously expressed C_2H_2 zinc finger and another outlier in terms of high aggregate binding. Like TFAP2A, it binds more on CpG island transcripts and has a PWM with high-information G/C columns (positions 3 and 4 in Figure 3.10). Unlike TFAP2A, the original assay that led to the PWM was unbiased (Thiesen and Bach 1990), and since it consisted of only eight sequences, single pseudocounts reduced the information content of positions 3 and 4 far more than they could to the first three positions of TFAP2A, which had 146 counts. Additionally, SP1 has a known biological relationship with CpG islands (Bouwman and Philipsen 2002) and expression that is less tissue-specific (Schug and co-workers 2005). For transcription of at least some genes (Li and co-workers 2008), it is dependent on tissue-specific CpG demethylation. SP1 recruits TBP in genes lacking a TATA box (Sandelin and co-workers 2007). Additionally, researchers have identified the SP1 GC-box motif as preventing CpG

methylation (Brandeis and co-workers 1994). It also has been shown to promote transcription up to 1700 bp away from a TSS (Courey and co-workers 1989), consistent with the broad range of binding positions predicted by Sunflower across the genome.

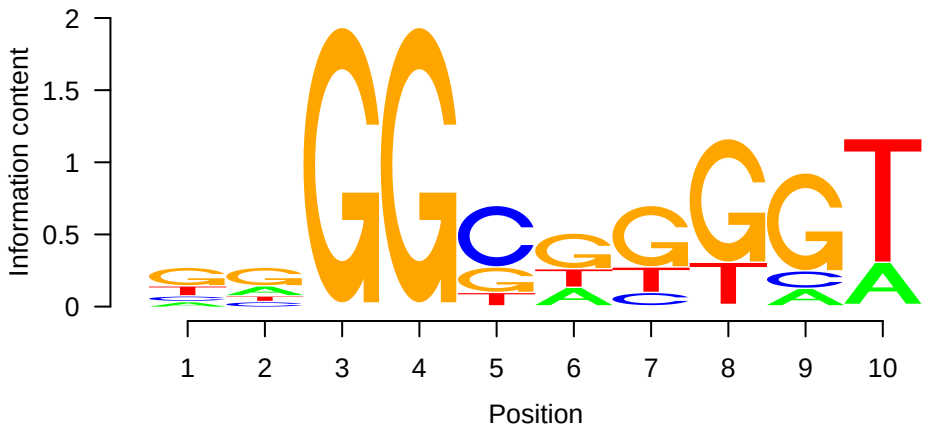


Figure 3.10: Sequence logo for SP1, generated by the seqLogo package of Bioconductor (Gentleman and co-workers 2004).

There are a number of forkhead box transcription factors included in this analysis, with aggregate binding profiles shown in Figure 3.11. The binding they show in aggregate is an order of magnitude less than the more promiscuous transcription factors discussed above. This is indicative of transcription factors which drive tissue-specific gene expression, such as the forkhead box TFs (Gilbert 2000).

Note that there are two different petals that seem to represent NF-κB, "NFKB1" (Figure 3.12) and "NF-kappaB" (Figure 3.13). One may see in Table 3.1 that this is because the former is based on SELEX experiments with the human protein (Kunsch and co-workers 1992), and the other manually compiled across all vertebrates (Grilli and co-workers 1993). Although JASPAR CORE is supposed to be non-redundant, this not always true if you take PWMs from large taxa such as the whole vertebrate subphylum. The observed motifs are very similar, and are indicative of how PWMs from one vertebrate species can be used to analyze another. The similar matrices may cause problems as they compete against each

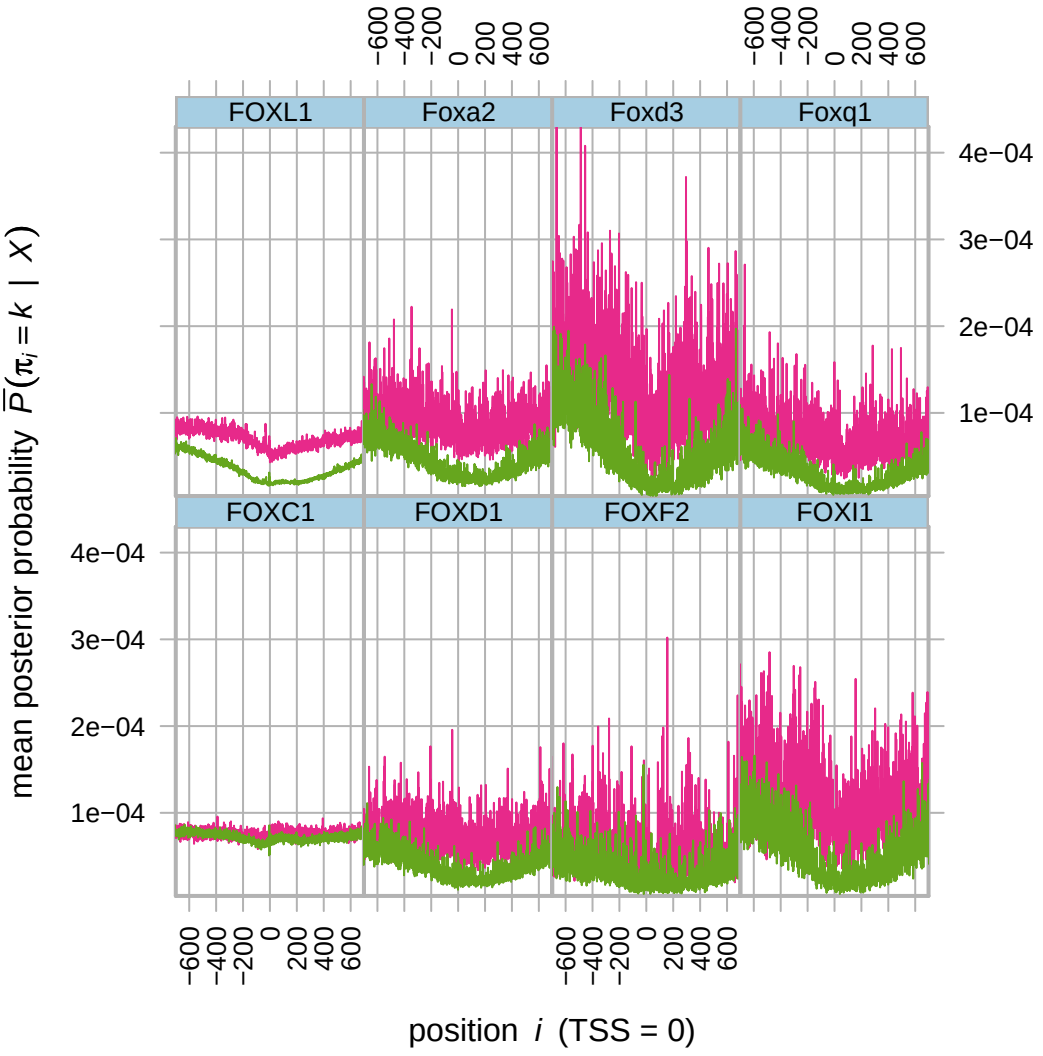


Figure 3.11: Mean posterior probability of forkhead box transcription factor binding, averaged over 17,600 human transcripts. There are two lines in each panel that represent the average of only promoters containing CpG islands (lime) or in CpG deserts (magenta).

other for similar binding sites. In principle, this problem could be solved by making a more non-redundant set of transcription factors.

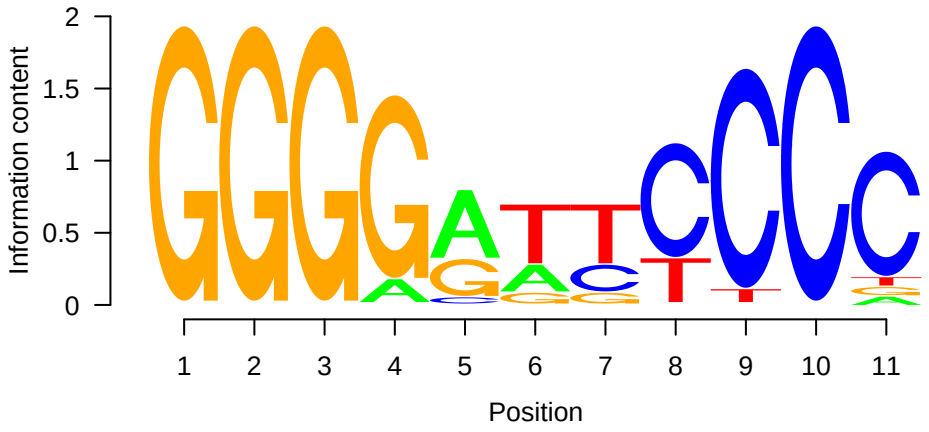


Figure 3.12: Sequence logo for NFKB1, generated by the seqLogo package of Bioconductor (Gentleman and co-workers 2004).

Figure 3.14 shows the unbound posterior probability once more, this time averaged over every transcription factor. Although the data has been smoothed by averaging, one can still observe some noise in the three lines. Once again, the noise is the least where $a_{\text{silent}\to\text{unbound}} = 0.999$, but since that line is always above 0.40, it reduces the likelihood of finding legitimate binding sites too much. The posterior probability is not necessarily equal to the prior probability of transitioning to the unbound state, especially when looking at regions such as promoters where the data generally suggest greater binding. The disconnect between the prior expectation when $a_{\text{silent}\to\text{unbound}} = 0.9$ and a median mean posterior probability 0.84, however is too great to accept for results that fit with the model's expectations. Using $a_{\text{silent}\to\text{unbound}} = 0.99$ appears to be the best choice for these reasons and those discussed in subsection 3.3.1.

The general pattern of the posterior probability of the unbound state is the additive inverse of total binding. Figure 3.14 indicates that on an idealized promoter, TF binding slowly increases to an absolute maximum at ~ -49, decreases until the TSS, where it then sharply decreases around the TSS, and continues to gradually decreasing towards downstream regions.

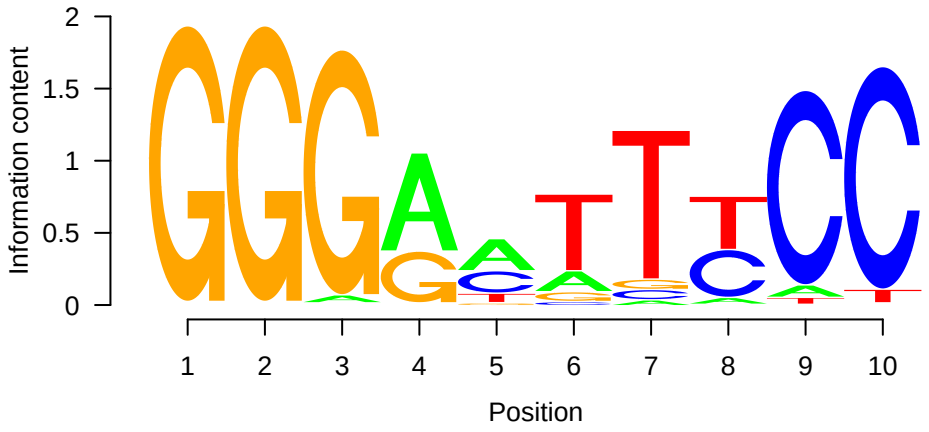


Figure 3.13: Sequence logo for NF-kappaB, generated by the seqLogo package of Bioconductor (Gentleman and co-workers 2004).

3.3.3 Whole chromosomes

Sunflower works well over large expanses of the genome. It produces data that can be used by wet-lab biologists who can look for potential TFBSs in their gene of interest without having to install, configure, or run Sunflower themselves. The data is available to any software that works with DAS, such as Ensembl. An example of data served by ProServer[4] and displayed with Ensembl ContigView can be seen in Figure 3.15. The chromosome-scale results reveal many peaks for various transcription factors outside previously identified promoter regions. This fits with the observations of ENCODE Project Consortium (2007) that the genome is transcribed pervasively.

For now, DAS access to data on all of chromosome 21 is available on request. I am planning to run Sunflower on the whole human genome and provide unrestricted public access soon.

[4] ProServer instance configured by Andy Jenkinson.

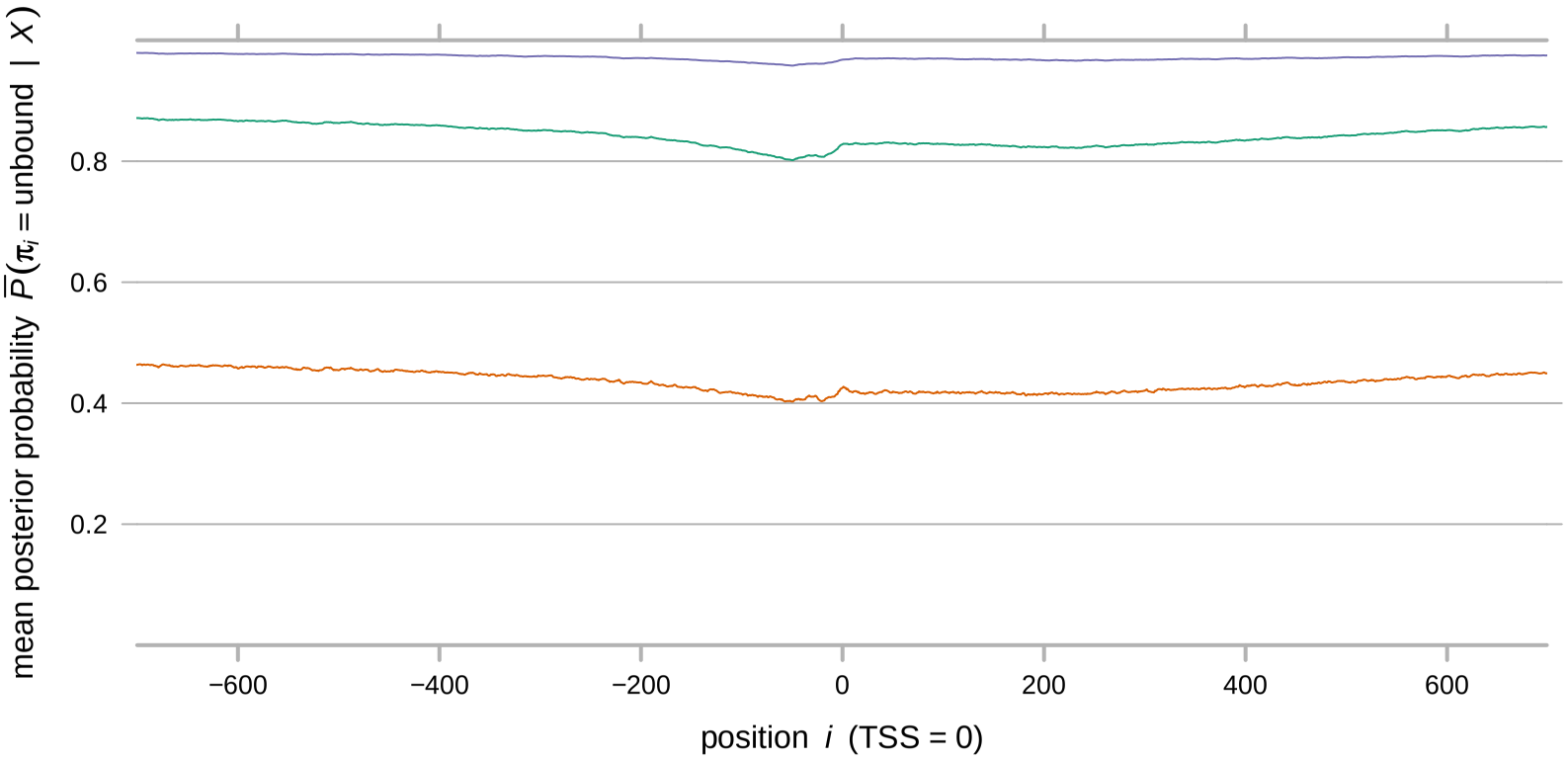


Figure 3.14: Mean posterior probabilities that the Sunflower model is in the unbound state at 1400 positions flanking the transcription start sites of 17,600 human transcripts, for three different models where $a_{\text{silent}\to\text{unbound}}$ is 0.9 (orange), 0.99 (teal), or 0.999 (purple).

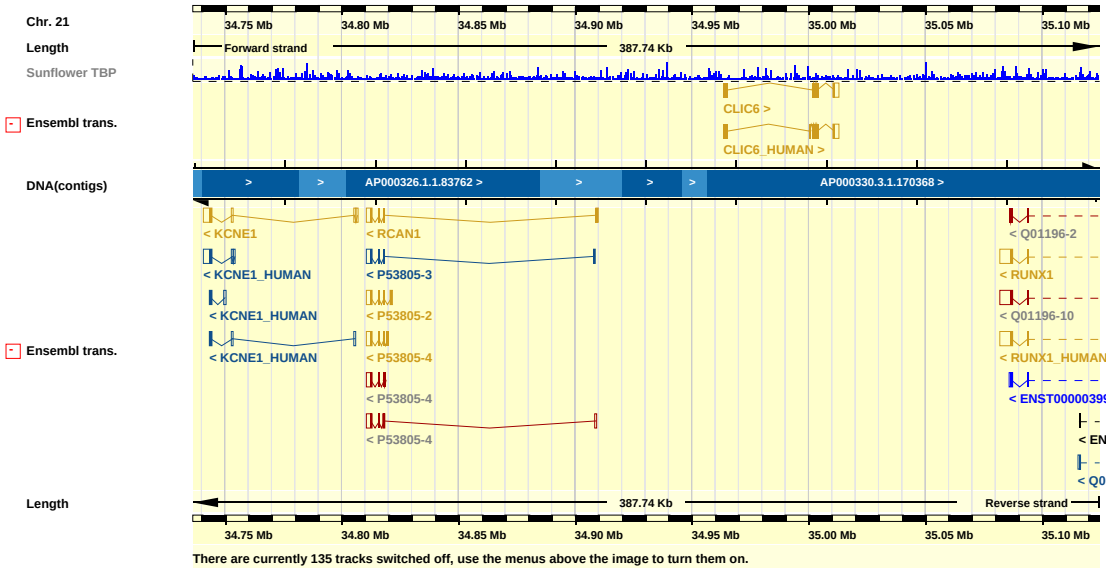


Figure 3.15: Sunflower results for TBP displayed in Ensembl ContigView for a region of human chromosome 21.

3.3.4 Performance

Time

On an IBM BladeCenter LS20 with a with an AMD Opteron 270 processor, Sunflower runs very quickly, producing the probabilities of a 2000 bp sequence on an 89-TF model in less than a second. For a 30,000 bp chunk of genome, it takes 13 s, as shown in Figure 3.16. This compares well with the performance of previous methods, such as Ahab (Rajewsky and co-workers 2002), which took two days to run the entire *Drosophila* genome on an 8-TF model. It is important for the processing to be fast because it is repeated many times per transcript in the applications shown in chapter 4.

Sunflower's worst-case running time for an arbitrary model is in the order of magnitude $O(m^2 n)$. But with the connection optimizations of subsection 3.2.3, the sparse models considered here run considerably faster in $O(mn)$ time.

Space

When producing HDF5 output, Sunflower produces a 64-bit floating point number for each state of interest for each position. While this produces 1424000 bytes per 2000 bp TSS-flanking region, it does not scale well to the whole genome.

