Abstract

Huang, Weichun. Computational methods for identifying and characterizing the human gene regulatory regions and *cis*-elements. (Under the direction of Bruce S. Weir)

The identification of functional regulatory regions and *cis*-elements is a preliminary step toward the reconstruction of gene regulatory networks. Comparative genomics has been demonstrated to be a powerful approach for motif discovery. However, the accurate alignment of complex genomic sequences, especially those of mammalians, remains a computational challenge. In chapter 2, we propose a novel pairwise alignment system, ACANA, to improve the alignment quality of genomic sequences. Compared with top competing alignment tools, ACANA achieves better alignment quality in aligning divergent sequences for both local and global alignments. When applied to the upstream sequences of human-mouse orthologs, ACANA is able to reliably detect the conserved functional regions containing most *cis*-elements.

Statistical motif modeling is another fundamental computational approach for motif prediction in large genome sequence. In chapter 3, we introduce the mixture of optimized Markov models to reduce false motif discovery rate in large genomic sequences. Our model is not only able to incorporate most dependency information within a motif by optimizing the arrangement of motif positions, but also flexible for adjusting model complexity limited by the size of training data. We implement the mixture model in our OMiMa system. Using OMiMa, we demonstrate that our model can improve motif prediction accuracy.

Although the reconstruction of complete human gene regulatory networks, at present, remains a distant hope, it is still possible to infer some distinct features of the networks from the available data. In chapter 4, we present an example of inferring major evolutionary features of human gene regulatory networks by combining information from both gene sequence data and functional annotations. We systematically analyze the association between gene function and upstream region conservation for human-rodent orthologs. Our study

shows that upstream regulatory regions of developmental transcription regulators, such as Hox genes, are extremely conserved while those of catalytic enzymes are significantly less conserved. We suggest that developmental and other important regulators constitute the central hub of human gene regulatory networks.

COMPUTATIONAL METHODS FOR IDENTIFYING AND CHARACTERIZING THE HUMAN GENE REGULATORY REGIONS AND *CIS*-ELEMENTS

BY

WEICHUN HUANG

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

BIOINFORMATICS

Raleigh
2005

APPROVED BY:

William R. Atchley	Leping Li
Jeffrey L. Thorne	Russell D. Wolfinger

Bruce S. Weir
CHAIR OF ADVISORY COMMITTEE

献给我的妻子,父母,姐姐和弟弟

To my wife, parents, sisters and brother

Biography

Weichun Huang was born at Qianjin, a small village in Yongkang, Zhejiang province, the People's Republic of China. In 1991, he entered Hunan Medical University (currently, Xiangya School of Medicine, Central South University), where he received his Bachelor of Medicine degree in June, 1996. After graduation, he was admitted to Peking Union Medical College & Chinese Academy of Medical Sciences, where he earned his M.S. in Biomedical Engineering in 1999. In August, 1999, Weichun, at the encouragements of his elder sister, flew to the United States to pursue his American dream. He spent the first year and a half in pursuing Ph.D. degree in Bioengineering while working as a teaching/research assistant at the University of Toledo. He then decided to follow his interests to do research in computational biology. He was enrolled in the Bioinformatics Program at North Carolina State University (NCSU) in 2001. In May, 2003, he received his M.S. in Bioinformatics and has since continued pursuing his Ph.D. degree at the same program under direction of Dr. Bruce S. Weir. During his study at NCSU, Weichun also worked as a research intern at Research Triangle Institute (RTI), RTP from 2001 to 2002. After that, he has been working as a guest researcher at the Biostatistics Branch of National Institute of Environmental Health Sciences (NIEHS), RTP, NC.

Acknowledgments

What a long and challenging journey it has been. Looking back, at the moment when I finished the last sentence of this dissertation, I was even amazed by myself that how I could tread through such a long journey filled with constant distractions, temptations, chaos, and dangerous pitfalls. However, this is impossible without the help, trust, encouragement, guidance, protection, and sacrifices from many people to whom I would be grateful in all my life.

I'm especially grateful to my adviser, Dr. Bruce S. Weir, for his constant supports and guidances through the past years. I would also like to extend my deepest gratitude to the other members of my committee, Drs. William R. Atchley, Leping Li, Jeffrey L. Thorne, and Russell D. Wolfinger for their valuable advice and suggestions. In addition, I want to thank RTI International, and National Institute of Environmental Health Sciences (NIEHS) for supporting my internships. In particular, I want to thank Drs. Leping Li, David M. Umbach, and Clarice R. Weinberg at Biostatistics Branch at NIEHS, for many beneficial discussions and lectures. My special thanks go to all other people in Bioinformatics Research Center, especially, to our Genomic Program director, Dr. Barbara Sherry, and staff members Lisa Barefoot, Juliebeth Briseno, Debra Hibbard, and Alexandra Rogers, for all their wonderful help and supports.

Finally, I am indebted to my family, especially, my wife, my elder sister, and my parents, for their support, encouragement, protection, and constant love throughout the entire journey.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Review

Background

The genetic programs controlling cell differential and embryonic development are long believed to be hardwired in the genomic sequence of an organism (Davidson et al., 2002b). The "hardware" underlying the genetic program are Gene Regulatory Networks (GRN) that precisely regulate gene expression in both spatial and temporal dimensions. Reconstructing GRN to understand gene expression regulation is one of the greatest challenges in modern molecular biology. An important step in this challenge is the ability to identify all *cis*-regulatory elements, notably transcription factor binding sites (TFBS) in the flanking non-coding regions of genes. The binding of transcription factors to specific *cis*-regulatory elements, typically those in upstream region of the Transcriptional Start Site (TSS) of a gene, is a primary mechanism for regulating gene transcription from DNA to mRNA. Although regulation of gene expression can occur at any point from DNA to protein, the transcriptional regulation is the central control step for many genes. While prokaryotes often use no more than a few transcription factors and *cis*-elements to regulate a gene transcription, eukaryotic gene transcription often involves the coordination of multiple transcription factors, whose binding sites are spatially clustered and form so-called Cis-Regulatory Modules (CRM). The increasing complexity of regulatory element modular architecture is likely to be the major factor responsible for the more versatile and robust regulatory networks in higher eukaryotes than those in lower eukaryotes and prokaryotes (Kirschner and Gerhart, 1998; Locascio et al., 2002; Lynch and Conery, 2003). Such a coordinated modulation mechanism effectively integrates many different signaling pathways to provide much more complex regulation networks found in higher eukaryotic organisms.

Traditional experimental techniques, which worked well for investigating function and regulation of individual genes in the past, and may continue to play a critical role in testing and validating newly identified network components and features, are tedious and not suitable for exploring large-scale complex networks. The recent breakthroughs in technologies

for whole genome sequences, genome-wide transcriptional profiling and location analysis, and other large-scale assessment of gene expression, in addition to the advances in development of statistical methods and computational tools, provide unprecedented resources and offer the promise of unraveling complex gene networks. Much progress has been made so far. For example, transcription regulation modules have been elegantly mapped genome-wide for Drosophila (Berman et al., 2002) and yeast (Vilo et al., 2000; Pilpel et al., 2001). Many, if not most, of regulatory elements in the yeast genome have been computationally and experimentally identified (Kellis et al., 2003; Doniger et al., 2005). Moreover, the networks involved in sea urchin endomesoderm specification have been comprehensively described and experimentally verified (Davidson et al., 2002a; Oliveri and Davidson, 2004). Genome sequence comparison of multiple mammals of different evolutionary distances has enabled the identification of many ultra-conserved elements involved in regulating developmental process (Bejerano et al., 2004). Despite these advances, we're facing many challenges in deciphering gene regulatory networks. For example, one of many challenges for computational biologists is to integrate and mine the exponentially increasing biological data, such as genome sequences and gene expression data. The ability to efficiently integrating large varieties of biological data is essential for decoding genomic languages.

In the past decade, a number of new approaches were developed for identifying DNA regulatory elements. Different from the traditional labor-intensive approaches, such as electrophoretic mobility shift assays and DNase footprinting, to locate TFBS on a gene-by-gene and site-by-site basis, the recently developed approaches are high-throughput and heavily focused on the computational analysis of large genomic data sets. In this review, we briefly introduce these recently developed experimental and computational approaches for genome-wide regulatory element identification and characterization, with more focus on their computational aspects. While we described these approaches individually, they are often used combinatorially to improve accuracy as they are complementary to each other.

Large-scale gene expression analysis

The mature technology of microarray makes it possible to study large-scale gene expression. A typical approach for regulatory element identification is to apply clustering algorithms to genome-wide expression data to find sets of co-expressed genes that are also potentially co-regulated (Eisen et al., 1998), then search upstream sequences of these genes to identify statistically over-represented common regulatory motifs (Zhang, 1999; Caselle et al., 2002; Sui et al., 2005). An underlying assumption of this approach is that the co-regulated genes have a similar expression pattern. This approach has been extended to identify regulatory modules and their condition-specific regulators of the co-regulated genes by using a more sophisticated algorithm (Segal et al., 2003). The algorithm makes additional assumption that the expression level of regulated genes will depend on the expression levels of regulators, which is a limitation in cases that the expression level of the regulator does not change appropriately (e.g., regulator is activated by post-transcriptional modification). The analysis of expression data also has been combined with other available information, such as shared DNA binding motifs and protein sequence categories, to identify gene regulatory modules (Pilpel et al., 2001; Ihmels et al., 2002; Berman et al., 2002).

Genome-wide location profiling

Combining a chromatin immunoprecipitation (ChIP) assay with DNA microarray analysis, genome-wide location analysis is capable for genome-wide mapping protein-DNA interactions *in vivo*, hence is the most powerful approach for identifying and verifying functional *cis*-regulatory elements (Buck and Lieb, 2004; Hanlon and Lieb, 2004; Ren and Dynlacht, 2004; Zhang et al., 2004; van Steensel, 2005). Briefly, location analysis is carried out by first cross-linking transcription factors to their DNA sites by fixing cells with formaldehyde. Second, the fixed cells are then disrupted, and DNA is shed into small fragments (about

500 *kb*) by sonication. Third, the DNA fragments cross-linked to a protein of interest are enriched by immunoprecipitation with a specific antibody. Fourth, reversing the cross-links of immunoprecipitated DNA and protein complex, the enriched and purified DNA then are amplified and labeled with a fluorescent dye (Cy5) by ligation-mediated-polymerase chain reaction (LM-PCR). As a control, an aliquot of the input chromatin DNA is processed using an identical procedure except a different fluorophore (Cy3) is used for DNA labeling. Both labeled DNA samples and controls are hybridized simultaneously to a single DNA microarray containing genomic DNA sequences. The enriched protein-DNA binding regions are identified as DNA spots in which there is significantly more fluorescence intensity with ChIP-enriched DNA than with the control input DNA.

Genome-wide location analysis has been successfully used to identify transcription factor binding sites in yeast genomes (Lieb et al., 2001; Lee et al., 2002; Liu et al., 2005) as well as in mammalian genomes (Ren et al., 2000). However, this procedure, which can only map the protein-DNA interaction loci within 1-2 *kb*, does not yet have the resolution to discriminate precise cis-regulatory elements, and cannot yet be readily applied to the entire human genome because of its large size (Weinmann et al., 2002). Using location analysis data, some computational algorithms have been developed to improved accuracy and resolution of TFBS identification (Liu et al., 2002; Garten et al., 2005). Furthermore, the accuracy of *cis*-regulatory module prediction can be improved by using information from both DNA-binding data of location analysis and gene expression profiles (Bar-Joseph et al., 2003).

Genome-wide DNase hypersensitivity analysis

It has been shown that functional DNA regulatory regions, such as promoters, enhancers, suppressors, insulators, and locus control regions, are hypersensitive to DNase I digestion (Gross and Garrard, 1988). Traditionally, mapping DNase hypersensitive sites was the gold-

standard experimental method for identifying gene regulatory elements. For example, the traditional method was used to detect the enhancer elements of epsilon-globin gene (Zaret and Yamamoto, 1984) and glucocorticoid-dependent regulatory elements (Tuan and London, 1984). This approach has the advantage of taking chromatin context into account by digesting nucleosome-free regions of the genome, allowing for identification of both ubiquitous and tissue-specific regulatory elements. However, this procedure is technically demanding and thus has not been previously considered applicable on a genome-wide scale until the last couple of years (Lu and Richardson, 2004). So far, Genome-wide DNase hypersensitive site mapping has been successfully used to identify active regulatory regions in human erythroid cells (Sabo et al., 2004) and CD4(+) T cells (Crawford et al., 2004).

Model-based prediction

Although degenerative in sequence composition, regulatory elements, such as TFBS, generally have distinct preference and are fairly conserved. Given a set of known motif instances, it is possible to construct a statistical model to describe the motif properties, then use the model to predict motif sites in genome sequences. Many different motif models, from the simplest consensus model, the widely used Position Weight Matrix model (PWM), to more complex Bayesian networks model (Werner, 2000; Stormo, 2000), have been developed so far. While each model may have different strengths and weaknesses (see Motif Modeling section), their prediction powers are all limited by the availability and abundance of experimentally verified functional sites. The good collection and alignment of known functional motif sites could certainly improve model-based prediction accuracy. For such purposes, a number of transcription factor binding site databases have been developed in the past decade, for example, PRODORIC (Münch et al., 2003) for Prokaryotic *cis*-acting promoter elements, RegulonDB (Salgado et al., 2004) for regulatory elements of Escherichia coli, Transfac database (Wingender et al., 2000; Wingender, 2004) for eukaryotic TFBS,

TRED (Zhao et al., 2005) for mammalian regulatory elements, JASPAR (Sandelin et al., 2004) for matrix-based TFBS profiles of multicellular eukaryotes. However, it has proved to be difficult to accurately predict all functional sites in genomic sequences, particularly those of mammalian, as all model-based approaches, at present, suffer from the high false positive rate problem (Pennacchio and Rubin, 2001). Nevertheless, with completion of genome sequencing of many species, and the increasing number of experimentally verified functional sites, model-based prediction has become the most efficient computational approach for genome-wide motif identification.

Statistical motif models are also used to *ab initio* motif discovery by finding the over-represented motif sites in a set of genes known to share characteristic regulatory control mechanisms, such as orthologous genes performing similar functions or those with similar expresstion profiles from large-scale gene expression analysis (Ohler and Niemann, 2001). In such cases, motif modeling is generally combined with heuristic searching algorithms (Workman and Stormo, 2000; Thompson et al., 2003; Sinha et al., 2003), such as Gibbs sampling or expectation-maximizing algorithm, to identify potential functional motifs. Many such motif discovery tools have been developed, such as MEME (Bailey and Elkan, 1995), AlignACE (Hughes et al., 2000), ANN-Spec (Workman and Stormo, 2000), MotifSampler (Thijs et al., 2001), YMF (Sinha and Tompa, 2003) and GLAM (Frith et al., 2004). While these tools work effectively on a small test set of sequences containing well-defined motifs, a more advanced assessment (Tompa et al., 2005) showed all these tools performed poorly with the highest site sensitivity at 0.22 and highest correlation coefficient at 0.20, which largely due to incomplete understanding of the underlying biology of regulatory mechanisms. Furthermore, it is difficult to use these tools for genome-wide or other large-scale motif discovery because they are extremely computationally expensive.

Motif modeling

Regulatory elements in higher eukaryotes are generally short DNA sequences, which can vary largely in both the degree of degenerative and position dependencies among different elements. Thus, selecting an effective motif model from many available models is critical in genome-wide motif identification in order to minimize the false discovery rate. The simplest model is a consensus sequence that refers to a sequence matching all known example sites closely but not necessary exactly in general (Schneider, 2002). Although it is somewhat arbitrary to define consensus, there are some methods that can make better trade-off between the number of mismatches allowed and the sensitivity and precision of the representation (Day and McMorris, 1992b,a). For a highly conserved regulatory element, a consensus model would be the most efficient. A better and commonly used model for motif identification is Weight Matrix Model (WMM) proposed by Staden (1984), also called Position Weight Matrix (PWM) or Mononucleotide Weight Matrix (MWM). PWM is usually generated by aligning sequences of known transcription factor binding sites, from which weights or scores in the matrix are calculated according to the observed frequencies of bases. Many different score conversion functions with only subtle variations have been used to calculate a position weight, including the negative log of a base frequency, information contents or relative information contents. The PWM inherently assumes that individual positions within motif are independent, which may be true for some regulatory elements. It has been shown that the score of a PWM, in some cases, is proportional to the binding energy contributed by each base at each position (Stormo and Fields, 1998). PWM has been used by many motif identification programs, e.g. MatInspector (Quandt et al., 1995) and Match (Kel et al., 2003), and performs reasonably good for motif identification in some cases. While a PWM can capture both nucleotide preferences at each position and different levels of position specificities, it does not contain information of correlations or dependencies between positions within a motif. Recent studies (Agarwal and Bafna, 1998; Benos et al., 2001; Bulyk

et al., 2002) indicate that there are important interactions or correlations between adjacent positions as well as non-adjacent positions within a motif in many cases. The inability of a PWM to capture such dependency information has limited its power to find true motif sites.

In the past few years, many models have been developed to incorporate such dependency information of positions. Motif models such as Dinucleotide Weight Matrix model (DWM) (Schneider et al., 1986), Weight Array Model(WAM) (Zhang and Marr, 1993), can incorporate dependencies between adjacent positions. To incorporate further dependencies of non-adjacent positions, Ponomarenko et al. (1999) extended DWM by introducing Oligonucleotide Weight Matrices model, which includes a comprehensive set of oligonucleotide matrices classified into 5 biological function categories. In principle, a WAM model could also be extended to high-order WAM, e.g. windowed second order WAM (Burge and Karlin, 1997). However, the exponentially increased number of parameters of these models makes it difficult to use them practically due to insufficient training data. To address the weaknesses of WAM in incorporating long-ranging interactions, Burge and Karlin (1997) proposed the Maximal Dependence Decomposition (MDD) model, which has binary tree structure formed by a set of conditional WAM models. While MDD model can capture non-adjacent dependencies through the conditional WAM models, it still requires a rather large number of training sequences, which are partitioned into smaller sub-sets to train all conditional WAM models. The idea of MDD is quite smart and its procedure is very similar to the recursive partitioning procedure for multiple variable analysis. First, a consensus is obtained from the original training set; then for each base C_i in the consensus, its dependencies on each base X_j ($i \neq j$) are calculated by Chi-Square statistics. The sum (S_i) of Chi-Square statistics is used as a measure of total dependency of C_i on other positions. Choosing C_i with maximum S_i, partition the training set into 2 complementary subsets, one set has C_i in the i^{th} position of all sequences and the other set doesn't. Using the same rule, each data set is recursively partitioned into smaller sets until one of the following conditions is satisfied: (1) Tree is

undividable, i.e., it has reached bottom level of the tree. (2) No significant dependency detected. (3) Reach the minimum number of sequences required to train model parameters. So each node in binary tree is a subset of the original training set, which is the root of tree. A series of WAM models are derived, one WAM model for each subset of training sequences. According to the path from root to leaf traveled by a sequence, the probability or logarithm of probability of the sequence given the motif model can be calculated. While MDD model can capture non-adjacent dependencies through the conditional WAM models, it requires a big training data, which can be partitioned into smaller but still adequate to properly train each conditional WAM model. To alleviate the requirement of a large training set, Cai et al. (2000) developed a Bayesian tree to model dependencies within RNA splicing sites; Ellrott et al. (2002) suggested a position-order optimized Markov chain model, which reorders motif positions to bring long-ranging dependent positions into the near neighbors. More recently, several other models have been developed, including Bayesian networks for modeling protein-DNA binding sites (Barash et al., 2003), Maximum Entropy Model (MEM) for splicing site identification (Yeo and Burge, 2003), Permuted Variable Length Markov Model (PVLMM) for finding TFBS and splicing sites (Zhao et al., 2004). While these models were proved to perform well for motif identification in some specific cases, it seems that there is still some room to improve.

Model selection

Selecting a proper motif model from the above large variety of models is equally important. Choosing the best model is to make a trade-off between advantages and disadvantages of succinct and more expressive models. Succinct models, such as PWM, can be robustly trained from a few examples, but are not able to account for correlation between positions; while expressive models, such as Bayesian networks models can account for most dependency information but, as mentioned above, could involve the dramatically increased number of

parameters hence require substantially large training data that may not available at present. Also, expressive models are more susceptible to over-fitting problem, which leads to inferior power in identifying new motif sites. When using a Bayesian model, the over-fitting problem is less likely to occur if using an appropriate prior. However, the choice of a sensible prior distribution is also a no-trivial problem (Zhou and Liu, 2004) for a complex Bayesian networks model. Comparing the different models above, PWM is the simplest and most robust model. Its easiness to compute and simplicity to train has made it the most widely used model. On the other side, its provision of overwhelming false positive binding sites was widely criticized. The DWM Model is a nice extension of PWM. While it is able to capture limited dependency information, DWM requires a much large training set for training. A model consisting of optimal subset of DWM may be helpful, but it is still questionable if it can compete with other models. Markov chain based models seem promising, particularly the MDD procedure and position-optimized Markov chain method. However, its application to discover new potential regulatory motif has not yet been fully explored. Bayesian networks models try to exploit information unreachable by other models. A good example of Bayesian networks models has been demonstrated by Barash et al. (2003). But it is still too early to say whether we can make profits from a Bayesian model because much of its gained power could be offset by its high cost in computation and requirement of a large training data. It will be interesting to see a study that compares all these models using a good benchmark data set. Although the performance of all motif models is strongly dependent on good training sets, in theory, the models incorporating position dependencies are generally more reliable and accurate in predicting functional elements. Particularly, it could help to reduce false positive rate.

Cross-species sequence comparison

Phylogenetic footprinting based on comparative genomics analysis has been demonstrated to be the most effective approach to identifying functional regulatory elements (Ureta-Vidal et al., 2003; Thomas et al., 2003). The evolutionary assumptions underlying such comparative analysis are: 1) mutations in functionally important genomic regions will accumulate more slowly due to negative selection forces than mutations in the regions subjected to neutral evolution or positive selection. 2) Genomic sequences under comparison are evolved from a common ancestor and their differences can be solely explained by evolutionary processes. Comparative genomics has been widely used to reveal conserved non-coding sequences (CNS) that likely contain most functional regulatory elements. In lower eukaryotes, genome comparison of several Saccharomyces species has led to the identification of most functional regulatory elements in the yeast genome (Kellis et al., 2003; Doniger et al., 2005). In higher eukaryotes, comparative analysis of divergent human and pufferfish genomes had revealed embryonic development related regulatory elements, while comparison of more closed species, human and mouse, had identified many putative TFBS (Lenhard et al., 2003; Herrmann et al., 2005). Furthermore, the recent comparison of multiple mammalian genomes has enabled identifying many highly conserved elements involved in regulating developmental process (Bejerano et al., 2004). Much of comparative genomics success has to attribute to the recent advances in developing more efficient computational methods and tools, particularly sequence alignment tools. On the other side, realizing the power of comparative genomics and the increasing availability of genome sequences of diverse species have motivated creation of new set of tools as well as related data resources (see Table 1.1). However, our understandings of molecular evolution are woefully inadequate, and we are still facing many problems and challenges in performing comparative genomics analysis. These challenges, notably, are: identification of promoter regions, particularly the location of transcription start sites, sequence alignment, choice of species for comparison, selection

of conserved regions, and so on. We select some of these problems to discuss below.

The choice of species

The first question for phylogenetic footprinting analysis is which species should be compared? The choice of species is important as the outcomes are largely determined by the choice. In general, if the species are closely related, the alignment of sequences is obvious but uninformative because the functional blocks are not sufficiently more conserved than non-functional blocks. On the other hand, if the species are separated too far away in evolutionary distance, it is difficult to align sequence accurately because sequences are so divergent and different that the true alignment does not have better score than other alternative alignments. For example, comparison of orthologous sequences of evolutionary distantly related species, such as human and pufferfish, which diverged approximately 450 million years ago, probably can reveal many conserved coding sequences but few regulatory elements (Aparicio et al., 2002). Comparative analysis of orthologous sequences of closely related species, such as human with chimpanzees or other nonhuman primates, may identify only genomic sequences that rearranged or changed in recent evolutionary history. These orthologous sequences are so evolutionarily close that even the non-functional sequences could be well conserved. Thus it could be difficult to separate the functional regions from non-functional regions by the observed conservation of sequences (Frazer et al., 2003). Comparison of orthologous sequences of human and mouse, which diverged about 40-80 million years ago, could uncover most conserved noncoding sequences, which are likely to be functional sites of genome, such as regulatory elements, splicing signals (Dubchak et al., 2000; Loots et al., 2000; Morgenstern et al., 2002). Nevertheless, it is difficult to tell whether most of conserved noncoding sequences are present because of functional constraints or as the results of lack of divergence time. Comparison of multiple genomes of different evolutionary distances will increase the power of footprinting to identify true functional elements.

Sequence alignment

A key step of the footprinting approach for identifying regulatory elements is to find the evolutionary relationship of the sequences via sequence alignment. In general, there are two different approaches for aligning sequences: local alignment and global alignment. Local alignment, which makes few assumptions about how similarity should be organized, is to find all local regions of sufficient similarity. Global alignment, assuming that important functional regions remain in the same order and orientation during evolution, asks for the optimal transformation of one sequence into the others that are assumed to be evolved from a common ancestor. Therefore, local alignment is generally better than global alignment in dealing with duplication and inversion problems in divergence sequences. On the other hand, global alignment has stronger capability to identify subtle similarities in weakly conserved regions between well-conserved blocks. Since Needleman and Wunsch (1970) elegantly described a dynamic programming algorithm for optimal global alignment, and Smith and Waterman (1981) and Gotoh (1982) extended it to local alignment, we witnessed the significant progresses in sequence alignment tool development featured by heuristic-based alignment tools that dramatically increase alignment speed. The adaptation of index-based heuristic alignment strategy is the major factor contributed to the huge success of BLAST program (Altschul et al., 1990, 1997), which is now routinely used for searching homologs in NCBI sequence database. The success of BLAST has boosted the development of many new faster heuristic-based alignment tools capable of efficiently aligning genome-size sequences. Many of these tools, including WABA (Kent and Zahler, 2000), MUMmer (Delcher et al., 2002), PatternHunter (Ma et al., 2002), BLAT (Kent, 2002), BLASTZ (Schwartz et al., 2003), AVID/MAVID (Bray et al., 2003; Bray and Pachter, 2003), and LAGAN/MLAGAN (Brudno et al., 2003), have been used for comparative analysis of human and mouse genomes. The approaches of these tools follow the similar three main steps. The first step is to use seeding strategies, such as index or suffix tree, to find identical or highly matched words

of fixed length as seeds. In second step, the seeds are extended in both directions to find local similar regions above a certain threshold. These similar regions are used as anchoring regions for global alignment if needed. Third, a dynamic programming-based algorithm is used to align sequence segments between anchoring regions. The detailed implements of three steps can be different among different alignment tools. The following three techniques, whose underlying principle is to increase the number of hit in actual homologous sequences while reducing the expect number of hit in random sequences, are generally employed by these tools to improve alignment sensitivity without compromising specificity:

1. Reduce the length of exact matching words, but requiring to match two words within a fixed distance instead of a single word in initial seeds searching. This technique is used by local aligning tools WU-BLAST, BLAT, and BLASTZ.

2. Look for highly similar words that allow a certain degree of degenerative instead of exact identical words. This technique is employed by CHAOS.

3. Use a non-consecutive searching model that has the same weight as the original consecutive model, e.g., BLAST uses the 11-base consecutive model 11111111111, while patternHunter employs 18-base non-consecutive model 111010010100111 0111 (1 indicating match, 0 indicating "don't care") with the same weight 11 of the BLAST consecutive model. This technique was first deployed in PatternHunter, and was also implemented in BLASTZ.

Different from above fast alignment tools, some more accurate but slower alignment tools are also available, such as SSEARCH (Pearson, 1991) based on the Smith-Waterman algorithm, and DiAlign (Morgenstern, 1999) using an interestingly segment-segment based approach. Both SSearch and DiAlign can be used to align both DNA and protein sequences. Other alternative approaches for alignment are also available. For example, Blanchette and Tompa (2002) proposed a motif finding method based on maximum parsimony of k-mer substring

sequences. The idea of this method is to calculate parsimony score of all k-mer substring according to a given phylogenetic tree of related species, and then output a set of k-mers with the lowest parsimony scores, in which parameter k and threshold of parsimony score are defined in advance. If several k-mers are overlapping in each of original sequences, they are merged to form a bigger conserved block. The method was implemented in a dynamic programming-based program called FootPrinter (Blanchette et al., 2002; Blanchette and Tompa, 2003). FootPrinter is much like the local alignment tools such as BLAST with respect to finding conserved blocks. The major difference is score schemes. BLAST or SSearch uses general score matrices such as BLOSUM matrices or PAM matrices, while FootPrinter uses parsimony scores based on a phylogenetic tree of related species, which make more sense if the included species have different relationships and divergent times. While FootPrinter can find all similar blocks as other local alignment tools, some of which may not be found in global alignment due to inaccurate alignment. It will be difficult to assess whether the similar blocks are real conserved blocks because there are possible random match, particularly for short substrings. In this sense, this method may have much high false positive rate than those based on global alignment though increasing the number of orthologous sequences may increase its power. Nevertheless, FootPrinter provides a nice alternative approach for footprinting. Also its idea of defining CNS using parsimony score makes more sense than the other criteria we will discuss below.

Despite these progresses, sequence alignment remains an open computational problem as researchers try to solve the old problems, such as biological meaningful scoring measures of sequence similarity, improving computational efficiency, and alignment of multiple sequences, while facing continually emerging challenges, such as aligning multiple genomes, improving alignment quality, dealing with sequence inversion, duplication and translocation. Comparative genomic analysis for identifying short regulatory elements is particularly demanding with respect to both accuracy and efficiency of alignment tools. While recently

developed heuristic-based tools, such as, AVID, LAGAN, and BLASTZ, are able to efficiently perform pairwise genome alignment, their alignment qualities are still questionable about whether they are good enough to identify most functional important genomic regions. Computational biologists are facing a dilemma about whether continue to focus on developing new index-based efficient alignment tools or pay more attention to the slower but more accurate dynamic programming algorithm-based tools. Further significant improvement of alignment tools would rely on more understanding of the underlying evolutionary mechanism of biological organism.