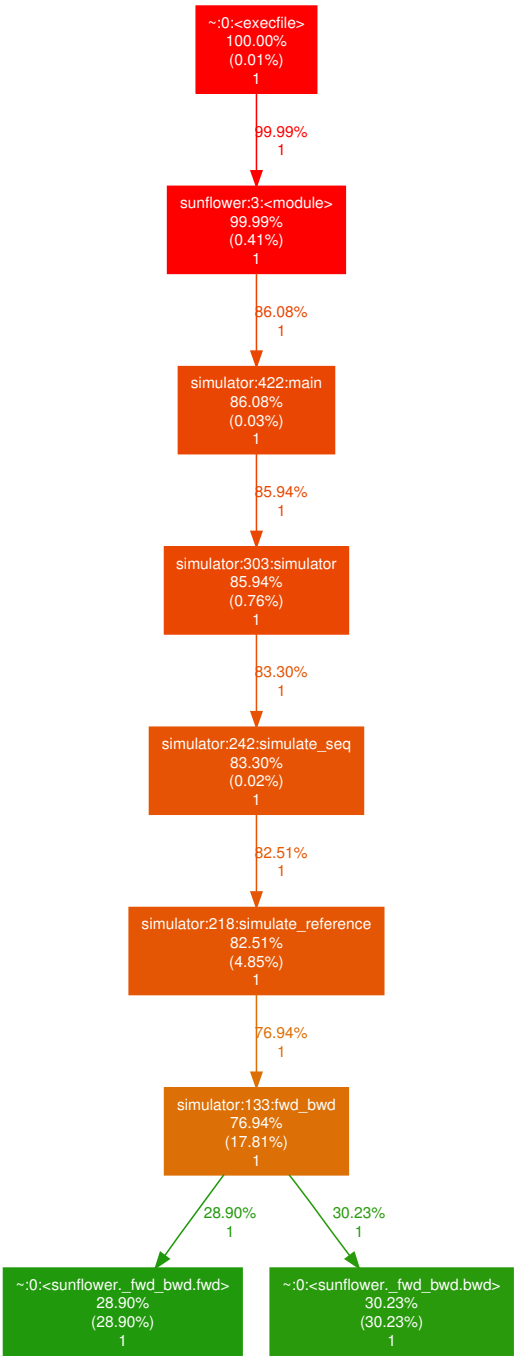


Figure 3.16: Call graph of a run of the Sunflower simulator on a 30,000 bp chunk of the human genome. Each box contains an identifier representing a unit of Python or C code, the cumulative time spent in that unit and units it calls, the time spent in that unit only without including further calls (in parentheses), and the number of times the unit was called during the run. Generated with Gprof2Dot (Fonseca 2007), and Graphviz (Ellson and co-workers 2004).

The run-length encoded MySQL database described in 3.2.6 is far more efficient in terms of space—for 33.8 Mbp of human chromosome 21 euchromatin the data required 1.31 GiB of disk space with an additional 2.15 GiB necessary for indexes.

3.4 Discussion

3.4.1 Transcription factor classes

The data in subsection 3.3.2 reveals that some of the transcription factors with the largest overall aggregate binding preferentially bind CpG island transcripts. Some of the TFs with the smallest binding, however, seem to prefer CpG desert transcripts. We can correct for the overall binding of a TF by taking the difference of the median for CpG island and CpG desert categories and dividing by overall binding. This is called differential CpG island binding. There is a linear correlation between differential CpG island binding and overall binding, as shown in Figure 3.17 (Spearman's rank-order correlation coefficient $r_S = 0.74$; $p < 2.2 \times 10^{-16}$).

One can explain the correlation of differential CpG island binding with overall binding through the correlation between PWM G+C content and overall binding (Figure 3.18; $r_S = 0.78$; $p < 2.2 \times 10^{-16}$). High G+C content PWMs obviously predict more binding to CpG island promoters, as well as promoters in general, which are GC-rich in human. Figure 3.18 also reveals that many ubiquitously expressed TFs such as SP1 have high G+C PWMs, whereas tissue-specific factors such as those with forkhead box domains have low G+C PWMs. Determining the predictive power of this observation would be an interesting topic for further study.

3.4.2 Advantages

Sunflower has several advantages over previous work. The most common programs used for finding regulatory elements are based on threshold approaches (Kel and co-workers 2003; Lenhard and co-workers 2003; Loots and co-workers 2002) which are only a crude approximation of biology and less relevant given the evidence for the importance of weak binding sites (ENCODE Project Consortium 2007). But even beyond available software packages with a probabilistic

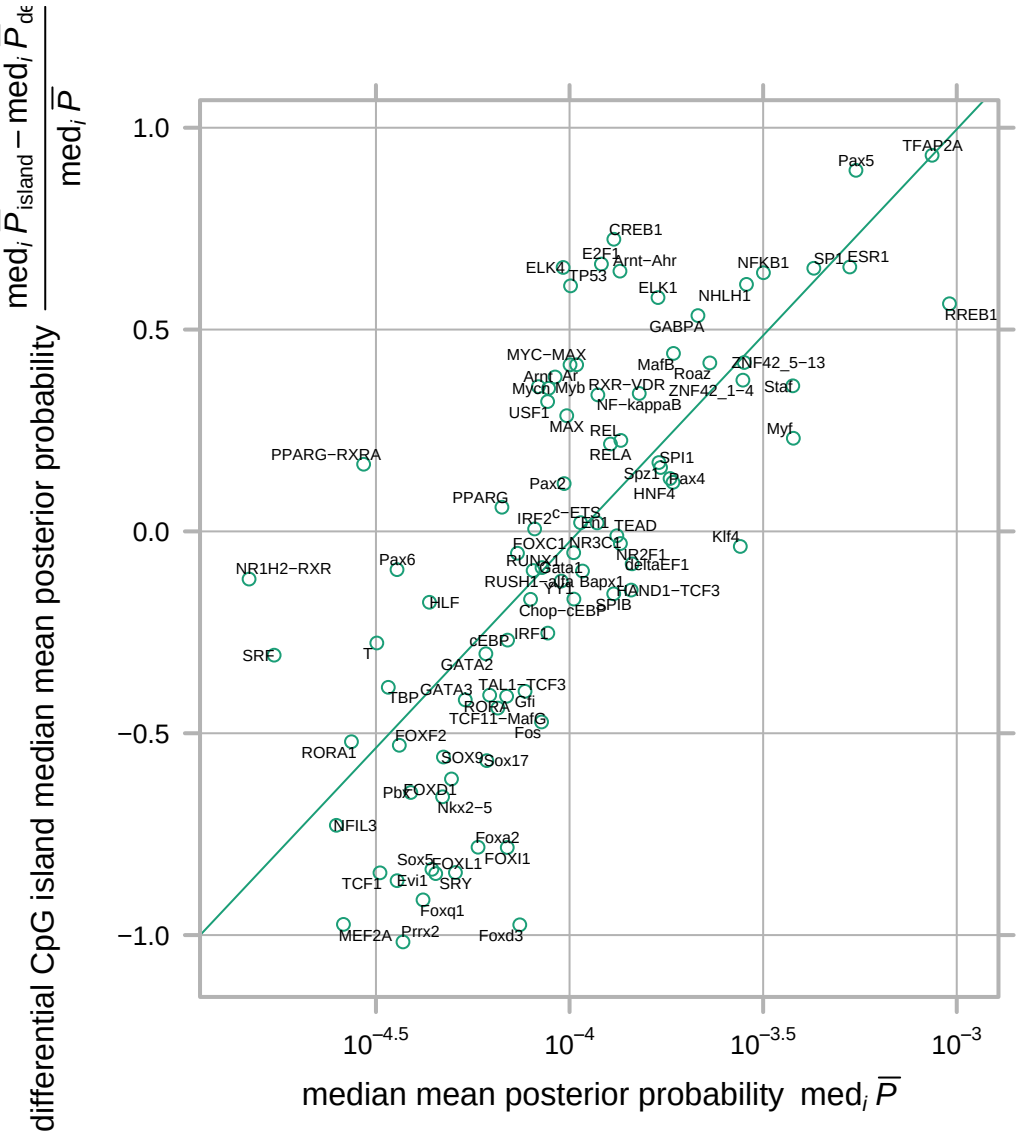


Figure 3.17: Scatterplot of the difference between the median mean probabilities for CpG island desert and CpG desert transcripts against the median mean probability for all transcripts for each of 89 transcription factors.

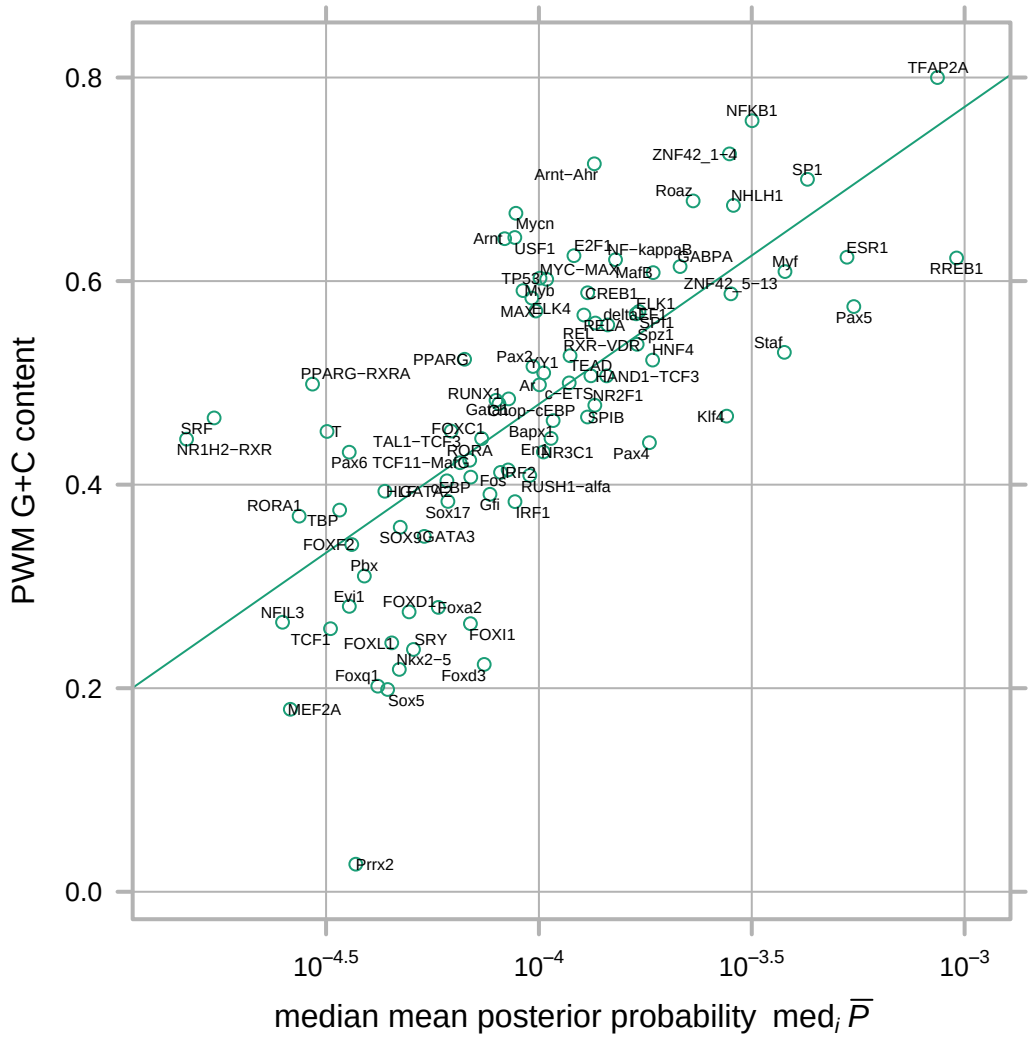


Figure 3.18: Scatterplot of position weight matrix G+C content against median mean posterior probability for 89 transcription factors.

or energetic approach to TFBS recognition, Sunflower has some advantages. First, I am making the results of Sunflower whole genome runs available to the public via the Ensembl genome browser, which allows biologists with an unsophisticated understanding of bioinformatics methods to use Sunflower results immediately (see subsection 3.4.3). Secondly, I designed the Sunflower software to be easily extensible by others so that it could be used for a wide range of HMM-based TFBS recognition activities.

The most important advantage of Sunflower is that it was designed as the core of an evolutionary model, which I discuss in chapter 4.

3.4.3 Applications for biologists

One of the aims of Sunflower was to be useful to individual biologists focused on single genes or sets of genes, who may look at data it generates as a series of DAS tracks in the Ensembl genome browser. Biologists can thereby determine which transcription factors are most likely to play an important role in the promoter of a gene of interest. This is one of the most common questions bioinformatics researchers receive from their wet-lab colleagues (data not shown).

Researchers who feel comfortable installing and running the Sunflower software themselves can investigate which TFs play an important role with respect to individual sets of concentration data by adjusting the prior transition probability matrix **a**.

The use of probabilistic output rather than an arbitrary score eases the interpretation of the results, although the range of the output values is highly influenced by the priors chosen for the model. Using a model with $a_{\text{silent}\rightarrow\text{unbound}} = 0.9$ generates TF binding probabilities approximately 10 times that of $a_{\text{silent}\rightarrow\text{unbound}} = 0.99$.

3.4.4 Extensibility

Shown in Figure 3.19 is a model implemented by Alison Meynert that includes states that represent cooperativity between TFs. The cooperativity states model an increased probability of binding when potential binding sites for two cooperative transcription factors are near each other. While I do not discuss further results obtained with this model, its implementation demonstrates the flexibility

of the Sunflower framework. It can successfully use HMMs of arbitrary and unforeseen complexity. It is limited primarily by the imagination of the user rather than the package's initial designer. Due to Sunflower's modular design, Meynert was able to easily implement her model without modifying the package's core programming.

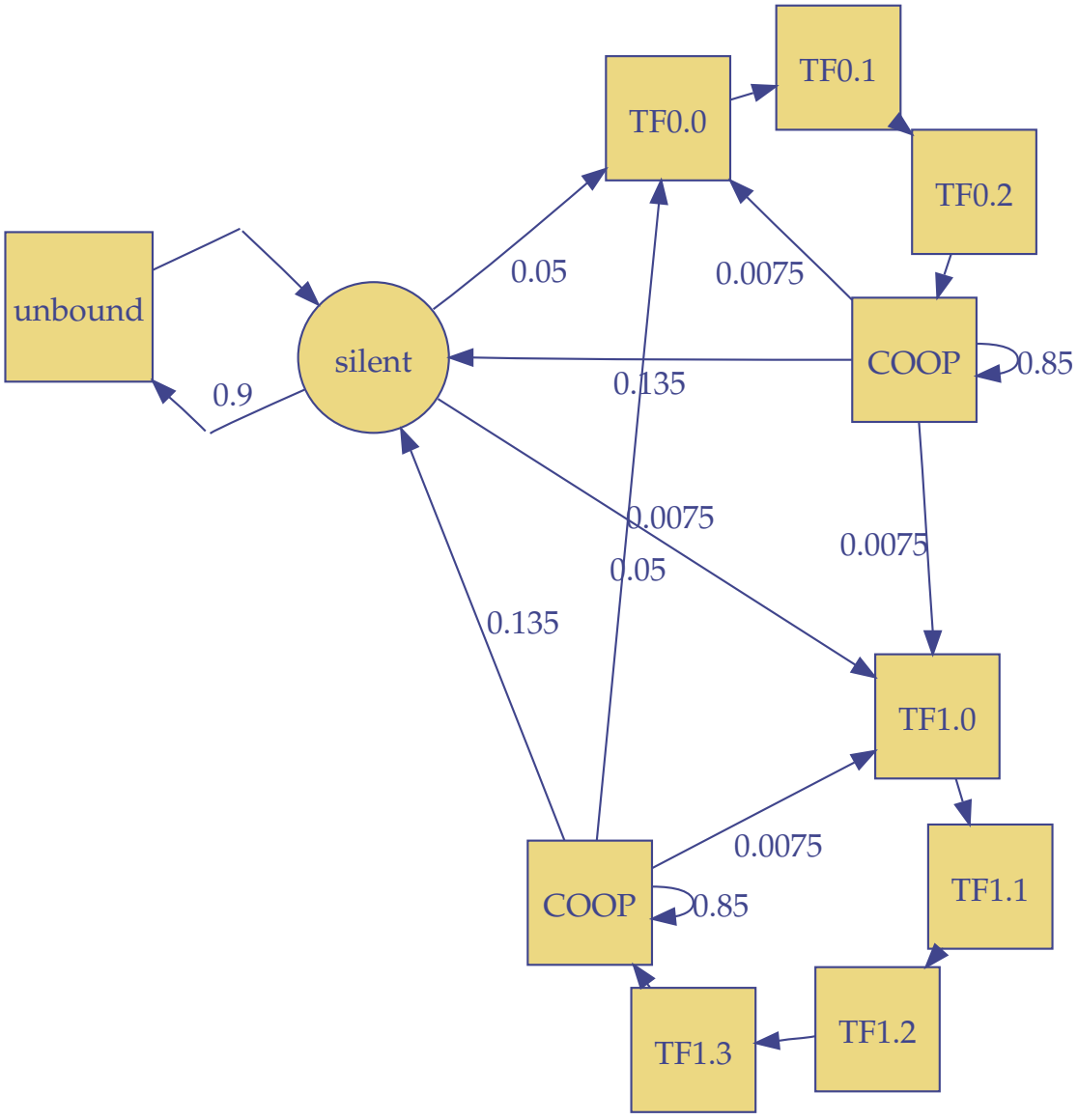


Figure 3.19: Toy example schematic of a model designed by a Sunflower user using its extensibility facilities. Alison Meynert modified a figure I created to make this figure.

Table 3.4: Aspects to be considered in a more complete model of transcription factor binding.

Aspect	Data source
affinity	included in Sunflower
competition	included in Sunflower
cooperativity	see subsection 3.4.4
epigenetic effects	DNAse I hypersensitivity data
genome methylation	bisulfite sequencing
concentration	gene expression data
factor-factor interactions	protein-protein interaction data

3.4.5 Completeness of model

Unfortunately the ensemble of transcription factors bound to a region of sequence cannot be predicted entirely by affinity data for a small number of factors. For more accurate prediction, it may be necessary to include information from a number of other sources, some of which are listed in Table 3.4.

Partially because of this incomplete model, I have focused on TSS-flanking regions, as we know that transcription factor binding has a phenotypic effect in those regions. In the future, models of cooperativity in Sunflower (subsection 3.4.4) and of epigenetic effects such as open chromatin (Segal and co-workers 2006; ENCODE Project Consortium 2007; Mikkelsen and co-workers 2007) and DNA methylation (Frommer and co-workers 1992) can be added to the current features of affinity and competition.

The accuracy of the model will improve as the data it relies on become more complete. While there are at least $\sim$1500 transcription factors in human (Messina and co-workers 2004), I only incorporate the 89 for which JASPAR CORE vertebrate PWMs are available into my model. Obviously, a 10-fold increase in the number of transcription factors, or an improvement of the data quality would lead to better predictions. I hope that protein binding microarrays (Bulyk 2006) and SELEX automation technology (Cox and co-workers 2002) will lead to more high-quality TF affinity data in the near future.

The simple model used in most of this work has the advantage of few unknown parameters, beyond the PWMs that can be taken directly from biological assays. The default parameters have a flat prior, which allows the results to remain concentration-agnostic. Users who have gene expression data on the TFs

that they want to simulate may do so by setting the prior transition probabilities themselves.

Additionally, the parameters used here have implications for the method used to elucidate evolutionary properties introduced in chapter 4. The transition probabilities to the initial bound states are set uniformly because there is no sensible way to decide accurately how much more prevalent one transcription factor is than another, given the variation in expression levels in different tissues. Since transcription changes in any one of these tissues could affect the fitness of the organism, they are all important in an evolutionary sense. However, it would still be possible to look only at the phenotype as expressed in certain tissues by starting with concentrations from gene expression data in those tissues.

While Sunflower can certainly find both strong and weak signals in DNA sequence that can be matched by the PWM for a transcription factor, this alone may not be sufficient for changes in gene expression. In some cases it has been shown that binding of a single TF was insufficient to cause transcription (Martone and co-workers 2003). In others, ChIP-chip experiments have shown that the Ste12 transcription factor can bind to sites in one yeast species, yet not bind to identical sequences in another species (Borneman and co-workers 2007). It is believed that this is because binding at these sites is dependent on the presence of the Tec1 transcription factor (M. Snyder, personal communication). In addition to synergistic effects, some transcription factors can repress the activity of others in *trans* (Courey and Jia 2001). Factor-factor interactions not modeled by Sunflower may be necessary for changes in protein expression. Still, it is possible to model these in the Sunflower framework after the biological mechanism is sufficiently understood.

3.4.6 Ambiguous nucleotides

At the moment, Sunflower does not accept sequence with ambiguous nucleotide codes, for good reason. It was originally designed for use on the human genome where high quality and sequence coverage mean that ambiguous nucleotides signal an area of heterochromatin that is unlikely to be of use. As Sunflower is applied to genomes of lower coverage such as the portions of the zebrafish genome (Wellcome Trust Sanger Institute 2007) sequenced to only $5.5\times$ coverage, ambiguous nucleotides will become more important.

The best thing to do in these cases would be to extend the emission probability matrix **E** so that it contains values for each ambiguous nucleotide summed for its unambiguous constituents. For example, the emission probability of the purine ambiguous code `R` would be summed from the unambiguous purine codes `A` and `G`. This violates the condition that the emission probabilities should add to 1 but gives appropriate probabilities that a nucleotide represented ambiguously would be bound in a certain region of a TF's motif, assuming that there is no additional information about which of the possible corresponding unambiguous nucleotides is more likely.

3.4.7 Transcription start site annotation

I used a publicly available Ensembl gene set for the first analyses of Sunflower. While CAGE data would make it possible to better estimate actual transcripts, making a better gene set would have been out of the scope of this project, and is more suited to professional gene annotators. Indeed, during the course of this project, the Ensembl human TSS predictions were revised and improved and it is likely that they will do so again to incorporate CAGE data (F. Kokocinski, personal communication).

Additionally CAGE studies have cast doubt on the classical concept of a transcript with a single TSS, finding that transcription may initiate at a broad range of positions for what previously would have been considered one alternative transcript (Carninci and co-workers 2006; Frith and co-workers 2008). The Sunflower algorithm itself is TSS-agnostic, so the implications of these studies do not affect its results on arbitrary sequence. The new TSS paradigm does affect analyses performed specifically on putative promoters. It does mean that one should find sharper peaks in an aggregated view of a CpG desert promoter, just as one does in subsection 3.3.2 for TBP. It may also explain the broader peaks of the promoters found most often in CpG islands.

One of the problems in studying the evolution of transcription factor binding sites is that evolutionary signals from coding sequence and UTRs confound the ability to elucidate signals of transcriptional regulation downstream of the TSS (see section 3.1). This could be addressed by limiting the examined transcripts to ones with much larger UTRs, with the cost of analyzing a smaller number of transcripts.

3.4.8 Unbound model

I set the unbound background model to be a static state emitting a single nucleotide at a time. The default emission frequencies are the observed frequencies of each nucleotide across the whole genome. This allows one to run Sunflower anywhere on the genome and get a sensible result. Having the default setting based on the frequency only within putative promoters would violate a design goal that Sunflower results should not depend on prior information about the location of transcription start sites or promoters (see subsection 3.4.7).

Users may use the Sunrecompose program to change the unbound state composition to any set of static values they choose, such as the average composition of TSS-flanking regions. This differs from the approach of Rajewsky and co-workers (2002), which was to locally generate a third-order Markov chain for each region of interest. I disagree with that approach. While there is hidden local information that has an impact on TF binding (see subsection 3.4.5), the base composition of the surrounding area is not a good signal of this information. There is little reason to believe that local base composition should be eliminated as a confounding factor in predicting TF binding. It is natural to assume that, for example, TFs with high G+C content PWMs would bind preferentially to areas of high G+C content, and this may have biological importance. See, for example, the discussion of SP1 in subsection 3.3.2. I would argue that the same background distribution should be used for every analyzed region in a species in order to avoid bias.

It might be possible to expand the background distribution represented by the unbound state to include information about dinucleotide or trinucleotide frequency. But this would be difficult with the current implementation and increase its complexity further. It is probably more fruitful to focus further work on modeling transcription factors rather than the background sequence.

3.4.9 Conservation

For many years sequence conservation has been used as a key ingredient in determining whether apparent TFBSs are functional (Wasserman and co-workers 2000; Moses and co-workers 2004). I have eschewed this because the ultimate goal of Sunflower is to understand patterns of conservation and evolution by producing a model against which one may compare substitutions as seen in

subsection 4.3.4. These conclusions would be circular if conservation were incorporated into the initial model. Additionally, conservation is neither sufficient nor necessary for functional transcription factor binding sites (McGaughey and coworkers 2008), so it is useful to have a model that does not require conservation to predict binding sites.

Chapter 4

Examining the impact of mutations with Sunflower

4.1 Introduction

In this chapter, I develop a novel model for estimating the biological changes induced by mutations on noncoding sequence, specifically in promoters. This model works by examining the impact of these mutations on the binding profile estimated by Sunflower. One can identify mutations that have either outsize or minimal effects on transcriptional regulation, and quantify what lies between these extremes. This will lead to a better understanding of variation in *cis*-regulatory regions, an important source of gene expression variation (Stranger and co-workers 2007) and phenotypic evolution (Rockman and Wray 2002; Wray and co-workers 2003; Bird and co-workers 2006).

Sunflower can estimate the effects of a single mutation or a set of mutations such as an alternate haplotype block, a resequenced genomic region, or an alignment to a related species. The model can also identify, *a priori*, the nucleotides where substitutions would affect the binding profile the most. Previously, most models relating genotypic changes to the estimated magnitude of phenotypic changes have focused on protein-coding sequence. The models with the simplest assumptions are the d_N/d_S model discussed in subsection 1.1.2 and the model of McDonald and Kreitman (1991), which use a binary classification of changes into synonymous or nonsynonymous categories. In recent years, they have been augmented by more sophisticated models such as PolyPhen (Ramensky

and co-workers 2002), SIFT (Ng and Henikoff 2003), HybridMeth (Capriotti and co-workers 2006), and SNPs3d (Yue and co-workers 2006). These models give continuous rather than binary assignments of the potential for phenotypic change. I seek to establish Sunflower as a general model to do the same for sequence bound by transcription factors, such as promoters.

After estimating a baseline binding profile for the sequence using the SUNFLOWER-REFERENCE algorithm in subsection 3.2.4, I can do the same for an aligned sequence, and summarize the total change in the probability landscape using a function that reduces the total change in the binding profile to a single number. Because of the underlying FORWARD-BACKWARD formulation of SUNFLOWER-REFERENCE, this simultaneously estimates the change over all possible paths through the hidden Markov model. I can exhaustively estimate the effects of change at each nucleotide by simulating each possible point mutation in the promoter and observing how the binding profile changes. I call this exhaustive estimation the SUNFLOWER-MUTATE algorithm.

In order to reduce total change into a single number, I use an accumulator function on the reference and mutated binding profiles at each position of the sequence. This function produces a number I call the binding shift t. The function I use in this chapter is the relative entropy from the reference binding profile to the mutated binding profile $t = H(P \,\|\, P')$. Relative entropy is commonly used to examine the difference between probability distributions (Benos and co-workers 2001; Durbin and co-workers 1998).

The SUNFLOWER-REFERENCE method produces continuous scale output rather than discrete hits based on a threshold. This is essential for the operation of SUNFLOWER-MUTATE because it allows continuous-scale accumulation of the changes due to mutations and the detection of small changes.

There is wide variation in the effects of simulated mutations at different positions. Some positions are neutral and mutations there do not significantly change the profile of probable bound transcription factors. Mutations in certain positions, however, cause large changes in the binding profile. This may be because they result in a significant alteration of a transcription factor's binding probability, which may cause a domino effect as other nearby positions become hindered or unhindered from binding to other transcription factors and so on.

4.2 Methods

4.2.1 Algorithm

The algorithm used to investigate the effects of mutations can be described simply. First, run the SUNFLOWER-REFERENCE algorithm from subsection 3.2.4 with a Sunflower model and a nucleic acid sequence to get the posterior probability matrix P. Then use the SUNFLOWER-MUTATE algorithm to calculate the relative entropy t:

SUNFLOWER-MUTATE($\mathbf{A}, \mathbf{E}, \boldsymbol{X} = (x_1 \dots x_n), \mathbf{F} = (f_{k,x})_{m \times n}, \mathbf{B}, \mathbf{P}$)

1 $\boldsymbol{X}' = (x'_1 \dots x'_n) \leftarrow \boldsymbol{X}$
2 $\mathbf{F}' = (f'_{k,x})_{m \times n} \leftarrow \mathbf{F}$
3 **for** $i \leftarrow 1$ **to** n
4 **do for each** a **in** $\mathcal{A}$
5 **do if** $a = x_i$
6 **then** $t_{i-1,a} \leftarrow 0.0$
7 **else** $x'_i \leftarrow a$
8 $\mathbf{P}' \leftarrow$ SUNFLOWER-REFERENCE($\mathbf{A}, \mathbf{E}, \boldsymbol{X}', \mathbf{F}', \mathbf{B}, i$)
9 $t_{i-1,a} \leftarrow H(P \parallel P')$
10 $x'_i \leftarrow x_i$
11 $\mathbf{f}'_{i-1} \leftarrow \mathbf{f}_{i-1}$
12 **return** $\mathbf{T} = (t_{i,x})_{n \times |\mathcal{A}|}$

This algorithm includes a significant optimization over the naive implementation, because it uses the three extra arguments in SUNFLOWER-REFERENCE to avoid re-running the whole FORWARD-BACKWARD algorithm each time. Only those columns j of the forward matrix where $j >= i$ and the backward matrix where $j <= i$ are recalculated, as the left and right partitions of these two matrices, respectively, would have the same value as when calculated from the reference sequence.

4.2.2 Model and reference sequences

In this chapter, I use the same model and reference genomic sequences discussed in chapter 3.

4.2.3 Alignments

In order to examine trends in how the relative entropy of simulated mutations matched with substitutions found in nature, I compared the TSS-flanking regions in human with aligned sequence from Ensembl Compara 47 (Flicek and co-workers 2008; http://oct2007.archive.ensembl.org/) for chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005; build NCBI 36), rhesus (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; build MMUL 1.0), and dog (Lindblad-Toh and co-workers 2005; build CanFam 2.0) using Compara's AlignSlice adaptors. Gaps were inserted only in the non-human sequence, and human deletions were ignored. I used these alignments to factor every simulated mutation into one of five categories: observed (the mutation was observed in the other species), unobserved (the mutation was not observed in the other species), insertion (aligned to a gap), unaligned, and ambiguous (aligned to an ambiguous nucleotide). Most of the mutations are unobserved against the chimp sequence, since a change in an aligned reference nucleotide can correspond to two or three unobserved substitutions, but only zero or one observed substitutions.

4.2.4 Single nucleotide polymorphisms

For each human 1400 bp TSS-flanking region, I located the single nucleotide polymorphisms (SNPs) mapped to that region in Ensembl 47 (build NCBI36) that were genotyped in the HapMap (International HapMap Consortium 2007) YRI population. This population consists of 30 mother-father-child trios drawn from the Yoruba people in Ibadan, Nigeria. I then associated each SNP with the Sunflower-simulated mutation that measured the effect of a change between the two different alleles at the same position. I did not include null alleles.

4.3 Results

4.3.1 Single transcript

We can first examine the results for a single transcript, such as ENST00000344265, the same transcript of *POU1F1* first discussed in subsection 3.3.1. These results are shown in Figure 4.1, where one can see several types of characteristic patterns

in a plot of the change due to a simulated mutation. In many places, there is a uniformly low effect of simulated mutations on transcription factor binding. In other positions, some or all of the different mutations may have a large effect. The larger changes are where, in effect, the model is predicting the domino effect postulated in section 4.1. The least transcriptionally synonymous sites indicate that there may be multiple competing transcription factors likely to bind in that area, and runs of these sites are a natural consequence of the information-rich parts of transcription factor binding motifs also occurring in runs. It might be the case that only two of the potential mutations have an effect if an important transcription factor binding motif at a site is likely to bind two different nucleotides (one of which would be the reference nucleotide), but not the other two.

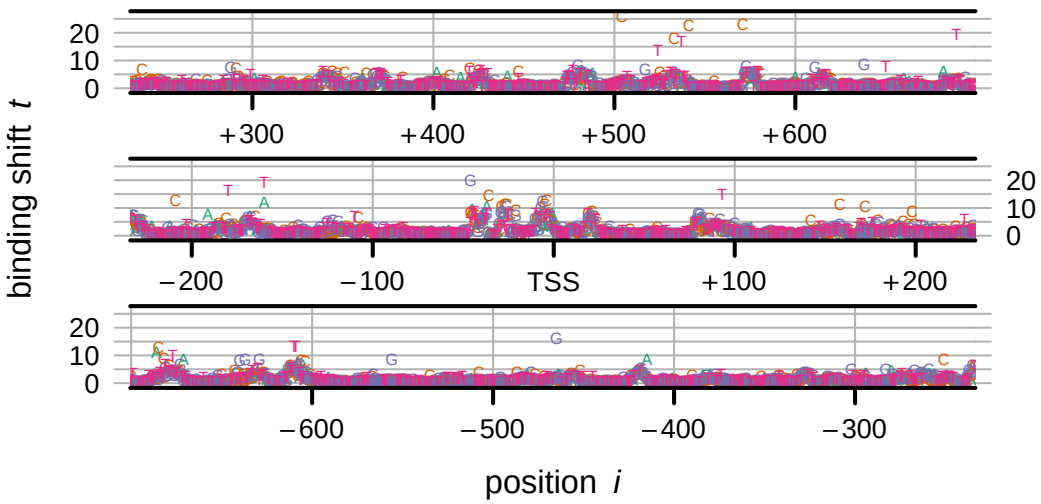


Figure 4.1: Relative entropy between the reference posterior probability and the posterior probabilities due to simulated mutations of each nucleotide along the sequence of ENST00000344265, the Ensembl transcript for *POU1F1*, POU class 1 homeobox 1, a gene on human chromosome 3p11.2, with the longest 5′ untranslated region. At each point along the 2000-nucleotide sequence, I change the reference nucleotide to each of the other three possible nucleotides. The relative entropy is plotted for the central 1400 bp as a single letter corresponding to the mutated nucleotide. Green A: adenine; orange C: cytosine; blue G: guanine; red T: thymine.

4.3.2 Aggregation

I aggregated the kind of results shown in Figure 4.1 over all the analyzed transcripts by performing a position-wise average. The results are shown in Figure 4.2. The most interesting parts of these results are closest to the TSS. They are not artifacts of the dynamic programming algorithms, since they still appear if the boundaries of the simulation region are shifted so that they are on the edge rather than in the center. At around position -35, there is a peak in the effects of the simulated mutations, the biggest in the entire plot. There is also a smaller peak at approximately -45. These peaks represent the aggregation of thousands of TATA boxes, as TATA binding protein is included in the analyzed set of transcription factors. Most of the other peaks identifiable on individual transcripts are smoothed out when averaged over thousands.

On average, mutations in the region between -100 and the TSS cause the most disruption to predicted TFBSs, although there is also significant disruption downstream of the TSS. This is consistent with an understanding of the most important transcription factors binding near the TSS. The fact that the most disruption occurs $5'$ of the called TSS could either mean that upstream factors are more likely or important biologically, or that many untranslated regions are miscalled, which would mean that many transcripts actually initiate upstream of the currently called TSS. CAGE data indicates that present UTR predictions may not reach the full extent of potential transcription initiation for all transcripts (Carninci and co-workers 2006), so the UTR miscalling interpretation has experimental evidence behind it for some transcripts.

A gentle rise from the TSS to around $+200$ occurs primarily in CpG island promoters, and indicates a higher density of TFBSs around that position in those transcripts.

4.3.3 CpG effects

When we cluster the various transcripts by their CpG-island–association status (see subsection 3.2.7), and then aggregate only amongst these groups, we find that the peaks at -35 and -45 are pronounced relative to their surroundings in areas associated with CpG deserts (Figure 4.3). This is consistent with the observation that TATA boxes are more important in promoters lacking CpG islands (Carninci and co-workers 2006). The CpG-island promoters seem to have

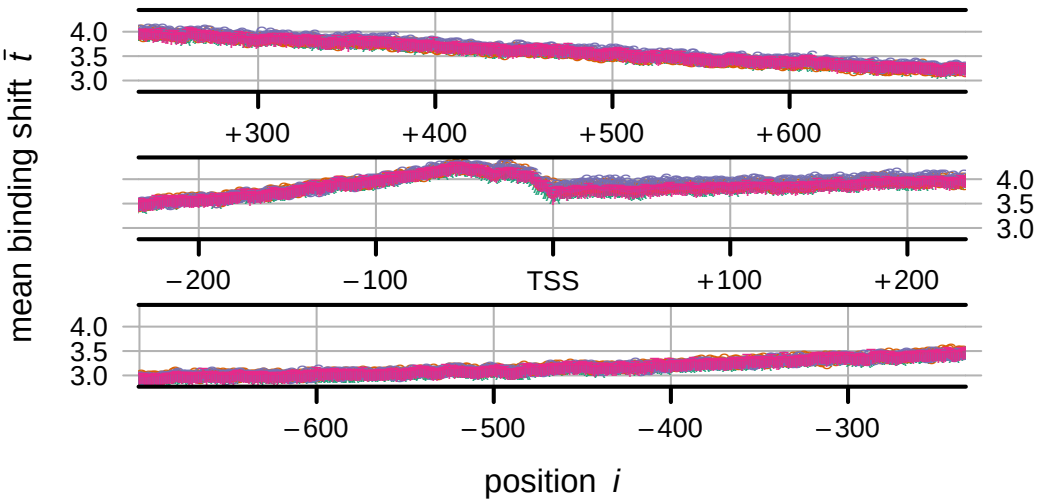


Figure 4.2: Sequence plot of mean relative entropy for the JASPAR CORE vertebrate model by position relative to the TSS averaged over 17,600 human transcripts. The mean t for each (position, mutation) pair is plotted as a single letter corresponding to the mutated nucleotide. Four colored letters are plotted at each position indicating the relative entropy from the reference sequence. Green A: adenine; orange C: cytosine; blue G: guanine; red T: thymine. Transcripts picked as discussed in Figure 3.5. Note that the y-axis scale is a few orders of magnitude smaller than previous figures examining one TSS-flanking region.

higher binding shifts overall than the CpG desert promoters. This fits well with findings from ChIP-chip experiments that aggregate transcription factor binding is much higher at CpG island TSSs than CpG desert TSSs (ENCODE Project Consortium 2007).