Criteria for defining CNS

The criteria of CNS need to be defined before any CNS can be identified in phylogenetic footprinting analysis. It is an important question about how should we define CNS as different criteria are likely to produce different results. Ideally CNS should include all conserved functional binding sites and exclude all non-functional sites. In reality, the CNS is defined arbitrarily based on some observed regulatory region or functional sites, such as Locus Control Region (LCR) of globin, conserved regulatory regions neighboring interleukin genes (IL-4, IL-5, and IL-13) in human chromosome 5, both of which have been well studied and functionally characterized. Some commonly used criteria are: 1) No gap segment; 2) Percent identity within defined window size in aligned orthologous sequence is larger than an arbitrarily defined cutoff value that is larger than expected from neutrally evolved sequences; 3) Contiguous segments satisfying the above 2 conditions are merged together to form larger CNS. For example, CNS of human and mouse is usually defined as sequence regions with percent identity larger than 70% over at least 100 bps stretch (Loots et al., 2000). Several similar criteria were proposed by other researchers , such as over 60% identity, gap free, and with at least 50 bps in length (Fickett and Wasserman, 2000), and 70% identity cut-off over 50 *bp* window (Lenhard et al., 2003). For the CNS of human and fish, different cutoff

values should be used. Using cutoff of more than 70% identity over at least 50 bps, Duret and Bucher (1997) showed that only 16% orthologous gene pairs have CNS in regulatory regions. Most of state-of-art tools to visualize and identify CNS are based on these cutoff values, such as VISTA (Mayor et al., 2000), PipMaker (Schwartz et al., 2000). The drawback is that these cutoff values are chosen too arbitrarily, usually based on common biological sense or rule of thumb. Through incorporating information of genome sequences of three species, human, mouse, and dog, Dubchak et al. (2000) suggested a better criterion to define CNS of human, which maximize the sum of all pairwise Intersection/Union (I/U) values, the ratio between the common CNS regions of three species and all CNS regions obtained from one species compared with two other species separately. The I/U analysis method is to choose conserved noncoding regions that maximize number of overlapping regions, and minimize number of unique regions among different species. Based on 200 kb orthologous sequences human (5q31), mouse (chromosome 11), and dog (chromosome 4), the sum of three I/U values is maximized at the following criteria: Human/Dog, 92% identity over 120 *bp*; Human/Mouse, 80% identity over 120 *bp*; and Mouse/Dog, 77% identity over 120 *bp*. However, it remains a question whether cutoff values chosen by the I/U method are better than those chosen arbitrarily or by other methods.

Castresana (2000) suggested a different method to a cutoff-value based criterion. The method was implemented in a program called Gblocks. Basically, this method searches for conserved blocks from aligned orthologous sequences that meet a set of requirements. It defines a total of five arbitrary thresholds, IS, FS, CP, BL1 and BL2, in the set of requirements, and proceeds step by step in the following order to identify conserved block in alignment: 1) Each position is classified into one of three classes, non-conserved (percent identity $<$ IS or there is a gap), conserved (percent identity $\geq$ IS but $<$ FS), highly conserved (percent identity $\geq$ FS). The default values of IS and FS in Gblocks are 50% and 80%, respectively. 2) All stretches of contiguous non-conserved positions with length $>$ CP are

rejected, in which alignment is likely ambiguous. The default value of CP is 8 positions. 3) Positions that are not highly conserved as defined in step 1 are removed from the flanks of all remaining blocks, so each block is surrounded by highly conserved franks. 4) Small blocks with length less than BL1 positions are rejected. The default value of BL1 is 15 positions. 5) All non-conserved positions adjacent to gaps (include gaps) are eliminated. 6) Only blocks with length larger than BL2 positions after gap cleaning are defined as conserved blocks. The original purpose of the method in the paper is to select conserved blocks thus with high confident in their alignment for phylogenetic tree reconstruction, but it can be applied to footprinting analysis with a little modification. The disadvantages of above methods are: 1) they are highly dependent on accurate alignment of orthologous sequences, which itself is a very difficult problem, particularly, a good alignment for divergent sequences is very difficult to obtain. 2) It may be difficult to obtain optimal values for the cutoff thresholds of parameters interested. For comparison of multiple species, additional information such as evolutionary distances and phylogenetic relationships between species can be used to more accurately identify regulatory elements or CNS.

Up to now, there is no gold standard to define CNS. However, a better standard should be based on not only similarity of sequences, evolutionary relationships between species, but also variation of substitution rates across different chromosomes and different regions, which can be used to more accurately distinguish true conserved region from false conserved region. But rate variation across different regions of genome is hard to estimate, which is part of the reason that none of above method considered it. With better understanding of evolution process of non-coding DNA sequences, particularly the regulatory region sequences, we could come out a better evolutionary model that would allow us to define CNS more appropriately.

Do all CNS contain regulatory elements?

Many studies have showed that regulatory elements are generally located within CNS due to their functional constraints. However, not all conserved blocks identified by phylogenetic footprinting contain functional elements (Dermitzakis et al., 2005). For example, conservation of some blocks of DNA may simply due to the short divergent time between species or low substitution rate at that region attributed to some unknown random factors; some of CNS may be just randomly matched blocks because of wrong alignment of orthologous sequences; and some CNS blocks may have functions in some species while do not function in other species. So, the question is what the proportion of non-function sites among all conserved regions is. The answer to this question really depends on what criteria is used to define conserved regions, and which reference species are used. In general, the stricter the criteria, the lower the proportion of non-function sites in defined conserved region, but more likely that the less conserved functional sites could be excluded. For example, suppose that same criteria are used, but the first set of CNS blocks of human are identified as conserved non-coding blocks of human and mouse, and the second set of CNS blocks are identified as common conserved non-coding blocks of human, mouse and dog. Then the second set of CNS blocks is more likely to have lower proportion of non-function sites than the first set because CNS blocks in the second set are more conserved than those in the first set. On the other side, the sites with functions only in human and mouse can be missed from the second set CNS blocks because they do not function in dog hence are no longer conserved. Also individual binding sites may exhibit relatively little conservation, either because of the degeneracy of the transcription factor binding requirements or because their small size makes it relatively likely that a new functional size will arise by chance. The proportion of non-functional sites in CNS blocks identified from distantly related species are less than that from closely related species because non-functional sites are less likely to be conserved in long evolutionary time. While distant interspecies comparison can reveal

highly conserved binding sites, which may have a radical effect on expression of genes, it makes other newly evolved or less conserved binding sites undetectable. On the other hand, comparisons of more closely related species are confounded by non-functional but low divergence sequences, which will result in high proportion non-functional sites in conserved regions identified. Therefore, in reality we should consider not only false positive rate, but also false negative rate in identifying functional conserved regions.

Since many factors (known or unknown) could affect the proportion of non-function sites within conserved regions, there is no general way to estimate false positive rate. However, using some well-known or experimentally verified data, it is possible to estimate the false positive rate and false negative rate for a given criteria of CNS and a fixed set of species. One of such studies was performed by Dermitzakis and Clark (2002). In the study, the authors performed comparative analysis transcription factor binding sites in the promoters of 51 human genes and their corresponding orthologous sequences in other primate species and rodents, of which transcription factors are all experimentally verified. Analysis of 20 orthologous genes with total 64 aligned binding sites between human and rodents showed that 33 binding sites shared function between human and rodents, and 14 are human specific and 17 are rodent specific. And the average divergence of sequences in shared binding sites is statistically significantly lower than that of sequences in species-specific binding sites. According to the result, the authors estimated that there are about 32%-40% of the human functional sites that have no function in rodents, which is proportional to the false negative rate of binding sites identified by phylogenetic footpinting approach. Direct estimate of false positive rate, the proportion of non-functional sites among conserved regions was not given, but it should be pretty straight forward to make such an estimate based on the same dataset. For human-rodent comparison, several studies have suggested that there are only about 17%-20% of non-coding regions conserved across entire genomes (Wasserman et al., 2000; Shabalina et al., 2001). So if we were able to estimate proportion of functional sites in

non-coding regions from known data, we could roughly estimate proportion of non-functional sites in conserved regions.

The power of footprinting

Although there are still some limitations of phylogenetic footprinting approach to identify regulatory elements, it offers many advantages over single genome approach. One important advantage is that it not only dramatically reduces false positive binding sites of single genome approach, but also has more power to detect or discover binding site over many single genome methods (Wasserman et al., 2000; Levy and Hannenhalli, 2002; Cliften et al., 2003; Kellis et al., 2003; Zhang and Gerstein, 2003). For example, in performance comparison of three computational methods for identifying human TFBS, Levy and Hannenhalli (2002) showed that specificity of the method based on footprinting of human and mouse is up to three-fold higher than the other two single genome methods, both of which are based on score of position weight matrices of motifs but with different filtering methods, the p-values of motif scores and modular clustering of motifs. Zhang and Gerstein (2003) studied a larger set of human-mouse gene pairs by phylogenetic footprinting and compared the results predicted by ConSite with the previously verified regulatory sites (Lenhard et al., 2003). The result suggested that on average, phylogenetic footprinting improved the selectivity of TFBS prediction by 85% compared to using matrix models alone, and could detect the majority of verified sites. These studies demonstrated that comparative genomics is an effective and powerful method for identifying conserved functional sites in genomic sequences.

Conclusions and future directions

The exponentially increasing biological data produced by high-throughput biotechnologies offer promises for deciphering gene regulatory mechanisms of each organism while at the same time presenting us enormous computational challenges. These challenges have mo-

tivated the development of many improved computational models and tools, which were demonstrated to be extremely valuable in deciphering gene regulatory networks. In particular, computational analysis genomics sequences of related species has been proved to be a powerful approach for identifying regulatory elements, especially when combined with the other approaches (Wasserman et al., 2000; Boffelli et al., 2003; Nardone et al., 2004). Despite these remarkable progresses, complete regulatory element identification and characterization in higher eukaryotes, particular human, remains a distant hope as too much unknown about the underlying mechanisms of sequence evolution and genome organization. Nevertheless, we are in the right direction, future work on the following emerging fields would certainly be helpful as we are facing the challenge.

First, comparison of multiple species of different evolutionary distances will increase the power of footprinting to identify functional important regulatory elements. For example, comparison of mammalian genomes with other higher eukaryotic genomes may help identify mammalian specific regulatory elements as well as other highly conserved development related elements. Multiple genome comparison are becoming more possible as the number of available genomes is fast growing. Second, in higher eukaryotes, transcription factor binding sites tend to organize into homotypic or heterotypic clusters, or CRM. In addition, different CRM may have different characteristic features, such as TFBS number, TFBS types and distances between TFBS. Incorporating these information into motif models will greatly increase the accuracy of regulatory element identification. For example, recent efforts (Berman et al., 2004; Gupta and Liu, 2005) showed that using CRM could dramatically reduce false positive rate in identifying TFBS. Third, current computation methods in the area are largely based on descriptions of regulatory regions rather than on an understanding of the fundamental molecular and evolutionary mechanisms (Claverie, 2000) because our knowledge about the details of the regulation of individual genes remains dramatically incomplete. The research on sequence evolution, particularly, on negative and positive selec-

tion, may provide essential theory that could be used to dramatically improve performance of many computation methods and tools, such as sequence alignment tools. Overall, effectively integrating large variety of comprehensive biological data, in addition to systematically combinatorial uses of available experimental and computational approaches including those discussed above, will be an essential key to deciphering human gene regulatory networks.

Table 1.1: Computational resources for regulatory element identification

Name	Authors	Website
Sequence Alignment		
DiAlign	Morgenstern 1999	http://bibiserv.techfak.uni-bielefeld.de/dialign
BLASTZ/PipMaker	Schwartz et al. 2000	http://bio.cse.psu.edu
AVID	Bray et al. 2003	http://baboon.math.berkeley.edu/mavid
LAGAN	Brudno et al. 2003	http://lagan.stanford.edu
Transcription Factor Binding Site database		
TRANSFAC	Wingender et al. 2000	http://www.gene-regulation.com/pub/databases.html
PRODORIC	Münch et al. 2003	http://prodoric.tu-bs.de
JASPAR	Sandelin et al. 2004	http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl
RegulonDB	Salgado et al. 2004	http://www.cifn.unam.mx/Computational_Genomics/regulondb
TRED	Zhao et al. 2005	http://rulai.cshl.edu/TRED
Orthologous Gene Database		
COGs	Tatusov et al. 2003	http://www.ncbi.nlm.nih.gov/COG
HOPS	Storm and Sonnhammer 2003	http://pfam.cgb.ki.se/HOPS
HomoloGene	Wheeler et al. 2004	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene
Motif Identification		
MatInspector	Quandt et al. 1995	http://www.genomatix.de/products/MatInspector
Cister	Frith et al. 2001	http://zlab.bu.edu/~mfrith/cister.shtml
eCis-analyst	Berman2004	http://rana.lbl.gov/cis-analyst
MSCAN	Alkema et al. 2004	http://tfscan.cgb.ki.se/cgi-bin/MSCAN
Motif Discovery		
MEME	Bailey and Elkan 1995	http://meme.sdsc.edu
AlignACE	Hughes et al. 2000	http://atlas.med.harvard.edu
YMF	Sinha and Tompa 2003	http://bio.cs.washington.edu/software.html
SeSiMCMC	Favorov et al. 2005	http://favorov.imb.ac.ru/SeSiMCMC

Bibliography

Agarwal, P. and Bafna, V. (1998). Detecting non-adjoining correlations with signals in dna. In *RECOMB '98: Proceedings of the second annual international conference on Computational molecular biology*, pages 2–8, New York, NY, USA. ACM Press.

Alkema, W. B. L., Johansson, O., Lagergren, J., and Wasserman, W. W. (2004). MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, 32(Web Server issue):W195–W198.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D. S., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J. K., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297(5585):1301–1310.

Bailey, T. L. and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3:21–29.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, 21(11):1337–1342.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-dna binding sites. In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology*, pages 28–37, New York, NY, USA. ACM Press.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.

Benos, P. V., Lapedes, A. S., Fields, D. S., and Stormo, G. D. (2001). SAMIE: statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.*, pages 115–126.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *PNAS*, 99(2):757–762.

Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B., and Celniker, S. E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol.*, 5(9):R61.

Blanchette, M., Schwikowski, B., and Tompa, M. (2002). Algorithms for phylogenetic footprinting. *J. Comput. Biol.*, 9(2):211–223.

Blanchette, M. and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, 12:739–748.

Blanchette, M. and Tompa, M. (2003). FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, 31(13):3840–3842.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299:1391–1394.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. *Genome Res.*, 13(1):97–102.

Bray, N. and Pachter, L. (2003). Mavid multiple alignment server. *Nucleic Acids Res.*, 31:3525–3526.

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

Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360.

Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.*, 30(5):1255–1261.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94.

Cai, D., Delcher, A., Kao, B., and Kasif, S. (2000). Modeling splice sites with Bayes networks. *Bioinformatics*, 16(2):152–158.

Caselle, M., Cunto, F. D., and Provero, P. (2002). Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics*, 3(1):7.

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

Claverie, J. M. (2000). From bioinformatics to computational biology. *Genome Res.*, 10(9):1277–1279.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76.

Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E. D., Wolfsberg, T. G., and Collins, F. S. (2004). From the Cover: Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *PNAS*, 101(4):992–997.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Rust, A. G., jun Pan, Z., Schilstra, M. J., Clarke, P. J. C., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., and Bolouri, H. (2002a). A genomic regulatory network for development. *Science*, 295(5560):1669–1678.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B., Lee, P. Y., Revilla, R., Schilstra, M. J., Clarke, P. J. C., Rust, A. G., Pan, Z., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., and Bolouri, H. (2002b). A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev. Biol.*, 246(1):162–190.

Day, W. H. and McMorris, F. R. (1992a). Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.*, 20(5):1093–1099.

Day, W. H. and McMorris, F. R. (1992b). Threshold consensus methods for molecular sequences. *J. Theor. Biol.*, 159(4):481–489.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, 30(11):2478–2483.

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites

in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, 19(7):1114–1121.

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, 6(2):151–157.

Doniger, S. W., Huh, J., and Fay, J. C. (2005). Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res.*, 15(5):701–709.

Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M., and Frazer, K. A. (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.*, 10:1304–1306.

Duret, L. and Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, 7(3):399–406.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868.

Ellrott, K., Yang, C., Sladek, F. M., and Jiang, T. (2002). Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, 18 Suppl 2:S100–S109.

Favorov, A. V., Gelfand, M. S., Gerasimova, A. V., Ravcheev, D. A., Mironov, A. A., and Makeev, V. J. (2005). A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–2245.

Fickett, J. W. and Wasserman, W. W. (2000). Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, 11(1):19–24.

Frazer, K. A., Chen, X., Hinds, D. A., Pant, P. V. K., Patil, N., and Cox, D. R. (2003).

Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.*, 13(3):341–346.

Frith, M. C., Hansen, U., Spouge, J. L., and Weng, Z. (2004). Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, 32(1):189–200.

Frith, M. C., Hansen, U., and Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889.

Garten, Y., Kaplan, S., and Pilpel, Y. (2005). Extraction of transcription regulatory signals from genome-wide DNA-protein interaction data. *Nucl. Acids Res.*, 33(2):605–615.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162(3):705–708.

Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem*, 57:159–197.

Gupta, M. and Liu, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *PNAS*, 102(20):7079–7084.

Hanlon, S. E. and Lieb, J. D. (2004). Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.*, 14(6):697–705.

Herrmann, D. C. C., Dieterich, C., Cunto, F. D., Provero, P., and Caselle, M. (2005). Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics*, 6(1):110.

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

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, 31(4):370–377.

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

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664.

Kent, W. J. and Zahler, A. M. (2000). Conservation, Regulation, Synteny, and Introns in a Large-scale C. briggsae-C. elegans Genomic Alignment. *Genome Res.*, 10(8):1115–1125.

Kirschner, M. and Gerhart, J. (1998). Evolvability. *PNAS*, 95(15):8420–8427.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804.

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

Levy, S. and Hannenhalli, S. (2002). Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, 13(9):510–514.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, 28(4):327–334.

Liu, X., Noll, D. M., Lieb, J. D., and Clarke, N. D. (2005). DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.*, 15(3):421–427.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, 20(8):835–839.

Locascio, A., Manzanares, M., Blanco, M. J., and Nieto, M. A. (2002). Modularity and reshuffling of Snail and Slug expression during vertebrate evolution. *PNAS*, 99(26):16841–16846.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–140.

Lu, Q. and Richardson, B. (2004). DNaseI hypersensitivity analysis of chromatin structure. *Methods Mol. Biol.*, 287:77–86.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047.

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., and Jahn, D. (2003). PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, 31(1):266–269.

Morgenstern, B. (1999). Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218.

Morgenstern, B., Rinner, O., Abdeddaim, S., Haase, D., Mayer, K. F. X., Dress, A. W. M., and Mewes, H.-W. (2002). Exon discovery by genomic sequence alignment. *Bioinformatics*, 18(6):777–787.

Nardone, J., Lee, D. U., Ansel, K. M., and Rao, A. (2004). Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. *Nat. Immunol.*, 5(8):768–774.

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

Ohler, U. and Niemann, H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, 17(2):56–60.

Oliveri, P. and Davidson, E. H. (2004). Gene regulatory network analysis in sea urchin embryos. *Methods Cell Biol.*, 74:775–794.

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivityand selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635–650.

Pennacchio, L. A. and Rubin, E. M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, 2(2):100–109.

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29(2):153–159.

Ponomarenko, M. P., Ponomarenko, J. V., Frolov, A. S., Podkolodnaya, O. A., Vorobyev, D. G., Kolchanov, N. A., and Overton, G. C. (1999). Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. *Bioinformatics*, 15(7):631–643.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, 23(23):4878–4884.

Ren, B. and Dynlacht, B. D. (2004). Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol.*, 376:304–315.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young,

R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309.

Sabo, P. J., Humbert, R., Hawrylycz, M., Wallace, J. C., Dorschner, M. O., McArthur, M., and Stamatoyannopoulos, J. A. (2004). Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *PNAS*, 101(13):4537–4542.

Salgado, H., Gama-Castro, S., Martínez-Antonio, A., Díaz-Peredo, E., Sánchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jiménez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martínez, C., and Collado-Vides, J. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res.*, 32(Database issue):D303–D306.

Sandelin, A., Wasserman, W. W., and Lenhard, B. (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*, 32(Web Server issue):W249–W252.

Schneider, T. D. (2002). Consensus sequence Zen. *Appl. Bioinformatics*, 1(3):111–119.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431.

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

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res.*, 10(4):577–586.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34(2):166–176.

Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A., and Kondrashov, A. S. (2001). Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.*, 17(7):373–376.

Sinha, S. and Tompa, M. (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 31(13):3586–3588.

Sinha, S., van Nimwegen, E., and Siggia, E. D. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:i292–i301.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.*, 12(1):505–519.

Storm, C. E. V. and Sonnhammer, E. L. L. (2003). Comprehensive analysis of orthologous protein domains using the HOPSdatabase. *Genome Res.*, 13(10):2353–2362.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23.

Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 23(3):109–113.

Sui, S. J. H., Mortimer, J. R., Arenillas, D. J., Brumm, J., Walsh, C. J., Kennedy, B. P., and Wasserman, W. W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, 33(10):3154–3164.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., Moor, B. D., Rouzé, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., Schwartz, S., Elnitski, L., Idol, J. R., Prasad, A. B., Lee-Lin, S.-Q., Maduro, V. V. B., Summers, T. J., Portnoy, M. E., Dietrich, N. L., Akhter, N., Ayele, K., Benjamin, B., Cariaga, K., Brinkley, C. P., Brooks, S. Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S.-L., Huang, M. C., Karlins, E., Laric, P. L., Legaspi, R., Lim, M. J., Maduro, Q. L., Masiello, C. A., Mastrian, S. D., McCloskey, J. C., Pearson, R., Stantripop, S., Tiongson, E. E., Tran, J. T., Tsurgeon, C., Vogt, J. L., Walker, M. A., Wetherby, K. D., Wiggins, L. S., Young, A. C., Zhang, L.-H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C. L., Jong, P. J. D., Lawrence, C. E., Smit, A. F., Chakravarti, A., Haussler, D., Green, P., Miller, W., and Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793.

Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003). Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, 31(13):3580–3585.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Réier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144.

Tuan, D. and London, I. M. (1984). Mapping of DNase I-hypersensitive sites in the upstream DNA of human embryonic epsilon-globin gene in K562 leukemia cells. *PNAS*, 81(9):2718–2722.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.*, 4(4):251–262.

van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.*, 37 Suppl:S18–S24.

Vilo, J., Brazma, A., Jonassen, I., Robinson, A., and Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:384–394.

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

Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H.-M., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.*, 16(2):235–244.

Werner, T. (2000). Identification and functional modelling of DNA sequence elements of transcription. *Brief Bioinform.*, 1(4):372–380.

Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004). Database resources of the National Center for Biotechnology Information: update. *Nucl. Acids Res.*, 32(90001):D35–D40.

Wingender, E. (2004). TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In. Silico. Biol.*, 4(1):55–61.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, 28(1):316–319.

Workman, C. T. and Stormo, G. D. (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, pages 467–478.

Yeo, G. and Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology*, pages 322–331, New York, NY, USA. ACM Press.

Zaret, K. S. and Yamamoto, K. R. (1984). Reversible and persistent changes in chromatin structure accompany activation of a glucocorticoid-dependent enhancer element. *Cell*, 38(1):29–38.

Zhang, M. Q. (1999). Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, 23(3-4):233–250.

Zhang, M. Q. and Marr, T. G. (1993). A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9(5):499–509.

Zhang, T.-Y., Kang, Zhang, Z.-F., and Xu, W.-H. (2004). Identification of a POU factor involved in regulating the neuron-specific expression of the gene encoding diapause hormone and pheromone biosynthesis-activating neuropeptide in Bombyx mori. *Biochem. J.*, 380(Pt 1):255–263.

Zhang, Z. and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.*, 2(2):11.

Zhao, F., Xuan, Z., Liu, L., and Zhang, M. Q. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res.*, 33(Database issue):D103–D107.

Zhao, X., Huang, H., and Speed, T. P. (2004). Finding short dna motifs using permuted markov models. In *RECOMB '04: Proceedings of the eighth annual international con-*

ference on Computational molecular biology, pages 68–75, New York, NY, USA. ACM Press.

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916.

Chapter 2

Accurate Anchoring Alignment of Divergent Sequences

Abstract

Comparative sequence analysis is a powerful approach for identifying conserved functional elements in orthologous sequences. The success of this approach depends on the accuracy of sequence alignment. Here we propose a novel pairwise sequence alignment algorithm, ACANA (ACcurate ANchoring Alignment), for aligning biological sequences at both local and global levels. Like many fast heuristic methods, ACANA uses an anchoring strategy. However, unlike others, ACANA uses a Smith-Waterman-like dynamic programming algorithm to *recursively* identify near optimal regions as anchors for a global alignment. Performance evaluations using a simulated benchmark dataset and real promoter sequences suggest that ACANA is accurate and consistent, especially for divergent sequences. Specifically, we use a simulated benchmark dataset to show that ACANA has the highest sensitivity to align constrained functional sites compared to BLASTZ, CHAOS and DIALIGN for local alignment and compared to AVID, ClustalW, DIALIGN, and LAGAN for global alignment. Applied to 6,007 pairs of human-mouse orthologous promoter sequences, ACANA identified the largest number of conserved regions (defined as over 70% identity over 100 *bp*) compared to AVID, ClustalW, DIALIGN and LAGAN. In addition, the average length of conserved region identified by ACANA was the longest. Thus, we suggest that ACANA is a useful tool for identifying functional elements in cross-species sequence analysis, such as predicting transcription factor binding sites in non-coding DNA. ACANA is publicly available at http://raga.statgen.ncsu.edu/ACANA.

Introduction

Discovering the function of genes and revealing gene regulation networks are important tasks in decoding genome sequences. The conserved protein domains and functional regulatory sites can provide valuable information for inferring gene functions and regulatory controls. Comparative analysis of homologous sequences from related species is an efficient way to reveal such functional domains or regulatory elements (Xie et al., 2005). With the increasing availability of genome sequences from related species, such cross-species comparative analysis has become a powerful and popular approach for decoding information in genome sequences. However, the success of a comparative analysis is largely dependent on the alignment accuracy of sequences.

In the past few decades, many alignment algorithms and corresponding tools have been developed. The standard pairwise alignment algorithms can be traced back to the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) for global alignment, and the Smith-Waterman algorithm (Smith and Waterman, 1981) for local alignment. Both the Needleman-Wunsch algorithm and the Smith-Waterman algorithm use dynamic programming techniques to find the optimal global and local alignments, respectively. Gotoh (1982) extended the algorithms by making the penalty for gap extension different from gap opening. The introduction of the affine gap cost model allows one to assign a more appropriate penalty for a long insertion/deletion, thus making alignments more biologically meaningful. For short and highly similar sequences, these standard deterministic algorithms work well. However, it is very difficult to assign biologically meaningful gap penalties. Thus the standard dynamic programming algorithms are unreliable in aligning divergent homologous sequences with long insertions or deletions. A large gap penalty forces mismatched alignments instead of inserting appropriate gap segments, whereas a small gap penalty results in spurious matching of unrelated regions. Furthermore, the computation time used by either the Needleman-Wunsch or the Smith-Waterman algorithm is proportional to the product of

lengths of two sequences and can increase by a factor of about three when the affine gap cost model is applied. Hence, many heuristic algorithms (Altschul et al., 1990, 1997; Batzoglou et al., 2000; Morgenstern et al., 1998; Morgenstern, 1999; Tatusova and Madden, 1999; Kent and Zahler, 2000; Bray et al., 2003; Brudno et al., 2003b; Schwartz et al., 2003) have been developed to increase alignment speed and/or make alignment more biologically meaningful. BLAST (Altschul et al., 1990, 1997), a standard tool to search for homologous segments, has impressive speed. MUMmer (Delcher et al., 2002) and AVID (Bray et al., 2003), both of which locate anchors using a suffix tree algorithm, are two well-known fast global alignment tools. To increase alignment speed, many of these algorithms search for highly matched gap-free segments of fixed length as alignment seeds. As a result, the obtained alignments may not be (biologically) optimal, especially when sequences are highly divergent.

In many comparative sequence analysis tasks, one is interested in finding short conserved protein domains or non-coding functional elements. In such analyses, one usually deals with sequences with moderate length, i.e., a few thousand to tens of thousand base pairs that can be aligned quickly even by the relatively slow and intricate dynamic programming algorithms. However, quality of alignment remains a challenging issue, particularly when sequences are highly divergent. To address this issue, we developed a new pairwise alignment algorithm, ACANA (ACcurate ANchoring Alignment), to improve quality of alignment of sequences that share only moderate similarity and/or have many long insertions/deletions.

Results

Outline of ACANA algorithm

ACANA, like some fast alignment tools, such as MUMmer (Delcher et al., 2002), AVID (Bray et al., 2003), and LAGAN (Brudno et al., 2003b), uses the anchoring approach for global alignment. However, unlike others, ACANA uses a new strategy for selecting anchoring regions. An anchor-based alignment algorithm has advantages in reducing computation

time and/or improving quality of global alignment. The latter feature is largely dependent on the accuracy of anchor selection. To ensure an accurate alignment, ACANA weights local similarity, regional conservation, and overall similarity of sequences, and chooses the best set of anchoring regions, from which a biologically meaningful global alignment can be constructed. The simplified ACANA alignment algorithm is illustrated in Figure 2.1. The four essential parts of ACANA algorithm are: a heuristic dynamic programming algorithm primarily based on the Smith-Waterman algorithm for calculating matrices and tracing local alignment pathes using affine gap cost model; a hash based algorithm for identifying non-overlapping local alignments in a single pass of calculation of matrices; a method of selecting anchoring regions from local alignments; an algorithm for avoiding unnecessary calculation in recursively searching for the best anchoring regions. The details of these algorithms are described in Methods. For a pair of sequences, ACANA outputs the non-overlapping local alignments as well as a global alignment, so it is a both local and global alignment tool.

Evaluation of performance

To evaluate the performance of ACANA, one would ideally apply it to real sequence data in which true alignments are known. Although several sets of benchmark protein sequences are available for evaluating the performance of multiple sequence alignment programs (Thompson et al., 1999; Bahr et al., 2001; Lassmann and Sonnhammer, 2002), so far, no good benchmark data sets of real genomic sequences are available. True biological alignments of genomic sequences are more difficult to obtain than those of protein sequences. For this reason, we used a benchmark data set of simulated non-coding sequences from Pollard et al. (2004) to evaluate the performance of ACANA . Also, we used a real sequence data set of human-mouse orthologous promoter sequences to assess indirectly the performance of ACANA.

Evaluation on simulated sequences

The simulated data set was generated and used by Pollard et al. (2004) for evaluating the performance of alignment tools in aligning non-coding DNA sequences. The original data consist of four different sets simulated under four different regimes. Here we used only one set whose sequences were simulated under the regime with insetion/deletion evolution and constraint blocks. The sequences in this data set are more realistic and biologically relevant than those in other sets as the authors suggested. In short, the data set consists of sequences with eleven divergence distances ranging from 0.25 to 5.0 substitutions per site. At each divergence distance, there are 1000 pairs of 10 *Kb* sequences. For evaluation, both global and local alignments produced by ACANA were scored for the six measures: overall coverage, overall sensitivity, constraint coverage, constraint sensitivity, constraint specificity, and local constraint sensitivity. These measures were calculated by the exactly same methods used by Pollard et al. (2004). For each of these six measures, a mean and a coefficient of variation were calculated from 1000 replicates at each divergence distance.

Local alignment performance

For a local alignment tool, one would be interested in its ability to accurately align functional constrained sites. To assess this ability, we compared ACANA with DIALIGN (Morgenstern et al., 1996, 1998; Morgenstern, 1999), BLASTZ(Schwartz et al., 2003), and CHAOS (Brudno et al., 2003a)(author suggested settings were used for BLASTZ and CHAOS. The details are described in Pollard et al. (2004)). Figures 2.2 and 2.3 show that ACANA can detect constrained functional sites with a high sensitivity and a reasonable specificity. In particular, ACANA has a higher constraint sensitivity for sequences of intermediate ($1.25 - 3.0$ substitutions per site) or large divergence distances ($3.0 - 5.0$ substitutions per site) than the other three tools. The difference in constraint sensitivity between ACANA and its closest competing tool DIALIGN becomes larger and larger as divergence distance increases

while the difference in constraint specificity is relatively unchanged or even smaller. In addition, the coefficients of variation of all six measures for ACANA are relatively small across different divergence distances (Figure 2.7), which suggests that performance of ACANA is consistent in aligning different pairs of sequences. We also found that ACANA, BLASTZ and DIALIGN all have a high local constraint sensitivity (over 90%) across different divergence distances and show no significant difference from each other, whereas CHAOS has a lower local constraint sensitivity.

Global alignment performance

To assess the performance of ACANA for global alignment, we compared it with top competing alignment tools: AVID (Bray et al., 2003), LAGAN (Brudno et al., 2003b) and DIALIGN, as well as with the classic ClustalW (Thompson et al., 1994). To assess alignment accuracy of a global alignment tool, the most relevant measures are overall sensitivity, constraint sensitivity and specificity. ACANA outperforms all other four tools with regard to these measures, and this is particularly true for sequences of intermediate or larger divergence distances (Figures 2.4, 2.5 and 2.6). Interestingly, after an initial decrease, the overall sensitivity of ACANA increases as the divergence distance increases. This is not the case for the other competing tools of which overall sensitivity either stays relatively unchanged or decreases as divergence distance increases. Again, the performance of ACANA in these measures is highly consistent, as its coefficients of variation stay relatively smaller across different divergence distances than those of the other tools (Figure 2.8).

Assessment on real sequences

The test data set consists of 6007 pairs of human-mouse putative orthologous promoter sequences that we extracted from NCBI GenBank (see Methods). The list of human-mouse orthologous genes was from NCBI HomoloGene database. Each promoter sequence is of

length 4,500 *bp*: 3,500 *bp* upstream and 1,000 *bp* downstream of the transcription start site (TSS) as annotated in GenBank. Because we did not know the true alignments, we used indirect measures of alignment accuracy for assessing ACANA performance.

Local alignment sensitivity

To test whether ACANA local alignment can effectively detect conserved functional sites in homologous sequences, we assessed its sensitivity for aligning conserved transcription factor binding sites (TFBS) in the human-mouse orthologous promoter sequences. Since we do not know the number of truly conserved sites in our promoter sequences, we used the relative TFBS sensitivity described in Methods for assessing a tool sensitivity for local alignment. The performance of ACANA was compared with that of CHAOS (using its default parameter settings) for local alignment. Both ACANA and CHAOS output non-overlapping local alignments, which are anchoring regions for global alignments by ACANA and LAGAN, respectively.

We used the known instances of TFBS from 20 matrix records of Transfac database (Wingender et al., 2000) to find the exactly matched sites as putative functional sites in our sequences. Overall, a total of 135,179 putative sites are found in 6007 unaligned human promoter sequences, among which 57902 and 14602 are detected as conserved sites by ACANA and CHAOS, respectively. That is, on average, the relative TFBS sensitivity of ACANA is 42.83%, and that of CHAOS is 10.80%. The differences between ACANA and CHAOS are consistent across 20 different matrix records (Figure 2.9).

To computationally support our claim that the putative sites are likely true functional sites, we also investigated whether these putative sites have the experimentally verified properties, such as preferred locations, neighboring elements and so on. Here, we use E2F, a well studied transcription factor, as an example to demonstrate that most putative sites may be conserved functional sites. It is known that E2F binding sites are often located in

proximity to the transcription start sites, which is consistent with the proposed E2F function as an initiator binding protein to form a pre-initiation complex of transcription (Cartwright et al., 1998; Black and Azizkhan-Clifford, 1999; Helin, 1998; Hateboer et al., 1998). So if a putative E2F binding site is a true functional site, it is most likely located in regions near transcription start sites. Using mRNA start positions annotated in GenBank as the likely transcription start sites (TSS), all of the putative sites that are exactly matched to the known instances in the record *M00932* of Transfac database are mapped into the relative location to TSS. The results in Figure 2.10 show that most putative E2F sites are near the putative TSS. For a few sites that are located far from TSS, it is possible that the putative TSS are not the true TSS because many mRNA annotated in GenBank are not of full length.

Global Alignment Quality

Instead of directly measuring alignment accuracy, which is impossible when true alignments are unknown, we assessed global alignment quality by the relative length of all conserved regions aligned by different tools. From a biological point of view, an accurate global alignment should correctly align evolutionarily related regions, including the syntenically conserved regions. Therefore, the relative length of syntenically conserved regions aligned can be used as an indirect measure for assessing quality of global alignments by different tools.

For this evaluation, we compared ACANA with AVID, ClustalW, LAGAN and DIALIGN. For DIALIGN, we used the improved version — DIALIGN-2 (Morgenstern, 1999) in this comparison. First, each tool, with its default parameter settings, was used to align each pair of orthologous sequences. Second, for each pairwise global alignment, we used VISTA (Mayor et al., 2000; Frazer et al., 2004) to extract conserved regions (see Methods). Although methods used to define conserved regions are somewhat arbitrary, one of the most frequently used is based on percentage identity over a region of fixed length (Fickett and Wasserman, 2000; Loots et al., 2000). VISTA employs this method with a default cutoff

value of 70% identity over 100 *bp*, a value commonly used for human and rodent species.

Results (Table 2.1 and 2.2) show ACANA not only finds the largest number of orthologous pairs of sequences containing at least one conserved region but also the longest conserved region on average compared to the other two tools. To see the differences, we randomly picked 100 orthologous pairs of sequences from the data set. For each pair, we manually examined the three alignments by their Percent Identity Plots (PIP). In all cases, the PIP plots of alignments from the three algorithms are similar for sequence regions with high similarity, but may be different for regions with only moderate similarity. An example is given in Figure 2.11.

Computation Time

ACANA has a quadratic running time and can be slower in aligning large sequences than tools that use a sub-quadratic running time such as AVID and LAGAN. However, ACANA is reasonably fast for sequences of moderate lengths. For instance, it took about 3 h 27 min to finish aligning the 6,007 pairs of human-mouse orthologous promoter sequences on a Red Hat Linux PC with 2.4 GHz processor. The corresponding times used by AVID, LAGAN, and DIALIGN, are 45 min, 1 h 28 min, and 6 h 21 min, respectively. ClustalW took more than 12 h to finish the same job.

Discussion

A challenge in comparative sequence analysis is to obtain high quality sequence alignments while minimizing computational time. In the past two decades, significant progress has been made. The most important achievement is the dramatic reduction in computation time by heuristic algorithms coupled with faster computers, which makes it possible to align genome-size sequences. Despite this progress, many challenges remain, notably the quality of alignment. Except when applied to the smallest and simplest sequences, almost no two

current alignment algorithms regularly give the same alignment. It is very difficult, if not impossible, to reflect accurately evolutionary events such as point mutation, insertion, deletion, duplication, rearrangement, etc. in a scoring function for alignment. Nonetheless, the recent advances in alignment methodologies have made a great impact on modern biological research.

The introduction of anchoring has made genome-wide alignment feasible and fairly accurate. While the index or word-chaining based approaches are very efficient, these heuristic approaches are not guaranteed to find the near optimal local alignments as anchors, especially for divergent sequences. The ACANA algorithm uses the Smith-Waterman-like dynamic programming algorithm for local alignment, enabling it to identify the near optimal local alignments. Furthermore, ACANA uses a new strategy to select an anchoring region from a set of significant local alignments.

Performance evaluations suggest that ACANA is an accurate and consistent alignment tool for both local and global alignments. Using a set of simulated benchmark dataset, we found that ACANA has the highest constrained sensitivity in correctly aligning the known constrained functional sites embedded in the sequences compared to BLASTZ, CHAOS, and DIALIGN for local alignment and AVID, ClustalW, DIALIGN, and LAGAN for global alignment. ACANA performs best for sequences of moderate to large divergent distances. When tested on a set of paired putative human/mouse orthologous promoter sequences, ACANA found the largest number of orthologues that contained at least one conserved region (over 70% identity over 100 *bp*) compared to AVID, ClustalW, DIALIGN and LAGAN. In addition, the average length of the conserved regions identified by ACANA was the longest. We believe that ACANA shows some improvement over existing tools for aligning divergent sequences. We attribute the potential improvements partially to ACANA's recursive anchoring selection strategy.

We would like to point out that the current version of ACANA is not capable of dealing

with inversions in sequences. We think that such capability can be easily incorporated by aligning sequence segments in both directions in the recursive anchoring selection step. Such work is in progress. Lastly, ACANA may be combined with other faster local alignment tools such as CHAOS when significant improvement in speed is needed to align genome-size sequences.

In conclusion, we believe that ACANA is a novel and accurate alignment algorithm. Its new recursive anchoring selection strategy may represent an improvement over existing methods. ACANA's ability to align conserved functional sites and its robustness to large insertions/deletions make it particularly useful in comparative genomic analysis of promoter sequences for functional element discovery.

Methods

Algorithms

Calculating alignment matrices

As we know, to find the best local alignment from a pair of sequences A and B of length m and n, respectively, the Gotoh algorithm needs to fill three matrices of size $m \times n$ by the following recursion relations:

$$F_{i,j} = \max(F_{i,j-1} - g_e, H_{i-1,j} - g_o, 0)$$
$$G_{i,j} = \max(G_{i-1,j} - g_e, H_{i,j-1} - g_o, 0)$$
$$H_{i,j} = \max(H_{i-1,j-1} + \text{score}(A_i, B_j), F_{i,j}, G_{i,j}, 0)$$

where g_o is the gap opening penalty, g_e is the gap extension penalty, and $\text{score}(A_i, B_j)$ is the score from a substitution scoring matrix where base A_i is matched with B_j. Thus the computational cost of the standard Smith-Waterman algorithm using the affine gap cost

model is about three times of that using the constant gap cost model. Some improvements have been proposed to increase computational speed of the algorithm (Green, 1993; Trelles et al., 1998; Rognes and Seeberg, 2000). For example, SWAT (Green, 1993) can increase the speed of alignment by a factor of about two by reducing unnecessary calculations in matrices F and G. The essential functions of F and G are to keep information used to make decisions as to whether an inserted gap is to extend an existing gap segment or to open a new gap segment in alignment. That is, F and G make true contributions only in the locations where insertions or deletions occur. However, biologically significant local alignments or conserved sequence regions generally do not have many insertions and deletions, which means that most parts of matrices F and G are not needed. In our algorithm, we replace F and G by a single matrix I. Instead of recording alignment scores, matrix I keeps the path of local alignments. Since the score of a cell in S can only come from three previous cells, two bits of memory is enough for a cell of I to record three possible sources, which can save computation time and space. ACANA fills the score matrix S and path-tracing matrix I by a dynamic programming algorithm with the following recursion relations.

1. IF $i = 0$, set $S_{i,j} = 0$, $I_{i,j} = 0$, where $j = 0...n$

2. IF $1 \leq i \leq m$, calculate $S_{i,j}$ and $I_{i,j}$ by

$$S_{i,j} = \max \begin{cases} c_0 = \max(0, S_{i-1,j-1} + \text{score}(A_i, B_j)) \\ c_1 = S_{i,j-1} + \begin{cases} g_e & \text{if } (I_{i,j-1} = 1) \\ g_o & \text{otherwise} \end{cases} \\ c_2 = S_{i-1,j} + \begin{cases} g_e & \text{if } (I_{i-1,j} = 2) \\ g_o & \text{otherwise} \end{cases} \end{cases} \qquad I_{i,j} = \begin{cases} 1 & \text{if } (S_{i,j} = c_1) \\ 2 & \text{else if } (S_{i,j} = c_2) \\ 0 & \text{otherwise} \end{cases}$$

During matrix calculation, ACANA is able to efficiently track all non-overlapping local alignments. This is based on the essential observation that many local alignments can be

traced back to one common start position in the alignment matrix S. The key is that among local alignments with a common start position, the best one can be tracked by dynamically updating its end position when we fill scores of alignment matrices. The local alignment tracking process is implemented in ACANA using a hash structure, in which the keys are start positions and the values are stop positions. In this way, ACANA is able to identify all non-overlapping local alignments of scores above a certain threshold in a single pass of calculation of alignment matrices.

Tracing alignment path

A tricky part of ACANA algorithm is to trace the path of a local alignment. It is necessary to combine information in both S and I to correctly trace the path of a local alignment. Suppose that a local alignment ends at the position (c, d) in the alignment matrix, and (i, j) is the current position in the course of tracing back. Variables $gLen1$ and $gLen2$ are defined as the number of gaps inserted at the current position in the first and second sequence, respectively. ACANA traces the local alignment according to the following steps.

1. Initially set $i = c$, $j = d$, and $gLen1 = 0$, $gLen2 = 0$.
2. If $S_{i,j} > 0$, then go to next step, otherwise stop.
3. If $I_{i,j} = 1$, then decrease j by 1 and increase $gLen1$ by 1.
4. Else if $I_{i,j} = 2$, then decrease i by 1 and increase $gLen2$ by 1.
5. Otherwise, do the following steps.

    (a) If $gLen1 > 2$, then perform the following gap shift steps.

        i. Let $x = i, y = j$

        ii. Continuously increase both x and y by 1 and reduce $gLen1$ by 1 until $I_{x+1,y+1} \neq 0$.

        iii. If $x > i$ then calculate s_a, and s_b by

        $s_a = \sum_{k=1}^{x-i} \text{score}(seq1[i + k - 1], seq2[j + k + gLen1 - 1])$

$s_b = \sum_{k=1}^{x-i} \text{score}(seq1[i+k-1], seq2[j+k-1])$

iv. If $s_b > s_a$, move the gap segment downstream $x - i$ positions in the local alignment.

(b) Else if $gLen2 > 2$, then perform gap shift in the second sequence according to the similar rule described in the above step.

(c) Decrease both i and j by 1, and reset $gLen1$ and $gLen2$ to zero.

6. Go back to step 2.

Recalculating alignment matrix

ACANA uses recursive approach for anchor selection. At the beginning, ACANA searches the entire sequence region for a set of significant (an alignment score above a certain threshold) local alignments as anchor candidates, from which the best one is chosen as the first anchor for global alignment according to a set of criteria (described below). Then, ACANA searches for anchor candidates in two smaller regions that are separated by the fixed anchor. For example, suppose that ACANA has found the first anchor starting at position (a, b) and ending at (c, d) in the alignment matrix S. ACANA then searches for an anchor in the upstream region of the previously fixed anchor from rectangular region $(0, 0)$ to (a, b) of the matrix S. It also searches for an anchor in downstream sequence region from rectangular region (c, d) to (m, n) of the matrix S, where m and n are the length of two sequences, respectively. This recursive process continues until no anchor found.

To find the best local alignment in the upstream sequence region of the fixed anchor, it is obvious that there is no need to recalculate the scores of the cells in the corresponding region of the matrix S. However, for the sequence region in the downstream of the fixed anchor, it is necessary to recalculate scores in the corresponding region of the matrix S, otherwise there will be no guarantee of finding the best local alignment in the region. This score recalculation by the dynamic programming algorithm can be computationally costly

in the recursive searching process. Fortunately, ACANA can avoid recalculating scores of most cells in the region, which won't change during the current cycle of local alignment. The recalculation of scores in a rectangular region from (c, d) to (e, f) of matrix S, where $e > c$ and $f > d$, is performed by the following steps.

1. Initially set $i = c$.
2. Update the score $S_{c,j}$ (set $S_{c,j}$ to 0 if $S_{c,j} \neq 0$), where $j = d...f$, and record the maximum of j (denoted by j_{max}) among those $S_{c,j}$ whose values have changed after update.
3. Set $k = \min(j_{max} + 1, f)$.
4. Increase i by 1. Update $S_{i,j}$ in row i, where $j = d...k$, and record j_{max} among those $S_{i,j}$ whose value have changed after update.
5. If $k > j_{max}$, then set $k = j_{max} + 1$. Otherwise, continue to update the score $S_{i,j}$ of cells in row i where $j = (k + 1)...f$, until $S_{i,j}$ does not change; then set $k = j$.
6. If $i < e$, then go back to step 4; Otherwise stop.

Actually, this algorithm can be used to identify top local alignments in any rectangular region of the alignment matrix S.

Anchor selection

In each cycle of the recursion process described above, ACANA selects an anchoring region from the top local alignments identified in the cycle. First, ACANA calculates a biological relevance score ω of each local alignment, which is defined by

$$\omega = \begin{cases} v \times \log u & \text{if } u \geq 1 \text{ and } v \geq 5 \\ 0 & \text{otherwise} \end{cases}$$

where u and v are the alignment score and length (without counting gaps) of a local alignment, respectively. For a local alignment, its chance to be biologically relevant increases

as its length and alignment score increase. If evolutionary rates are homogeneous across different regions of sequences, the alignment scores and lengths of local alignments should be highly correlated. However, the correlation is not strong when evolutionary rates are significantly different. The score of a shorter local alignment can be larger than that of a much longer one, which is usually more biologically relevant. So it may be a better way to consider both alignment score and length in assessing the biological relevance of a local alignment. For the same increase in alignment scores, the chance of biological relevance increases more sharply for alignments with small alignment scores than for those with large scores. To approximate that pattern, we weigh ω by the logarithm of an alignment score. Also, as a reasonable choice, we make ω proportional to the length of a local alignment. So ω weighs more on alignment score when a local alignment is short, and more on alignment length when a local alignment is long. In this way, we believe, ω can appropriately assess the biological relevance of a local alignment. It helps ACANA to detect and distinguish biologically relevant local alignments from random matched segments, especially when aligning highly divergent sequences, where the alignment scores of local alignments are generally small while lengths can be quite long.

Second, ACANA ranks local alignments according to their ω values. If ω of the top ranked local alignment is larger than a certain threshold, ACANA retains all local alignments as anchor candidates whose ω values are not much different from the top ranked candidates; otherwise no anchor is chosen by ACANA. If there are several anchor candidates, ACANA further calculates a regional weight score for each one by

$$G_a = u_a + \sum_b u_b$$

where u_a is the alignment score of the anchor candidate a, and each b is a non-overlapping local alignment that does not intersect with a. G_a is the regional weight score of the anchor candidate a. The candidate with the highest regional weight score is chosen as an anchor

for global alignment.

Construction of global alignment

After fixing anchoring regions, ACANA uses the standard Gotoh dynamic programming algorithm to align the remaining sequence segments. These alignments are connected with the anchoring regions to form a global alignment. The default nucleotide substitution scoring matrix of ACANA is based on the alignment-scoring scheme derived by Chiaromonte et al. (2002), which is also used by alignment tools BLASTZ, AVID and LAGAN. The default amino acid substitution scoring matrix is BLOSUM62 from NCBI.

Evaluation

Alignment measures for simulated data

For each pairwise alignment of the simulated sequences, the six measures: overall coverage, overall sensitivity, constraint coverage, constraint sensitivity, constraint specificity, local constraint sensitivity are calculated by the exactly same software provided by Pollard et al. (2004). In brief, overall coverage is the fraction of un-gapped sites in a simulated alignment that are included in a tool alignment; overall sensitivity is the fraction of un-gapped sites in a simulated alignment that are aligned to the correct base in a tool alignment; constraint coverage is the fraction of un-gapped constrained sites in a simulated alignment that are included in a tool alignment; constraint sensitivity is the fraction of un-gapped constrained sites in a simulated alignment that are aligned to the correct base in a tool alignment; constraint specificity is the fraction of unconstrained sites in a simulated alignment that are gapped or not included in a tool alignment; local constraint sensitivity is the fraction of un-gapped constrained sites are aligned corrected among all those contained in a tool alignment. For each of these six measures, a mean and a coefficient of variation were calculated from 1000 replicates at each divergence distance. For ACANA, we calculated these measures

directly from its alignments. For the other tools used for comparison, these measures except coefficient of variation, which was derived from a sample mean and its standard deviation, were from Pollard et al. (2004).

Preparation of promoter sequence data

We first extracted all mRNA accession numbers of human genes that were in the RefSeq reviewed records (the quality of sequences is the best) according to the annotations of LocusLink downloaded from NCBI. Second, for each human gene, we searched for its mouse orthologous gene from HomoloGene database, which was also from NCBI. Only orthologous pairs were kept in the final list. Once we had the mRNA RefSeq accession numbers of orthologous pairs, we extracted their promoter sequences from NCBI human and mouse genome sequences, respectively. Each promoter sequence is of length 4500 *bp*: 3500 *bp* upstream and 1000 *bp* downstream of the transcription start site (TSS) as annotated in GenBank. In total, 6007 pairs of human-mouse putative orthologous promoter sequences are in our test data set. All known repetitive elements from Repbase database (Jurka, 1998, 2000) were masked in the sequences by Censor (ver. 4.1) (Jurka, 2000) and WU-BLAST (ver. 2) (Altschul and Gish, 1996) installed at our local server.

Identification of putative functional sites

The known instances of TFBS are from matrix records of Transfac database (version 7.4). We used 20 human TFBS matrix records, each of which contains at least 20 known binding sites. For each matrix, we counted all motif sites, which are exactly matched to one of the known instances of TFBS from the matrix, in both strands of the unaligned human promoter sequences as the total number (P_n) of putative TFBS sites. Among all these putative sites, the sites in the local alignments of human-mouse orthologous sequences aligned by an alignment tool are counted as the number (C_n) of conserved sites detected by the tool.

Relative TFBS sensitivity (RS) is defined as the fraction of all putative TFBS sites in full human promoter sequences that are detected by a tool as conserved sites, that is, $RS = C_n/P_n$.

Extraction of conserved regions

We used the VISTA tool (Mayor et al., 2000; Frazer et al., 2004) to extract conserved regions in each global alignment of orthologous promoter sequences. The 70% identity over 100 bp stretch was used as the cutoff value of a conserved region. For each global alignment, we added up the lengths of all conserved regions found by VISTA as the total length(ℓ) of conserved regions in an orthologous alignment. The average length of conserved regions per alignment is calculated by $\frac{\ell}{N}$, where N is the number of alignments each of which contains at least one conserved region. The summary statistics (Tables 2.1 and 2.2) were calculated using SAS.

Acknowledgments

We thank Bruce Weir and Jeffrey Thorne for critically reading the manuscript, and Clarice Weinberg and Stephen Heber for helpful comments.

Tables

Table 2.1: The summary statistics of the average length (*bp*) of conserved regions per pairwise alignment of 6,007 pairs of human-mouse promoter sequences by five different alignment tools: ACANA, AVID, ClustalW, DIALIGN, and LAGAN. A conserved region is defined as more than 70% identity over 100 bp stretch. N is the number of alignments that contain at least one conserved region. SD stands for standard deviation.

	N	Mean	SD	Min	Max
ACANA	4851	907	697	50	4402
AVID	4658	753	614	51	4294
CLUSTALW	4219	893	725	48	4488
DIALIGN	4710	753	621	44	4284
LAGAN	4805	862	677	50	4353

Table 2.2: The summary statistics of length differences of conserved regions detected by five different alignment tools: ACANA, AVID, ClustalW, DIALIGN, and LAGAN. Here the N is the number of pairs of orthologous genes, from which both alignment tools can find conserved regions. The p-value were computed from a paired t test. SD stands for standard deviation.

Difference	N	Mean	SD	$\Pr > \|t\|$
ACANA - AVID	4641	185	168	$< .0001$
ACANA - CLUSTALW	4190	67	213	$< .0001$
ACANA - DIALIGN	4699	177	167	$< .0001$
ACANA - LAGAN	4788	52	92	$< .0001$

Figures

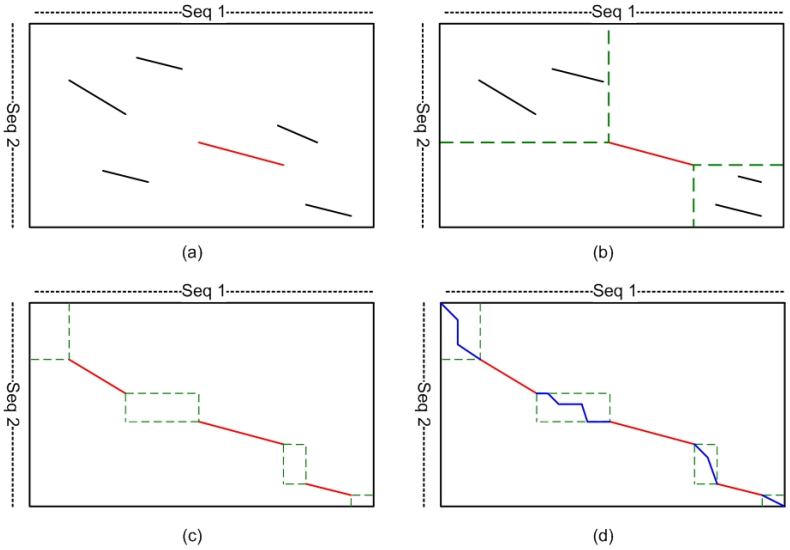


Figure 2.1: Illustration of the simplified ACANA algorithm. (a) Compute score of each cell in the matrix by a dynamic programming algorithm, and select the best anchor from local alignments with scores above a certain threshold. (b) Fixing the anchor, recursively select the best anchors in its up-left and down-right regions. (c) All selected anchors are fixed for the global alignment. (d) Finding optimal global alignment for each region between the fixed anchors by Gotoh algorithm, and connecting them with the fixed anchors to generate the global alignment.

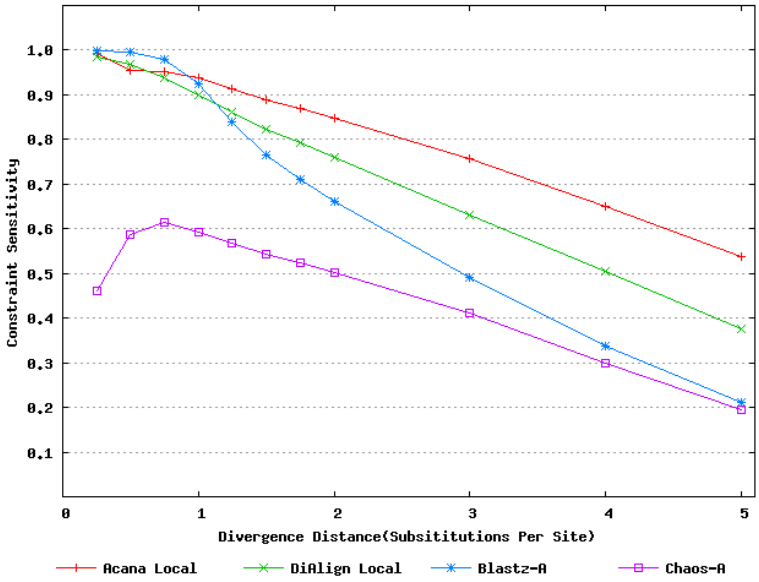


Figure 2.2: The constraint sensitivities of local alignment tools.

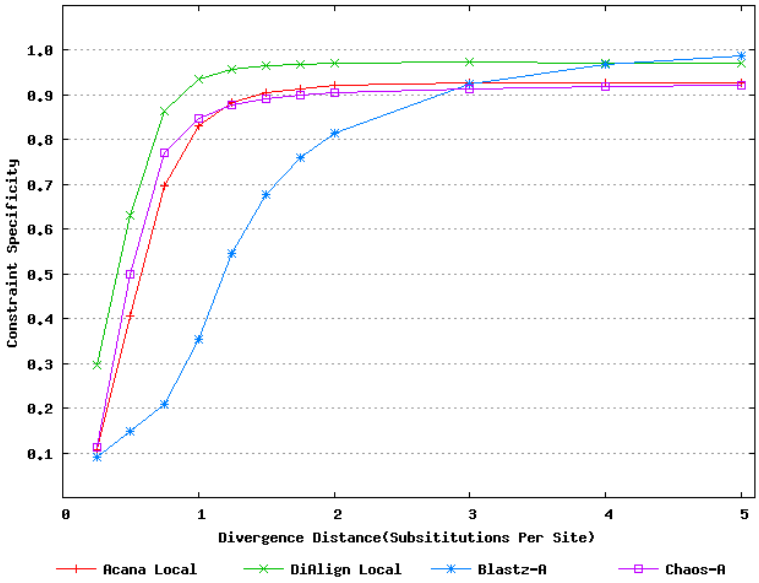


Figure 2.3: Constraint specificities of local alignment tools.