4.3.4 Alignments

In Figure 4.4 one can see that summed over thousands of promoters, there are a tiny number of substitutions from human to chimp, while the number of substitutions to rhesus is more and to dog is much more. Rhesus is frequently used in this thesis as a comparison species because it is still relatively closely related to human, but nonetheless exhibits enough sequence differences to reduce concerns of small-sample error. While there is a good deal of turnover of functional TFBSs between species (ENCODE Project Consortium 2007; Borneman and co-workers

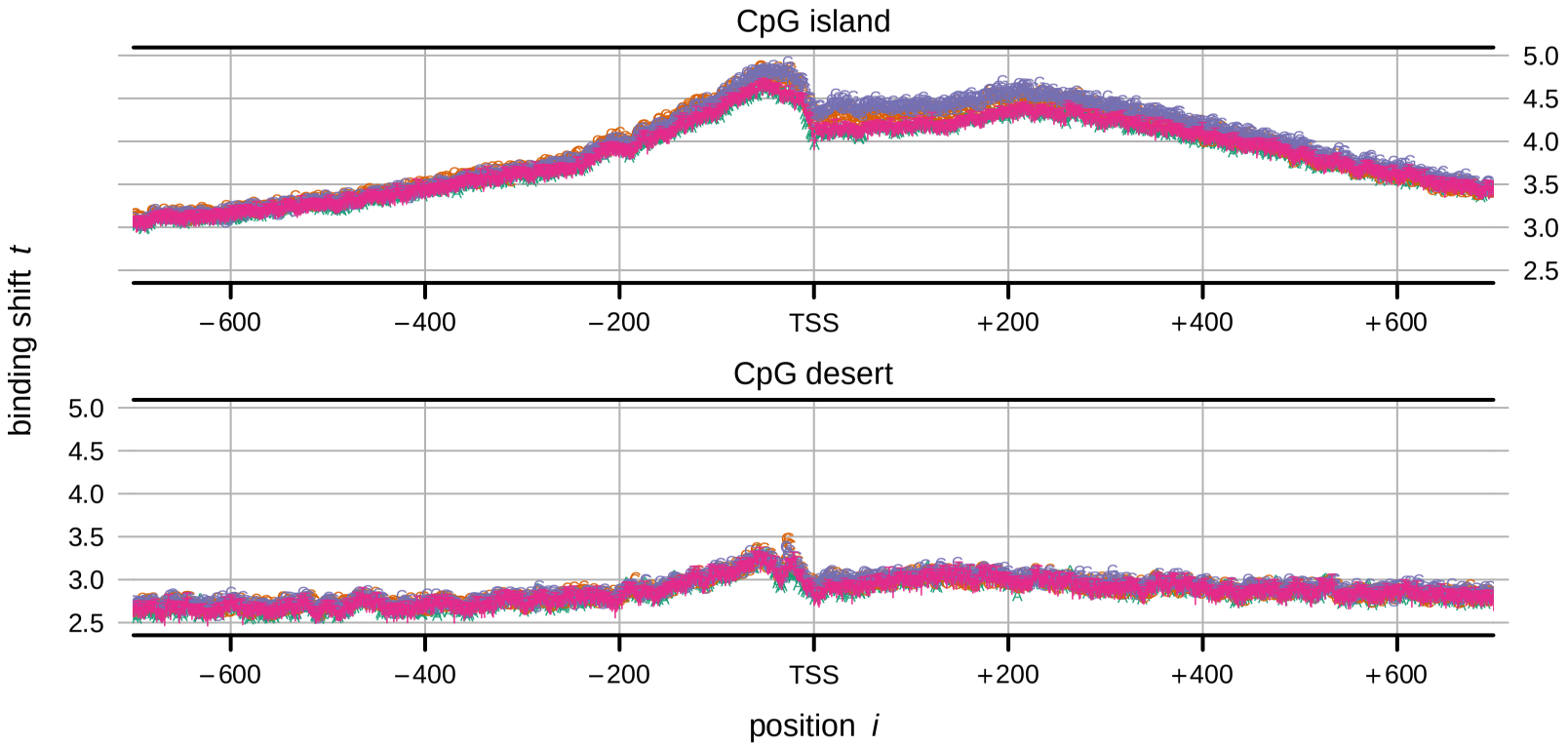


Figure 4.3: Mean binding shift against position for 11,822 CpG island genes and 5,778 CpG desert genes. The relative entropy is plotted for the central 1400 bp as a single letter corresponding to each mutated nucleotide. Green A: adenine; orange C: cytosine; blue G: guanine; red T: thymine.

2007), the majority are conserved between human and rodents (Dermitzakis and Clark 2002), so one expects even more conservation of binding sites at the closer evolutionary distances I examine.

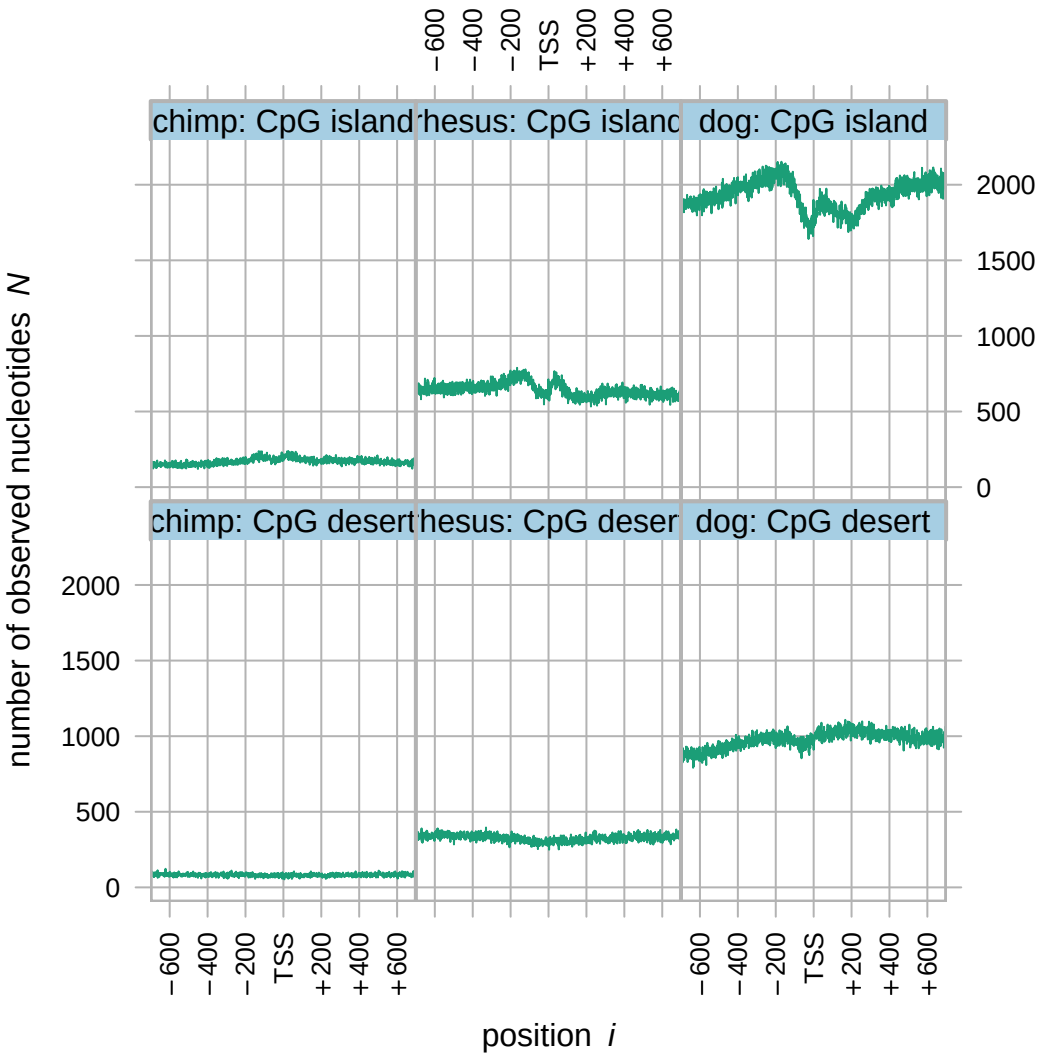


Figure 4.4: Total number of observed changes between the human gene and another species by position relative to the TSS summed over 17,600 human transcripts. Each column specifies a species and the rows specify the CpG island status of the transcript.

As can be seen in Figure 4.5, mean binding shift for observed mutations is equal to or less than the mean binding shift for unobserved mutations. This is especially visible in dog where there are enough observed data points to smooth

out the noise, and one can see that for CpG islands the observed mutation relative entropy is consistently less than the unobserved mutation relative entropy. This means that the sorts of substitutions observed in evolution are more likely to be those that disrupt the overall binding profile the least. Dermitzakis and Clark (2002) found that TFBS substitutions between human and rhesus were heterogeneously concentrated in a few positions. The heterogeneity in binding shift at different positions may partially explain this phenomenon.

Much of the difference in effects for particular nucleotides can be explained by the base composition of the underlying sequence. If we call M the number of positions where there is an alignment match between the human reference sequence and the other species, and call m the number of positions where there is a mismatch, we get the number of observed and unobserved nucleotides

$$N_{\text{observed}} = m$$
$$N_{\text{unobserved}} = 3M + 2m.$$

High G+C content around the TSS means that the number of sequences where a change results in C or G is lower than the number where a change results in A or T. This pattern changes as the sequence becomes A+T-rich as the distance from the TSS increases. These TSS flanks have an elevated G+C content of 55% when compared to an overall genomic G+C content of 41%.

The mean for a mutation from the human reference is aggregated over positions that are either observed or unobserved in the compared species. Since the number of unobserved substitutions is several orders of magnitude greater, the observed plots are much noisier than the unobserved plots—they do not have the smoothing inherent in thousands of positions being averaged together. This phenomenon is reduced for more distant species pairs such as human–dog.

4.3.5 Single nucleotide polymorphisms

Figure 4.6 shows how the binding shift relates to minor allele frequency for HapMap YRI SNPs. The YRI population has a greater sequence diversity than other HapMap populations (International HapMap Consortium 2005). In particular, I am examining the region (-200, 0), which has the largest average binding shift of analyzed regions (see subsection 4.3.3). Also, I am only examining those

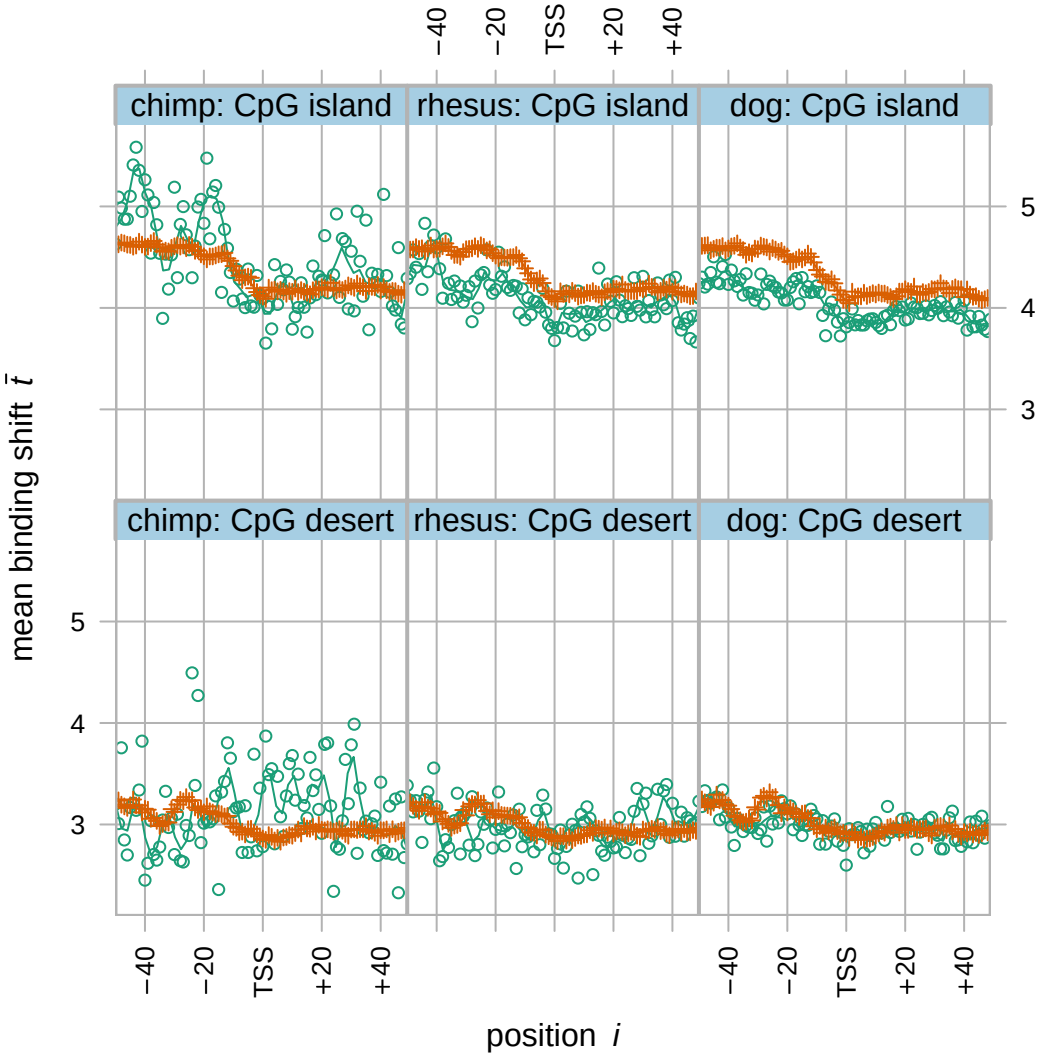


Figure 4.5: Mean binding shift against position for 17,600 human transcripts aligned against chimp, rhesus, and dog and segregated on CpG island association. Teal circles indicate the mean value at a position for mutations that have been observed. Orange crosses indicate the value for mutations that have not been observed. Lines are loess (Cleveland and Devlin 1988) approximations of the identically-colored points.

data points in CpG deserts and in the 90th percentile of binding shift values ($t > 5.94$). In these extreme values there is a weak negative correlation (Spearman's rank-order correlation coefficient $r_S = -0.14$; $p = 0.0854$). This might indicate that the largest disruptions in binding profile are being selected against weakly and are therefore less prevalent in the population. Unfortunately, there are only 3153 genotyped SNPs in this region. This makes it difficult to thoroughly investigate the correlation at present. Studying published resequencing data (Venter and co-workers 2001; Levy and co-workers 2007; Egholm and co-workers 2008) and forthcoming projects (Kaiser 2008) may provide more insight into the prevalence of variation and heterozygosity at positions of high binding shift.

4.3.6 Performance

Figure 4.7 shows a typical run of Sunflower in simulated mutation mode. On an IBM BladeCenter LS20 with an AMD Opteron 270 processor, this takes 99 min to run the 89-factor model on a sequence of 2000 bp with output on the inner 1400 bp.

4.4 Discussion

4.4.1 Data limitations

Throughout this analysis, one must be aware that changes in simulated transcription factor binding obviously reflect only the factors analyzed. It is possible that a simulated mutation that seems to be transcriptionally synonymous is only so with regard to the transcription factors I analyzed, and it might greatly affect one of the other myriad transcription factors.

In positions where Sunflower reports large impacts of mutations, however, it seems extremely unlikely that the report would change if there more TFs had been included in the model. This is at least partially borne out by results of large impacts near transcription start sites that seem to be biologically relevant, as shown above. It would require a complicated and unlikely set of coincidences for a site to appear to be susceptible to enormous changes in binding profile when probed with a model of 89 TFs, when if, in reality, unmodeled TFs would cause the change to be quite small. If such coincidences were common, most of the

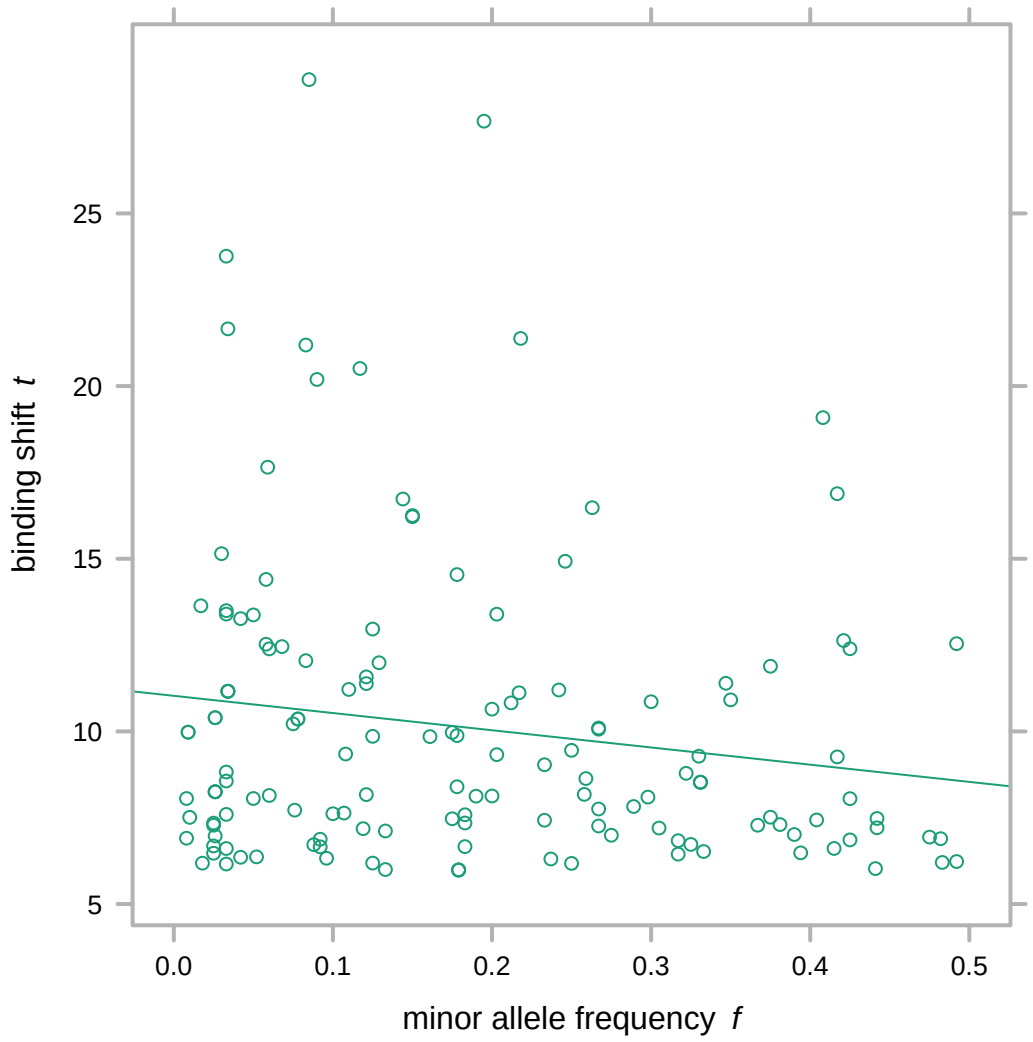


Figure 4.6: Binding shift against minor allele frequency for 145 HapMap phase II YRI SNPs in CpG desert transcripts. Only points in the 90th percentile of t values ($t > 5.94$) are shown.

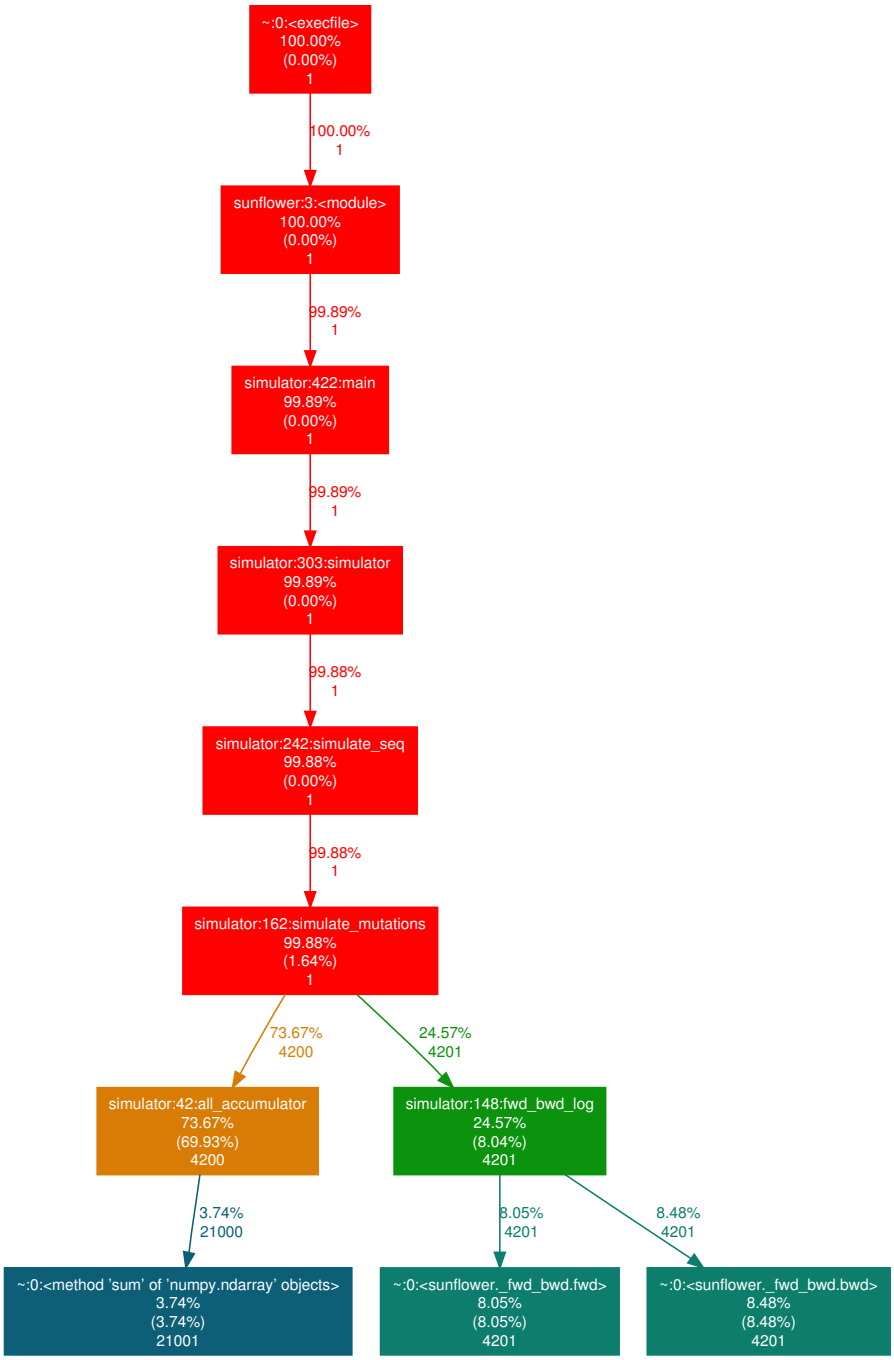


Figure 4.7: Call graph of a typical run of the Sunflower simulator. Each box contains an identifier representing a unit of Python or C code, the cumulative time spent in that unit and units it calls, the time spent in that unit only without including further calls (in parentheses), and the number of times the unit was called during the run. Only functions taking at least 0.5% of the total run time are shown. Generated with Gprof2Dot (Fonseca 2007), and Graphviz (Ellson and co-workers 2004).

methods for computational detection of TFBSs and CRMs would be invalidated, as one would never be able to even partially trust results produced with an incomplete set of TFs. As positive results from other TFBS finding methods are well-established and accepted, one should also accept Sunflower reports of higher binding shifts.

4.4.2 Haplotypes

The general framework discussed in this chapter can be used not only for exhaustive mutations but also for individual haplotypes or aligned sequence in other species, or indel-alignment approximations of ancestral sequence (Paten and co-workers 2008). This works considerably more quickly, because it only requires two passes of the SUNFLOWER-REFERENCE algorithm rather than the $3n$ required by the full SUNFLOWER-MUTATE algorithm. It also requires significantly less storage space for the results.

4.4.3 Heuristic drop-off approach

The approach of exhaustively considering the effects of each possible mutation on the unbound state of every position in the examined window is inefficient, and does not scale well. Furthermore, we know that a single substitution ceases to have an effect detectable above error rates a few hundred base pairs downstream of the substitution, and it stops having a significant effect at an even smaller distance.

The optimization discussed in subsection 4.2.1 speeds up the algorithm over the naive approach by half, by eliminating the need to recalculate parts of the forward and backward matrices that would be the same even after the mutation. In order to allow the scaling of the mutation-mode Sunflower to the entire genome, I will need to add heuristics to the algorithm that avoid recalculating the points of these matrices that have an undetectable difference from the matrices for the reference sequence. This will afford a considerable increase in speed when multiplied by thousands of mutation simulations per window.

In future work, I plan to introduce a drop-off heuristic approach common in bioinformatics, used in programs such as BLAST (Altschul and co-workers

1997). By periodically comparing the recalculated matrices against the reference matrices, Sunflower will stop recalculating when the difference gets to an undetectably small magnitude.

4.4.4 Applications for biologists

Just as in subsection 3.4.3, one of the goals of Sunflower's mutation analysis mode was that it could be used by individual biologists. By analyzing genes in a way similar to that in Figure 4.1, biologists can identify positions and mutations that affect the binding profile the most. These mutations can become candidates for experiments to determine whether they affect the regulation of gene expression or function.

One of the most important applications of the method is as a stand-in for d_N for the purposes of detecting positive selection in promoters. I discuss this further in chapter 5. It can also be used to pinpoint and confirm specific explanations for putative selection which have been previously identified using methods that can only narrow down the area of selection to broad regions (Smith and co-workers 2005; Lunter and co-workers 2006).

Chapter 5

A novel measurement of promoter evolution

5.1 Introduction

Identifying functional regions in noncoding sequence is a difficult but important aspect of understanding the biology of the genome and targets for experimental investigation. One way of doing this is to identify regions associated with signals of positive or purifying selection. Many heritable phenotypes are associated with genes that have no change to their coding sequence, so the phenotype is probably associated with mutations in non-coding sequence (Ponting and Lunter 2006; Cretekos and co-workers 2008). One of the chief aims of developing Sunflower was to use the exhaustive mutations model to determine potential areas of functional change. Sunflower provides a new technique in identifying the character of selection—positive, neutral, or purifying—within promoter regions.

The inspiration for the Sunflower model is the realization that the use of d_N and d_S (see subsection 1.1.2) is a special case of a general model relating changes in genotype to changes in phenotype. While the d_N/d_S regime has been established for decades (Miyata and Yasunaga 1980; Nei and Kumar 2000; Graur and Li 2000; Felsenstein 2004), the underlying model is used unquestioned by many. This model embodies an assumption that any change in the amino acid sequence is subject to natural selection and equally so, and that DNA changes in degenerate protein coding positions are neutral. These assumptions could only be

correct if synonymous changes had zero effect on the phenotype of the organism, which is sometimes false, as discussed in section 2.1. Nonetheless, as evidence for the problems with the underlying model mount up, so too does evidence of its usefulness (Waterston and co-workers 2002; Gibbs and co-workers 2004; International Chicken Genome Sequencing Consortium 2004; Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). The explosion of genomic sequence available demands a simple method for discriminating between genes on the basis of selective pressure. This has been useful in finding key gene families to explain the pattern of evolution in the amniotes and unique traits in the vertebrate phylogenetic tree. I seek to expand tests of natural selection that compare potential function-altering mutations to noncoding sequence, and specifically to promoters.

Since noncoding sequence has been revealed to be of increasing importance in understanding the genome (Ponting and Lunter 2006; ENCODE Project Consortium 2007), it seems necessary to find a model analogous to d_N/d_S for aspects of non–protein-coding biology, relying on a simple relationship for translation of genotypic changes into phenotype much like the genetic code used to translate nucleotide codons into amino acid residues. Just as that code is degenerate, so too would be a transcription factor binding code (Benos and co-workers 2002). While certain positions in a TF's binding motif are extremely likely to contain certain bases for binding, others can change freely without any effect on the transcription factor's affinity. A position that is truly degenerate in this way can be likened to the 4D sites in protein-coding sequence, which can also change freely without affecting function under the d_N/d_S model, if not in reality. These positions can be seen in sequence logos (see subsection 1.2.2) as positions of zero or near-zero height.

Where the Sunflower model provides an accurate mapping of genotypic changes to certain changes in phenotype, positions where mutations cause small effects are akin to synonymous coding sites. One would expect these positions to be under little selective pressure with regard to regulation by the transcription factors included in the model. I regard these positions as neutrally evolving. In certain positions, however, mutations can cause large changes in the binding profile. These positions are under selective pressure with regard to regulation by the transcription factors included in the underlying binding

model. A mutation can only cause the largest changes if it engages a *domino effect*, where changes in the binding site of one TF cause its probability of binding to change greatly, thereby lessening or increasing competition with other TFs at other potential TFBSs. While other factors in the same sequence may influence selective constraint, the positions with the largest potential binding shifts will have the most predicted disruption, and therefore the most constraint with respect to the Sunflower model of TF binding.

As Sunflower allows quantitative estimates of how substitutions at these positions affect an organism's phenotype in the modeled aspects, I construct a measurement d_T that does for this model what d_N does for a model of phenotype that considers only amino acid substitutions (see subsection 1.1.2). The quantities discussed in this chapter, usually being derived from averages over a whole promoter, are meant to be gene- or transcript-specific (as in chapter 2) as opposed to the position-specific measurements in chapter 4.

5.2 Methods

5.2.1 Proportions and distances

One can think of the binding shift measurement t introduced in chapter 4 as a measurement of the synonymity of a particular nucleotide. To get a measurement of the potential disruption in TF binding for a gene T, similar to the total number of nonsynonymous nucleotides N, one first must select a region of interest. I limit my inspection to only those nucleotides I am most sure have an effect on transcripts by selecting the region $[-100, +100)$ relative to the TSS. These are the nucleotides where t is highest on average (see subsection 4.3.2). If the value of T is used for further comparisons to an aligned sequence, then I exclude positions that do not align. I call the set of included positions in the region of interest $\mathcal{P}$.

Inspired by the logic used by Nei and Gojobori (1986) to assign a fractional synonymity to protein-coding nucleotides that are only partially degenerate, I consider the average binding shift from the reference nucleotide to all other possibilities as a measurement of the potential disruption for that nucleotide.

Summing the values for all these nucleotides, we get

$$T = \frac{1}{3} \sum_{i \in \mathcal{P}} \sum_{a \in \mathcal{A}} t_{i,a}.$$

To compare a human promoter with sequence $\boldsymbol{X} = (x_0 \dots x_n)$ with the promoter in a related species, I use the same alignments produced in subsection 4.2.3, but limit to only those alignments with fewer than 25% gap columns. I call the sequence in the other species $\boldsymbol{Y} = (y_0 \dots y_n)$, and the two positions align at each position i. With $\boldsymbol{Y}$, we can define the amount of observed binding profile disruption

$$T_{\mathrm{d}} = \sum_{i \in \mathcal{P}} t_{i,y_i}.$$

Remember that $t_{i,y_i} = 0$ whenever $x_i = y_i$, so T_{d} is nonzero only at positions where the two sequences differ. While t_{i,y_i} may be larger than the average t for any given position, this is unlikely to be true across the whole gene.

Using T and T_{d}, we can calculate a proportion of binding profile disruption

$$p_T = \frac{T_{\mathrm{d}}}{T},$$

analogous to p_N and p_S in subsection 1.1.2, and p_I in subsection 2.2.6. I use the Jukes-Cantor transformation from subsection 1.1.3 to arrive at a distance measurement

$$d_T = -\frac{3}{4} \ln(1 - \frac{4}{3} p_T).$$

5.2.2 Gene ontology

I used FUNC (Prüfer and co-workers 2007) to find enriched Gene Ontology (Gene Ontology Consortium 2006) terms at the high and low tails of continuous variables. FUNC estimates a raw significance value p and a false discovery rate (FDR; Pounds 2006) for each term, and also estimates a global significance value for each analysis. I used the GO annotations from Ensembl 48 (http://dec2007.archive.ensembl.org/Homo_sapiens/).

5.2.3 d_I values

Values for d_I were taken from the results in chapter 2.

5.3 Results

5.3.1 Estimating potential binding disruption with T

Figure 5.1 shows the distribution of T in human transcripts. The minimum, quartiles, and maximum of T are (205.5, 598.5, 782.9, 972.6, 1919.1). The distribution has a long right tail, much like the distribution of the individual values of t used to calculate T.

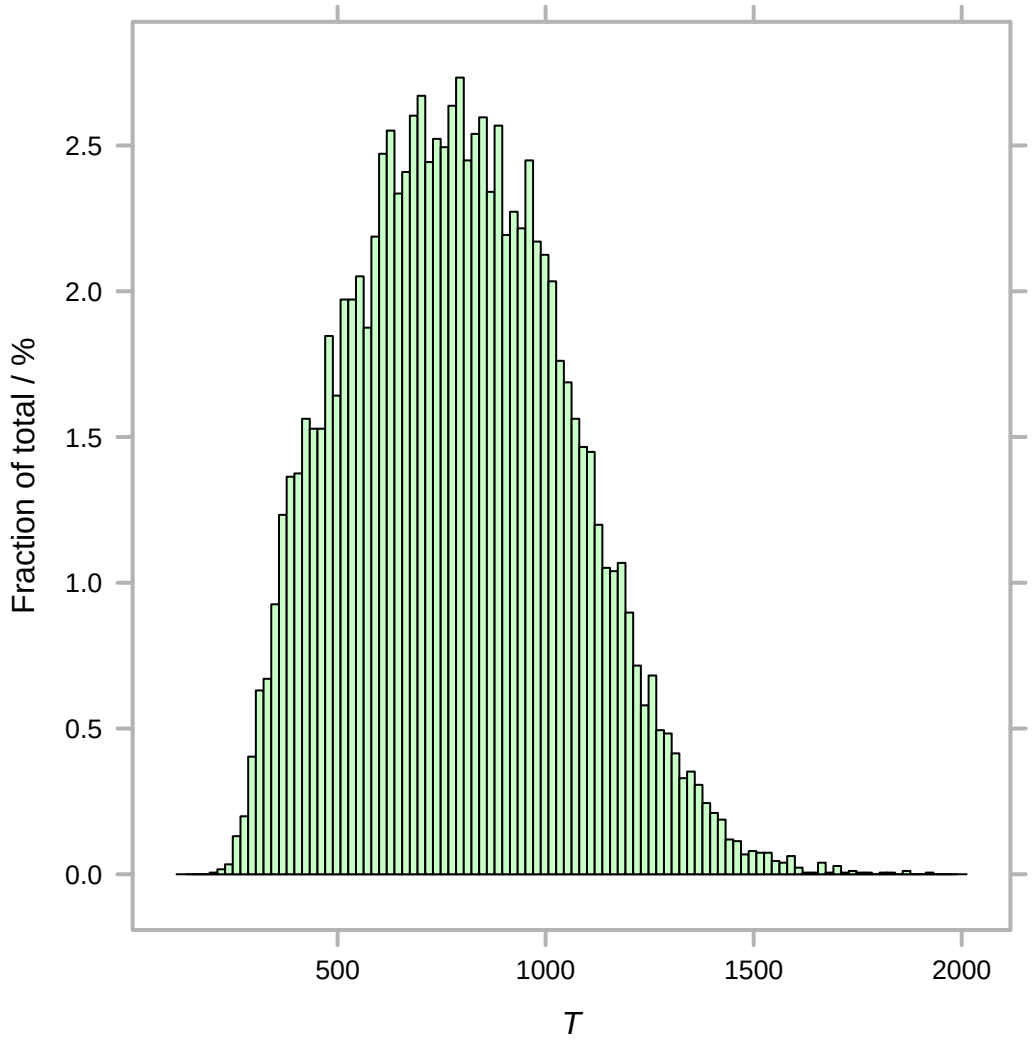


Figure 5.1: Histogram of T for 17,600 human transcripts.

Table 5.1 shows GO biological process terms enriched in genes with high-T TSS-flanking regions. The kinds of genes most enriched for high T are associated with post-translational protein modulation, including genes associated with the ubiquitin cycle or protein amino acid phosphorylation. In fact, seven of the significant terms are ancestors of these two terms. Other highly enriched terms include neuron differentiation (and two ancestors), regulation of cellular process (and two ancestors), and intracellular signaling cascade. Table 5.2 shows the enriched molecular function terms, the majority of which are associated with phosphorus transfer. Magnesium ion binding activity is also enriched, which may relate to the enriched neuronal biological processes.

GO term	p	FDR
post-translational protein modification	$< 2 \times 10^{-308}$	$< 1 \times 10^{-4}$
protein modification process	1×10^{-16}	$< 1 \times 10^{-4}$
biopolymer modification	1×10^{-16}	$< 1 \times 10^{-4}$
protein amino acid phosphorylation	9×10^{-12}	$< 1 \times 10^{-4}$
regulation of biological process	3×10^{-11}	$< 1 \times 10^{-4}$
biological regulation	3×10^{-11}	$< 1 \times 10^{-4}$
regulation of cellular process	2×10^{-10}	$< 1 \times 10^{-4}$
phosphorus metabolic process	5×10^{-10}	$< 1 \times 10^{-4}$
phosphate metabolic process	5×10^{-10}	$< 1 \times 10^{-4}$
ubiquitin cycle	3×10^{-9}	$< 1 \times 10^{-4}$
phosphorylation	1×10^{-8}	$< 1 \times 10^{-4}$
biopolymer metabolic process	3×10^{-8}	$< 1 \times 10^{-4}$
generation of neurons	3×10^{-7}	$< 1 \times 10^{-4}$
intracellular signaling cascade	6×10^{-7}	$< 1 \times 10^{-4}$
neuron differentiation	6×10^{-7}	$< 1 \times 10^{-4}$
cell development	6×10^{-7}	$< 1 \times 10^{-4}$
primary metabolic process	1×10^{-6}	$< 1 \times 10^{-4}$

Table 5.1: GO biological process terms enriched in genes with high-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$).