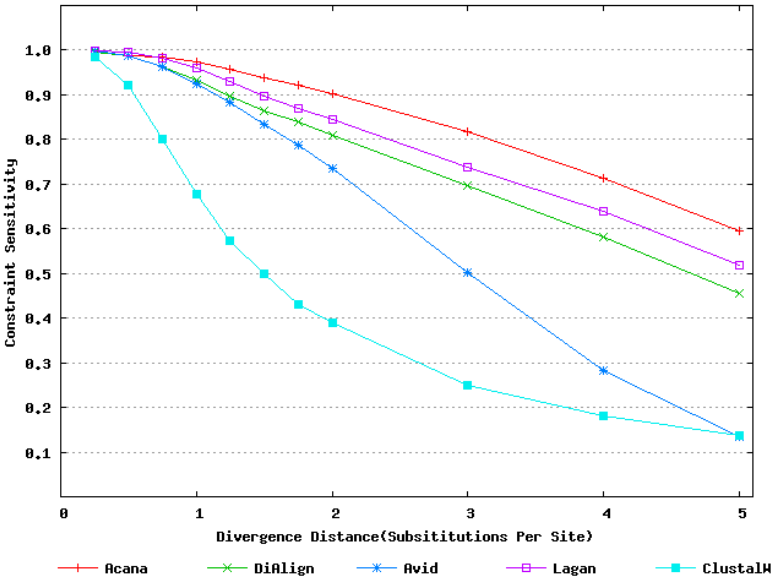


Figure 2.4: Constraint sensitivities of global alignment tools

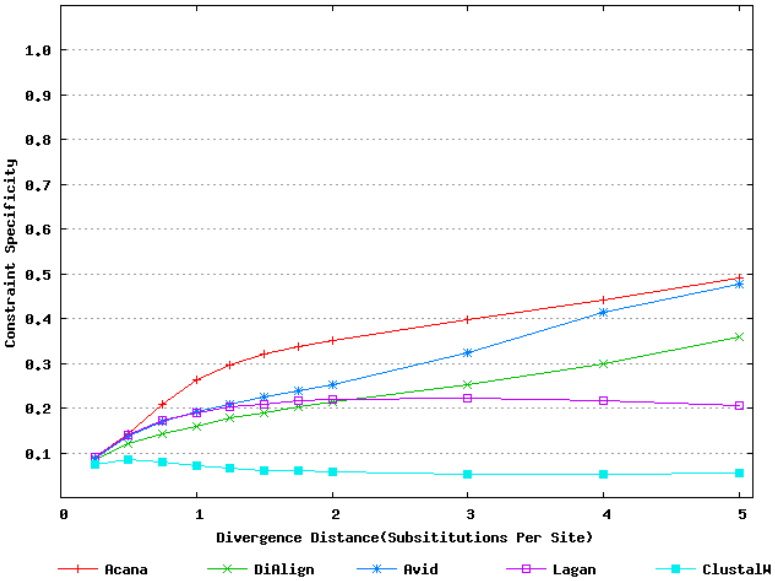


Figure 2.5: Constraint specificities of global alignment tools

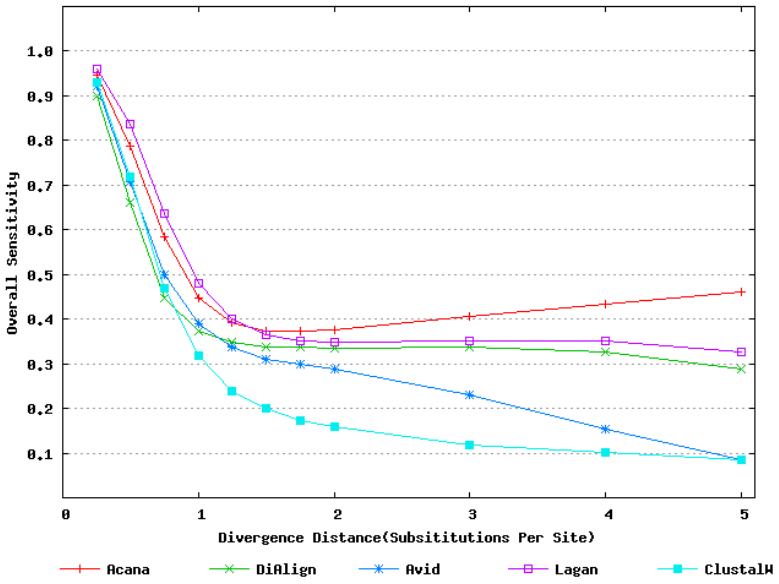


Figure 2.6: Overall alignment sensitivities of global alignment tools

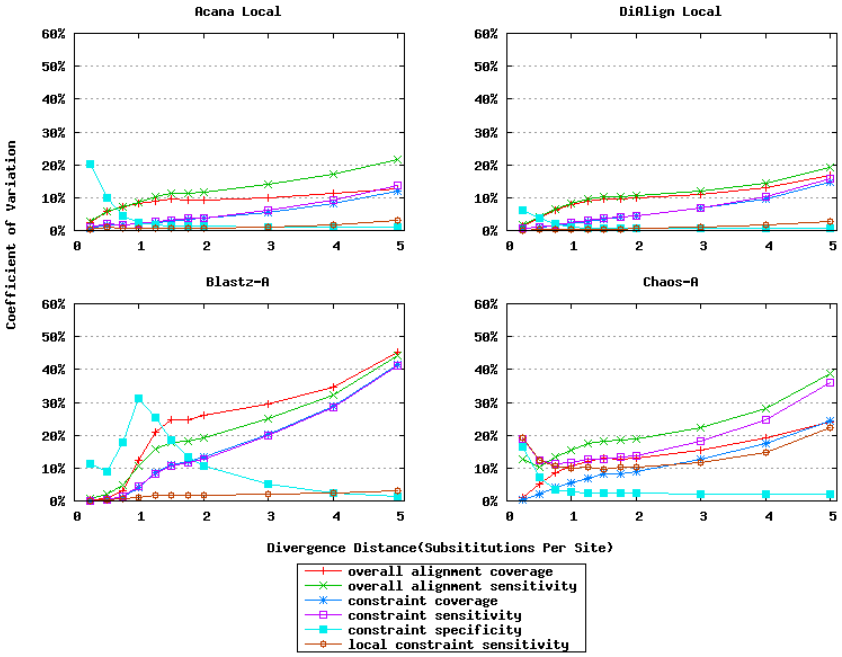


Figure 2.7: Coefficients of Variation of local alignment tools

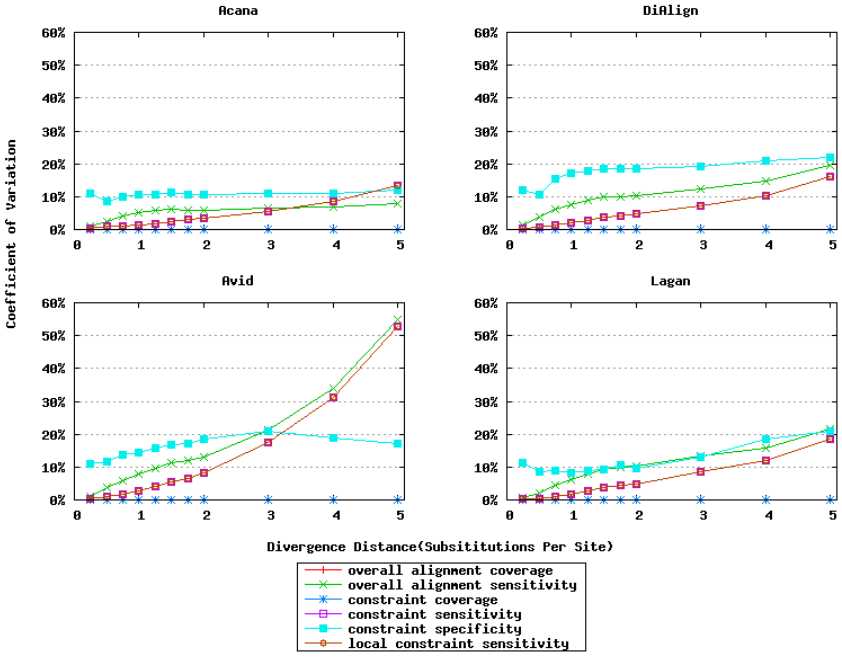


Figure 2.8: Coefficients of Variation of global alignment tools

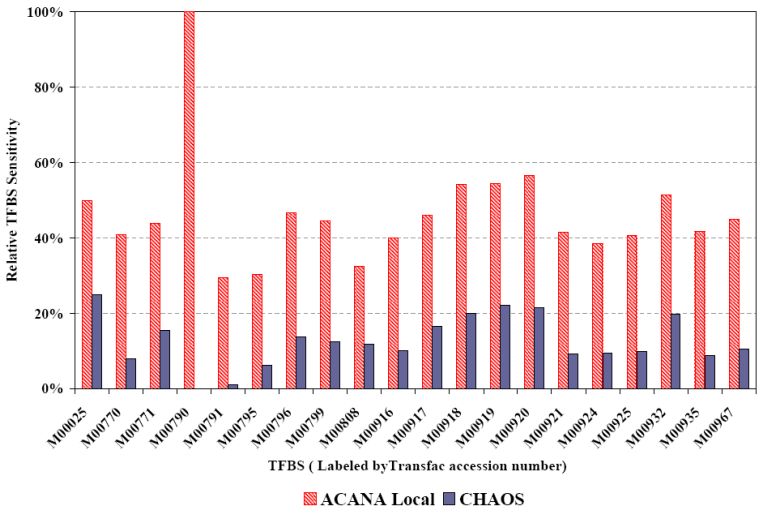


Figure 2.9: Relative TFBS sensitivity 6007 pairs of putative human-mouse orthologous promoter sequences were aligned by ACANA(local alignment) and CHAOS. All putative sites that are exactly matched to one of the known instances of TFBS from Transfac database in the unaligned human sequences were counted as the total number of sites. Relative TFBS sensitivity is defined as the fraction of all putative TFBS sites in full human promoter sequences that are detected by a tool as conserved sites.

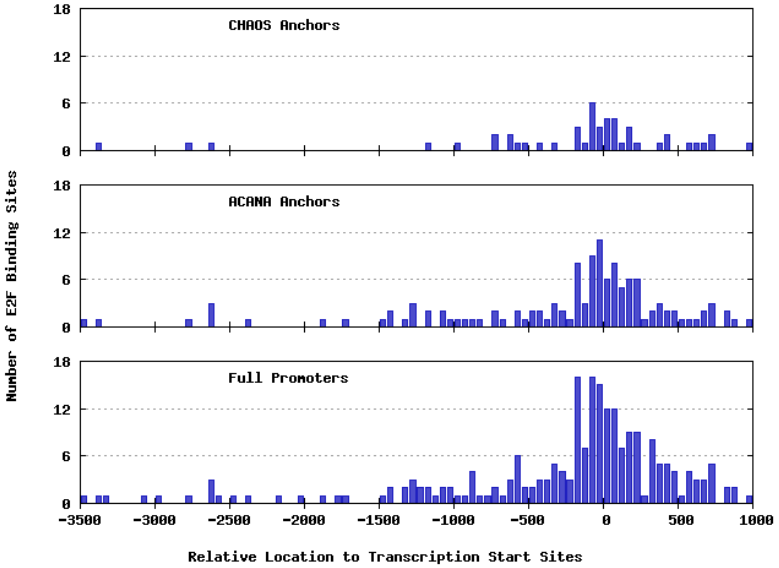


Figure 2.10: Location distribution of E2F binding sites 6007 pairs of putative human-mouse orthologous promoter sequences were aligned by ACANA (local alignment only) and CHAOS. All known instances of functional site of E2F in the record *M00932* of Transfac database 7.4 were used to find the exactly matched sites in the unaligned human sequences. The bottom plot is the location distribution of all E2F putative sites. The middle plot is the distribution of the putative sites in ACANA local alignments (Anchors for ACANA global alignment). The top plot is the distribution of the putative sites in CHAOS alignments (Anchors for LAGAN global alignment).

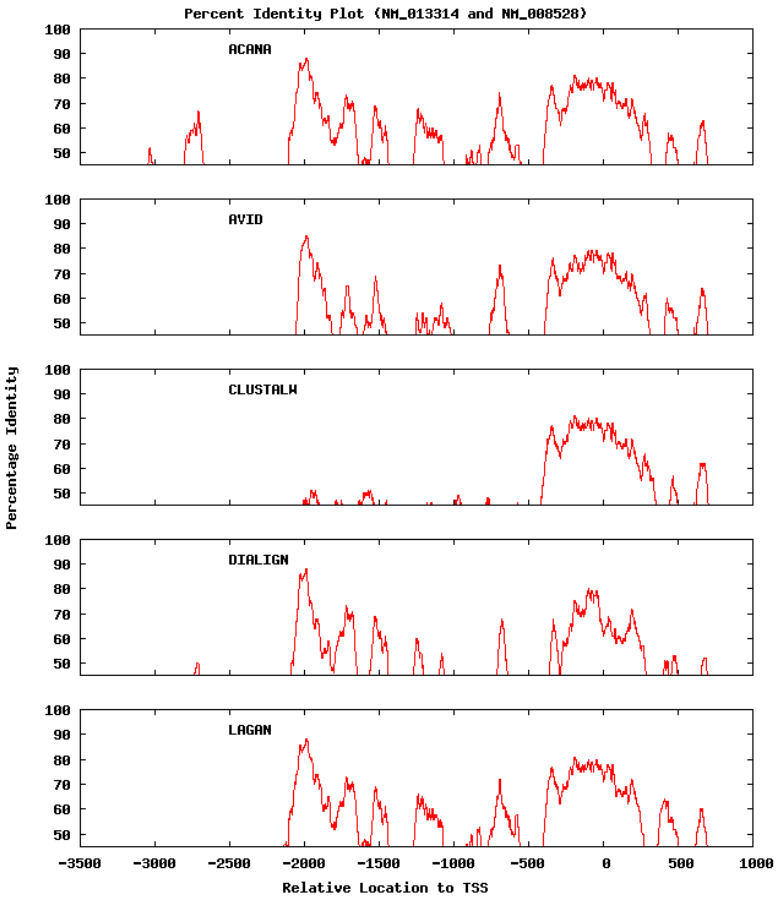


Figure 2.11: Percent Identity Plots of global alignments The promoter sequences are of human (NM_013314) and mouse (NM_008528) orthologous genes encoding B-cell linker protein. The plots show that only ACANA is able to detect a conserved region between positions [-3000,-2500] relative to the transcription start site of the human gene.

Appendix: a short manual of ACANA

A Short Manual of ACANA

```
INSTALLATION
    Just extract all files from one of ACANA compressed files
    to a directory by one of the following commands:
            tar xvf ACANA_<System>.tar
            tar xvfz ACANA_<System>.tgz
            unzip ACANA_<System>.zip

    Enter the directory ACANA_DIR, and follow the usage
    instructions below to run ACANA.

USEAGE

    Usage: ./ACANA -I InputSeqFile
           ./ACANA -I InputSeqFile -O outAlignfile.fa
           ./ACANA -I InputSeqFile -S ScoreMatrix -O outAlignfile.fa

    ========Required Parameters with Values========

    -I InputSeqFile Input sequence file in FASTA format

    ========Optional Parameters with Values========
    The default values are in brackets

    -S ScoreMatrix[dnaMatrix]    The scoring matrix file name
    -O OutputFile                The output file name for alignments,
                                 see OUTPUT for the default value
    -G Open_Cost[-500]           The opening gap penalty
    -A Ext_Cost[-25]             The gap extension penalty
    -T ScoreFactor[10]           The factor used to increase
                                 the cutoff of local alignments
    -C CutOff_len[100]           The minimum length of anchoring
                                 segments

    ========Optional Indicator Parameters========

    -L Only output non-overlapping local alignments
    -P The input is protein sequences. The default is DNA sequences
    -H Print the usage information of ACANA
    -V Print the version information of ACANA

INPUT SEQUENCE FORMAT
    The input sequence File must be in FASTA format. A sequence file
    can contain multiple pairs of sequences, for example, if the
```

```
    file contains following sequences:

    >specie1Seq1
    aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
    >specie2Seq1
    aaaaaaaaaaaaaaaaaaaaaaaaccccccccccccc
    >specie1Seq2
    ccccccccccccccccccccccccccccccccccccc
    >specie2Seq2
    aaaaaaaaccccccccccccccccccccccccccccc
    >specie1Seq3
    ttttttttttttttttttttttttttttttttttttt
    >specie2Seq3
    ttttttttttttttttttttttttaaaaaaaaaaaaaa

    ACANA will align 3 pairs of sequences: specie1Seq1 and specie2Seq1,
    specie1Seq2 and specie2Seq2, specie1Seq3 and specie2Seq3

OUTPUT
    By default, ACANA outputs three files with the following file
    extensions:
    1) *.ACANA: contains all pairwise global alignment in FASTA
       format

    2) *.ACANA.loc: contains alignments of anchoring segments
       for global alignments with following format

       >seq1ID  seq2ID  score_Local_align  score_global_align
       1 start_pos    end_pos    score    sequence
       2 start_pos    enc_pos    score    sequence

       For example,
       >seq1    seq2   1616    -3179
       1 557    777    100    ggggtgacgccctggccaggc...
       2 141    361    100    gaggtgatatcccggacacgc...

    3) *.ACANA.score: contains the alignment score of conserved
       region and the score of a global alignment

    If no output filename specified, ACANA uses the input sequence
    filename as the prefix of output filename and output files will
    be in the same directory of the input sequence file.
    For exmaple, if the input sequence file is ./dir/SEQ.fa,
    then the output files are:
     (1) ./dir/SEQ.fa.ACANA
     (2) ./dir/SEQ.fa.ACANA.loc
     (3) ./dir/SEQ.fa.ACANA.score
```

```
OTHERS
    + ACANA is default program.
    + ACANA2 runs a little bit faster and more memory efficient, but
      it may be not as accuracy as the default ACANA program in
      aligning sequences with high similarity but having many
      insertions or deletions.
```

Bibliography

Altschul, S. F. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol.*, 266:460–480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402.

Bahr, A., Thompson, J. D., Thierry, J. C., and Poch, O. (2001). BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, 29(1):323–326.

Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B., and Lander, E. S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, 10(7):950–958.

Black, A. R. and Azizkhan-Clifford, J. (1999). Regulation of E2F: a family of transcription factors involved in proliferation control. *Gene*, 237(2):281–302.

Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. *Genome Res.*, 13(1):97–102.

Brudno, M., Chapman, M., Göttgens, B., Batzoglou, S., and Morgenstern, B. (2003a). Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, 4(1):66.

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

Cartwright, P., Müller, H., Wagener, C., Holm, K., and Helin, K. (1998). E2F-6: a novel member of the E2F family is an inhibitor of E2F-dependent transcription. *Oncogene*, 17(5):611–623.

Chiaromonte, F., Yap, V. B., and Miller, W. (2002). Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.*, pages 115–126.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, 30(11):2478–2483.

Fickett, J. W. and Wasserman, W. W. (2000). Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, 11(1):19–24.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, 32(Web Server issue):W273–W279.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162(3):705–708.

Green, P. (1993). http://www.genome.washington.edu/uwgc/analysistools/swat.cfm.

Hateboer, G., Wobst, A., Petersen, B. O., Cam, L. L., Vigo, E., Sardet, C., and Helin, K. (1998). Cell cycle-regulated expression of mammalian CDC6 is dependent on E2F. *Mol. Cell. Biol.*, 18(11):6679–6697.

Helin, K. (1998). Regulation of cell proliferation by the E2F transcription factors. *Curr. Opin. Genet. Dev.*, 8(1):28–35.

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol*, 8(3):333–337.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, 16(9):418–420.

Kent, W. J. and Zahler, A. M. (2000). Conservation, regulation, synteny, and introns in a large-scale c. briggsae-c. elegans genomic alignment. *Genome Res.*, 10:1115–1125.

Lassmann, T. and Sonnhammer, E. L. L. (2002). Quality assessment of multiple alignment programs. *FEBS Lett.*, 529(1):126–130.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–140.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047.

Morgenstern, B. (1999). Dialign 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218.

Morgenstern, B., Dress, A., and Werner, T. (1996). Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA*, 93:12098–12103.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998). Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics*, 14:290–294.

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

Pollard, D. A., Bergman, C. M., Stoye, J., Celniker, S. E., and Eisen, M. B. (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, 5(1):6.

Rognes, T. and Seeberg, E. (2000). Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics*, 16(8):699–706.

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

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197.

Tatusova, T. A. and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, 174(2):247–250.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680.

Thompson, J. D., Plewniak, F., and Poch, O. (1999). BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88.

Trelles, O., Andrade, M. A., Valencia, A., Zapata, E. L., and Carazo, J. M. (1998). Computational space reduction and parallelization of a new clustering approach for large groups of sequences. *Bioinformatics*, 14(5):439–451.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, 28(1):316–319.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345.

Chapter 3

Optimized Mixed Markov Models for Motif Identification

Abstract

Identification of biological motifs, such as transcription factor binding sites, protein domains, and RNA splicing sites, is an important task in functional genomics. Because of the large variation of the degree of motif conservation and position dependency structures within a motif, an effective model for motif prediction can have complexity ranging from the one as simple as a consensus sequence to that as complicated as a fully connected network model. In this paper, we present a flexible and effective mixture of optimized Markov models to allow adjustment of model complexity for different motifs. Also we present the Directed Neighbor-Joining (DNJ) method for efficiently optimizing the arrangement of motif positions for a Markov chain to incorporate both adjacent and non-adjacent dependencies. We then describe the method for efficiently training a mixture of Markov models, as well as the method for calculating the effective number of parameters of the mixture to enable a correct model to be selected. We implemented these methods into our OMiMa system, an efficient tool for motif identification in biological sequences. Finally, we demonstrate, from different aspects in several examples, that optimized mixture of Markov models can improve prediction accuracy for motif identification in both DNA and protein sequences. We also show that OMiMa can correctly select the best model, and is efficient in terms of computation time and memory usage. OMiMa was developed with C++ language and is publicly available at http://raga.statgen.ncsu.edu/omima.

Introduction

Biological sequences, including DNA, RNA and proteins, contain many conserved short motifs, such as transcription factor binding sites (TFBS), RNA splicing sites, and protein domains, which are important for gene function and regulation. With increasing availability of homologous sequences, identification of such functional motifs has become an important approach for gene finding, gene function prediction, and understanding the mechanisms of their regulation and evolution (Burge and Karlin, 1997; Wray et al., 2003; Kellis et al., 2003; Negre et al., 2005; Xie et al., 2005).

A most commonly used model for motif identification is the Weight Matrix Model (WMM) proposed by Staden (1984), also called Position Weight Matrix (PWM) or Mononucleotide Weight Matrix (MWM). In this paper, we use PWM to refer to such a model. PWM is usually generated by aligning sequences of known transcription factor binding sites, from which weights or scores in the matrix are calculated according to the observed frequencies of bases. The score function is defined somehow arbitrarily, such as the negative log of a base frequency, information content or relative information content. Although PWM is simple, biologically it does make some sense. It has been shown that the score of a PWM is proportional to the binding energy contributed by each base at each position (Stormo and Fields, 1998). PWM has been used by many motif identification programs, e.g. MatInspector (Quandt et al., 1995) and Match (Kel et al., 2003), and performs reasonably well for motif identification in some cases. While a PWM can capture both nucleotide preferences at each position and different levels of position specificities, it does not contain information of correlations or dependencies between positions within a motif. Recent studies (Agarwal and Bafna, 1998; Benos et al., 2001; Bulyk et al., 2002) indicate that there are important interactions or correlations between adjacent positions as well as non-adjacent positions within a motif in many cases. The inability of a PWM to capture such dependency information has limited its power to find true motif sites.

Many models have been developed to incorporate dependency information of positions. Motif models, such as Dinucleotide Weight Matrix Model (DWMM) (Schneider et al., 1986), Weight Array Model(WAM) (Zhang and Marr, 1993), can incorporate dependencies between adjacent positions. To incorporate further dependencies of non-adjacent positions, Ponomarenko et al. (1999) extended DWMM by introducing the Oligonucleotide Weight Matrices model, which includes a comprehensive set of oliogonucleotide matrices classified into 5 biological function categories. A WAM model could also be extended to high-order WAM in principle, e.g. windowed second order WAM (Burge and Karlin, 1997). However, the exponentially increased number of parameters of these models makes it difficult to use them practically due to insufficient training data. To address the weaknesses of WAM in incorporating long-ranging interactions, Burge and Karlin (1997) proposed the Maximal Dependence Decomposition (MDD) model, which has binary tree structure formed by a set of conditional WAM models. While MDD model can capture non-adjacent dependencies through the conditional WAM models, it still requires a rather large number of training sequences, which are partitioned into smaller sub-sets to train all conditional WAM models. To alleviate the requirement of a large training set, Cai et al. (2000) developed a Bayesian tree to model dependencies within RNA splicing sites; Ellrott et al. (2002) suggested a position-order optimized Markov chain model, which reorders motif positions to bring long-ranging dependent positions into the near neighbors. More recently, several other models have been developed, including Bayesian networks for modeling protein-DNA binding sites (Barash et al., 2003), Maximum Entropy Model (MEM) for splicing site identification (Yeo and Burge, 2003), Permuted Variable Length Markov Model (PVLMM) for finding TFBS and splicing sites (Zhao et al., 2004). While these models proved to perform well for motif identification in some specific cases, it seems that there is still some room for improvement for motif model as well as for the corresponding software implementation.

An equally important but less-well addressed issue is model selection for motif identifica-

tion. Choosing the best model is to make a trade-off between advantages and disadvantages of succinct and more expressive models. Succinct models, such as PWM, can be robustly trained from a few examples, but are not able to account for correlation between positions; while expressive models, such as Bayesian network models, can account for most dependency information but, as mentioned above, could involve the dramatically increased number of parameters and hence require substantially large training data that may not be available at present. Also, expressive models are more susceptible to over-fitting problem, which leads to inferior power in identifying new motif sites. When using a Bayesian model, the over-fitting problem is less likely to occur if using an appropriate prior distribution. However, the choice of a sensible prior distribution is also a non-trivial problem (Zhou and Liu, 2004) for a complex Bayesian network model.

In this paper, we introduce the Optimized Mixture of Markov models (OMiMa) for all kinds of biological motif identification. We describe the Directed Neighbor-Joining (DNJ) method, an efficient method for Markov chain optimization. We then describe and discuss the methods for our mixture model selection. Finally, we present users with our OMiMa system, which efficiently implements the mixture model and other methods reported in this paper, and is freely available to public.

Methods

Mixed Markov models

Let X_i be the discrete random variable associated with position i in a biological motif X of length w. For DNA sequences, X_i takes values from set $\chi = \{A, C, G, T\}$; and for protein sequences, X_i takes values from 20 different amino acids. Naturally, X_i follows a multinomial distribution. Let $X_i^k = X_{i-k} \dots X_{i-1}$ and $x_i^k = x_{i-k} \dots x_{i-1}$, where upper case X (X_i) is a random variable and lower case x (x_i) is a particular value. Also let $X_j = X_{w+j}$ and $x_j = x_{w+j}$ if $j \leq 0$. If one uses a k-order Markov model(M_k), the probability of observing

a motif sequence x is just the product of conditional/transition probabilities. Let M_k^L be a k^{th}-order Markov model of a linear chain, and M_k^C be a k^{th}-order Markov model of a circular chain. The probability of a motif sequence is given by equation (3.1) for a linear chain and equation (3.2) for a circular chain, respectively.

$$\Pr(x|M_k^L) = \Pr(X_1 = x_1) \prod_{i=k+1}^{w} \Pr(X_i = x_i|X_i^k = x_i^k) \tag{3.1}$$

$$\Pr(x|M_k^C) = \prod_{i=1}^{w} \Pr(X_i = x_i|X_i^k = x_i^k) \tag{3.2}$$

Compared to a linear Markov chain, a circular Markov chain incorporates additional dependency information, which could be a crucial signal for distinguishing false and true motifs, especially when false motifs are very similar to true motifs. Both equation (3.1) and (3.2) can be expressed in terms of probabilities of oligomers of motif bases, therefore can be easily calculated (see Appendix).

Suppose a motif X can be divided into m independent sub-motifs, that is

$$X = X1, \cdots, Xm$$

and each sub-motif is modeled as an independent Markov chain, that is $M = M1, \cdots, Mm$, then probability of a motif sequence $x = x1, \cdots, xm$ is the product of probabilities of sub-motifs:

$$\Pr(x|M) = \prod_{j=1}^{m} \Pr(Xj = xj|Mj) \tag{3.3}$$

These independent Markov models, each of which is position-optimized for its corresponding sub-motif, form an Optimized Mixture of Markov models (OMiMa). An example of OMiMa is illustrated in Figure 3.1. For short motifs such as transcription factor binding sites, we use a mixture of Markov models consisting of only a zero order chain and a higher order chain

(Figure 3.2). For convenience, we use term 'a 0-k mixture model' or 'mixed 0-k Markov models' to refer to a mixture of Markov models consisting of only a zero order chain and a k^{th} order chain. If the k^{th} order chain is linear, the model is also referred as a 0-k mixed linear model, and if it is circular, the model is referred as a 0-k mixed circular model.

As we know, the different parts of a motif could have distinct roles in the interaction with their partners. Motif positions involved in the same role usually are highly dependent on each other as a module, while those playing different roles could be quite independent of each other. A mixture of Markov models allows different models to fit different signals of a motif. A simple zero order Markov chain can effectively model some strong signals such as highly conserved positions in which the probability of a certain base occurring is almost one. Also noise, such as positions whose bases themselves provide no biological functions hence their compositions are not different from background, does not need a more complex model than a zero order Markov model. On the other hand, a higher order Markov model is better than a zero order Markov model in detecting subtle but important signals such as long-range dependencies.

Motif dissection and Markov chain optimization

Motif dissection

To apply the mixture of Markov models to a motif, the first step is to dissect a motif into several independent sub-motifs, each of which is modeled as a Markov chain. In motif dissection, we employ chi-square tests to find significant pairwise dependencies between positions. According to pairwise dependencies, motif positions are grouped into independent sets, each of which forms a Markov chain. The outline of our procedure for grouping motif positions is described in the following steps.

1. Calculate base frequencies for each position, and find perfectly conserved positions where the frequency of a certain base is 1. These conserved positions then are put into

set H as defined below.

$$H = \{\, i : \max_{x \in \chi} f(i,x) = 1\}$$

where $f(i,x)$ is the frequency of base x at position i, and χ is the bases set of motif sequences, for DNA sequences $\chi = \{\ A, C, G, T\ \}$.

2. Put remaining positions in another set M, and calculate pairwise chi-square values (χ^2) for every pair of positions in M.

$$\chi^2_{i,j} = \sum_{x_i \in \chi_i} \sum_{x_j \in \chi_j} \frac{(O(x_i,x_j) - E(x_i,x_j))^2}{E(x_i,x_j)} \tag{3.4}$$

where χ_i and χ_j are the sets of bases observed in position i and j, respectively; $O(x_i,x_j)$ and $E(x_i,x_j)$ are the observed and expected counts of pair (x_i,x_j), respectively. The degree of freedom of test is $(|\chi_i|-1) \times (|\chi_j|-1)$, where $|\chi_i|$ and $|\chi_j|$ are the number of different bases in sets χ_i and χ_j, respectively.

3. According to the above χ^2 tests, find all positions that are not significantly dependent on any other positions in M, and move them to the set I, as defined by

$$I = \left\{ i : \min_{i \neq j; i,j \in M} p_{i,j} > 0.05 \right\}$$

Here $p_{i,j}$ is the p-value of $\chi^2_{i,j}$ test.

4. The remaining positions in M are further grouped into subsets according to the following rules:

   (a) Calculate $D_i = \sum_{i,j \in M, i \neq j} I(p_{i,j} < 0.05)$ for each position i in M. Find the largest D_i, and move position i and all positions j that $p_{i,j} < 0.05$ from M into a new set C_{D_i}.

   (b) For each remaining position, check if it significantly depends on any position in

C_{D_i}. If it does, then move it from M into C_{D_i}.

(c) If M is not empty, go back to step (a)

Markov chain optimization

The next step is to arrange the positions in each independent set into a Markov chain. Since the positions in sets H and I are independent of each other, they can be arranged in their natural order to form a zero order Markov chain (the positions in H can be treated differently from those in set I for motif identification by requiring a perfect match for a true site). However, this is not the case for sets C_{D_i}, whose positions need to be arranged in a particular order, so that a Markov model can account for most dependencies. For a given set C_{D_i}, we use the median (K_{di}) of D ($D = \{D_j, j \in C_{D_i}\}$) as the maximum order of its potential Markov model. We then optimize position arrangement for the k^th-order Markov chain ($k = 0, \ldots, K_{di}$) by the Directed Neighbor-Joining (DNJ) method described below. A series of Markov models whose orders range from 0 to K_{di} are fitted for each of K_{di} different chains. The best Markov model then is selected according to the criteria described in Model Selection.

The neighbor-joining (NJ) method proposed by Saitou and Nei (1987) is a well-known distance method for phylogenetic tree reconstruction. The principle of the NJ method is to find pairs of operational taxonomic units that minimize the total branch length at each stage of clustering. Our Directed Neighbor-Joining (DNJ) method is based on the exactly same principle. The only major difference is that DNJ needs to consider the direction in joining two nearest neighbors to form a new node while NJ does not. Instead of producing a phylogenetic tree as the NJ method does, DNJ method creates a chain structure, which arranges closely dependent positions in the nearest neighbors, The DNJ method for constructing a K^{th}-order Markov chain from a given subset (C_{D_i}) of motif positions is described in the following steps.

1. For a given set C_{D_i}, put each position in the set into a different vector. Here a vector is represented by a letter, a harpoon at the top of the letter may be used to indicate the direction of a vector, e.g. a stands for either $\vec{a}$ or $\overleftarrow{a}$. If $\vec{a} = (1, 2, 3)$, then $\overleftarrow{a} = (3, 2, 1)$, $\vec{a}\overleftarrow{a} = (1, 2, 3, 3, 2, 1)$, and $\vec{a}\vec{a} = (1, 2, 3, 1, 2, 3)$. Initially, there is only one position per vector.

2. Create an initial distance matrix whose elements are $d(u, v) = p_{i,j}$, where i is the position in vector u, j is the position in vector v, and $p_{i,j}$ is the p-value of chi-square test for the hypothesis that positions i and j are independent.

3. Calculate transformed distance $D(u, v)$ by equation (3.5). The transformed distance matrix is called matrix T.

$$D(u, v) = d(u, v) - \frac{(r_u + r_v)}{(N - 2)} \qquad \text{where} \quad r_u = \sum_{u \neq z, u, z \in C_{D_i}} d(u, z) \tag{3.5}$$

4. Find the minimum $D(u, v)$ in T. Then a new vector x is formed by joining vector u and v according to the Algorithm 1 on page 88 for a K^{th}-order Markov chain.

5. Calculate a new distance matrix by replacing u and v by x. The distance of x to each (y) of the other remaining vectors is defined by equation (3.6).

$$d(x, y) = (d(u, y) + d(v, y) - d(u, v))/2 \tag{3.6}$$

6. Go back to step 4 if the number of vectors in C_{D_i} is larger than 2, otherwise join the last two vectors according to the Algorithm 1 on page 88.

7. The order of positions in the final vector is the optimized linear chain for Markov model. If the first position is joined to the last position in the vector, it forms a circular chain.

A linear chain could be further optimized by forming a circular chain first from the final vector, then break the circular chain between positions with the weakest dependency, e.g. between position i and j where $p_{i,j}$ is the largest or the log-likelihood of the corresponding linear chain model is maximized. DNJ not only optimizes position order for linear chain models, but also makes circular chain models better, particularly when the order of Markov model is low, e.g. the first or second order Markov models.

Model selection

Many different mixtures of Markov models can be formed from the combination of different Markov chains. It is essential to choose the best model that could give us the least prediction error. In model selection, we first fit each model using maximum likelihood smoothed by a Dirichlet prior, then compute either the Akaike information criterion (AIC) (Akaike, 1974) or the Bayesian information criterion (BIC) (Schwarz, 1978). The model with the minimum value of AIC or BIC is selected as the potential best model. In model selection, minimizing AIC is the same as choosing the model with the minimum prediction error or loss, while minimizing BIC is equivalent to choosing the model with the largest posterior probability. Nonetheless, AIC and BIC have a similar form in calculation as shown in formula (3.7).

$$-2 \cdot loglik + \lambda \cdot DF \tag{3.7}$$

where $\lambda = 2$ for AIC and $\lambda = \log(N)$ for BIC (N is the number of motifs to fit the model); DF is the degree of freedom of a model. We replace DF with the effective degree of freedom (EDF) in calculating AIC or BIC of the mixture of Markov models, which enables an appropriate model to be selected (see sub-section *Effective number of free parameters* for explanation and calculation of EDF). There is no clear better choice between AIC and BIC for model selection. AIC tends to choose more complex model as $N \to \infty$, while BIC tends to choose a model too simple when N is small. In our test on 61 different regulatory motif

Algorithm 1: Algorithm for joining the two nearest nodes in the directed neighbor-joining method for K^{th} order Markov chain

`// Start with two nearest neighbors represented by vectors` $\vec{u}$ `and` $\vec{v}$`, respectively. The harpoon at top of` u `and` v `indicate the order of positions in` u `and` v`. For example, if` $\vec{u} = (2, 1, 6, 5)$ `and` $\vec{v} = (7, 3, 4, 8)$`, then` $\overleftarrow{u} = (5, 6, 1, 2)$`,` $\overleftarrow{v} = (8, 4, 3, 7)$`, and` $\vec{u}\overleftarrow{v} = (2, 1, 6, 5, 8, 4, 3, 7)$`.` $\vec{u}_i$ `denotes` i^{th} `position in vector` $\vec{u}$`.`

Data: $\vec{u}$ and $\vec{v}$

M = number of positions in $\vec{u}$;

N = number of positions in $\vec{v}$;

if $M < N$ **then**

  Swap$(\vec{u}, \vec{v})$;

  Swap(M, N) ;

endif

if $M == 1$ *and* $N == 1$ **then**

  `// Directly join` $\vec{u}$ `and` $\vec{v}$ `to form a new node`

  $\vec{x} = \vec{u}\vec{v}$;

else if $N == 1$ *and* $M <= K$ **then**

  $m = \text{int}(M/2)$;

  `// The distance of` $\vec{v}$ `to left side of` $\vec{u}$

  $L_{Left} = \sum_{i=1}^{m} d(\vec{u}_i, \vec{v}_1)$;

  `// The distance of` $\vec{v}$ `to right side of` $\vec{u}$

  $L_{Right} = \sum_{i=m+1}^{M} d(\vec{u}_i, \vec{v}_1)$;

  $$\vec{x} = \begin{cases} \vec{v}\vec{u} & \text{if } D_{Left} < D_{Right} \\ \vec{u}\vec{v} & \text{if } D_{Left} >= D_{Right} \end{cases};$$

else

  $m = \min(M, K)$ and $n = \min(N, K)$;

  $L_{\overleftarrow{v},\vec{u}} = \sum_{j=1}^{n} \sum_{i=j}^{m} d(\vec{u}_i, \vec{v}_j)$;

  $L_{\vec{u},\vec{v}} = \sum_{j=1}^{n} \sum_{i=j}^{m} d(\vec{u}_{M-i+1}, \vec{v}_j)$;

  $L_{\vec{v},\vec{u}} = \sum_{j=1}^{n} \sum_{i=j}^{m} d(\vec{u}_i, \vec{v}_{N-j+1})$;

  $L_{\vec{u},\overleftarrow{v}} = \sum_{j=1}^{n} \sum_{i=j}^{m} d(\vec{u}_{M-i+1}, \vec{v}_{N-j+1})$;

  $L_{min} = \min\left(L_{\overleftarrow{v},\vec{u}}, L_{\vec{u},\vec{v}}, L_{\vec{v},\vec{u}}, L_{\vec{u},\overleftarrow{v}}\right)$;

  $$\vec{x} = \begin{cases} \overleftarrow{v}\vec{u} & \text{if } L_{\overleftarrow{v},\vec{u}} = L_{min} \\ \vec{u}\vec{v} & \text{if } L_{\vec{u},\vec{v}} = L_{min} \\ \vec{v}\vec{u} & \text{if } L_{\vec{v},\vec{u}} = L_{min} \\ \vec{u}\overleftarrow{v} & \text{if } L_{\vec{u},\overleftarrow{v}} = L_{min} \end{cases};$$

endif

return $\vec{x}$

data sets, of which sizes range from 20 to 130, we found AIC was better than BIC to pick an appropriate model. However, for a data set of large size, we suggest that BIC, instead of AIC, should be used in model selection.

Cross validation

In K-fold cross validation, we first randomly split motif data into K roughly equal-sized parts, then fit a model with the data of $k-1$ parts, and calculate prediction error of the fitted model for the k^{th} part data. Let M_{-k} be the model fitted with data set of which the k^{th} part is removed, and p_{-k} be the cutoff threshold for the model M_{-k}. Also let E_k be the prediction error when using model M_{-k} to predict the k^{th} part data. For model selection, we choose the model with the smallest E_k as the final model. The prediction error of M_{-k} for a sample (training or testing data) of size m is given by equation (3.8).

$$\mathrm{E_k} = \frac{1}{m}\sum_{i=1}^{m}\delta(\mathrm{S}(x_i|M_{-k}) > s_{-k}) \tag{3.8}$$

where δ is a $0-1$ loss function, which equals 1 when $\mathrm{S}(x_i|M_{-k})$, the score of sequence x_i given model M_{-k}, is larger than a certain threshold s_{-k}, and 0 otherwise.

Effective number of free parameters

Let χ be the base set of biological sequences, e.g. for DNA sequences $\chi = \{A, C, G, T\}$ (In this paper, we use $|\chi|$ to denote the number of different bases in χ). For a motif of length w, the number of free parameters for a k-order Markov model is $(|\chi|^k - 1) \times (w - k)$ for a linear Markov chain; and $(|\chi|^k - 1) \times w$ for a circular chain model. That is, the number of parameters increases exponentially as the order of Markov chain increases. Using either AIC or BIC criterion for model selection, a simpler model can always be selected because of the exponential increase in number of parameters, especially when $|\chi|$ is large, e.g. for protein sequences, $|\chi| = 20$. Tested on 61 human regulatory motifs from Transfac database

(Wingender et al., 2000), we found that both AIC and BIC selected the zero order Markov models for all 61 DNA regulatory motifs when using the number of free parameters defined above. To avoid such a problem, we use the effective number of free parameters described below to calculate AIC and BIC for our model selection.

Generally, only a subset of bases from χ appears in a particular position of a set of biological motifs. The more conserved a position, the fewer bases are in the subset. The actual number of free parameters for a model is related to the observed bases in a training data set. For example, suppose that one would like to estimate nucleotide frequencies occurred in a position of a set of DNA training motifs. If only a base `A` is observed in the position, then one need estimate only the frequency of `A`, the remaining parameters, i.e. the frequencies of `C,G,T`, can be derived. Therefore, the actual number of free parameters is one in this case. For our mixture of Markov models, the effective number of parameters is defined as the number of parameters that are direct estimates of the observed bases in a training motif set. Let b_i be the base set observed in a position i of a training set of motifs, w be motif length, h_k be the sequence of motif positions in the k-order Markov chain of a mixture model and $\sum |h_k| = w$ ($|h_k|$ is the number of positions in h_k), then the effective number of free parameters or effective degree of freedom (EDF) for a mixture of Markov models is calculated by

$$\begin{aligned} \text{EDF} = & \sum_{i=1}^{|h_0|} \max(b_{h_0[i]} - 1, 1) + \sum_{i=2}^{|h_1|} \max(b_{h_1[i-1]} \times b_{h_1[i]} - 1, 1) \\ & + \cdots + \sum_{i=k+1}^{|h_k|} \max(b_{h_k[i-k]} \times \cdots \times b_{h_k[i]} - 1, 1) \end{aligned} \tag{3.9}$$

The above formula assumes each chain in the mixture of Markov models is linear. If a

k^{th}-order chain is circular, the effective number of free parameters of the chain is

$$\sum_{i=1}^{|h_k|} \max(b_{h_k[i-k]} \times \cdots \times b_{h_k[i]} - 1, 1)$$

where $h_k[i-k] = h_k[|h_k| - i - k]$ if $i - k \leq 0$.

Results

We used real motif sequence data to test the effectiveness of our DNJ method for Markov chain optimization, and assessed the capability of our optimized mixture of Markov models for motif identification. For prediction results, we use the following abbreviations for empirical quantities: TP (# true positives), TN (# true negatives), FP (# false positives), FN (# false negatives), Ac (Accuracy), Sn (sensitivity), Sp (specificity), and Mc (Matthews correlation coefficient). Sn, Sp, Ac, and Mc are defined by

$$\begin{aligned} Sn &= \frac{TP}{TP + TN} \\ Sp &= \frac{TN}{TN + FP} \\ Ac &= \frac{TP + TN}{TP + FP + TN + FN} \\ Mc &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \end{aligned}$$

Sensitivity is the percentage of the correctly predicted real sites among all real sites, and specificity is the percentage of the correctly predicted false sites among all false sites. Accuracy is the percentage of correctly predicted sites among all sites. Matthews correlation coefficient (Matthews, 1975), also called Phi (correlation) coefficient, has a value between -1 and 1, with 1 indicating a perfect prediction, 0 indicating a random prediction, values below 0 indicating a worse than random prediction.

OMiMa can use two different ways to score a motif site x: log-likelihood score and

log-likelihood ratio score, which are defined by

$$\text{log-likelihood} = \log \Pr(x|M_s) \qquad \text{log-likelihood ratio} = \log \frac{\Pr(x|M_s)}{\Pr(x|M_b)}$$

where M_s is the signal model that is fitted with true motif sites, and M_b is the background model or false signal model which has the exactly same structure as M_s but its parameters are fitted with background sequences or false motif sites. A sequence x is predicted as a positive site if the score of x is larger than a certain threshold. Depending on what kind of criterion is used, the best threshold for prediction can be different. Results presented here are based on the following three criteria: balanced sensitivity and specificity, the maximum prediction accuracy, and the maximum Matthews correlation coefficient. For each potential threshold, there is an associated true positive rate and a false positive rate. The plot of true positive rates against false positive rates generates a Receiver Operating Characteristic (ROC) curve, which is used for comparing and selecting the best model.

To distinguish a variety of different models and the score methods used by the models, we used the following two notations to refer to a model and its score method. The first notation uses two digits and a letter with format 'digit-letter-digit', where the first digit is k for a mixture of 0-k Markov models, the letter in the middle is either 'L' or 'C' to indicate whether the k^{th} order chain is linear ('L') or circular ('C'), and the second digit is either 0 or 1 to indicate whether log likelihood score (0) or log-likelihood ratio score (1) is used. For example, '1-L-1' stands for a 0-1 mixture of linear Markov models that uses log-likelihood ratio to score a motif site. The second notation, which has format 'digit-digit-letter-digit', is the extension of the first notation with an additional digit adding to the first notation as a prefix. The additional digit indicates which part of testing data is used in k-fold cross-validation, and has value between 0 to k-1.

Effectiveness of DNJ method for optimization

To assess the ability of our DNJ method for optimizing a Markov chain, we compared the effectiveness of DNJ method with that of random permutation method for optimizing Markov chains. In this evaluation, we used a 0-k mixture model (where k takes an integer value from 0 to 2) (Figure 3.2) to model transcription factor binding sites (TFBS) from Transfac database (Wingender et al., 2000). For each TFBS, we first fitted a 0-k mixture model (denoted as M_{DNJ}), of which the k^{th} order Markov chain is optimized by our DNJ method and calculated the log-likelihood of the data given the model M_{DNJ} ($\log \Pr(data|M_{DNJ})$). Second, we fitted a new 0-k mixture model (denoted as M_R), which is same as M_{DNJ} except that the position in its k^{th} order chain is ordered by random permutation, with the same data and calculated $\log \Pr(data|M_R)$. This step was repeated 1,000 times, so we have 1,000 log-likelihoods of the randomly permuted models ($M_{R_1}, \ldots, M_{R_{1000}}$). We then calculated the rank (Rk) of the DNJ optimized model in the 1,000 randomly permuted models as follows:

$$Rk = 1 - \frac{\sum_{i=1}^{1000} \mathrm{I}(\log \Pr(data|M_{DNJ}) < \log \Pr(data|M_{R_i}))}{1000} \tag{3.10}$$

where I is an indicator function with value 1 if condition is true, and 0 if condition is false. So the higher the rank, the better the DNJ optimization is. Also the value of $(1 - Rk)$ approximates the probability of observing $\log \Pr(data|M_{R_i})$ larger than $\log \Pr(data|M_{DNJ})$.

Fifty-three human transcription factors, of which binding site contains at least 4 dependent positions by the χ^2 test given by equation (3.4), are used for this evaluation (Table 3.1). The assessment were performed for four mixture models, of which the only differences are the higher order chains of the mixture models . The higher order chains of the four mixture models are 1^{st} order linear chain model, 1^{st} order circular chain model, 2^{nd} order linear chain model, and 2^{nd} order circular chain model, respectively.

Results suggest that DNJ method performs remarkably well in optimizing the first order

linear Markov chains, as shown in figure 3.3, that most ranks are 1s or close to 1. The optimization for the second order linear chains is not as good as that for the first order linear chains. This could partly due to DNJ method only relies on the pairwise dependencies between two single positions. Nevertheless, it still reasonably good, as shown in figure 3.4 that most ranks are close to 1. DNJ method is designed for optimizing linear Markov chains, but results suggest that it works well in optimizing the first order circular Markov chains (Figure 3.5). In optimizing the second order circular Markov chains, DNJ method does not perform as well though, as we can see from figure 3.6 that quite a few ranks are smaller than 0.5. However, the result should be no surprising considering DNJ method is for linear chain optimization and uses only simple pairwise dependent information. Also the number of motif sites in the training data set and the dependent structure of motif positions can affect the performance of DNJ method. Although we have not performed a statistical test, it appears that the larger the number of sites and the more distinct the differences among pairwise dependencies, the better DNJ performs for optimization.

For each of the mixture models, we demonstrated, through an example of transcription factor V$AP1_Q4_01, how significant the improvement of such a mixture model can be made by DNJ optimization. We plotted the histogram of the log-likelihood per instance given a model M_{R_i}, $\log \Pr(data|M_{R_i})/N$, where N is the number of binding sites in the data set, and $i = 1, \ldots, 1000$ for 1,000 mixture models of random permuted Markov chains. The histogram can be treated as a simulated Non-distribution of log-likelihood per instance given a mixture model. Then we mapped the location of the likelihood per instance given DNJ optimized model, $\log \Pr(data|M_{DNJ})/N$, in the histogram. For the transcription factor V$AP1_Q4_01 (Figures 3.7–3.10) and many others, we found that for the 0-1 mixture model of either linear or circular structure, DNJ optimized models are better than any models from $1,000$ random permutations (Figures 3.7 and 3.9).

Theoretically, the optimal model can be found by searching exhaustively all possible

models. However, this is not always possible in practice due to very large searching space. In this case, the number of possible Markov chains is the factorial of the length of the Markov chain and dramatically increases as the length of chain increases. For example, for a linear chain of length 5, the number of possible models to fit is $5! = 120$, which can be done quickly. However, for a chain of length 10 ($10! = 3,628,800$), it will take a lot of computational time to fit all possible models , and for length as short as 15 ($15! = 1.307674e + 12$), its computational time becomes practically unacceptable. So it is not surprising that for the transcription factor V$AP1_Q4_01, which has 8 ($8! = 40,320$) dependent positions, the optimal model may not be found from $1,000$ random permutations as shown in the above examples (Figures 3.7 and 3.9). Approximation algorithms such as simulated annealing (Kirkpatrick et al., 1983) and genetic algorithm (Koza, 1998) could be used to reduce computational time for this NP-hard problem though no guarantee to find the optimal model.

Splicing site recognition

Data collection

The collection of human RNA splicing sites, which is from http://www.fruitfly.org/seq_tools/datasets/Human/GENIE_96/splicesets/, is used by the GENIE system to detect splicing signal. The data sets consist of real and fake 'GT-AG' canonical donor and acceptor sites. A donor site is of length 15 bases from -7 to +8 around the conserved 'GT' dinucleotide. An acceptor site is of length 90 bases, with 70 bases in the intron (ending with AG) and 20 bases in the following exon. In total, there are 6246 donor sites, of which 1324 are real, 4922 are false, and 6877 acceptor sites, of which 1324 are real and 5553 are false. The data sets, which are divided into training and testing data, with about $\frac{5}{6}$ of them for training and remaining for testing, were used to assess performance of NNSplice (Reese et al., 1997). NNSplice is a splicing site predictor based on neural network model (http://www.fruitfly.org/seq_tools/splice.html)

Model selection

OMiMa uses a 0-k mixture model, of which k^{th} order chain is selected by either AIC or BIC criterion, for motif identification. To test whether OMiMa can pick the correct model, we fitted a set of 0-k mixture models, of which the k^{th} order chains are either linear or circular and k ranges from 0 to 3, with the above training data, and used the fitted models to predict splice sites in testing data. The performances of different models were compared and evaluated by ROC curves. Results are shown in figures 3.11–3.14. We also calculated the maximum accuracy (Ac) and the maximum Matthews correlation efficient (Mc) archived by each model (data not shown here). Overall, for both donor and acceptor sites, the 0-3 mixture models perform best for training data, as we can see from the ROC curves. However, they are over fitted and perform worst for testing data. When using the log-likelihood score, the best models for both donor and acceptor sites are the 0-1 mixture models. When applying the log-likelihood ratio score, the selected models are different: the best model for donor site is still the 0-1 mixture model (Figure 3.12) while for acceptor site is the 0-0 mixture model, which is just a zero order Markov model (Figure 3.14). For all 0-1 mixture models, the performance is much better when a log likelihood ratio score is used, while there is not much difference between a linear chain model and the corresponding circular chain model. The results of model selection by ROC curves are consistent with those by OMiMa itself. For both donor and acceptor sites, OMiMa, which selects signal model (M_s) based on log-likelihood scores, correctly picks the 0-1 mixture models as the best models according to AIC criterion. The selected models were further confirmed by a six-fold cross validation, of which part of the results are shown in figures 3.15 and 3.16.

Performance comparison with NNSplice

To test if the performance of OMiMa is comparable to the existing leading splicing site predictors, we compared OMiMa with NNSplice, one of the top available splicing site pre-

dictors. Since OMiMa is trained and tested with the exactly same data used by NNSplice, their prediction results can be directly compared. NNSplice is based on the more complex neural network model, and trained by both true sites and false sites. In the comparison, we compared only the prediction results of donor site because the acceptor sites apparently are not exactly the same data used to train and test NNSplice as described in the authors' orginal paper (Reese et al., 1997). The results of OMiMa are reported by the best model (1-L-1) selected by AIC criterion, and those of NNSplice are reported at the NNSplice Web site at http://www.fruitfly.org/seq_tools/splice.html. The comparison results are shown in table 3.2, which also includes the results from a competing model 1-C-1. We found that both 1-L-1 and 1-C-1 models perform comparably with NNSplice neural network model with regard to prediction accuracy. However, when considering computation time, OMiMa is much more efficient than NNSplice. In this case, OMiMa took only a couple of seconds to train and test a mixture of Markov models while NNSplice need longer time (about 25 seconds) to report the testing results. Also, NNSplice limits the total length of input sequence to 10 *kb* while OMiMa has no discernible limit for sequence length. Another distinct difference between NNSplice's neural network model and OMiMa's 0-1 mixture model is the shapes of their score distributions (Figures 3.17 and 3.19). The 0-1 mixture model appears more natural to separate true sites from false sites than the neural network model does.

Comparison of mixture models with standard Markov models

Does a mixture of 0-k Markov models perform better than a zero order Markov model, which is the same as PWM model, or a k^{th} order model? To answer the question, we compared performance of 0-1 mixture models with both the zero order and the first order Markov models for donor site. The optimized position arrangements for the 0-1 Markov model of donor site are

zero order chain: 7-8

first order chain: 5-12-6-10-11-4-0-1-2-3-14-13-9

A six-fold cross validation was used in this assessment. Also, since there are two different scoring functions: log-likelihood and log-likelihood ratio, the comparisons were made for each case separately. Results (Table 3.3) showed that when using the log-likelihood score, DNJ optimized 1-L-0 model performs better than either the zero order or the first order Markov model. However, when the log-likelihood ratio score is used (Table 3.4), the first order Markov model performs slightly better than the 0-1 mixture model (1-L-1). This can be explained by two main reasons: first, the false sites in training data are not random sequences, and positions within a false site are not totally independent of each other, therefore the first order Markov chain can be better than the mixed 0-1 Markov chains to model false sites; second, DNJ method does not consider background sequence or false sites when optimizing 0-1 mixture model.

Identification of transcription factor binding sites

In this subsection, we assessed DNJ optimized 0-1 mixture model by comparing it with the full first order Markov model. Normally, the mixture model consists of a zero order and a first order Markov chains, both of which lengths are larger than zero. In this case, the 0-1 mixture model can be treated as the reduced model (though not in perfect sense) compared to the full first order Markov model. Therefore we can test the hypothesis that the reduced model performs at least as well as the full model, hence is adequate for TFBS prediction. For a special case, a 0-1 mixture model can be reduced to the first order Markov model when there are no independent positions, a quite common case for a transcription factor binding site. The special case allows us to investigate the effects of DNJ optimization on performance without being confounded by the factor of different model structures when directly comparing DNJ optimized first order model with un-optimized first order model. Here we used the transcription factor binding sites from Transfac database to assess performance of DNJ optimized 0-1 mixture model for both cases by 10-fold cross validation. In the following

assessment, the true binding sites are directly extracted from Transfac matrix records, and false sites are simulated from the uniform distribution of four nucleotides (the simulated sites matched to true sites are removed). The number of false sites simulated is exactly 50 times the number of true sites. Both true sites and false sites are randomly divided into 10 roughly equal-sized parts for 10-fold cross validation.

Reduced *vs* full model

For the normal case, we used the transcription factor V$ATF_01 as an example to demonstrate that the 0-1 mixture model performs better than the full first order Markov model. For V$ATF_01, there are only 25 known binding sites from Transfac database. According to our pairwise χ^2 tests (see equation 3.4), V$ATF_01 binding sites have only 8 dependent positions among total 14 positions. The 0-1 mixture model, which is chosen as the best model by OMiMa (AIC criterion), has two balanced Markov chains whose position orders are given as

zero order chain: 3-4-5-6-7-8

first order chain: 1-0-10-9-11-2-13-12

Therefore the 0-1 mixture model of V$ATF_01 forms a legitimate reduced model to the full first order Markov model. The motif positions of the full first order Markov chain are in the natural order. Since a non-zero order Markov chain can be either linear or circular, we compared the reduced model with the full model for both cases. Results are shown in table 3.6. We found that the 0-1 mixture model performed as well as or slightly better than the full first order Markov model. Therefore, we accepted the hypothesis that the 0-1 mixture model as the reduced model is adequate compared to the full first order Markov model.

Optimized *vs* unoptimized model

For the special case, we used the transcription factor V$AP1_Q4_01 as an example to investigate whether DNJ optimization improves the performance of the first order Markov model. By our χ^2 test, there is no independent position within a binding site V$AP1_Q4_01, thus its 0-1 mixture Markov model, of which the length of the 0 order chain is zero, is reduced to the first order Markov model. Based on AIC criterion, the optimized first order Markov model is chosen by OMiMa as the best model. Here we compared two different first order Markov models: one with DNJ optimization, the other without and positions are in the natural order. The Markov chains of the two models have different arrangements of motif positions as given in the following:

with optimization:	7-3-1-2-0-6-5-4
without optimization:	0-1-2-3-4-5-6-7

Comparisons between the two models were made when both Markov chains are linear as well as when both chains are circular. Results in table 3.5 show that the model with DNJ optimization is better than the one without in both cases. The score distributions (Figures 3.20-3.21) also clearly indicate the optimized models are better than unoptimized models to separate true sites from false sites.

Detection of protein domains

OMiMa is a versatile motif prediction tool for biological sequences, including DNA, RNA and protein sequences. In this evaluation, we tested OMiMa's capability for motif identification in protein sequences. The functional motifs in protein sequences are also known as function domains that are usually well-conserved across different members in the same protein family. The data for this evaluation are the alignment seeds of Pfam protein domains (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Pfam-A.seed.gz). We used a domain named *"A Propetide"*

(accession # PF07966), of which the length is 29 bases. In total, there are 86 true domain sites, and for each true domain site, 50 false domain sites are simulated by randomly shuffling the amino acids of the true site. For the 0-1 mixture model, OMiMa optimized position arrangement for the first order chain is given by

```
5-3-15-7-10-12-26-9-28-13-1-24-21-16-20-0-2-14-11-6-25-17-27-8-22-18-23-4-19
```

Although no position is independent of all others for this domain moitf (the length of zero order chain is 0), the best model chosen by OMiMa (AIC criterion) is the zero order Markov model. We compared the following three Markov models: zero order, un-optimized first order, and DNJ optimized first order models by 10-fold cross validation. The log-likelihood ratio $\frac{\log \Pr(x|M_s)}{\log \Pr(x|M_b)}$ were used to score a domain site, and the first order Markov chains were set to linear. Results (not given here) show that both the zero order, and DNJ optimized Markov models have perfect prediction (both Sn and Sp are 1), while the un-optimized first order Markov model does not.

Discussion

The prediction accuracy of a probabilistic model is largely determined by the effectiveness of the model in characterizing a biological motif. Since the large variation of the signals embedded in a biological motif, an effective model can be as simple as a consensus sequence, or as complex as a fully connected network model. In this paper, we described a mixture of Markov models to allow adjustment of model complexity for different motifs. Also, we extended the traditional linear chain Markov model to the circular chain Markov model, which can better represent position dependencies within a motif in some cases. We presented a novel method to efficiently optimize position arrangement for a non-zero order Markov chain to effectively incorporate most dependency information into the Markov model. We then

explained how to efficiently train a mixture of Markov models, as well as how to calculate the effective number of free parameters of a mixture model to enable a correct model to be selected. We implemented these methods into our OMiMa system, a comprehensive tool for motif identification in biological sequences. Finally, we demonstrated from different aspects in several examples, that an optimized mixture of Markov models can improve prediction accuracy for motif identification in both DNA and protein sequences. We also showed that OMiMa can correctly select the best model, and is efficient in terms of computation time and memory usage.

The interaction of biological macromolecules, such as transcription factors bound to DNA sites, usually involves several contact positions, which depend on each other to form a functional entity. While some highly conserved or distinguishable motifs can be identified by a simple consensus or PWM, the accurate prediction of many biological motifs, especially those highly degenerated, relies on the models that can account for the subtle dependency signals embedded within a motif. Many methods including Markov chain, tree, neutral networks, and Bayesian networks have been used to model dependency structures of motif. Among these models, the Markov model is the simplest yet can be very powerful for motif prediction when it is optimized. In this paper, we demonstrated that Markov models can perform at least as well as the neural network models for splicing site prediction. The beauty of a Markov model is its simplicity and efficiency in computation, yet it is still able to account for most dependency information. Moreover, the effectiveness of a Markov model can be improved by optimizing position order in the Markov chains so that the closely related positions are brought to the nearest neighbors. Such an optimized Markov chain can incorporate both local and non-local dependencies into the model, which enables it to compete against tree or network models in predicting short biological motifs. In this paper, we demonstrated that optimized Markov model can be an excellent motif predictor.