There are a number of possible explanations for the association of high T with certain GO terms. One is that transcripts associated with these terms may have more binding sites for the TFs analyzed here. This may be because binding sites of the TFs specifically in this set are selected for in these transcripts, or because there is more selective pressure in these categories to evolve a highly sophisticated transcriptional activation system with many TFBSs.

GO term	p	FDR
phosphotransferase activity, alcohol group as acceptor	2×10^{-14}	$< 1 \times 10^{-4}$
kinase activity	2×10^{-13}	$< 1 \times 10^{-4}$
protein serine/threonine kinase activity	2×10^{-13}	$< 1 \times 10^{-4}$
transferase activity, transferring phosphorus-containing groups	3×10^{-13}	$< 1 \times 10^{-4}$
protein kinase activity	9×10^{-13}	$< 1 \times 10^{-4}$
protein-tyrosine kinase activity	1×10^{-12}	$< 1 \times 10^{-4}$
protein binding	4×10^{-11}	$< 1 \times 10^{-4}$
transferase activity	8×10^{-11}	$< 1 \times 10^{-4}$
magnesium ion binding	6×10^{-7}	$< 1 \times 10^{-4}$

Table 5.2: GO molecular function terms enriched in genes with high-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$).

I posit that the explanation involving the necessity of a complex transcriptional regulation system is most likely. One would expect that transcripts involved in complex post-translational protein modification and signal transduction systems would have some of the most complicated inputs in transcriptional initiation. Kinases and ubiquitinases might be associated with the cell cycle, which uses complex transcriptional activation and translational modification (Alberts and co-workers 2002). While many terms primarily relating to transcriptional regulation are not found in this set of the most highly enriched terms, the genes associated with the regulation of transcription term still have significantly higher T than other genes ($p = 3 \times 10^{-4}$; FDR $= 6 \times 10^{-3}$).

Another possibility is that transcripts annotated with these terms have reached a local maximum in the fitness landscape, and that any further modifications would result in binding profile changes so disruptive that purifying selection would prevent them from fixing. This explanation would require the model's predictions to be absolutely correct much of the time, with little error. I do not think this is possible because the Sunflower model is of necessity currently incomplete, as discussed in subsection 3.4.5.

Table 5.3 contains GO biological process terms enriched in genes with low-T TSS-flanking regions. Many of these terms have known associations with positive selection and recent duplication in mammals (Yang and Bielawski 2000), including immune and defense response, and sensory perception (especially

olfactory perception). The bulk of the molecular function terms in Table 5.4 are also related to sensory perception.

GO term	p	FDR
immune response	1×10^{-19}	$< 1 \times 10^{-4}$
defense response	4×10^{-18}	$< 1 \times 10^{-4}$
response to stimulus	7×10^{-16}	$< 1 \times 10^{-4}$
sensory perception of chemical stimulus	2×10^{-15}	$< 1 \times 10^{-4}$
sensory perception of smell	9×10^{-15}	$< 1 \times 10^{-4}$
immune system process	3×10^{-12}	$< 1 \times 10^{-4}$
antigen processing and presentation	6×10^{-12}	$< 1 \times 10^{-4}$
defense response to bacterium	2×10^{-9}	$< 1 \times 10^{-4}$
multi-organism process	5×10^{-9}	$< 1 \times 10^{-4}$
inflammatory response	2×10^{-8}	$< 1 \times 10^{-4}$
response to other organism	3×10^{-8}	$< 1 \times 10^{-4}$
response to bacterium	5×10^{-8}	$< 1 \times 10^{-4}$
response to wounding	2×10^{-7}	$< 1 \times 10^{-4}$
response to biotic stimulus	2×10^{-7}	$< 1 \times 10^{-4}$
response to external stimulus	5×10^{-7}	$< 1 \times 10^{-4}$
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	5×10^{-7}	$< 1 \times 10^{-4}$

Table 5.3: GO biological process terms enriched in genes with low-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p = 3 \times 10^{-3}$).

GO term	p	FDR
olfactory receptor activity	9×10^{-19}	$< 1 \times 10^{-4}$
peptide receptor activity	9×10^{-10}	$< 1 \times 10^{-4}$
peptide receptor activity, G-protein coupled	9×10^{-10}	$< 1 \times 10^{-4}$
peptide binding	8×10^{-9}	$< 1 \times 10^{-4}$
rhodopsin-like receptor activity	1×10^{-8}	$< 1 \times 10^{-4}$
G-protein coupled receptor activity	4×10^{-7}	$< 1 \times 10^{-4}$

Table 5.4: GO molecular function terms enriched in genes with low-T TSS-flanking regions (Wilcoxon rank sum test; $p < 10^{-6}$). The global analysis is significant (Kolmogorov-Smirnov–like test; $p < 10^{-4}$).

There are several possible explanations for the particular set of terms enriched in low-T genes, which broadly are the inverse of the potential explanations for the high T enriched set. First, transcripts in these categories may have fewer

binding sites for the TFs in the model. This may be because the transcripts tend to have binding sites for unexamined TFs instead. Fewer binding sites for these TFs might also be explained if they were somehow selected against on these transcripts. These genes could be too new to have many well-developed TFBSs. Or, the systems involving these genes may exert less selective pressure to evolve a highly sophisticated transcriptional activation system with many TFBSs.

An alternative explanation to all of the previous explanations involving the density of TFBSs in the current model is that transcripts in these categories may have recently evolved, so they are unlikely to have reached a local maximum in the fitness landscape. I would consider this unlikely for the same reasons that caused me to dismiss the fitness landscape hypothesis for high T.

The most likely of these explanations is that these genes are too new to have developed many TFBSs, although it is not the inverse of my hypothesis for the high T set. Many of these transcripts do not have any alignment whatsoever to dog in the ± 100 region, which supports an argument that they are relatively new genes. Even so, this result is difficult to untangle from confounding factors.

5.3.2 The distance measurement d_T and other distance measurements

It would be possible to get a rough estimate of the amount of change induced by observed mutations using T_{d}. This quantity is not scaled, however, and different numbers of gaps in the alignments can by themselves yield great differences in T_{d}. Since I correct T in these cases to only include aligned columns, the proportion p_T, and therefore the distance d_T, are normalized for the number of gaps. I use d_T when measuring because it is transformed in the same way as measurements such as d_N and d_S.

In Figure 5.2, one can see the relationship between the transcription factor binding distance measurement d_T and the local neutral mutation rate as estimated by d_S, and one can also compare it with the relationships of d_N and d_I to d_S. In many ways, d_T behaves more like d_N in the comparison. There is a leading edge where high d_T and d_N values are not observed when coupled with low d_S, but at sufficiently high d_S there is little relationship between the two. Despite this there is only weak correlation ($r_{\mathrm{S}} = 0.22$; $p < 2.2 \times 10^{-16}$) between d_N and d_T. Values of d_I, however, are well-correlated with values of d_S (see

section 2.3).

5.3.3 Using the rate ratio ψ to correct for the local neutral mutation rate

To get a true idea of the amount of substitution affecting transcription factor binding, without the confounding factor of neutral mutation, one must correct for the local mutation rate. Therefore, I use the rate ratio $\psi = {}^{d_T}/_{d_S}$ to estimate the magnitude and direction of selection with regard to transcription factor binding.

Figure 5.3 shows the distribution of ψ values in this study. It is a simple unimodal distribution, with 68% of ψ values less than 1. This means that the majority of the promoters analyzed are under purifying selection, much as most protein-coding genes are under purifying selection (see subsection 1.1.1).

The relationship between ψ and the protein-coding estimate of selection $\omega = {}^{d_N}/_{d_S}$ is shown in Figure 5.4. The data points can be separated into three clusters. One cluster is characterized by $\omega, \psi > 10$. The extremely large values are due to very small d_S, including values where $S = 0$, but PAML (Yang 2007) reports $d_S = 0.0001$ (A. Vilella, personal communication). There are also 13 values where $d_S = 0$, and therefore ψ and ω are infinite. The cluster where $\omega < 0.005$ is due to very small d_N. The bulk of the data between these clusters consists in 9581 data points.

The variables ψ and ω are uncorrelated ($r_S = 0.076$; $p = 6.56 \times 10^{-14}$). This matches the conclusions of Taylor and co-workers (2006), who found that selective coding constraint in promoters and protein-coding sequence have little correlation.

The relationship between ψ and its underlying T is shown in Figure 5.5. This reveals that variation in ψ is not due wholly to variation in T. There is no significant correlation between the two variables ($r_S = 0.007$; $p = 0.4871$). Instead, the variation must come primarily from T_d ($r_S = 0.70$; $p < 2.2 \times 10^{-16}$), and therefore from the pattern of differences between human and rhesus, rather than the inherent properties of these promoter sequences under the Sunflower model.

Figure 5.6 displays the GO biological process terms enriched in genes with high-ψ TSS-flanking regions in either human–rhesus or human–dog comparisons. Note that the terms are dramatically different from the sets of terms most

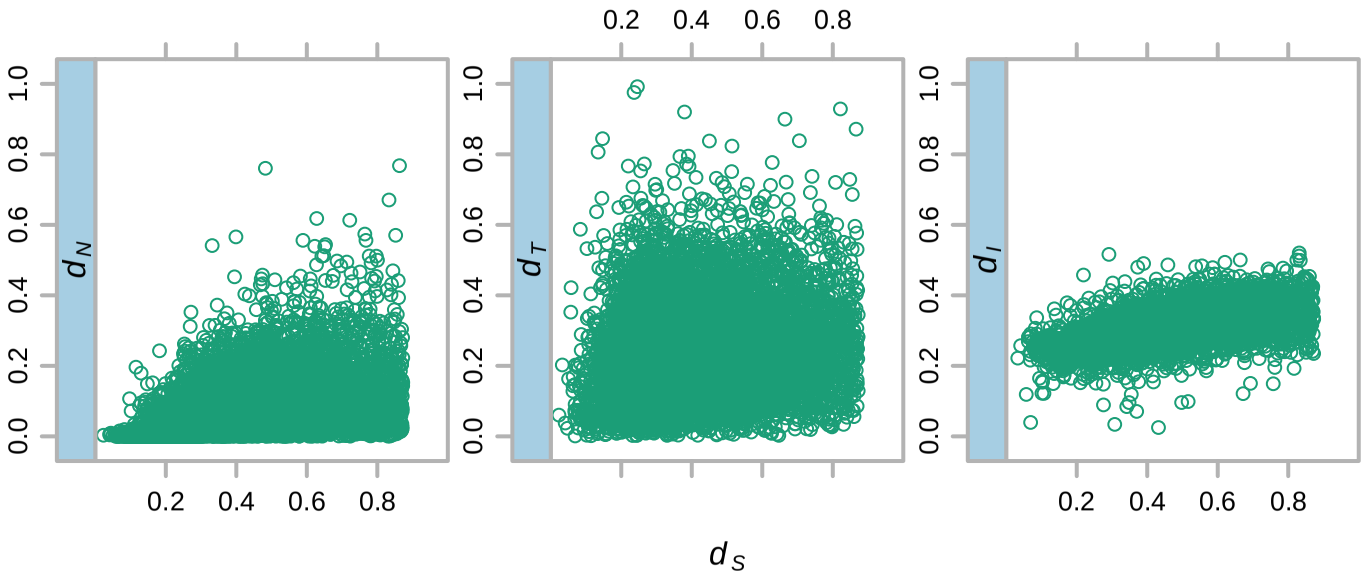


Figure 5.2: Scatterplot of d_S against other human–dog distance measurements. The panels show d_S against 10,242 values of d_N, 9058 values of d_T, and 6893 values of d_I. Only values in the range 0.0 to 1.0 are shown.

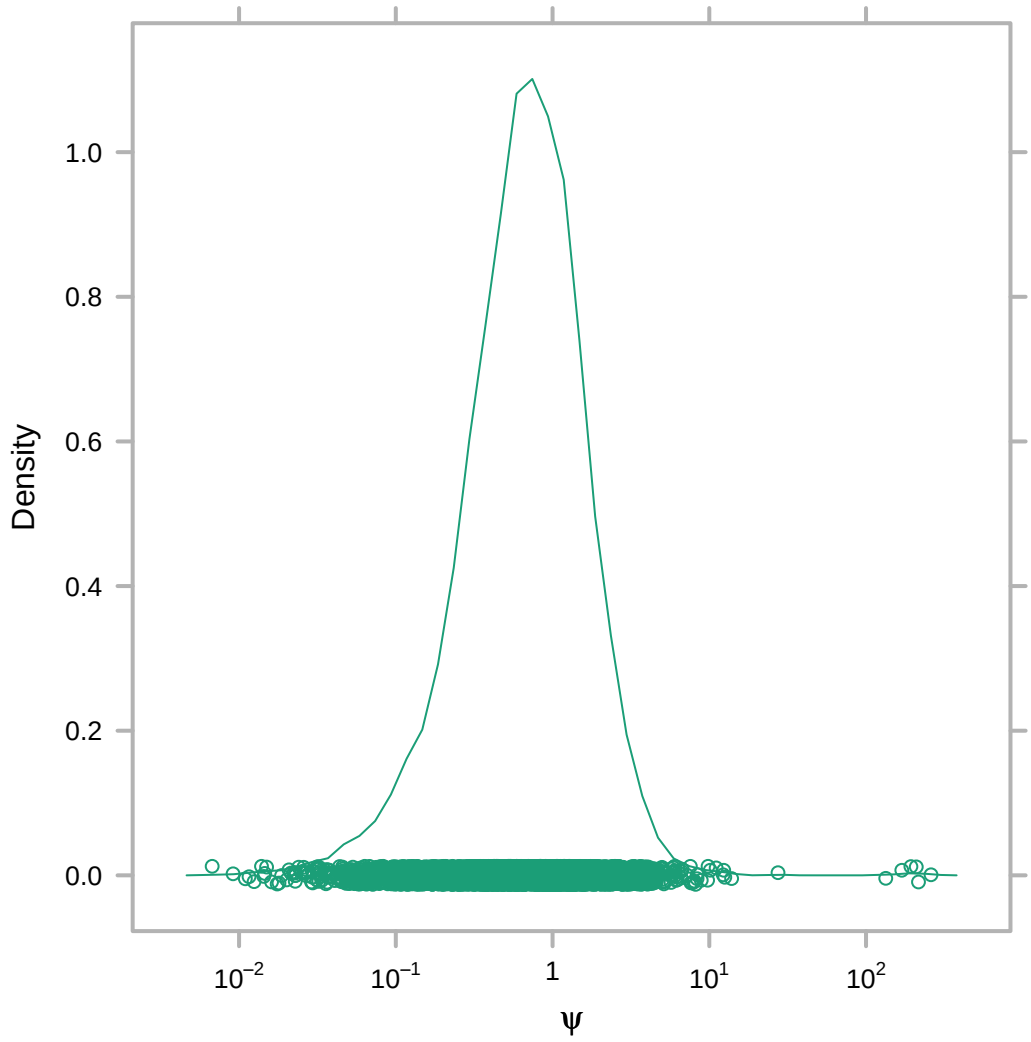


Figure 5.3: Kernel density plot of the distribution of 7930 human–rhesus ψ values. The x-coordinates of points indicate individual values of ψ while the y-coordinates are randomly jittered to add visibility.

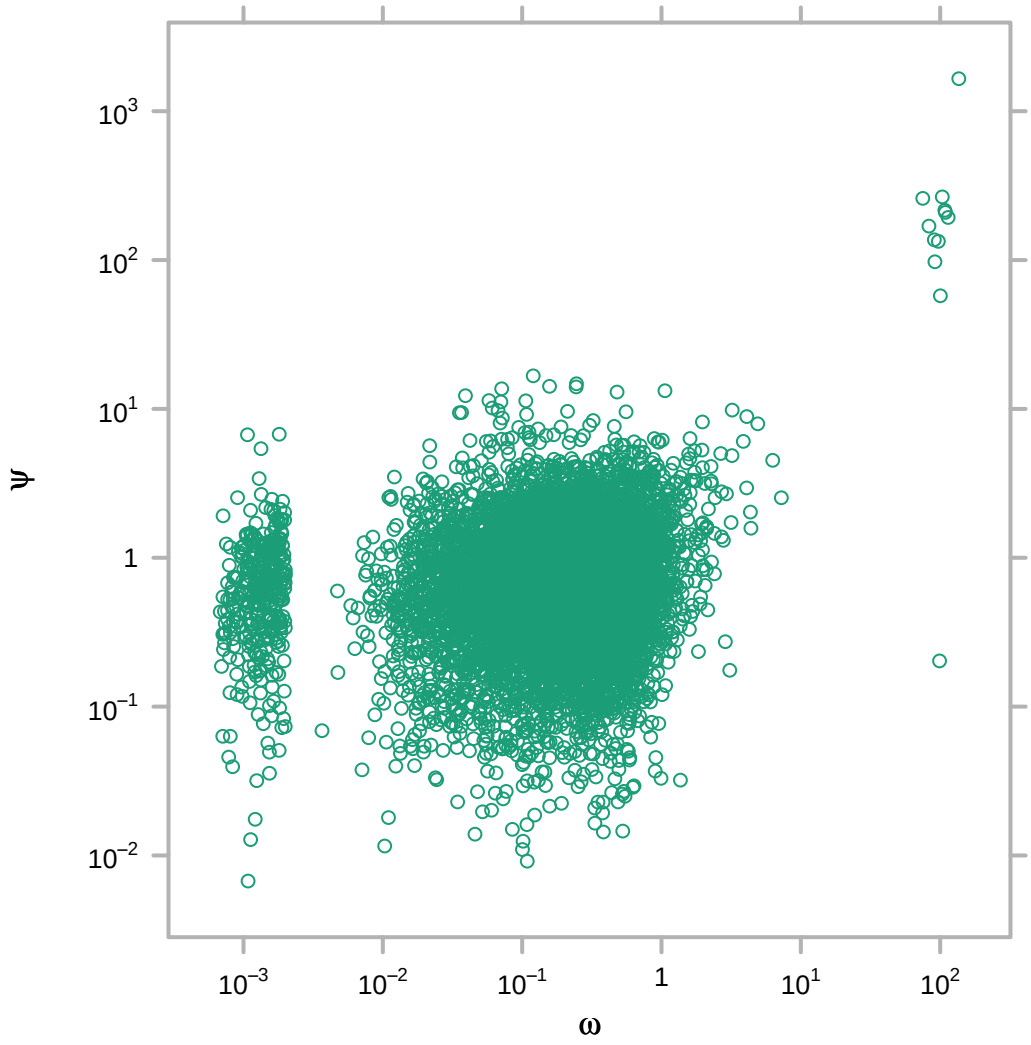


Figure 5.4: Log-log scatterplot of human–rhesus ψ values versus the corresponding ω values for 17,587 transcripts. Not included are 13 transcripts where $d_S = 0$.