An important issue in modeling biological models is the number of parameters of a model

or model complexity. The more complex a model, the more data are needed for training. However, for many biological motifs, the number of known (experimentally verified) sites is usually relative small in the currently available biological databases. This limits the usage of complex models, such as higher order Markov models, tree or tree-like model, and complicated network models, even though these models in some cases can perform better given enough training data. For a standard Markov model, the number of parameters grow exponentially with the order of the Markov chain. As a result, while a low order Markov model may perform poorly due to failing to incorporate potentially more distant dependencies, it is often impractical to train higher order Markov model, which can also easily be over fitted. However, these problems have been addressed by several approaches or methods. The first is to use a variable length Markov model (VLMM) (Rissanen, 1986; Bühlmann and Wyner, 1999), of which the context lengths [1] can vary among different positions. VLMM can effectively reduce Markov model complexity when the variation of actual context lengths is large. However, VLMM may still not be able to account for long-range dependencies. The second is to optimize Markov chain by arranging most correlated positions in the nearest neighbors (Ellrott et al., 2002). The optimization enables a Markov model to incorporate important distant dependencies without increasing Markov chain order. However, the effectiveness of this method largely depends on the optimization routine. The third is the permuted variable length Markov model (PVLMM) (Zhao et al., 2004), which tries to combine advantages of the first and second methods. The disadvantage of PVLMM is the number of possible permutation is the factorial of motif length, which makes it computationally expensive even using heuristic methods, e.g. simulated annealing, to search for optimal/sub-optimal solution when the length of a motif is moderate (e.g. 10-15 *bp*) or longer. The optimized mixture of Markov models we presented here tries to inherit advantages of these models while avoid

[1] Context lengths is defined as the number of previous positions $(X_{i-1} \cdots X_j)$ on which the current position X_i depends. In terms of a conditional probability, it can be expressed as $\Pr(X_i|X_{i-1} \cdots X_j)$, and context length is $i - j$.

their disadvantages. In OMiMa, we replace VLMM with a mixture of more homogeneous [2] Markov models. As we know, when there are close correlations between each neighboring pair of positions within a Markov chain, a low order Markov model can perform as well as the higher order chain. Although we have not tested it, we expect that a VLMM and a mixture of homogeneous Markov models have similar performance. Also, the optimization method we proposed for a Markov chain is different from that proposed by Ellrott et al. (2002). While the results suggest that our method can perform well for the low order Markov chains, we have not directly compared it with the method by Ellrott et al. (2002), of which software is not available.

AIC and BIC are the standard criteria for model selection. For a small training data set, AIC is generally better than BIC to pick the correct model. We found (results not shown here) that when using the number of all potential parameters of the mixture model, neither criteria work well to select the correct mixture of Markov models. However, when applying our effective number of free parameters defined in method section, AIC generally can pick the correct model in case of small or moderate size of training data. For large training data, BIC is better than AIC for model selection. Another important way for model selection is k-fold cross-validation, which is also used to validate the model selected by AIC, BIC or other criteria. In our evaluation of OMiMa, we found that the model selected by AIC or BIC is consistent with the model selected by cross-validation.

In conclusion, our optimized mixture of Markov models presents a comparable approach alternative to the existing methods to model dependent structures within a biological motif. The simplicity, effectiveness and computational efficiency of the optimized mixture of Markov models can give it some advantages over other competing models in a large scale motif prediction. The optimized mixture of Markov models is implemented in our computational tool OMiMa, which is able to use a variety of mixture models for motif prediction. OMiMa,

[2]Homogeneous means that the dependent structures among neighboring positions are similar. In the other words, the actual context lengths are similar for each position in a motif.

of which most parameters are configurable (see Appendix for OMiMa configuration), is freely available at http://raga.statgen.ncsu.edu/omima.

Tables

Table 3.1: The optimized arrangement of dependent positions within TFBS for the first order Markov model

No	Name	Len	#Dep	Pos. order
1	V$AP1_Q4_01	8	8	7-3-1-2-0-6-5-4
2	V$AP1_Q6_01	9	8	2-3-1-4-5-7-8-6
3	V$AP1_Q2_01	12	9	4-3-5-6-7-10-11-9-1
4	V$CDPCR1_01	10	9	3-4-2-9-6-5-7-8-1
5	V$ATF_01	14	8	1-0-10-9-11-2-13-12
6	V$CHOP_01	13	10	5-4-6-7-9-10-0-8-11-12
7	V$CDPCR3_01	15	10	3-0-1-8-9-13-4-6-2-5
8	V$CDPCR3HD_01	10	5	1-8-9-2-7
9	V$CREB_Q2_01	14	8	1-11-12-0-2-3-9-8
10	V$CREB_Q4_01	11	6	7-6-1-8-9-10
11	V$CREB_Q3	6	4	4-5-1-0
12	V$CEBP_Q3	12	9	8-9-5-6-4-11-3-2-10
13	V$CEBPB_01	14	4	0-13-11-3
14	V$E2F_Q4_01	11	4	1-8-7-0
15	V$E2F_Q6_01	12	8	8-3-7-0-2-11-9-10
16	V$E2F1DP1_01	8	5	3-4-0-6-7
17	V$E2F1DP2_01	8	5	5-6-7-3-4
18	V$E2F4DP1_01	8	4	3-4-0-1
19	V$E2F4DP2_01	8	5	4-3-7-1-0
20	V$ETS_Q4	12	8	11-2-5-10-4-3-0-1
21	V$ELK1_02	14	4	10-11-2-3
22	V$FAC1_01	14	12	12-6-10-11-13-4-9-8-5-1-0-7
23	V$FOXD3_01	12	11	1-3-8-7-9-10-11-2-0-4-6
24	V$FOXO1_02	14	11	8-9-10-12-7-6-2-0-11-1-3
25	V$HNF4_Q6	9	7	4-3-2-6-8-1-7
26	V$HNF1_Q6	18	15	3-11-12-1-4-8-13-5-9-0-6-16-14-2-10
27	V$HNF3_Q6	13	11	1-10-7-5-3-4-12-9-0-2-8
28	V$E2F1DP1RB_01	8	5	1-7-3-0-4
29	V$IRF7_01	18	13	3-2-0-16-15-17-1-7-6-8-14-9-12
30	V$LUN1_01	17	8	8-9-10-7-12-11-14-13
31	V$MZF1_01	8	4	0-1-4-5
32	V$MYC_Q2	7	4	4-5-3-1
33	V$NFAT_Q4_01	10	4	6-8-9-5
34	V$NFKAPPAB_01	10	4	5-7-9-2
35	V$NKX22_01	10	6	9-8-6-1-0-7
36	V$OCT_Q6	11	10	8-2-0-10-5-3-9-6-4-7
37	V$PAX_Q6	11	10	10-6-7-0-9-3-1-5-4-2
38	V$PAX6_01	21	21	15-17-16-18-6-8-19-13-11-3-2-1-0-20-7-10-4-9-5-14-12
39	V$PBX1_02	15	10	6-12-2-0-3-1-11-13-14-4
40	V$RSRFC4_Q2	17	6	6-7-0-13-2-3
41	V$RSRFC4_01	16	8	6-7-9-1-2-13-12-8
42	V$STAT5A_01	15	7	8-12-1-13-0-4-5
43	V$SOX9_B1	14	9	1-13-0-2-11-5-3-10-4

Continue at next page

Table 3.1 (continue)

No	Name	Len	#Dep	Pos. order
44	V$SRY_01	7	4	4-6-0-1
45	V$SRY_02	12	4	1-3-11-4
46	V$STAT5A_02	24	16	7-12-20-15-16-17-18-19-22-1-21-13-5-6-9-23
47	V$SP1_Q2_01	10	7	7-3-8-0-4-9-5
48	V$SP1_Q4_01	13	13	0-2-11-12-6-1-3-10-9-8-7-4-5
49	V$SP1_Q6_01	10	10	3-5-8-9-0-7-4-2-6-1
50	V$USF_Q6_01	12	8	3-11-4-5-7-2-1-8
51	V$XBP1_01	17	9	13-5-3-4-15-11-10-12-0
52	V$ZID_01	13	8	6-7-4-8-12-10-9-11
53	I$DRI_01	10	7	6-9-8-7-0-1-2

Table 3.2: Performance comparison for donor site prediction by NNSplice's neural network model, OMiMa's 1-L-1 and 1-C-1 mixed Markov models

		Network	1-L-1	1-C-1
Ac maximized	Ac	0.951	0.955	0.954
	Sn	0.904	0.928	0.947
	Sp	0.963	0.962	0.955
Mc maximized	Mc	0.857	0.869	0.869
	Sn	0.942	0.938	0.952
	Sp	0.951	0.959	0.954

Table 3.3: Comparison of three different Markov models: the zero order (0-L-0) model, the mixed zero-first order (1-L-0) model, and the first order model, for RNA donor site recognition, in which the score of a site is a log-likelihood ($\log \Pr(x|M_s)$).

	zero order		mixed zero-first order		first order	
	$\max(Ac)$	$\max(Mc)$	$\max(Ac)$	$\max(Mc)$	$\max(Ac)$	$\max(Mc)$
Part 0	0.935	0.811	0.932	0.800	0.940	0.823
Part 1	0.919	0.755	0.926	0.780	0.919	0.764
Part 2	0.936	0.806	0.947	0.844	0.937	0.827
Part 3	0.934	0.791	0.941	0.818	0.938	0.812
Part 4	0.929	0.790	0.931	0.794	0.920	0.770
Part 5	0.925	0.780	0.928	0.789	0.935	0.809

Table 3.4: Comparison of three different Markov models: the zero order (0-L-1) model, the mixed zero-first order (1-L-1) model, and the first order model, for RNA donor site recognition, in which the score of a site is a log-likelihood ratio ($\log \frac{\Pr(x|M_s)}{\Pr(x|M_b)}$).

	zero order		mixed zero-first order		first order	
	max(Ac)	max(Mc)	max(Ac)	max(Mc)	max(Ac)	max(Mc)
Part 0	0.946	0.842	0.955	0.869	0.959	0.882
Part 1	0.937	0.806	0.947	0.841	0.957	0.870
Part 2	0.949	0.846	0.962	0.890	0.958	0.877
Part 3	0.955	0.854	0.961	0.883	0.971	0.909
Part 4	0.936	0.817	0.947	0.843	0.946	0.847
Part 5	0.931	0.808	0.939	0.820	0.948	0.843

Table 3.5: Identification of V$AP1_Q4_01 binding sites Performance comparison of the first order Markov models with and without DNJ optimization by 10-fold cross validation. The prediction results are based on log-likelihood ratio scores of motif model and background model.

		without optimization			with optimization		
Chain	Data	Mc	Sn	Sp	Mc	Sn	Sp
Linear	part 0	0.855	0.818	0.998	0.907	0.909	0.998
	part 1	0.814	0.667	1.000	0.814	0.667	1.000
	part 2	0.915	0.917	0.998	0.915	0.917	0.998
	part 3	0.814	0.667	1.000	0.779	0.750	0.997
	part 4	0.814	0.667	1.000	0.864	0.750	1.000
	part 5	0.816	0.917	0.993	0.845	0.917	0.995
	part 6	0.878	0.917	0.997	0.924	1.000	0.997
	part 7	0.878	0.917	0.997	0.790	1.000	0.988
	part 8	0.779	0.750	0.997	0.816	0.917	0.993
	part 9	0.915	0.917	0.998	0.924	1.000	0.997
	average	0.848	0.815	0.998	0.858	0.883	0.996
Circular	part 0	0.855	0.818	0.998	0.907	0.909	0.998
	part 1	0.814	0.667	1.000	0.814	0.667	1.000
	part 2	0.957	0.917	1.000	0.957	0.917	1.000
	part 3	0.814	0.667	1.000	0.779	0.750	0.997
	part 4	0.818	0.750	0.998	0.864	0.750	1.000
	part 5	0.797	0.833	0.995	0.845	0.917	0.995
	part 6	0.878	0.917	0.997	0.915	0.917	0.998
	part 7	0.878	0.917	0.997	0.812	1.000	0.990
	part 8	0.779	0.750	0.997	0.816	0.917	0.993
	part 9	0.915	0.917	0.998	0.924	1.000	0.997
	average	0.851	0.815	0.998	0.863	0.874	0.997

Table 3.6: Identification of V$ATF_01 binding sites Performance comparison of the full model (the standard first order Markov model) and the reduced model (the 0-1 mixture model optimized by DNJ method) by 10-fold cross validation. The prediction results are based on log-likelihood ratio scores of motif model and background model.

		full model			reduced model		
Chain	Data	Mc	Sn	Sp	Mc	Sn	Sp
Linear	part 0	1.000	1.000	1.000	1.000	1.000	1.000
	part 1	0.814	0.667	1.000	0.863	1.000	0.993
	part 2	1.000	1.000	1.000	1.000	1.000	1.000
	part 3	0.863	1.000	0.993	1.000	1.000	1.000
	part 4	1.000	1.000	1.000	1.000	1.000	1.000
	part 5	0.574	0.333	1.000	1.000	1.000	1.000
	part 6	1.000	1.000	1.000	1.000	1.000	1.000
	part 7	1.000	1.000	1.000	1.000	1.000	1.000
	part 8	1.000	1.000	1.000	1.000	1.000	1.000
	part 9	1.000	1.000	1.000	1.000	1.000	1.000
	average	0.925	0.900	0.999	0.986	1.000	0.999
Circular	part 0	1.000	1.000	1.000	1.000	1.000	1.000
	part 1	0.814	0.667	1.000	0.863	1.000	0.993
	part 2	0.660	0.667	0.993	1.000	1.000	1.000
	part 3	0.863	1.000	0.993	1.000	1.000	1.000
	part 4	1.000	1.000	1.000	1.000	1.000	1.000
	part 5	0.574	0.333	1.000	1.000	1.000	1.000
	part 6	1.000	1.000	1.000	1.000	1.000	1.000
	part 7	1.000	1.000	1.000	1.000	1.000	1.000
	part 8	1.000	1.000	1.000	1.000	1.000	1.000
	part 9	1.000	1.000	1.000	1.000	1.000	1.000
	average	0.891	0.867	0.999	0.986	1.000	0.999

Figures

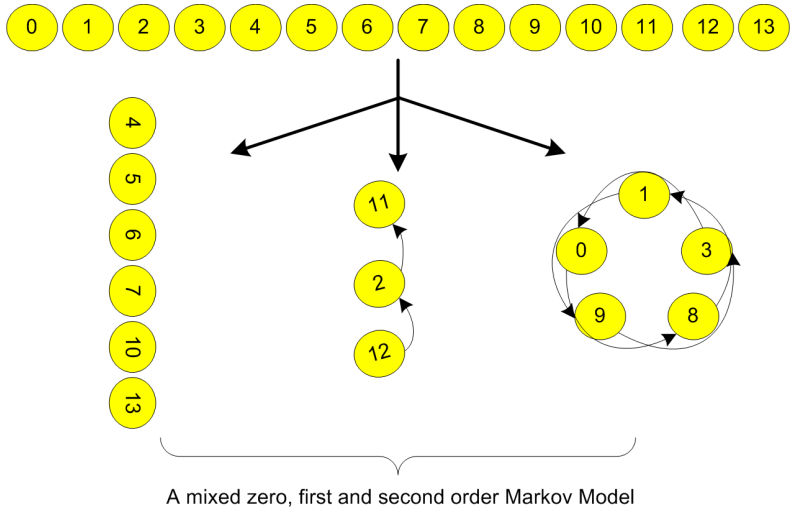


Figure 3.1: A graphic representation of a mixture of Markov models. On the top is a motif of length 14 bases. On the left, 6 positions, which are independent of each other and all other positions, form a zero order Markov chain; In the middle, 3 positions form a linear chain of first order Markov model; and on the right, the remaining positions that closely depend on each other form a circular chain of the second order Markov model.

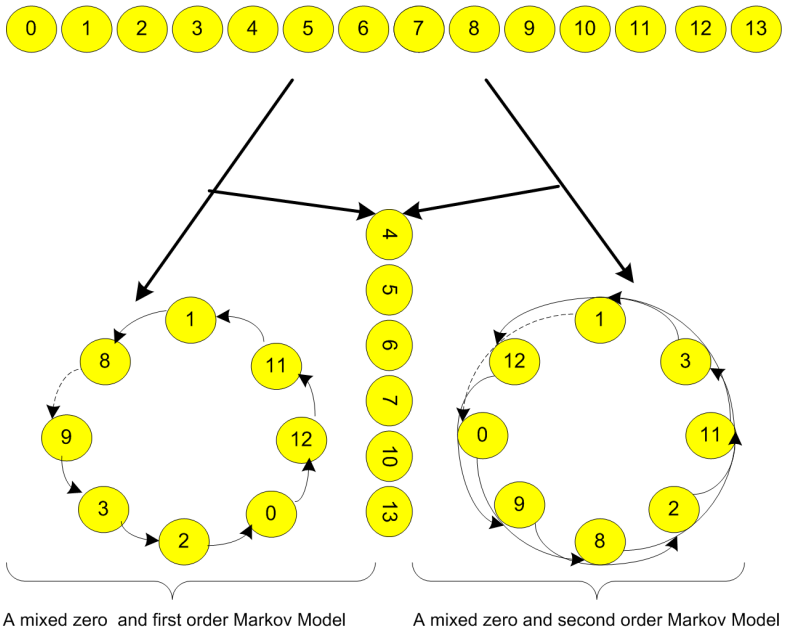


Figure 3.2: The mixture of Markov models for Transcription Factor Binding Sites(TFBS). Because the TFBS are usually only about 5-16 base pairs, a mixture model consisting of the 0 order and $1^{st}/2^{nd}$ order Markov chains is generally adequate for predicting new binding sites. We employ χ^2 test to find the positions that are independent of all other positions. The sub-motif formed by these independent positions is modeled by a zero-order Markov order model or position weight matrix model. The sub-motif consisting of the remaining positions is modeled by either 1^{st} or 2^{nd} order Markov chain, and a Markov chain could be either linear (break at dotted arrows) or circular. The positions in the 1^{st} or 2^{nd} order Markov chain are arranged by our directed neighbor-joining method, so that closely related positions are in the nearest neighbors.

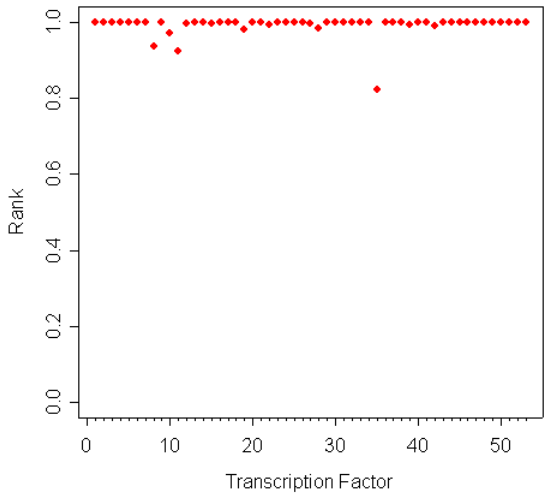


Figure 3.3: Ranks of DNJ optimized 0-1 linear models of 53 different TFBS. The values of transcription factors on x-axis correspond to the numbers in the first column of table 3.1

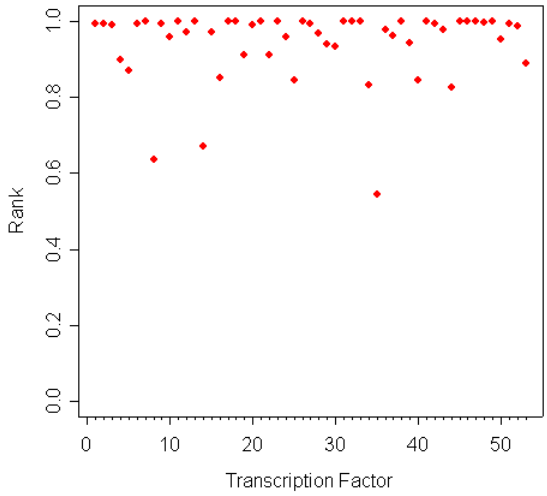


Figure 3.4: Ranks of DNJ optimized 0-2 linear models of 53 different TFBS. The values of transcription factors on x-axis correspond to the numbers in the first column of table 3.1

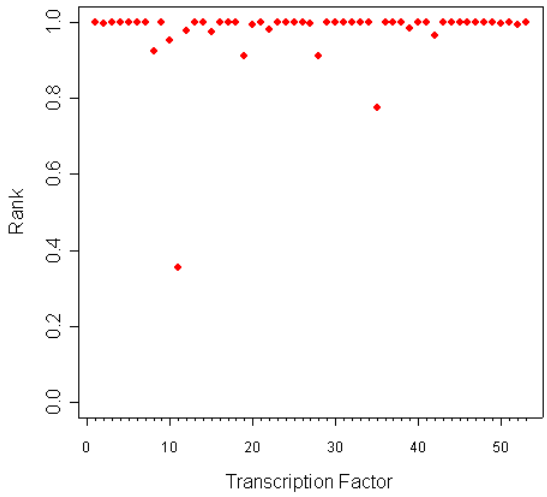


Figure 3.5: Ranks of DNJ optimized 0-1 circular models of 53 different TFBS. The values of transcription factors on x-axis correspond to the numbers in the first column of table 3.1

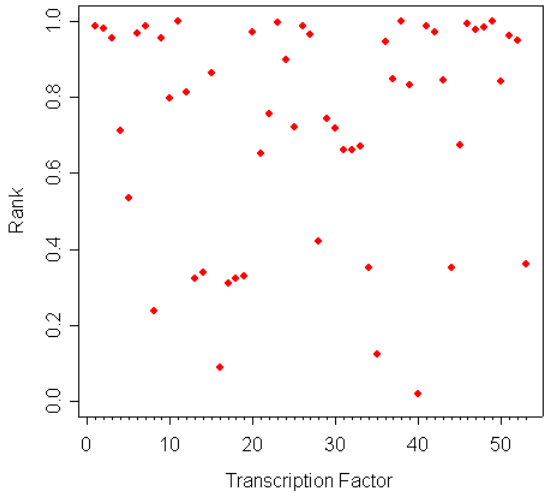


Figure 3.6: Ranks of DNJ optimized 0-2 circular models of 53 different TFBS. The values of transcription factors on x-axis correspond to the numbers in the first column of table 3.1

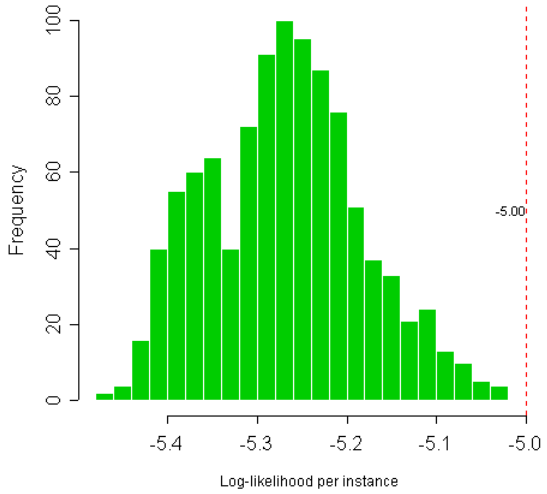


Figure 3.7: The histogram of log-likelihood scores from 1,000 randomly permuted mixed 0-1 linear models for the transcription factor V$AP1_Q4_01. The location of the score from the DNJ optimized model is indicated by the red reference line.

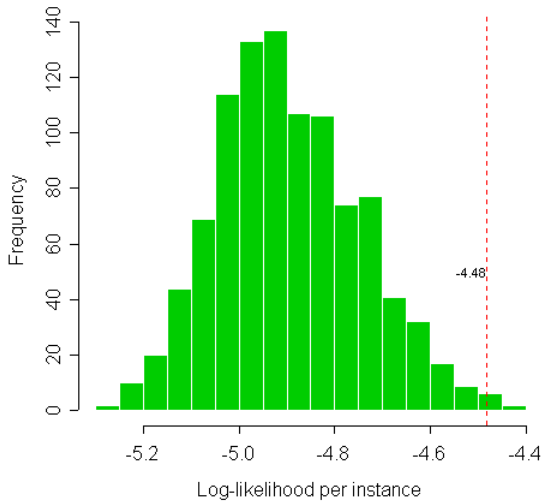


Figure 3.8: The histogram of log-likelihood scores from 1,000 randomly permuted mixed 0-2 linear models for the transcription factor V$AP1_Q4_01. The location of the score from the DNJ optimized model is indicated by the red reference line.

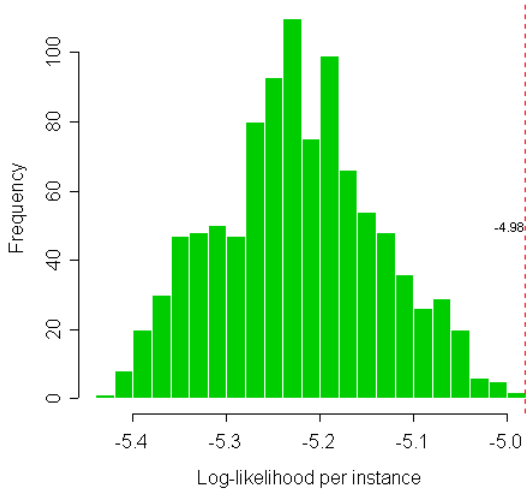


Figure 3.9: The histogram of log-likelihood scores from 1,000 randomly permuted mixed 0-1 circular models for the transcription factor V$AP1_Q4_01. The location of the score from the DNJ optimized model is indicated by the red reference line.

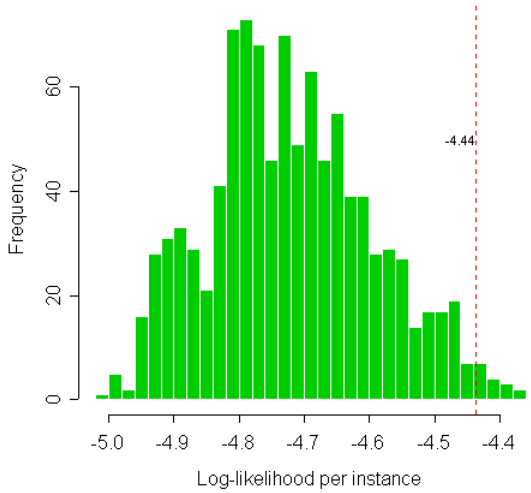


Figure 3.10: The histogram of log-likelihood scores from 1,000 randomly permuted mixed 0-2 circular models for the transcription factor V$AP1_Q4_01. The location of the score from the DNJ optimized model is indicated by the red reference line.

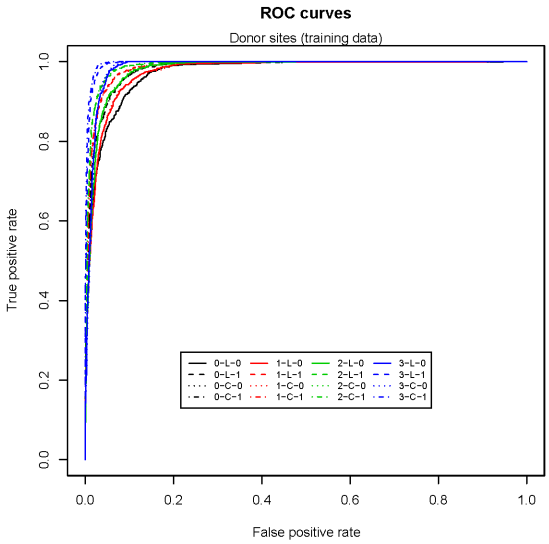


Figure 3.11: Donor site prediction (training data) Performance comparison of different 0-k mixture models by ROC curves. The plot shows that the best models are 3-L-1 and 3-C-1 while the worst model is 0-L-0 (same as 0-C-0) according to the Area Under Curve (AUC) criterion.

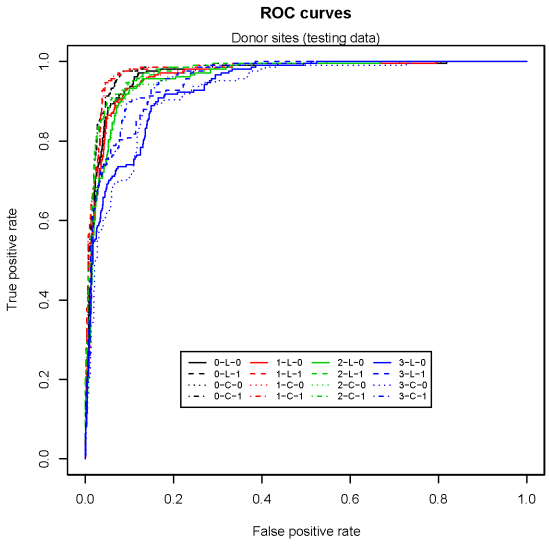


Figure 3.12: Donor site prediction (testing data) Performance comparison of different 0-k mixture models by ROC curves. The plot shows that the best models are 1-L-1 and 1-C-1 while the worst models are 3-L-0 and 3-C-0 according to the AUC criterion.

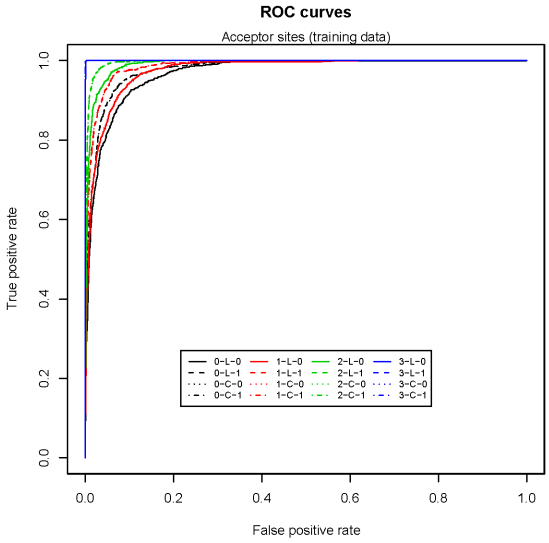


Figure 3.13: Acceptor site prediction (training data) Performance comparison of different 0-k mixture models by ROC curves. The plot shows that all 0-3 mixture models (3-L-0, 3-L-0, 3-C-0 and 3-C-1) have almost pefect prediction, while the model 0-L-0 (same as 0-C-0) has the worst prediction according to the AUC criterion.

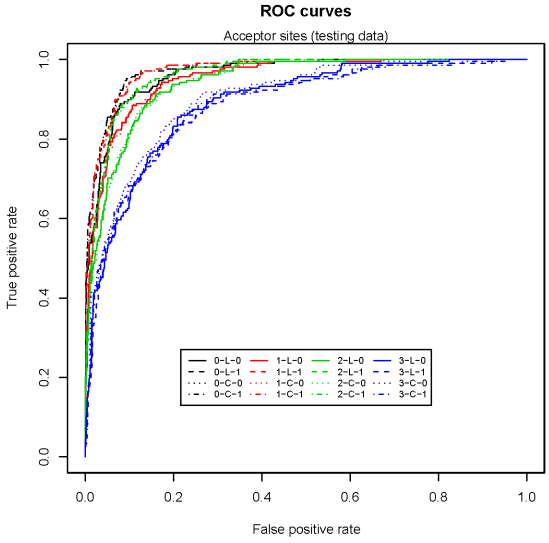


Figure 3.14: Acceptor site prediction (testing data) Performance comparison of different 0-k mixture models by ROC curves. The plot shows that the best model is 0-L-1 (same as 0-C-1) while the worst model is 3-C-1 according to the AUC criterion.

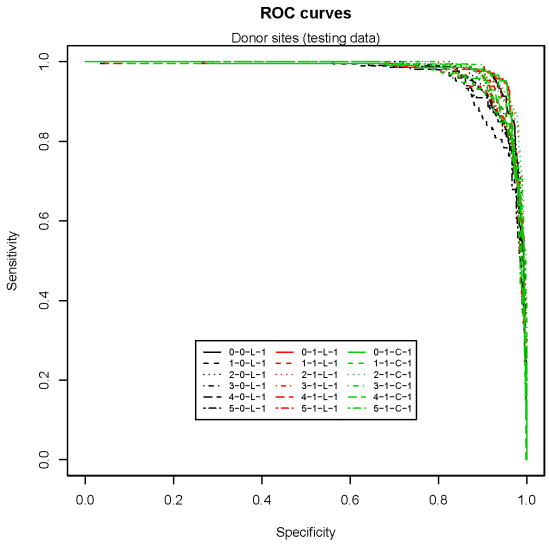


Figure 3.15: Model selection for donor site Six-fold cross validation confirms that the models 1-L-1 and 1-C-1, both of which are better than the model 0-L-1, have similar prediction performance as shown by the ROC curves in this plot.