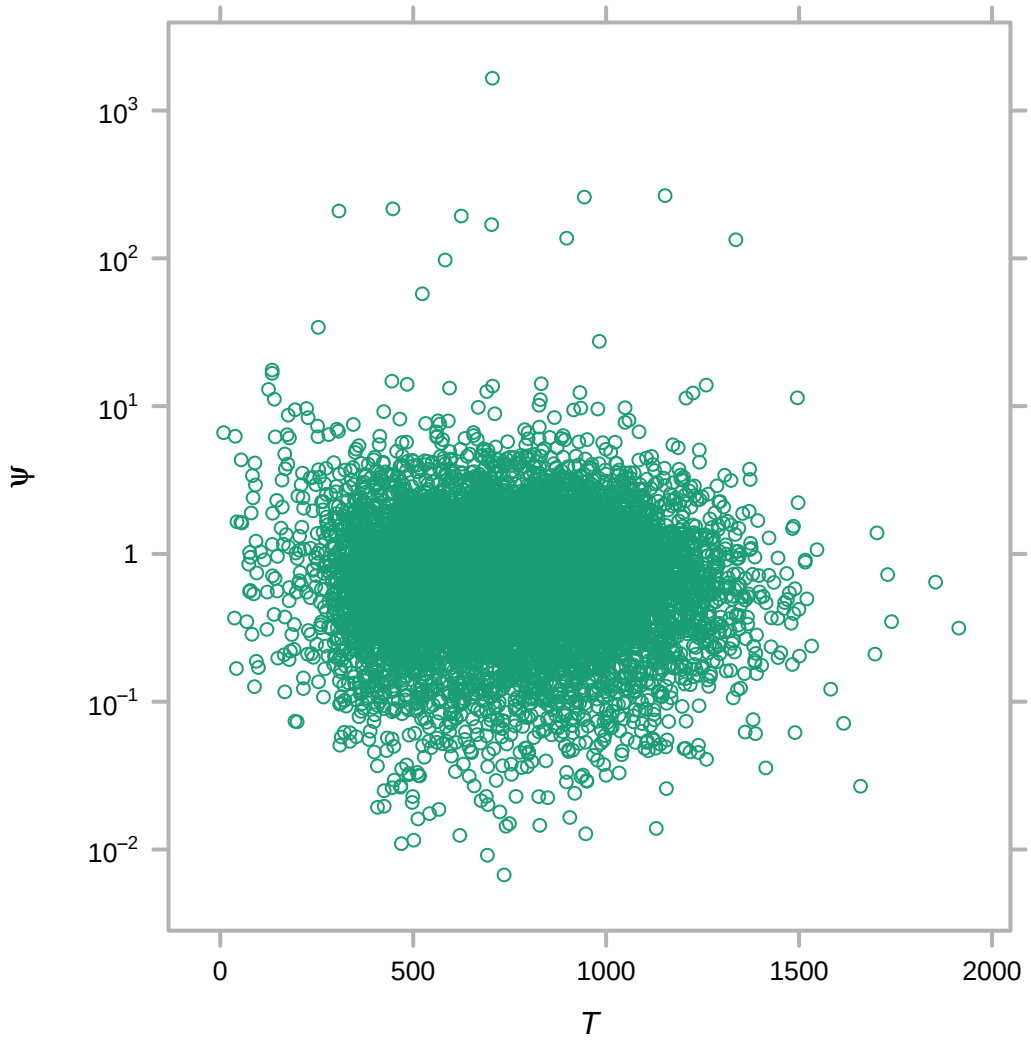


Figure 5.5: Scatterplot of human–rhesus ψ values versus the corresponding T values for 17,600 transcripts. Values of ψ are plotted on a logarithmic scale.

highly enriched for extreme values of T. Most significant in human–dog are a variety of terms related to the metabolism of proteins, nucleic acid, and lipids, while these terms are less important in human–rhesus. Figure 5.7 shows the corresponding molecular function terms, which broadly consist of the metabolic functions underlying the enriched biological processes.

Specifically enriched in human–rhesus for high ψ (as well as in the human–dog to some extent) are terms related to the Golgi apparatus. Figure 5.8 shows the distribution of ψ values associated with Golgi vesicle transport. The median ψ of these genes is 1.16. This is over 1, so most of the Golgi vesicle transport genes are under positive transcriptional selection. There are also outliers at ψ values of 3.70, 11.06, and 13.88.

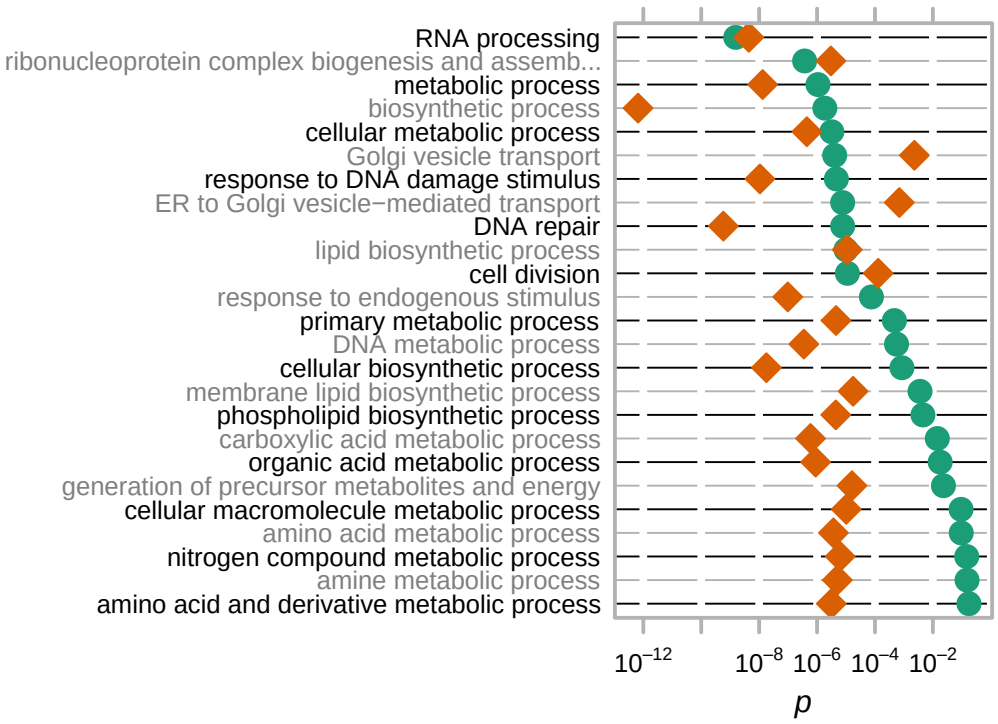


Figure 5.6: GO biological process terms enriched in genes with high-ψ TSS-flanking regions in either rhesus (teal circles) or dog (orange diamonds). The 25 most significant terms pooled from both species are displayed, sorted by rhesus significance values (Wilcoxon rank sum test). The global analyses are significant (Kolmogorov-Smirnov–like test; rhesus $p = 10^{-3}$; dog $p = 10^{-3}$).

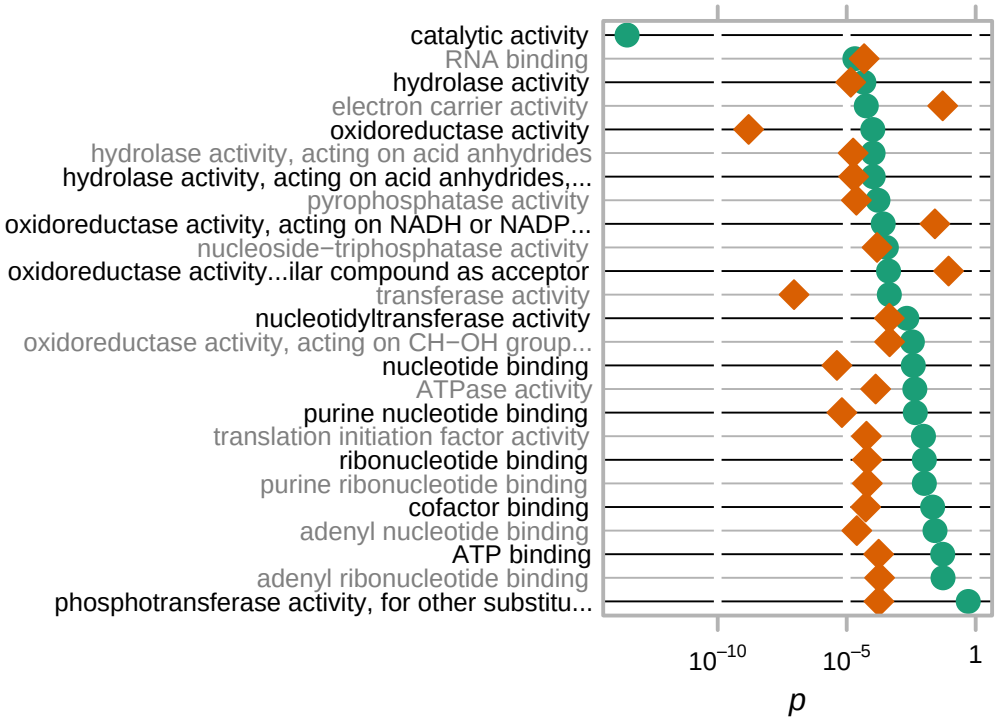


Figure 5.7: GO molecular function terms enriched in genes with high-ψ TSS-flanking regions in either rhesus (teal circles) or dog (orange diamonds). The 25 most significant terms pooled from both species are displayed, sorted by rhesus significance values (Wilcoxon rank sum test). The global analyses are significant (Kolmogorov-Smirnov–like test; rhesus $p = 3 \times 10^{-3}$; dog $p < 10^{-4}$).

Figure 5.9 shows the GO biological process terms enriched for genes with low-ψ TSS-flanking regions. These terms predominantly fit into three categories—development, regulation (especially of nucleic acids), and communication. It makes sense that developmental and regulatory genes would, in general, experience lower change in their transcription factor binding sites than other genes, because of their importance in complex signal transduction networks as master or intermediate regulators. This result matches previous reports that genes involved in complex processes such as development have conserved upstream regions (Lee and co-workers 2005). This is analogous to the situation with Hox genes, which display some of the most remarkable conservation across species in their amino acid sequence (Ruddle and co-workers 1994), ostensibly because

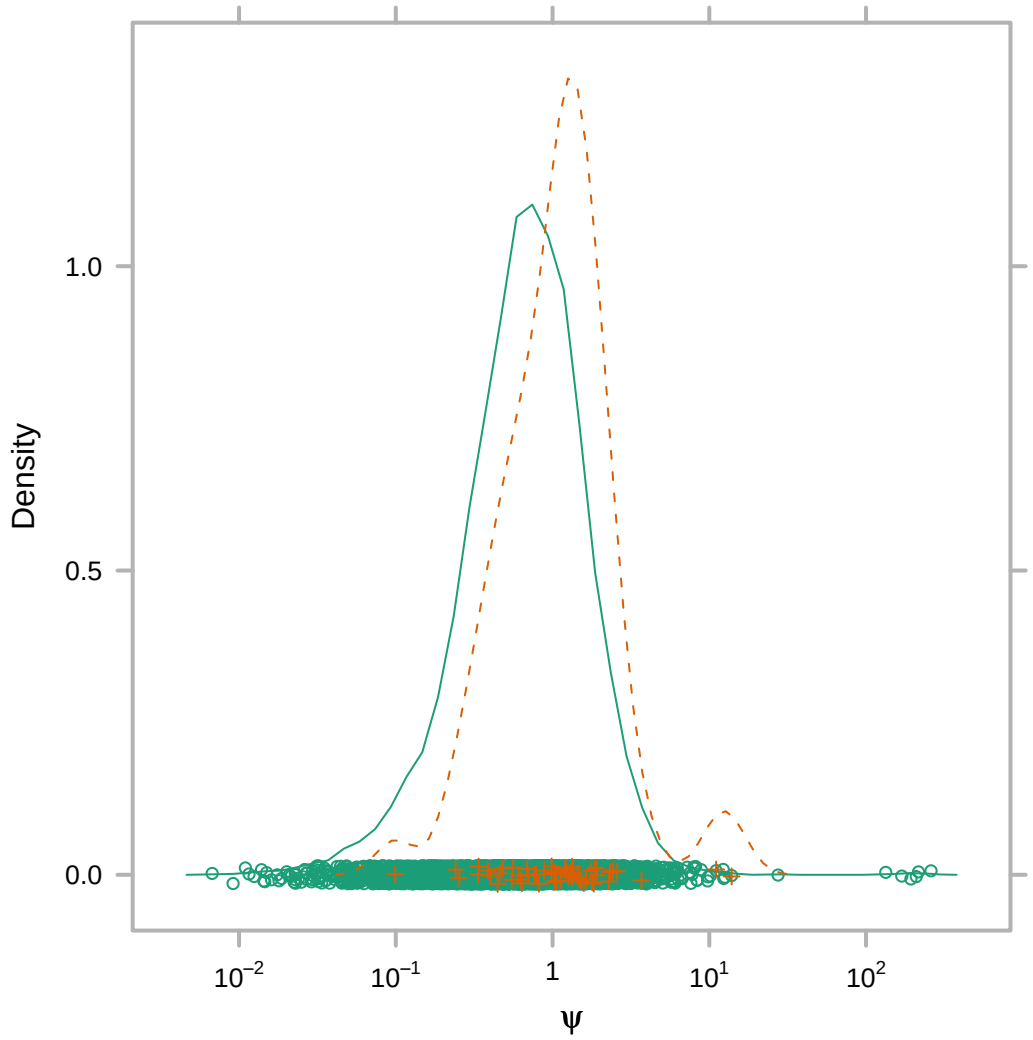


Figure 5.8: Kernel density plot of the distribution of all 7930 human–rhesus ψ values (teal points and solid line) and the distribution of ψ in only the 56 transcripts associated with the Golgi vesicle transport GO term (orange points and dashed line). Points in the Golgi vesicle transport subset are also plotted in the parent distribution.

disruption in these genes also disrupts a number of downstream systems.

Figure 5.10 shows the GO molecular function terms enriched for genes with low-ψ TSS-flanking regions, which includes many terms relating to binding of macromolecules and simple solutes. It is possible that these functions are important to the classes of biological processes also enriched for low ψ.

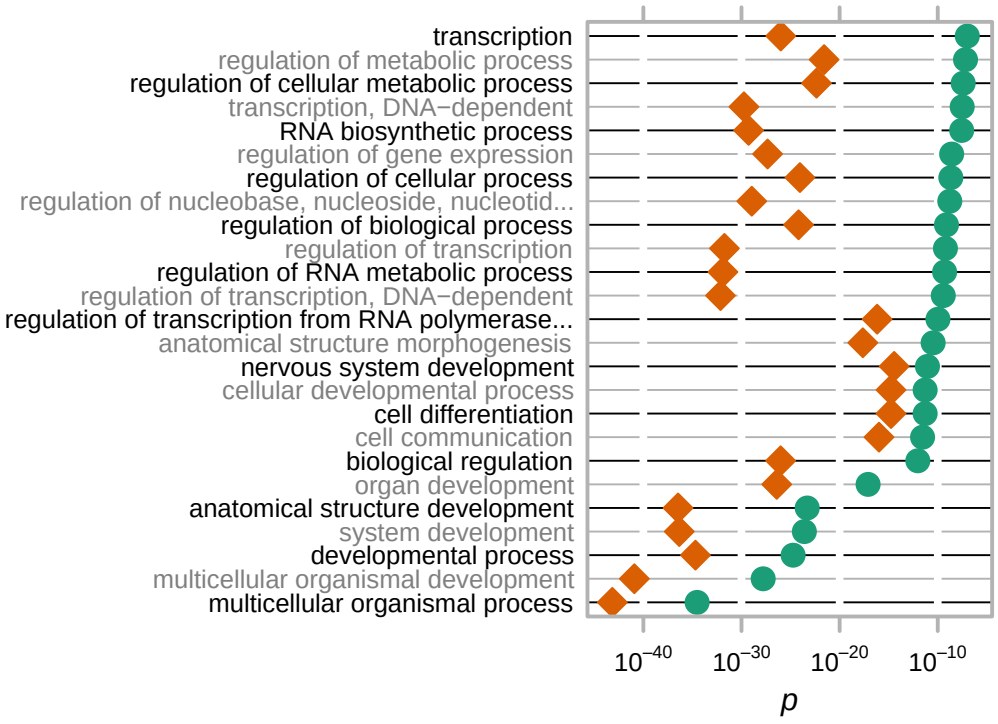


Figure 5.9: GO biological process terms enriched in genes with low-ψ TSS-flanking regions in either rhesus (teal circles) or dog (orange diamonds). The 25 most significant terms pooled from both species are displayed, sorted by rhesus significance values (Wilcoxon rank sum test). The global analyses are significant (Kolmogorov-Smirnov–like test; rhesus $p < 10^{-4}$; dog $p < 10^{-4}$).

5.4 Discussion

5.4.1 Contrasting the d_T/d_S model with the d_N/d_S model

The proportion p_T has a similar range, [0, 1], as p_N and p_S. Therefore d_T is defined using the Jukes-Cantor model in the same range [0, 0.75) as d_N and d_S. The use

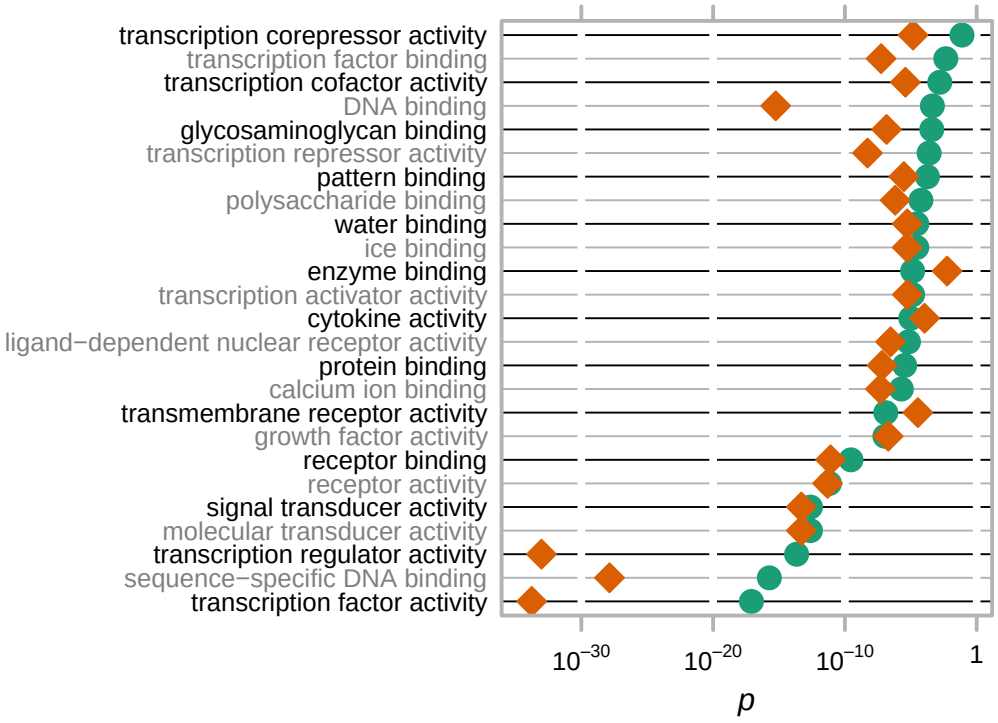


Figure 5.10: GO molecular function terms enriched in genes with low-ψ TSS-flanking regions in either rhesus (teal circles) or dog (orange diamonds). The 25 most significant terms pooled from both species are displayed, sorted by rhesus significance values (Wilcoxon rank sum test). The global analyses are significant (Kolmogorov-Smirnov–like test; rhesus $p < 10^{-4}$; dog $p < 10^{-4}$).

of ratios of continuous variables does not really fit the assumptions of the Jukes-Cantor model, which involve measurements of discrete nucleotide substitution, but I am merely continuing the extension of the Jukes-Cantor model to partial synonymity (such as in Nei and Gojobori 1986) in protein-coding sequence onto noncoding sequence. Here, the model provides a similar transformation as that used for d_N and d_S to supply some measure of correction for multiple mutations in an evolutionary lineage.

Strictly speaking, d_T is not a metric as d_N, d_S, and d_I are in some of their formulations. This is because it is derived from relative entropy calculations and is therefore asymmetrical. However, none of these satisfy the mathematical definition of a distance because they do not satisfy the triangle inequality

(Felsenstein 2004).

I have demonstrated how one can use ψ to identify putative directional selection. While the underlying model is admittedly incomplete (see subsection 3.4.5), there are arguably more problems with the simplistic model inherent in ω. For every question one asks about the underlying assumptions of ψ, one finds the same problems with ω. Nevertheless, this simple model has proven its worth over time, despite the existence of newer and more sophisticated models for characterizing the phenotypic impact of amino acid substitutions.

5.4.2 Heterogeneous promoter composition over time

One of the problems with using ψ for comparisons is that transcription factor binding sites may have been intermittent. Approximately one third of TFBS in a human–rodent comparison are functional in only one of the species (Dermitzakis and Clark 2002). Furthermore, mammalian promoters have varied significantly in evolutionary rates over time (Taylor and co-workers 2006). Heterotachy is a problem in many evolutionary models, but becomes more significant here if it is particularly pronounced in the regions under analysis.

5.4.3 Comparison with other noncoding models

This work relies on a generalized model for molecular evolution that can work for any model of phenotype and genotype, of which existing models are a special case. Despite the limitations of this model and the data used to parametrize it, it is a useful proof of concept that will become more useful as more data and better models become available. In comparison, the neutral indel model of Lunter and co-workers (2006), does not presume a model that gives separate measurements of neutral and non-neutral mutation. Instead, it assumes that indels occur primarily in neutral areas. While it can identify neutrally evolving regions of the genome, it lacks the explanatory power at the level of individual nucleotides that my transcription factor model has. The neutral indel model relies on looking at which indels were fixated *a posteriori* rather than deciding which kinds of mutations affect phenotype more *a priori*, like the transcription factor model in this work, or commonly used synonymous coding sites models.

When there are so many methods for determining selection in noncoding

sequence based on population genetics (Sabeti and co-workers 2006), one might ask what the point is of a method based on a model of function? Methods of determining selection based on functional changes can detect moderate selection over much longer timescales—millions of years rather than hundreds of thousands of years (Sabeti and co-workers 2006). Additionally, collecting variation data is resource-intensive and we may not have it in the same abundance for other species as we will soon have for humans.

Methods of determining selection based on functional models will continue to be important in the years to come. I hope that Sunflower is only one of the first of many.

Bibliography

Adkins RM, Gelke EL, Rowe D, Honeycutt RL. 2001. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol Biol Evol* 18(5):777–791.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002. *Molecular biology of the cell*. New York: Garland, 4th edition.

Altet F, Vilata I, Prater S, Mas V, Hedley T, Valentino A, Whitaker J, and coworkers. 2007. PyTables: hierarchical datasets in Python. ⟨http://www.pytables.org/⟩.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.

Benos PV, Lapedes AS, Fields DS, Stormo GD. 2001. SAMIE: statistical algorithm for modeling interaction energies. In *Pac Symp Biocomput*, 115–126.

Benos PV, Lapedes AS, Stormo GD. 2002. Is there a code for protein-DNA recognition? Probab(ilistical)ly... *Bioessays* 24(5):466–475.

Berg OG, von Hippel PH. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193(4):723–750.

Bernardi G. 1995. The human genome: organization and evolutionary history. *Annu Rev Genet* 29:445–476.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241(1):3–17.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74(6):1111–1120.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. 2005. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 15(2):116–124.

Bird CP, Stranger BE, Dermitzakis ET. 2006. Functional variation and evolution of non-coding DNA. *Curr Opin Genet Dev* 16(6):559–564.

Birney E. 2002. Wise2. ⟨http://www.ebi.ac.uk/Wise2/⟩.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Gräf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlić A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ. 2006. Ensembl 2006. *Nucleic Acids Res* 34(Database issue):D556–D561.

Blake JA, Eppig JT, Richardson JE, Davisson MT. 2000. The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Res* 28(1):108–111.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* 317(5839):815–819.

Bouwman P, Philipsen S. 2002. Regulation of the activity of Sp1-related transcription factors. *Mol Cell Endocrinol* 195(1-2):27–38.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20(18):3710–3715.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H. 1994. Sp1 elements protect a CpG island from *de novo* methylation. *Nature* 371(6496):435–438.

Brown TA. 2006. *Genomes*. Oxford: Wiley-Liss, 3rd edition.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13(4):721–731.

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. 2008. The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res* 36(Database issue):D445–D448.

Bulyk ML. 2006. DNA microarray technologies for measuring protein–DNA interactions. *Curr Opin Biotechnol* 17(4):422–430.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* 32(Database issue):D262–D266.

Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22(22):2729–2734.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38(6):626–635.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3(4):285–298.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. *Nature* 401(6751):344.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.

Castresana J. 2002. Estimation of genetic distances from human and mouse introns. *Genome Biol* 3(6):research0028.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7(2):98–108.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68(2):444–456.

Chiaromonte F, Yap VB, Miller W. 2002. Scoring pairwise genomic sequence alignments. In *Pac Symp Biocomput*, 115–126.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69–87.

Chomez P, De Backer O, Bertrand M, De Plaen E, Boon T, Lucas S. 2001. An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res* 61(14):5544–5551.

Cleveland WS. 1993. *Visualizing data*. Summit, NJ: Hobart Press, 1st edition.

Cleveland WS, Devlin SJ. 1988. Locally-weighted fitting: an approach to fitting analysis by local fitting. *J Am Stat Assoc* 83:596–610.

Courey AJ, Holtzman DA, Jackson SP, Tjian R. 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell* 59(5):827–836.

Courey AJ, Jia S. 2001. Transcriptional repression: the long and the short of it. *Genes Dev* 15(21):2786–2796.

Cox JC, Hayhurst A, Hesselberth J, Bayer TS, Georgiou G, Ellington AD. 2002. Automated selection of aptamers against protein targets translated *in vitro*: from gene to aptamer. *Nucleic Acids Res* 30(20):e108.

Cretekos CJ, Wang Y, Green ED, NISC Comparative Sequencing Program, Martin JF, Rasweiler JJ, Behringer RR. 2008. Regulatory divergence modifies limb length between mammals. *Genes Dev* 22(2):141–151.

Das MK, Dai HK. 2007. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 8 Suppl 7:S21.

Dermitzakis ET, Bergman CM, Clark AG. 2003. Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. *Mol Biol Evol* 20(5):703–714.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19(7):1114–1121.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. *Biological sequence analysis*. Cambridge: Cambridge University Press, 1st edition.

Eckert D, Buhl S, Weber S, Jäger R, Schorle H. 2005. The AP-2 family of transcription factors. *Genome Biol* 6(13):246.

Egholm M, Srinivasan M, Wheeler DA, McGuire A, He W, Chen YJ, Makhijani V, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Chinault AC, Yuan Y, Jiang H, Song X, Liu Y, Nazareth L, Scherer S, Lupski JR, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The genome of James Dewey Watson. In preparation.

Ellegren H, Smith NGC, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 13(6):562–568.

Ellington AD, Szostak JW. 1990. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818–822.

Ellson J, Gansner ER, Koutsofios E, North SC, Woodhull G. 2004. Graphviz and Dynagraph—static and dynamic graph drawing tools. In Junger M, Mutzel P, eds., *Graph drawing software*, 127–148. Berlin: Springer-Verlag.

Emes RD, Beatson SA, Ponting CP, Goodstadt L. 2004. Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Res* 14(4):591–602.

Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418(6900):869–872.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799–816.

Eyre-Walker A. 1992. The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. *Genet Res* 60(1):61–67.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol* 21(10):569–575.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.

Felsenstein J. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer Associates, 1st edition.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). ⟨http://evolution.genetics.washington.edu/phylip.html⟩.

Finn RD, Stalker JW, Jackson DK, Kulesha E, Clements J, Pettett R. 2007. ProServer: a simple, extensible Perl DAS server. *Bioinformatics* 23(12):1568–1570.

Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlić A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S. 2008. Ensembl 2008. *Nucleic Acids Res* 36(Database issue):D707–D714.

Fonseca J. 2007. Gprof2Dot. ⟨http://code.google.com/p/jrfonseca/wiki/Gprof2Dot⟩.

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. 2008. A code for transcription initiation in mammalian genomes. *Genome Res* 18(1):1–12.

Frommer M, McDonald L, Millar D, Collis C, Watt F, Grigg G, Molloy P, Paul C. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89(5):1827–1831.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol* 22(3):650–658.

Fullerton SM, Carvalho AB, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol* 18(6):1139–1142.

Gene Ontology Consortium. 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 34(Database issue):D322–D326.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, de Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R,

Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428(6982):493–521.

Gilbert SF. 2000. *Developmental biology*. Sunderland, MA: Sinauer Associates, 6th edition.

Gillespie JH. 2004. *Population genetics: a concise guide*. Baltimore: John Hopkins University Press.

Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, Smith CW. 2006. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol* 7(1):R1.

Graur D, Li WH. 2000. *Fundamentals of molecular evolution*. Sunderland, MA: Sinauer Associates, 2nd edition.

Grilli M, Chiu JJ, Lenardo MJ. 1993. NF-κB and Rel: participants in a multiform transcriptional regulatory system. *Int Rev Cytol* 143:1–62.

Hamblin MT, Rienzo AD. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66(5):1669–1679.

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 13(1):13–26.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22(2):160–174.

HDF Group. 2007. *HDF5 user's guide: HDF5 release 1.6.6*. ⟨http://www.hdfgroup.org/HDF5/doc1.6/UG/⟩.

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* 72(6):1527–1535.

Hoffman MM. 2006. xyddplot. ⟨http://www.ebi.ac.uk/~hoffman/software/xyddplot/⟩.

Hoffman MM, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. *Mol Biol Evol* 24(2):522–531.

Hoffman MM, Khrapov MA, Cox JC, Yao J, Tong L, Ellington AD. 2004. AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res* 32(Database issue):D174–D181.

Holmes I, Durbin R. 1998. Dynamic programming alignment accuracy. *J Comput Biol* 5(3):493–504.

Huber PJ. 1981. *Robust statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley, 1st edition.

Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 296(5):1205–1214.

Hurst LD. 2002. The *Ka*/*Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet* 18(9):486.

Iida K, Akashi H. 2000. A test of translational selection at "silent" sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* 261(1):93–105.

Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol* 40(2):190–226.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018):695–716.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437(7063):1299–1320.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

Jabbari K, Bernardi G. 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* 333:143–149.

Jareborg N, Birney E, Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res* 9(9):815–824.

Jukes TH, Cantor CR. 1964. Evolution of protein molecules. In Munro HN, Allison JB, eds., *Mammalian protein metabolism*, 21–132. New York: Academic.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* 94(5):1872–1877.

Kaiser J. 2008. A plan to capture human diversity in 1000 genomes. *Science* 319(5862):395.

Kamal M, Xie X, Lander ES. 2006. A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci U S A* 103(8):2740–2745.

Karn RC, Laukaitis CM. 2003. Characterization of two forms of mouse salivary androgen-binding protein (abp): implications for evolutionary relationships and ligand-binding function. *Biochemistry* 42(23):7162–7170.

Karn RC, Nachman MW. 1999. Reduced nucleotide variability at an androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. *Mol Biol Evol* 16(9):1192–1197.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 14(1):160–169.

Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 31(13):3576–3579.

Kernighan BW, Ritchie DM. 1988. *The C programming language*. Englewood Cliffs, NJ: Prentice Hall, 2nd edition.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217(5129):624–626.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16(2):111–120.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Pääbo S. 2007. The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr Biol* 17(21):1908–1912.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80(4):727–739.

Kunsch C, Ruben SM, Rosen CA. 1992. Selection of optimal κB/Rel DNA-binding motifs: interaction of both subunits of NF-κB with DNA is required for transcriptional activation. *Mol Cell Biol* 12(10):4412–4421.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng

JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human genome. *Genomics* 13(4):1095–1107.

Lee S, Kohane I, Kasif S. 2005. Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics* 6:168.

Leisch F. 2002. Sweave: dynamic generation of statistical reports using literate data analysis. In Härdle W, Rönz B, eds., *Compstat 2002—proceedings in computational statistics*, 575–580. Heidelberg: Physica Verlag.

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2(2):13.

Lenhard B, Wasserman WW. 2002. TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics* 18(8):1135–1136.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* 18(7):337–340.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AWC, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. 2007. The diploid genome sequence of an individual human. *PLoS Biol* 5(10):e254.

Li D, Da L, Tang H, Li T, Zhao M. 2008. CpG methylation plays a vital role in determining tissue- and cell-specific expression of the human cell-death-inducing DFF45-like effector A gene through the regulation of Sp1/Sp3 binding. *Nucleic Acids Res* 36(1):330–341.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* 5(1):182–187.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SM, Sutter NB, Thomas R, Webber C, Baldwin J, Abebe A, Abouelleil A, Aftuck L, Ait-Zahra M, Aldredge T, Allen N, An P, Anderson S, Antoine C, Arachchi H, Aslam A, Ayotte L, Bachantsang P, Barry A, Bayul T, Benamara M, Berlin A, Bessette D, Blitshteyn B, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Brown A, Cahill P, Calixte N, Camarata J, Cheshatsang Y, Chu J, Citroen M, Collymore A, Cooke P, Dawoe T, Daza R, Decktor K, DeGray S, Dhargay N, Dooley K, Dooley K, Dorje P, Dorjee K, Dorris L, Duffey N, Dupes A, Egbiremolen O, Elong R, Falk J, Farina A, Faro S, Ferguson D, Ferreira P, Fisher S, FitzGerald M. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819.

Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* 12(5):832–839.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* 2(1):e5.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* 19(6):330–338.

Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the drosophila embryo. *Proc Natl Acad Sci U S A* 99(2):763–768.

Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. 2003. Distribution of NF-κB–binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* 100(21):12247–12252.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652–654.

McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res* 18(2):252–260.

McPherson LA, Weigel RJ. 1999. AP2α and AP2γ: a comparison of binding site specificity and *trans*-activation of the estrogen receptor promoter and single site promoter constructs. *Nucleic Acids Res* 27(20):4040–4049.

Messina DN, Glasscock J, Gish W, Lovett M. 2004. An orfeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res* 14(10B):2041–2047.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448(7153):553–560.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol* 16(1):23–36.

Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol* 5(12):R98.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443–453.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3(5):418–426.

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. Oxford: Oxford University Press.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31(13):3812–3814.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* 86(Pt 6):641–647.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* 8(11):857–868.

Ogurtsov AY, Sunyaev S, Kondrashov AS. 2004. Indel-based evolutionary distance and mouse-human divergence. *Genome Res* 14(8):1610–1616.

Page RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 12(4):357–358.

Paten B, Beal K, Birney E. 2008. Pecan: large-scale consistency alignment. Submitted.

Ponting CP, Lunter G. 2006. Signatures of adaptive evolution within human non-coding sequence. *Hum Mol Genet* 15(Review Issue 2):R170–R175.

Pounds SB. 2006. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* 7(1):25–36.

Prlić A, Down TA, Kulesha E, Finn RD, Kähäri A, Hubbard TJP. 2007. Integrating sequence and structural biology with DAS. *BMC Bioinformatics* 8:333.

Prüfer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 8:41.

R Development Core Team. 2007. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. ⟨http://www.R-project.org⟩.

Rajewsky N, Vergassola M, Gaul U, Siggia ED. 2002. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3:30.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 30(17):3894–3900.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222–234.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276–277.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4(8):651–657.

Rockman MV, Wray GA. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol Biol Evol* 19(11):1991–2004.

Rogozin IB, Lyons-Weiler J, Koonin EV. 2000. Intron sliding in conserved gene families. *Trends Genet* 16(10):430–432.

Ruddle FH, Bartels JL, Bentley KL, Kappen C, Murtha MT, Pendleton JW. 1994. Evolution of hox genes. *Annu Rev Genet* 28:423–442.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312(5780):1614–1620.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 8(6):424–436.

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103(5):1412–1417.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18(20):6097–6100.

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 6(4):R33.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human–mouse alignments with BLASTZ. *Genome Res* 13(1):103–107.

Seeburg PH. 2002. A-to-I editing: new and old sites, functions and speculations. *Neuron* 35(1):17–20.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang JPZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* 442(7104):772–778.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 100(26):15776–15781.

Smit AFA, Hubley R, Green P. 1996. RepeatMasker. ⟨http://www.repeatmasker.org/⟩.

Smith AV, Thomas DJ, Munro HM, Abecasis GR. 2005. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* 15(11):1519–1534.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* 147(1):195–197.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A* 100(3):1056–1061.

Stoltzfus A, Logsdon JM Jr, Palmer JD, Doolittle WF. 1997. Intron "sliding" and the diversity of intron positions. *Proc Natl Acad Sci U S A* 94(20):10739–10744.

Stormo GD. 1990. Consensus patterns in DNA. *Methods Enzymol* 183:211–221.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET. 2007. Population genomics of human gene expression. *Nat Genet* 39(10):1217–1224.

Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E. 2007. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS ONE* 2(8):e807.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10(3):512–526.

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet* 2(4):e30.

Thiesen HJ, Bach C. 1990. Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* 18(11):3203–3209.

Trinklein ND, Aldred SJF, Saldanha AJ, Myers RM. 2003. Identification and functional analysis of human transcriptional promoters. *Genome Res* 13(2):308–312.

Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968):505–510.

Tufte ER. 2001. *The visual display of quantitative information*. Cheshire, CO: Graphics Press, 2nd edition.

Uemura H, Koshio M, Inoue Y, Lopez MC, Baker HV. 1997. The role of Gcr1p in the transcriptional activation of glycolytic genes in yeast *Saccharomyces cerevisiae*. *Genetics* 147(2):521–532.

van Rossum G. 2006. *Python reference manual*. Python Software Foundation. ⟨http://docs.python.org/ref/⟩.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman

C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. *Science* 291(5507):1304–1351.

Vlieghe D, Sandelin A, Bleser PJD, Vleminckx K, Wasserman WW, van Roy

F, Lenhard B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 34(Database issue):D95–D97.

Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 26(2):225–228.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5(4):276–287.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

Webber C, Ponting CP. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res* 15(12):1787–1797.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol* 23(6):1203–1216.

Wellcome Trust Sanger Institute. 2007. *Danio rerio* sequencing project. ⟨http://www.sanger.ac.uk/Projects/D_rerio/⟩.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards

YJK, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3(1):e7.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 20(9):1377–1419.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34(Database issue):D187–D191.

Yandell M, Mungall CJ, Smith C, Prochnik S, Kaminker J, Hartzell G, Lewis S, Rubin GM. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput Biol* 2(3):e15.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15(12):496–503.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17(1):32–43.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 7:166.

Yuh C, Bolouri H, Davidson E. 2001. *Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control. *Development* 128(5):617–629.