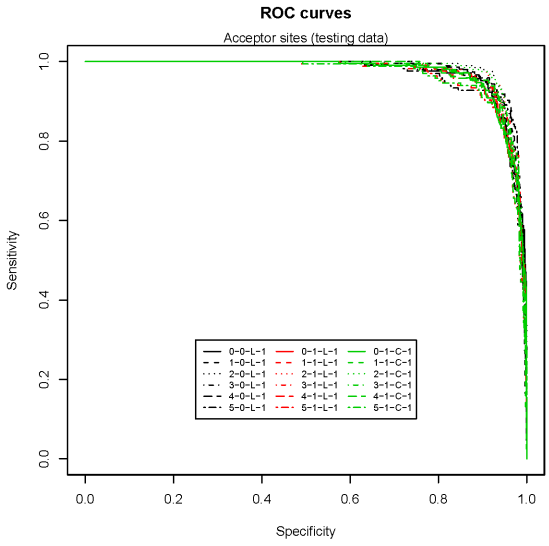


Figure 3.16: Model selection for acceptor site Six-fold cross validation confirms that the model 0-L-1 is the best model though its performance is very close to those of the models 1-L-1 and 1-C-1 as shown by the ROC curves in this plot.

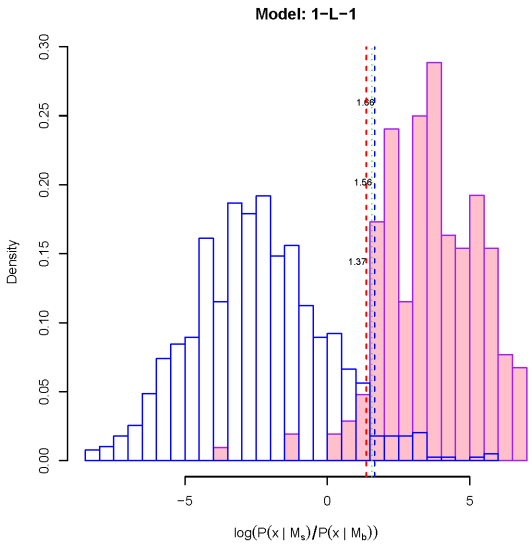


Figure 3.17: Donor site recognition (Model 1-L-1) The score distributions of false (blue) and true (pink) sites. A site is predicted as a positive site if its score is larger than a defined cutoff. In this plot, three different cutoffs are chosen as labeled by 3 reference lines to achieve three different goals: the balanced Sn and Sp (red dashed line), the maximum of Mc (green dotted line), and the maximum of Ac (blue dashed line).

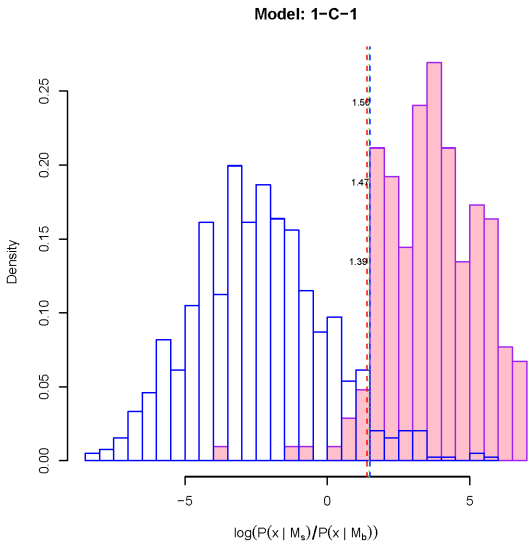


Figure 3.18: Donor site recognition (Model 1-C-1) The score distributions of false (blue) and true (pink) sites. A site is predicted as a positive site if its score is larger than a defined cutoff. In this plot, three different cutoffs are chosen as labeled by 3 reference lines to achieve three different goals: the balanced Sn and Sp (red dashed line), the maximum of Mc (green dotted line), and the maximum of Ac (blue dashed line).

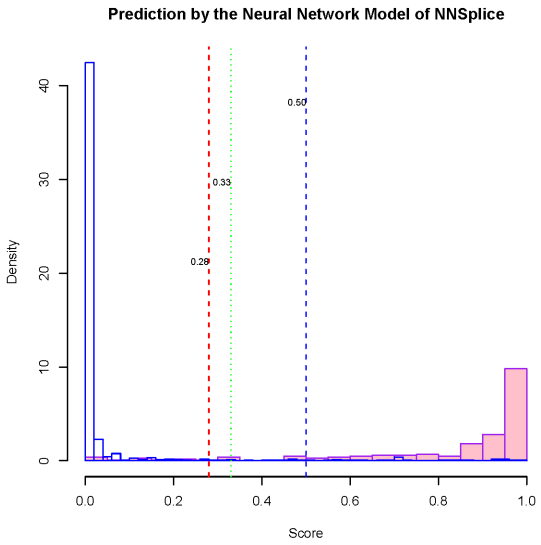
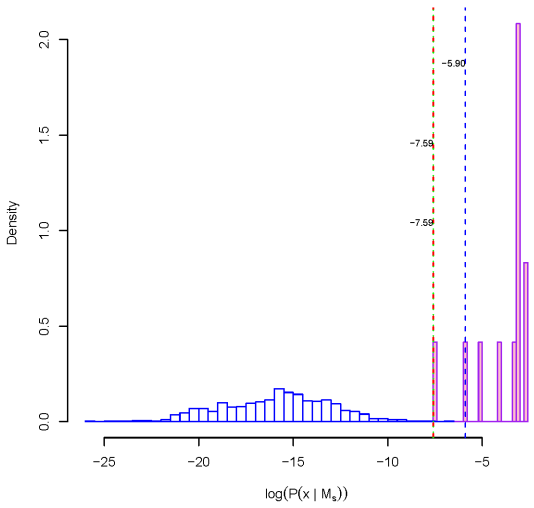
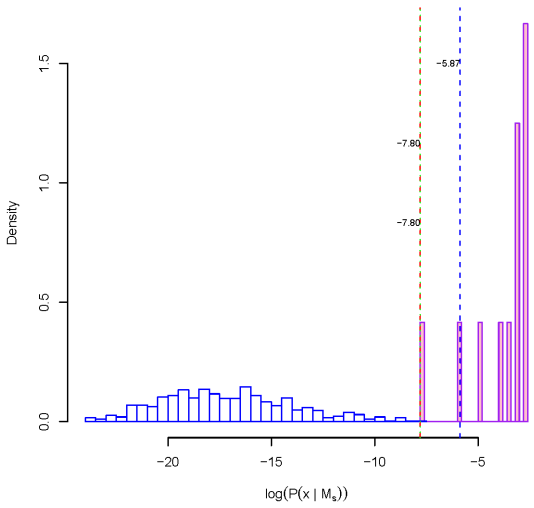


Figure 3.19: Donor site recognition (NNSplice's neural network model) The score distributions of false (blue) and true (pink) sites. A site is predicted as a positive site if its score is larger than a defined cutoff. In this plot, three different cutoffs are chosen as labeled by 3 reference lines to achieve three different goals: the balanced Sn and Sp (red dashed line), the maximum of Mc (green dotted line), and the maximum of Ac (blue dashed line).

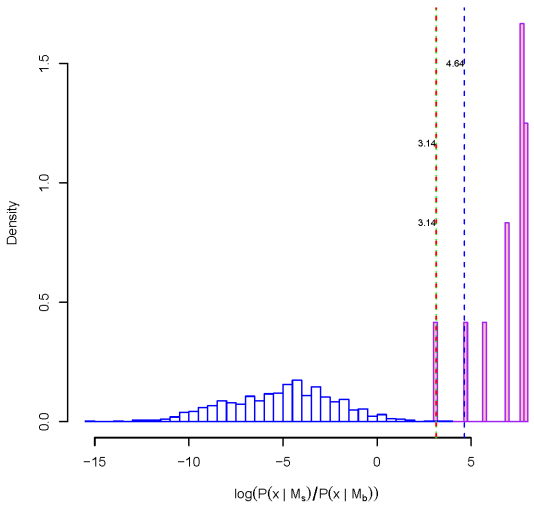


(a) Unoptimized $1st$ order Markov model of linear chain

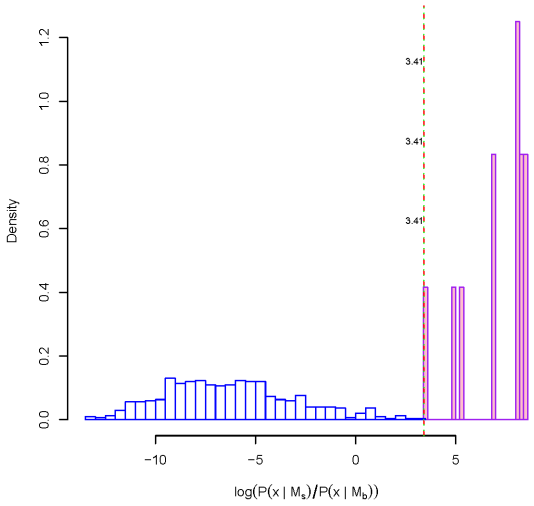


(b) Optimized $1st$ order Markov model of linear chain

Figure 3.20: Comparison of the score (**log-likelihood**) distributions of true sites and false sites of transcription factor V$AP1_Q4_01. The score distributions of true sites are in pink and those of false sites are in blue. The figure shows that DNJ-optimized model are better than unoptimized model to separate true sites from false sites. Three reference lines show score cutoffs chosen to achieve three different goals: the balanced Sn and Sp (red dashed line), the maximum of Mc (green dotted line), and the maximum of Ac (blue dashed line).



(a) Unoptimized $1st$ order Markov model of linear chain



(b) Optimized $1st$ order Markov model of linear chain

Figure 3.21: Comparison of score (**log-likelihood ratio**) distributions of true sites and false sites of transcription factor V$AP1_Q4_01. The score distributions of true sites are in pink and those of false sites are in blue. The figure shows that DNJ-optimized model are better than unoptimized model to separate true sites from false sites. Three reference lines show score cutoffs chosen to achieve three different goals: the balanced Sn and Sp (red dashed line), the maximum of Mc (green dotted line), and the maximum of Ac (blue dashed line).

Appendix

Expansion of equation (3.1) and (3.2)

In real applications, it may be easier to calculate $\Pr(x|M_k^L)$ or $\Pr(x|M_k^C)$ in terms of probabilities of oligomers. Also it is interesting to see the difference of the probabilities of a motif sequence given the Markov models of different orders. For such purposes, we expanded the equation (3.1) and (3.2) as the following:

1. The expansion of linear chain models:

$$\Pr(x|M_0^L) = \Pr(x|M_0^C) = \prod_{i=1}^{w} \Pr(x_i) \tag{3.11}$$

$$\begin{aligned}
\Pr(x|M_1^L) &= \Pr(x_1)\Pr(x_2|x_1)\Pr(x_3|x_2)\cdots\Pr(x_w|x_{w-1}) \\
&= \frac{\Pr(x_1,x_2)\Pr(x_2,x_3)\cdots\Pr(x_{w-1},x_w)}{\Pr(x_2)\cdots\Pr(x_{w-1})} \\
&= \Pr(x_1,x_2)\prod_{i=2}^{w-1}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)} \\
&= \prod_{i=1}^{w}\Pr(x_i) \times \prod_{i=1}^{w-1}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})} \\
&= \Pr(x|M_0^L) \times \prod_{i=1}^{w-1}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})}
\end{aligned} \tag{3.12}$$

$$
\begin{aligned}
\Pr(x|M_2^L) &= \Pr(x_1,x_2)\Pr(x_3|x_1,x_2)\Pr(x_4|x_2,x_3)\cdots\Pr(x_w|x_{w-2},x_{w-1}) \\
&= \frac{\Pr(x_1,x_2,x_3)\Pr(x_2,x_3,x_4)\cdots\Pr(x_{w-2},x_{w-1},x_w)}{\Pr(x_2,x_3)\cdots\Pr(x_{w-2},x_{w-1})} \\
&= \Pr(x_1,x_2,x_3)\prod_{i=2}^{w-2}\frac{\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})} \\
&= \prod_{i=1}^{w}\Pr(x_i)\times\prod_{i=1}^{w-1}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})}\times\prod_{i=1}^{w-2}\frac{\Pr(x_{i+1})\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})\Pr(x_{i+1},x_{i+2})} \\
&= \Pr(x|M_1^L)\times\prod_{i=1}^{w-2}\frac{\Pr(x_{i+1})\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})\Pr(x_{i+1},x_{i+2})} \\
&\approx \Pr(x|M_1^L)\times\prod_{i=1}^{w-2}\frac{\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i)\Pr(x_{i+1},x_{i+2})}
\end{aligned}
\tag{3.13}
$$

$$
\begin{aligned}
\Pr(x|M_3^L) &= \Pr(x|M_2^L)\times\prod_{i=1}^{w-3}\frac{\Pr(x_{i+1},x_{i+2})\Pr(x_i,x_{i+1},x_{i+2},x_{i+3})}{\Pr(x_i,x_{i+1},x_{i+2})\Pr(x_{i+1},x_{i+2},x_{i+3})} \\
&\approx \Pr(x|M_2^L)\times\prod_{i=1}^{w-3}\frac{\Pr(x_i,x_{i+1},x_{i+2},x_{i+3})}{\Pr(x_i)\Pr(x_{i+1},x_{i+2},x_{i+3})}
\end{aligned}
\tag{3.14}
$$

$$\vdots$$

$$
\begin{aligned}
\Pr(x|M_k^L) &= \Pr(x|M_{k-1}^L)\times\prod_{i=1}^{w-k}\frac{\Pr(x_{i+1},\ldots,x_{i+k-1})\Pr(x_i,\ldots,x_{i+k})}{\Pr(x_i,\ldots,x_{i+k-1})\Pr(x_{i+1},\ldots,x_{i+k})} \\
&\approx \Pr(x|M_{k-1}^L)\times\prod_{i=1}^{w-k}\frac{\Pr(x_i,x_{i+1},\ldots,x_{i+k})}{\Pr(x_i)\Pr(x_{i+1},x_{i+2},x_{i+k})}
\end{aligned}
\tag{3.15}
$$

2. Let $x_i = x_{i-w}$ if $i > w$. The expansion of circular chain models:

$$\begin{aligned}
\Pr(x|M_1^C) &= \Pr(x_2|x_1)\Pr(x_3|x_2)\cdots\Pr(x_w|x_{w-1})\Pr(x_1|x_w) \\
&= \frac{\Pr(x_1,x_2)\Pr(x_2,x_3)\cdots\Pr(x_{w-1},x_w)\Pr(x_1,x_w)}{\Pr(x_1)\Pr(x_2)\cdots\Pr(x_{w-1})\Pr(x_w)} \\
&= \frac{\Pr(x_1,x_w)}{\Pr(x_w)}\prod_{i=1}^{w-1}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)} = \prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)} \\
&= \prod_{i=1}^{w}\Pr(x_i)\times\prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})} \\
&= \Pr(x|M_0^C)\times\prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})} = \Pr(x|M_0^C)\times\prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)^2}
\end{aligned} \tag{3.16}$$

$$\begin{aligned}
\Pr(x|M_2^C) &= \Pr(x_3|x_1,x_2)\cdots\Pr(x_w|x_{w-2},x_{w-1})\Pr(x_1|x_{w-1},x_w)\Pr(x_2|x_w,x_1) \\
&= \frac{\Pr(x_1,x_2,x_3)\cdots\Pr(x_{w-2},x_{w-1},x_w)\Pr(x_{w-1},x_w,x_1)\Pr(x_w,x_1,x_2)}{\Pr(x_1,x_2)\Pr(x_2,x_3)\cdots\Pr(x_{w-2},x_{w-1})\Pr(x_{w-1},x_w)\Pr(x_w,x_1)} \\
&= \prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})} \\
&= \prod_{i=1}^{w}\Pr(x_i)\times\prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1})}{\Pr(x_i)\Pr(x_{i+1})}\times\prod_{i=1}^{w}\frac{\Pr(x_{i+1})\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})\Pr(x_{i+1},x_{i+2})} \\
&= \Pr(x|M_1^C)\times\prod_{i=1}^{w}\frac{\Pr(x_{i+1})\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i,x_{i+1})^2} \\
&\approx \Pr(x|M_1^C)\times\prod_{i=1}^{w}\frac{\Pr(x_i,x_{i+1},x_{i+2})}{\Pr(x_i)\Pr(x_{i+1},x_{i+2})}
\end{aligned} \tag{3.17}$$

$$\begin{aligned}\Pr(x|M_3^C) &= \Pr(x|M_2^C) \times \prod_{i=1}^{w} \frac{\Pr(x_{i+1}, x_{i+2}) \Pr(x_i, x_{i+1}, x_{i+2}, x_{i+3})}{\Pr(x_i, x_{i+1}, x_{i+2})^2} \\ &\approx \Pr(x|M_2^C) \times \prod_{i=1}^{w} \frac{\Pr(x_i, x_{i+1}, x_{i+2}, x_{i+3})}{\Pr(x_i) \Pr(x_{i+1}, x_{i+2}, x_{i+3})} \end{aligned} \tag{3.18}$$

$$\vdots$$

$$\begin{aligned}\Pr(x|M_k^C) &= \Pr(x|M_k^C) \times \prod_{i=1}^{w} \frac{\Pr(x_{i+1}, \ldots, x_{i+k-1}) \Pr(x_i, \ldots, x_{i+k})}{\Pr(x_i, \ldots, x_{i+k-1})^2} \\ &\approx \Pr(x|M_{k-1}^C) \times \prod_{i=1}^{w} \frac{\Pr(x_i, x_{i+1}, \ldots, x_{i+k})}{\Pr(x_i) \Pr(x_{i+1}, x_{i+2}, x_{i+k})} \end{aligned} \tag{3.19}$$

Estimation of model parameters

As shown in above expansions, the probability of a motif sequence for a given model can be expressed in terms of probabilities of oligomers. Therefore, we only need estimate the probabilities of oligomers (the length of oligomer is from 1 to $k+1$, where k is the order of Markov model) as model parameters when we fit a Markov model with motif data. Let $Y = Y_1, \ldots, Y_v$ be a random vector associated with oligomers that consist of bases from some certain positions of a motif, then Y follows multinomial distribution:

$$\Pr(Y|\theta) = \begin{pmatrix} N \\ y_1, \ldots, y_v \end{pmatrix} \prod_{i=1}^{v} \theta_i^{y_i} \tag{3.20}$$

where N is the total number of motif sequences, v is the number of all possible y_i, y_i is the number of oligomer i and $\sum_{i=1}^{v} y_i = N$, and θ_i is the probability of oligomer i occurred and $\sum_{i=1}^{v} \theta_i = 1$. Then according to the Bayes' theorem, the probability of θ given Y can be expressed as

$$\Pr(\theta|Y) = \frac{\Pr(Y|\theta) \cdot \Pr(\theta)}{\Pr(Y)} \propto \Pr(Y|\theta) \cdot \Pr(\theta) \tag{3.21}$$

Suppose prior distribution of θ is a Dirichlet distribution as given by equation (3.22), then we can get the estimate of posterior mean of θ by maximum likelihood (equations 3.23 and

3.24).

$$\Pr(\theta_i|\alpha) = \frac{\Gamma(\sum_{i=1}^{v}\alpha_i)}{\prod_{i=1}^{v}\Gamma(\alpha_i)}\prod_{i=1}^{v}\theta_i^{\alpha_i-1} \tag{3.22}$$

$$\hat{\theta} = \text{argmax}_\theta \log\left(Pr(Y|\theta)\Pr(\theta|\alpha)\right) \tag{3.23}$$

$$\hat{\theta_i} = \frac{y_i + \alpha_i - 1}{N + \sum_{i=1}^{v}\alpha_i - v} \tag{3.24}$$

where α is the parameter of Dirichlet distribution. In the context of Markov motif models, α is equivalent to pseudo counts of oligomers.

Usage of OMiMa

OMiMa can simultaneously search for multiple different motifs. The outputs of OMiMa include the scores and locations of motif sites, and distributions of both scores and locations. For DNA sequences, OMiMa is able to scan both strands of sequences in two directions. There are two standard usages of OMiMa as given by

1. **OMiMa input-seq-file configure-file** The configure-file is used to change the default parameter values of OMiMa. The example of OMiMa configure file is given in the following.

2. **OMiMa input-seq-file** In this case, the standard configure file named `OMiMa.conf` must exist.

An example of OMiMa configure file

```
# 1= scan orignial sequence from left to right, 2=scan two strands
# of DNA, 4=scan two directions of 2 strands of DNA. default=2
WAY 2

# The directory of training motif, each motif file with suffix ".mf",
# and first line must be ">motifName", with remaining lines are the
# alignment of motif sequences
motifDir /To/Train/motif/directory
```

```
# The file contains false motifs or background sequences, of which the bases
# for masking repeats or other sequences should be removed. The file
# should be in the Fasta format
bgseqFile /Dir/background_seq.fa

# The output directory. All output file except the main score
# files are put in this directory
outDir /output/directory

#main output file
outFile /directory/motif_score.out

# The order of the Markov model. In case of requiring the best model,
# this is the maximum Markov order. default=2
MaxMcOrder 1

# The percentage of the sample size of training data. This count is
# used to calculate flat Dirichlet prior of model. default=0.1
mcPrior   0.1

# Prediction accuracy for training data (containing true motif).
# It is used to select cutoff threshold for positive sites.
# default=1.00
testRate 1.00

# Markov model structure: 1=linear, 2=circle, 0=select the best
# among all, default=0
model  1

# 1=use log(P(s|M_s)/P(s|M_b)), log likelihood ratio score.
# 0= use logP(s|M_s), log likelihood score. default=1
LogRatio 1

# Print out model selection information, such as the Chi-square
# tests, AIC, BIC, and log likelihood of models. default=0
printModel 1

# The model selection criterion: A=AIC, B=BIC. default=A
criteria

# The interval of the histogram of motif scores default=0.1
hisInterval

# The minimum cutoff value of logP(s|M_s), default=-15.00
minLogProb

# The minimum cutoff value of log(P(s|M_s)/P(s|M_b)), default=2.00
minLogRatio
```

```
# Indicate whether prediction outputs include motif sites
# having different bases from those perfectly conserved
# in training data. 0-No output, 1-output with indicator (0/1),
# 2-output with score setting to the threshold, Default=0
IdkPC

#------------------------------------------------------------------------
#The score thresholds for different motifs, with the following format
motif_name_1 cutoff_1
motif_name_2 cutoff_2
```

Motif logo

Figure 3.22: The logo of human RNA 3' splicing site (donor site)

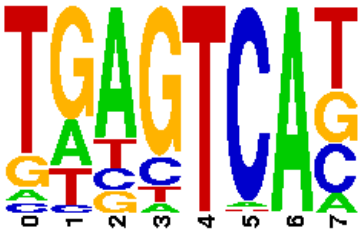


Figure 3.23: The logo of human transcription factor binding site V$AP1_Q4_01

Figure 3.24: The logo of human transcription factor binding site V$AP1_Q4_01

Bibliography

Agarwal, P. and Bafna, V. (1998). Detecting non-adjoining correlations with signals in dna. In *RECOMB '98: Proceedings of the second annual international conference on Computational molecular biology*, pages 2–8, New York, NY, USA. ACM Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19(6):716–723.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling dependencies in protein-dna binding sites. In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology*, pages 28–37, New York, NY, USA. ACM Press.

Benos, P. V., Lapedes, A. S., Fields, D. S., and Stormo, G. D. (2001). SAMIE: statistical algorithm for modeling interaction energies. *Pac Symp Biocomput*, pages 115–26.

Bühlmann, P. and Wyner, A. J. (1999). Variable length markov chains. *Ann. Statist.*, 27(2):480–513.

Bulyk, M. L., Johnson, P. L. F., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.*, 30(5):1255–1261.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94.

Cai, D., Delcher, A., Kao, B., and Kasif, S. (2000). Modeling splice sites with Bayes networks. *Bioinformatics*, 16(2):152–158.

Ellrott, K., Yang, C., Sladek, F. M., and Jiang, T. (2002). Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, 18 Suppl 2:S100–S109.

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

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Koza, J. R. (1998). Genetic programming. In Williams, J. G. and Kent, A., editors, *Encyclopedia of Computer Science and Technology*, volume 39, pages 29–43. Marcel-Dekker.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405(2):442–451.

Negre, B., Casillas, S., Suzanne, M., Sanchez-Herrero, E., Akam, M., Nefedov, M., Barbadilla, A., de Jong, P., and Ruiz, A. (2005). Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome Res.*, 15(5):692–700.

Ponomarenko, M. P., Ponomarenko, J. V., Frolov, A. S., Podkolodnaya, O. A., Vorobyev, D. G., Kolchanov, N. A., and Overton, G. C. (1999). Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. *Bioinformatics*, 15(7):631–643.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, 23(23):4878–4884.

Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. *J. Comput. Biol.*, 4(3):311–323.

Rissanen, J. (1986). Complexity of strings in the class of markov sources. *IEEE Trans. Inform. Theory*, 32(4):526–532.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.*, 12(1):505–519.

Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 23(3):109–113.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, 28(1):316–319.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., and Romano, L. A. (2003). The Evolution of Transcriptional Regulation in Eukaryotes. *Mol Biol Evol*, 20(9):1377–1419.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345.

Yeo, G. and Burge, C. B. (2003). Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology*, pages 322–331, New York, NY, USA. ACM Press.

Zhang, M. Q. and Marr, T. G. (1993). A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9(5):499–509.

Zhao, X., Huang, H., and Speed, T. P. (2004). Finding short dna motifs using permuted markov models. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 68–75, New York, NY, USA. ACM Press.

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916.

Chapter 4

Computational Analysis of the Association Between Gene Function and Regulatory Region Conservation

Abstract

Cross-species sequence comparisons have shown that Non-coding sequences, particularly, Conserved Non-coding Sequences (CNS), play functional roles as important as protein-coding sequences for an organism. In this study, with aim of elucidating the evolutionary function constraints of CNS, we systematically analyzed the conservation of upstream sequences of human-rodent orthologs. We found that the genes with highly conserved upstream regions are significantly associated with important regulation functions. In particular, the developmental process-related transcription regulators, such as Hox genes and POU genes, showed extreme upstream region conservation. Furthermore, we found the degree of upstream region conservation is highly correlated with the density of transcription factor binding sites in the upstream regions. Our results suggested that the key developmental process-related transcription regulators are under the sophisticated control, both temporal and spatial, of their upstream regulatory regions, which are subjected to stringent evolutionary constraints, hence are highly conserved in both human and rodent species.

Introduction

Comparative genomics provides an effective approach to identifying conserved functional elements, e.g. *cis*-elements, in the non-coding genomic regions of related species. In lower eukaryotes, genome comparison of several *Saccharomyces* species has led to the identification of most functional regulatory elements of yeast (Kellis et al., 2003; Doniger et al., 2005). In high eukaryotes, comparison of multiple mammalian genomes has enabled the identification of many ultra-conserved elements involved in regulating developmental processes (Bejerano et al., 2004). In addition, comparative genomic analysis gains more power in identifying regulatory elements when combined with the experimental approaches, such as genome-wide chromatin immuno-precipitation (ChIP), genome-wide transcriptional profiling analysis. By using both computational and experimental approaches, for example, transcription regulation modules have been elegantly mapped genome-wide for Drosophila (Berman et al., 2002) and yeast (Vilo et al., 2000; Pilpel et al., 2001).

Despite remarkable advances in the identification and characterization of *cis*-regulatory regions, deciphering the whole genetic regulatory networks for higher eukaryotes, particularly human, remains a distant hope and aspiration as insufficient evidence is currently available to construct fully connected networks. In this study, with the aim of elucidating some features of human genetic regulatory networks, we systematically investigated the upstream sequence conservation and its relationships with gene functions using nearly $6,000$ and $1,000$ pairs of human-mouse and human-rat orthologous genes, respectively.

To construct gene regulatory networks, it is necessary to discover all constituent *cis*-regulatory elements and their genomic locations. Unfortunately, computational identification of all such elements is difficult, even with a powerful comparative genomic approach, as our current knowledge of their syntax or grammar is limited. By contrast, vast amounts of biological data related to gene function annotation have been accumulated in a variety of databases, as the result of well-established protein-coding gene prediction methods and

the advance of experimental technologies for genome-wide gene function analysis. Gene function annotation, however, can facilitate the identification and characterization of non-coding regulatory elements. In this study, we combined upstream sequence information of human-rodent orthologs with gene function annotations from Gene Ontology to elucidate evolutionary features of upstream regulatory regions. The Gene Ontology (GO), which provides consistent gene function and sequence annotations by integrating many biological data sources, is one of the most widely used ontologies (Harris et al., 2004).

While most regulatory elements for gene expression control are expected to locate in the proximity of transcription start sites (TSS), some elements, such as enhancers, can locate as far as several *kb* upstream TSS in higher eukaryotes (Waterston et al., 2002). To select an upstream region that can appropriately reflect the features of evolutionary conservation of human-mouse orthologs, we estimated the degree of conservation for four different upstream regions. Also, as a precautionary measure, in function association analysis we removed the orthologous pairs whose upstream sequences may not be from the same region relative to the corresponding TSS possibly due to poor annotation.

We report here, as results mainly inferred from 4.5 *kb* upstream regulatory regions of human-mouse orthologous genes, that the genes involved in key regulation processes, such as developmental control and transcriptional regulation, generally have longer or more conserved regulatory regions, while genes performing housekeeping functions, such as catalytic enzymes, have shorter or less conserved regulatory regions.

Materials and methods

Ortholog collection and upstream sequence extraction

The lists of human-mouse and human-rat orthologous genes were obtained from the NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/homol_seq_pairs). Now it is a part of HomoloGene database. We removed all orthologous pairs whose official gene symbols are

not available. The upstream sequences of genes are extracted from the NCBI (Build 34) genome assembly and annotation. For each gene, we extracted sequence regions of four different lengths: 5 *kb* upstream, 4.5 *kb* (3.5 *kb* upstream and 1.0 *kb* downstream), 2.5 *kb* upstream and 1*kb* upstream, relative to the transcription start sites(TSS) [1]. It may not be possible to obtain all four upstream regions for some genes located near the end or beginning of a sequence contig. In our analysis, we ignored the orthologous pairs if the upstream sequence of either gene in the pair is less than the length required.

Generating false orthologous pairs

We randomly shuffled the order of human genes in a two-column list of orthologous pairs, so that each human gene forms a false orthologous pair with a rodent gene. This allows us to generate a set of false orthologous pairs having same genes of the original orthologs data set.

Sequence alignment and conserved region identification

We first used Censor (version 4.1) (Jurka, 2000) and WU-BLAST (version 2) (Altschul and Gish, 1996) to mask all known repetitive elements in the upstream sequences. We then aligned each pair of sequences using our ACANA system (Huang et al., 2005) with the default parameter setting. ACANA uses a recursively dynamic programming algorithm to align sequences, and has shown some improvement over other alignment tools in aligning genomic sequences. We obtained both a global alignment and non-overlapping local alignments for each pair of sequences. We used ACANA local alignment score, the sum of scores of non-overlapping local alignments identified in the upstream sequences of an orthologous pair, to assess the upstream region conservation of the orthologous pair.

Using 70% identity over 100 *bp* stretch as the cutoff value for a conserved segment, we then used the VISTA tool (Mayor et al., 2000; Frazer et al., 2004) to extract conserved

[1] We used the 5' end of mRNA (assume full length) as the transcription start site of a gene.

regions from each global alignment. VISTA outputs all the conserved segments found in a global alignment. We added up the lengths of these conserved segments to obtain the total length(ℓ) of conserved region in the upstream sequences of the pair. ℓ is used as an alternative measure of upstream sequence conservation.

Gene and GO term association

We first associated genes with GO terms according to the GO annotation of NCBI Entrez Gene database (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go). We then categorized these genes into corresponding GO categories. A GO category is described by a GO term, e.g. "transcription factor activity (GO:0003700)". A gene is assigned to a GO category if it is associated with the GO term of the GO category or any of its descendant GO terms in the GO hierarchy structure.

Association test

We removed the orthologous pairs whose human gene is not associated with any GO term. We then divided the remaining orthologous pairs into two groups: those associated with the category form the first group and the others form the second group. Since ACANA local alignment scores for both groups do not necessarily follow a normal distribution (an assumption for two-sample t-test), we used the Wilcoxon rank sum test [2] to calculate the p-value of the hypothesis test that the ACANA alignment score distributions are same for both gene groups.

[2]The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is used to test whether two samples are drawn from the same population when the distribution of population violates the normality assumption for two-sample t-test. The test first ranks the combined data set and divides the ranks into two sets according to the group membership of the original observation. It then uses the pooled variance estimate to calculate a two sample z statistic and the corresponding p-value.

Enrichment test

We sorted the human-mouse orthologous genes by their local alignment scores of upstream 4.5 *kb* regions, and obtained the top n conserved genes. We counted the total number (M) of human genes assigned to a certain GO category among all human-mouse orthologs (N), and the corresponding number (m) of these M genes in the top list. Let X be the observed number of genes assigned to the GO category in the top n list. Assuming that these M genes are uniformly distributed in the sorting list of all human-mouse orthologs, then X follows a binomial distribution (M, f), where $f = \frac{M}{N}$. We calculated the probability of observing m or more genes of the GO category in the top list as follows:

$$\Pr(X >= m) = \sum_{x=m}^{M} \Pr(X = x) = \sum_{x=m}^{M} \left(\begin{array}{c} M \\ x \end{array} \right) f^x (1-f)^{M-x}$$

If $\Pr(X >= m)$ is less than a significant level α, e.g. $\alpha = 0.01$, then the GO category is overrepresented in the top list.

Identification of putative transcription factor binding sites

We developed a motif-search program called *cisFinder* (available from the author upon request) that uses Position Weight Matrix (PWM) model to identify the potential motif sites in upstream sequences. The scoring method of *cisFinder* is based on the entropy-related information measure proposed by Quandt et al. (1995). For a DNA motif of length w, the score of a potential site is given by

$$Score = \frac{\sum_{i=1}^{w} C_i \cdot f_i^b}{\sum_{i=1}^{w} C_i \cdot \max_{b \in B}(f_i^b)}$$

where $C_i = 1 + \frac{1}{\ln 4} \cdot \sum_{b \in B} f_i^b \ln f_i^b$ and $B = \{A, C, G, T\}$. The $Score$ has a value ranging from 0 to 1, with 0 and 1 indicating the most unlikely and likely sites, respectively. To reduce

the number of false positive sites, we used a stringent threshold, 0.95, as the cutoff value of a putative binding site. In this study, we used *cisFinder* to identify all putative binding sites of 106 transcription factors, whose PWMs were from Transfac database (version 7.4), in the upstream sequence of human genes.

Results

Upstream sequence data

For both human-mouse and human-rat orthologs, we obtained four sequence data sets of different upstream region lengths: 1.0, 2.5, 4.5 and 5 *kb*, respectively. Each human-mouse data set contains about 5,900 non-redundant orthologous pairs, of which about 4,700 human genes are associated with at least one GO term[3]. Each human-rat data set contains the same 933 human-rat orthologous pairs, of which 905 human genes are associated with at least one GO term.

To estimate the expected local alignment score of upstream sequences of a random pair of human-rodent genes, we generated a corresponding data set of false orthologous pairs by random shuffling for each above data set. We then aligned the upstream sequences of the random pairs and computed the 95th percentile of the local alignment scores as the estimate of expected 95th percentile ($E_{.95}$) from non-orthologugous pairs.

From each sequence data set of orthologous pairs, we extracted two subsets: the first subset contains only orthologous pairs whose human gene is associated with at least one GO term; the second subset is the subset of the first one, with further removing the orthologous pairs whose local alignment scores (S) are less than $E_{.95}$. Summary of these data sets and subsets is given in Table 4.1. We used the second subset to analyze the association of gene function with their upstream region conservation.

[3]Because we removed the genes that do not have enough upstream sequence available, the exact numbers of orthologous pairs in different data sets are not the same.

The main results reported in this study are based on the analyses of 4.5 kb (3.5 kb upstream and 1.0 kb downstream of TSS) regulatory regions of human-rodent orthologous pairs (datasets 1 and 5 in Table 4.1). We think that most functional regulatory elements are likely to locate within such 4.5 kb regions, from which more accurate results could be inferred than from the other three regions. The results from the 4.5 kb region are further verified by those from the other three regions as they are consistent with each other.

Regulatory region conservation

We used either ACANA local alignment score or the length of conserved region (over 70% identity) to assess the conservation of upstream regulatory sequences of a pair of orthologous genes. We found that the number of orthologous pairs exponentially decreased approximately after an initial increase as the degree of conservation increased (Figures 4.1 and 4.2).

To find out whether there is any correlation between the conservation of different upstream regions, we computed the Pearson's correlation coefficient of the local alignment scores of different regions. We found that conservation of different regions are significantly correlated (Table 4.2). It appears that a gene conserved in the near upstream regions (within 2.5 kb) of TSS is also likely to have a conserved far upstream region (over 2.5 kb from TSS) (Figures 4.3 and 4.4).

For both human-mouse and human-rat orthologous upstream sequences, the distributions of local alignment scores appear to have two modes (Figures 4.5 and 4.6). While the larger mode includes most orthologous pairs whose local alignment scores are larger than $E_{.95}$, the smaller mode almost spans the same score region as the mode of the score distribution of false orthologous pairs. We suspected that most pairs of upstream sequences in the smaller mode are "false" orthologs. Several possibilities could lead to such "false" orthologous upstream sequences: (i) the human-rodent pairs are not real orthologs; (ii) or the sequence

regions extracted from the upstream of an orthologous pair are not well matched due to poor annotation of transcription start sites; (*iii*) the orthologous genes do not perform similar functions any more, e.g. one gene is active, the other is a pseudo-gene without function. In the following further analysis, we removed those "false" orthologous pairs.

Association of gene function with upstream region conservation

As we learned from above results that there are significant variations among different genes in the degree of upstream region conservation. We would like to see if there are any associations between the degree of upstream region conservation and gene function. To elucidate the association relationships, we systematically categorized orthologous pairs, based on their GO function annotation, into GO categories in the 1^{st}, 2^{nd}, and 3^{rd} level of GO hierarchical structure. We then tested whether the genes of a GO category have significantly more or less conserved upstream regions than other genes. From the results (Table 4.3), we found that genes with important regulation functions, such as development control (Figure 4.7), regulation of biological process, regulation of physiological process, transcription regulation (Figure 4.8), and signal transduction, had significantly more conserved upstream regions than all other genes. Furthermore, the genes encoding transcription factors and those controlling organ development or morphogenesis showed the most significant (p-value ≈ 0) association with highly conserved upstream regions. In contrast, the genes involved in physiological process, or encoding the components of protein complex or catalytic enzymes are significantly associated with the less conserved upstream regions. Particularly, the catalytic enzyme genes were the least conserved in the upstream regions. While we observed similar associations between the degree of upstream conservation and gene function for the four different upstream regions, such associations were generally more significant (smaller p-values) when using the 4.5 *kb* upstream region (Supplement Table 4.6). In addition, we found that results from the human-mouse orthologous data were consistent with those from

the human-rat orthologous data (Supplement Table 4.7).

Top upstream conserved genes and functional categories

We further focused on genes that have the highest degree of upstream region conservation. We ranked the human-mouse orthologous genes by their local alignment scores of upstream 4.5 *kb* regions. Percent identity plots (Figure 4.9) showed that the top ranked human-mouse orthologs had extremely conserved upstream regions. For example, for the top 10 orthologous genes, 3,780 *bp* out of 4,500 *bp* upstream sequence, on average, are conserved with 81.70% identity (Table 4.4). To find the important biological functions that may significantly associated with the top ranked genes, we obtained the top 50 orthologous genes and used the enrichment test to find overrepresented GO functional categories in the top list. As a results, we found 9 GO categories whose genes were most significantly enriched (p-value $< 5 \times 10^{-5}$) in the top list (Tables 4.5). As we expected, the genes involved in the developmental process (GO:0007275), transcription regulation (GO:0030528), and regulation of biological process (GO:0050789) overwhelmingly dominate the top list. Notably, four of the top five genes are involved in all the three regulation processes. Also, the top upstream conserved genes involved in the developmental process are mostly for regulating organ development (GO:0009653) and/or morphogenesis (GO:0048513). Similarly, among the top upstream conserved genes involved in transcription regulation, most are from the subcategory GO:0003700, transcription factor encoding genes. So, it is not surprising that the nucleic acid binding genes (GO:0003676) are also significantly enriched in the top list, as transcription factors or other transcription regulation proteins generally bind to nucleic acids. In the top list, we also found 12 among all 14 genes, which do not belong to any of the 9 categories, are important cell membrane components (GO:0016020) involved in cell communication, receptor binding, and signal transduction. For example, the top ranked gene KIAA0319L, as well as PLXNC1, PCDH7, PCDH8, and TRHDE, is a membrane component

involved in cell communication process.

TFBS map in upstream regions

The upstream conservation of human-rodent orthologs suggested evolutionarily functional constraints in regulatory regions. We would expect that the higher the degree of conservation, the more functional elements, e.g. transcription factor binding sites (TFBS), would be in the regulatory upstream regions. This is likely to be true even if we count only the known TFBS, a significant subset of all functional elements, as many of them may not be discovered yet. To verify that, we counted the number of all putative transcription factor binding sites in the 4.5 *kb* upstream sequences of human genes. For each gene, we calculated its density of TFBS, which is defined as the number of sites, on average, found in the 51 genes centered around this gene in the ranks of their local alignment scores. We then plotted the density of TFBS as the function of gene rank. We found that the density of TFBS generally decreased as the degree of upstream conservation decreased though there were some fluctuations (Figure 4.10). Top ranked genes showed high density of TFBS in their upstream sequence regions (Figure 4.11).

Discussion

Variation of regulatory regions

This study has shown that there is large variation in the degree of upstream region conservation, which may be largely due to the variation of the stringency of evolutionary constraints imposed on different genes. Furthermore, the study showed that the conservation of different upstream regions could be highly correlated. Particularly, genes with more conserved immediate region of TSS are more likely to have more conserved far upstream regions. We suggested that there are possible large variations of the length of upstream regions which directly regulate transcription of individual genes. Genes that perform important regulation

functions, such as developmental control, are usually under delicate transcription control by many transcription factors. As each transcription factor may occupy different binding sites, a longer upstream regulatory region becomes necessary. This is supported by our results that showed the highest density of TFBS in the upstream regions of key developmental transcription factors (Figure 4.11). It has been shown that protein-coding genes are asymmetrically distributed in the human genome, of which, as a consequence, 25% regions form gene deserts (Lander et al., 2001; Venter et al., 2001). Futhermore, recent studies (Nobrega et al., 2003; Ovcharenko et al., 2005) found that the genes in stable gene desert regions are mostly related to developmental control and transcription regulation. These evidences support the hypothesis that genes with key regulation functions, hence under stringent purifying selection, are more likely to have longer and more conserved regulatory regions than the general housekeeping genes.

The central hub of gene regulatory networks

Our findings, which fit well with previous observations from published studies (Bergman and Kreitman, 2001; Halligan et al., 2004), indicate that the developmental process related genes are under the strongest evolutionary selection pressure and highly conserved in upstream regulatory regions. Particularly, we found that developmental transcription factors, such as the Homeotic (Hox) genes and POU genes, have the most conserved upstream regions. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies. The Hox family of transcription factors is believed to be activated in a precise temporal and spatial sequence following their chromosomal order, which is referred as the “Hox clock" (Ferrier and Minguillón, 2003; Kmita and Duboule, 2003; Garcia-Fernàndez, 2005). Many previous studies showed Hox gene clusters were flanked by the high density of CNS that contain distal enhancers (Santini et al., 2003; Negre et al., 2005; Kmita et al., 2005). The POU (Pit-Oct-

Unc) family of transcription factors is also involved in the temporal and spatial regulation of expression of many genes, and has been shown to play critical roles in the development and functioning of the nervous system of mammals (Wegner et al., 1993; Schonemann et al., 1998; Ilia, 2004; Zhang et al., 2004). These developmental process-related genes have been found to govern the formation of various tissues and organs in nematodes, flies, sea urchins, frogs, and mammals (see review by Levine and Davidson, 2005). This evidence suggests that the Hox and POU transcriptional factors possibly function as the central hub of gene regulatory networks, and they, together with the key regulators of signal transduction, cell communication, such as *KIAA0319L, PLXNC1, PCDH7, PCDH8*, and *TRHDE*, constitute the major network components of the whole complex regulatory networks.

Positive selection for catalytic enzymes

By contrast, this study show that the upstream regions of genes encoding catalytic enzymes are significantly less conserved than those of the other genes. These enzymes are generally involved in the synthesis or degradation of biomacromolecules, for example, helicase, isomerase, cyclase, and transferase for synthesis, and lyase, deaminase, and hydrolase for degradation. Since the genes perform housekeeping functions, they are much less likely under the complex regulatory control of their upstream regions than the gene regulating critical processes. This may not, however, fully account for much less conserved upstream regions, even when the region is short (1 kb) and at the immediate upstream of TSS. Our TFBS density map found there were still reasonable number of TFBS at the immediate upstream regions of these genes. We suspected, therefore, there may be some positive selection force that drives the divergence of TFBS in the upstream region of these genes. Such evolutionary selection force could come from the fast changing of environments and the resultant striking physiological differences. However, more evidences would be needed to support such a hypothesis.

Summary

In this study, we have inferred some possible features of the complex gene regulatory network via cross-species sequence comparison and utilization of function annotation from GO database. The reliability of our computational results was confirmed by some experimental evidences reported in published studies. Based on our results and existing evidences, we proposed the hypothesis that the developmental transcription regulators constitute the center hub of mammalian gene regulatory networks. Regulation of the center hub is subjected to strong purifying selection pressure, thus is highly conserved even between highly divergent species, such as Human and pufferfish. By contrast, the regulation of catalytic enzymes is subjected to the natural positive selection forces, such as fast changing of environments, thus can be very different even between closely-related species, such as human and mouse. We suggested that, therefore, the regulation of catalytic enzymes may constitute local peripheral networks near the terminals of the complex gene regulatory networks of mammalians. However, these hypotheses need more supporting evidence from further studies.

Tables

Table 4.1: The sequence data sets of orthologous human-rodent pairs. For both human-mouse and human-rat orthologous genes, we obtained four datasets of different upstream regions. Each data set has two subsets: the first subset contains only orthologous pairs whose human genes are associated with at least one GO term; the second subset is the subset of the first one, keeping only the orthologous pairs whose local alignment scores (S) are larger than the 95th percentile of expected score ($E_{.95}$) of a random pair. The table shows the number of orthologous pairs in each dataset and its subsets.

Dataset #	Sequence len(bp)	# orthologous pairs total	GO Ann.	$S > E_{.95}$
Human-mouse orthologs				
1	4500	5943	4697	3944
2	1000	5972	4721	3330
3	2500	5947	4701	3415
4	5000	5923	4681	3417
Human-rat orthologs				
5	4500	933	905	759
6	1000	933	905	652
7	2500	933	905	679
8	5000	933	905	694

Table 4.2: Correlation coefficients (r) of the degree of conservation between different upstream regions. Here the variables LS1k, LS2k5, LS4k5, and LS5k are the local alignment scores for four upstream regions of length 1 kb, 2.5 kb, 4.5 kb, and 5.0 kb, respectively. The variable D2k5-1k is the difference of LS2k5 and LS1k, and the similar interpretation applies to D5k-1k, D5k-2k5. For all correlation coefficients, the probability is less than 0.001 for hypothesis testing under H_0: r=0.

	LS1k	LS2k5	LS5k	LS4k5	D2k5-1k	D5k-1k	D5k-2k5
LS1k	1.00	0.83	0.66	0.75	0.53	0.46	0.34
LS2k5	0.83	1.00	0.87	0.91	0.91	0.77	0.55
LS5k	0.66	0.87	1.00	0.89	0.84	0.97	0.88
LS4k5	0.75	0.91	0.89	1.00	0.83	0.82	0.67
D2k5-1k	0.53	0.91	0.84	0.83	1.00	0.83	0.58
D5k-1k	0.46	0.77	0.97	0.82	0.83	1.00	0.94
D5k-2k5	0.34	0.55	0.88	0.67	0.58	0.94	1.00

Table 4.3: Statistical tests of the association between gene function and the degree of upstream region conservation of human-mouse orthologs. The table showed the test results for 4.5 *kb* upstream region (3.5k *bp* upstream and 1k *bp* downstream of TSS). The Wilcoxon rank sum test p-values are for the two tails tests of hypothesis that there is no difference in the distribution of alignment scores between the genes in a GO category and those not. In the table, the rows in bold font are level 2 GO terms, and the remaining rows are level 3 GO terms. The last column is the median of ACANA alignment scores of genes in a GO category (Yes) and those not (No).

GO Term	Term Definition		Test	#genes		Median	
Accession#	Type	Name	p-value	Yes	No	Yes	No
GO:0007610	**BP**	**behavior**	**1.4e-02**	**89**	**3855**	**7.20**	**6.25**
GO:0007631		feeding behavior	1.4e-01	10	3934	8.73	6.26
GO:0007611		learning and/or memory	7.3e-01	6	3938	7.27	6.26
GO:0007626		locomotory behavior	5.2e-02	69	3875	6.81	6.26
GO:0009987	**BP**	**cellular process**	**3.7e-02**	**3089**	**855**	**6.33**	**6.08**
GO:0030154		cell differentiation	2.1e-08	118	3826	8.63	6.20
GO:0007154		cell communication	0.0e+00	921	3023	7.29	5.97
GO:0050794		regulation of cellular process	6.2e-15	807	3137	7.36	6.04
GO:0050875		cellular physiological process	6.0e-02	2671	1273	6.17	6.48
GO:0007275	**BP**	**development**	**0.0e+00**	**559**	**3385**	**7.85**	**6.00**
GO:0007389		pattern specification	2.5e-02	9	3935	14.37	6.26
GO:0009790		embryonic development	2.6e-03	15	3929	11.52	6.26
GO:0007498		mesoderm development	7.3e-02	8	3936	9.01	6.26
GO:0030154		cell differentiation	2.1e-08	118	3826	8.63	6.20
GO:0048513		organ development	0.0e+00	295	3649	8.19	6.11
GO:0009653		morphogenesis	0.0e+00	373	3571	8.00	6.08
GO:0040029		regulation of gene expression, epigenetic	1.1e-01	9	3935	7.85	6.26
GO:0040007		growth	4.4e-02	54	3890	7.41	6.26
GO:0050793		regulation of development	2.5e-01	54	3890	6.82	6.26
GO:0048468		cell development	1.0e-00	7	3937	6.69	6.26
GO:0007548		sex differentiation	6.7e-01	11	3933	5.81	6.27
GO:0000003		reproduction	7.5e-02	46	3898	5.74	6.27
GO:0007582	**BP**	**physiological process**	**1.4e-03**	**2935**	**1009**	**6.14**	**6.61**
GO:0044419		interaction between organisms	1.0e-01	5	3939	10.05	6.26
GO:0048511		rhythmic process	8.1e-02	13	3931	8.00	6.26
GO:0050791		regulation of physiological process	3.9e-13	824	3120	7.22	6.04
GO:0043062		extracellular structure organization and biogenesis	4.6e-01	10	3934	7.22	6.26
GO:0050874		organismal physiological process	8.2e-01	465	3479	6.33	6.26
GO:0050896		response to stimulus	4.7e-01	491	3453	6.26	6.27
GO:0050875		cellular physiological process	6.0e-02	2671	1273	6.17	6.48
GO:0008152		metabolism	2.5e-03	2048	1896	6.12	6.46
GO:0051179		localization	2.5e-01	595	3349	5.99	6.33
GO:0016265		death	3.3e-01	145	3799	5.82	6.28
GO:0046903		secretion	2.2e-01	55	3889	5.68	6.27
GO:0050817		coagulation	1.1e-01	27	3917	5.00	6.27
GO:0042592		homeostasis	5.3e-02	25	3919	4.85	6.27
GO:0050789	**BP**	**regulation of biological process**	**3.3e-14**	**891**	**3053**	**7.26**	**6.02**
GO:0040029		regulation of gene expression, epigenetic	1.1e-01	9	3935	7.85	6.26
GO:0050794		regulation of cellular process	6.2e-15	807	3137	7.36	6.04
GO:0050791		regulation of physiological process	3.9e-13	824	3120	7.22	6.04
GO:0050793		regulation of development	2.5e-01	54	3890	6.82	6.26
GO:0050790		regulation of enzyme activity	4.8e-01	49	3895	6.64	6.26

Continue at next page

Table 4.3 (continue)

GO Term Accession#	Term Definition Type	Name	Test p-value	#genes Yes	#genes No	Median Yes	Median No
GO:0048518		positive regulation of biological process	4.4e-01	143	3801	6.34	6.26
GO:0048519		negative regulation of biological process	1.8e-01	177	3767	6.21	6.27
GO:0016032	**BP**	**viral life cycle**	**6.4e-01**	**9**	**3935**	**5.96**	**6.27**
GO:0019058		viral infectious cycle	9.6e-01	5	3939	5.96	6.27
GO:0005623	**CC**	**cell**	**2.0e-03**	**2800**	**1144**	**6.38**	**6.06**
GO:0009986		cell surface	3.4e-01	6	3938	6.53	6.26
GO:0016020		membrane	2.8e-02	1254	2690	6.42	6.17
GO:0042995		cell projection	3.2e-01	16	3928	6.46	6.26
GO:0000267		cell fraction	4.7e-01	283	3661	6.38	6.26
GO:0005622		intracellular	2.5e-01	1948	1996	6.14	6.33
GO:0031012	**CC**	**extracellular matrix**	**1.1e-01**	**115**	**3829**	**6.89**	**6.25**
GO:0005578		extracellular matrix (sensu Metazoa)	1.1e-01	115	3829	6.89	6.25
GO:0005576	**CC**	**extracellular region**	**2.2e-01**	**313**	**3631**	**6.62**	**6.25**
GO:0005578		extracellular matrix (sensu Metazoa)	1.1e-01	115	3829	6.89	6.25
GO:0005615		extracellular space	1.2e-01	89	3855	6.89	6.25
GO:0043226	**CC**	**organelle**	**9.1e-01**	**1641**	**2303**	**6.26**	**6.27**
GO:0043233		organelle lumen	4.1e-01	7	3937	8.47	6.26
GO:0043228		non-membrane-bound organelle	6.7e-01	315	3629	6.27	6.26
GO:0043229		intracellular organelle	9.1e-01	1641	2303	6.26	6.27
GO:0043227		membrane-bound organelle	9.8e-01	1466	2478	6.26	6.27
GO:0043234	**CC**	**protein complex**	**4.0e-03**	**463**	**3481**	**5.82**	**6.31**
GO:0008076		voltage-gated potassium channel complex	5.3e-06	20	3924	11.07	6.25
GO:0030880		RNA polymerase complex	2.1e-02	10	3934	9.14	6.26
GO:0016585		chromatin remodeling complex	5.8e-01	7	3937	7.70	6.26
GO:0016591		DNA-directed RNA polymerase II, holoenzyme	8.8e-01	28	3916	7.46	6.26
GO:0005667		transcription factor complex	5.7e-01	35	3909	7.31	6.26
GO:0005834		heterotrimeric G-protein complex	4.5e-01	9	3935	7.28	6.26
GO:0000786		nucleosome	5.6e-01	7	3937	6.80	6.26
GO:0005875		microtubule associated complex	5.6e-01	31	3913	6.23	6.27
GO:0043235		receptor complex	4.4e-01	21	3923	5.83	6.27
GO:0008180		signalosome complex	6.9e-01	5	3939	5.64	6.27
GO:0016010		dystrophin-associated glycoprotein complex	8.5e-01	6	3938	5.63	6.26
GO:0016282		eukaryotic 43S preinitiation complex	4.4e-01	10	3934	5.52	6.26
GO:0005941		unlocalized protein complex	5.5e-01	16	3928	5.48	6.27
GO:0016469		proton-transporting two-sector ATPase complex	7.4e-01	14	3930	5.35	6.27
GO:0000151		ubiquitin ligase complex	2.7e-01	68	3876	5.32	6.27
GO:0045259		proton-transporting ATP synthase complex	5.7e-01	6	3938	5.29	6.26
GO:0030894		replisome	3.5e-01	7	3937	4.88	6.27
GO:0000502		proteasome complex (sensu Eukaryota)	1.7e-02	20	3924	4.66	6.27
GO:0016012		sarcoglycan complex	7.9e-01	5	3939	4.66	6.27
GO:0016011		dystroglycan complex	7.9e-01	5	3939	4.66	6.27
GO:0030529		ribonucleoprotein complex	2.5e-05	99	3845	4.65	6.31
GO:0016209	**MF**	**antioxidant activity**	**9.9e-01**	**11**	**3933**	**7.49**	**6.26**
GO:0004601		peroxidase activity	6.0e-01	9	3935	7.51	6.26
GO:0005488	**MF**	**binding**	**1.1e-06**	**2303**	**1641**	**6.56**	**5.87**

Continue at next page

Table 4.3 (continue)

GO Term Accession#	Term Definition Type	Name	Test p-value	#genes Yes	#genes No	Median Yes	Median No
GO:0042165		neurotransmitter binding	6.9e-02	35	3909	7.60	6.26
GO:0043021		ribonucleoprotein binding	5.0e-01	6	3938	7.42	6.26
GO:0001871		pattern binding	6.3e-02	31	3913	7.40	6.26
GO:0005102		receptor binding	7.7e-04	161	3783	7.18	6.24
GO:0003676		nucleic acid binding	1.4e-06	778	3166	6.94	6.12
GO:0005515		protein binding	6.1e-05	779	3165	6.78	6.12
GO:0043176		amine binding	6.6e-01	12	3932	6.67	6.26
GO:0043167		ion binding	1.0e-00	579	3365	6.42	6.25
GO:0005496		steroid binding	9.1e-01	13	3931	6.31	6.26
GO:0000166		nucleotide binding	5.4e-02	504	3440	6.19	6.27
GO:0030246		carbohydrate binding	9.9e-01	65	3879	6.12	6.27
GO:0008289		lipid binding	1.4e-01	59	3885	5.61	6.27
GO:0042277		peptide binding	2.4e-01	42	3902	5.52	6.27
GO:0046906		tetrapyrrole binding	2.9e-01	8	3936	5.03	6.27
GO:0019840		isoprenoid binding	3.8e-01	6	3938	4.89	6.27
GO:0048037		cofactor binding	6.6e-03	20	3924	4.42	6.27
GO:0019842		vitamin binding	1.3e-03	16	3928	3.90	6.27
GO:0008144		drug binding	2.7e-02	5	3939	3.58	6.27
GO:0003824	**MF**	**catalytic activity**	**4.0e-14**	**1403**	**2541**	**5.72**	**6.57**
GO:0008639		small protein conjugating enzyme activity	1.5e-01	21	3923	7.28	6.26
GO:0016740		transferase activity	4.1e-02	470	3474	6.10	6.30
GO:0016829		lyase activity	2.6e-02	49	3895	5.54	6.27
GO:0016787		hydrolase activity	1.7e-06	554	3390	5.59	6.39
GO:0009975		cyclase activity	5.1e-01	8	3936	5.46	6.27
GO:0016874		ligase activity	6.4e-02	133	3811	5.41	6.28
GO:0016491		oxidoreductase activity	3.5e-05	194	3750	5.20	6.32
GO:0016853		isomerase activity	2.5e-02	35	3909	5.08	6.28
GO:0004386		helicase activity	7.8e-02	37	3907	4.82	6.27
GO:0019239		deaminase activity	5.5e-02	7	3937	4.24	6.27
GO:0030234	**MF**	**enzyme regulator activity**	**7.6e-01**	**151**	**3793**	**6.13**	**6.27**
GO:0019207		kinase regulator activity	1.2e-01	11	3933	8.00	6.26
GO:0019208		phosphatase regulator activity	1.4e-01	12	3932	7.96	6.26
GO:0030695		GTPase regulator activity	6.0e-01	58	3886	6.76	6.26
GO:0004857		enzyme inhibitor activity	6.4e-01	63	3881	5.80	6.28
GO:0008047		enzyme activator activity	4.4e-01	62	3882	5.74	6.27
GO:0003774	**MF**	**motor activity**	**7.6e-01**	**32**	**3912**	**6.62**	**6.26**
GO:0003777		microtubule motor activity	6.1e-01	21	3923	6.60	6.26
GO:0004871	**MF**	**signal transducer activity**	**4.0e-06**	**623**	**3321**	**7.04**	**6.14**
GO:0005057		receptor signaling protein activity	1.6e-01	46	3898	7.23	6.26
GO:0005102		receptor binding	7.7e-04	161	3783	7.18	6.24
GO:0004872		receptor activity	3.1e-02	359	3585	6.79	6.24
GO:0005070		SH3/SH2 adaptor activity	2.7e-01	19	3925	6.72	6.26
GO:0005198	**MF**	**structural molecule activity**	**2.4e-01**	**157**	**3787**	**6.68**	**6.26**
GO:0008307		structural constituent of muscle	1.8e-02	6	3938	11.05	6.26
GO:0005200		structural constituent of cytoskeleton	3.1e-02	18	3926	9.35	6.26
GO:0005201		extracellular matrix structural constituent	4.1e-02	23	3921	8.13	6.26
GO:0005212		structural constituent of eye lens	5.5e-01	6	3938	7.21	6.26
GO:0003735		structural constituent of ribosome	6.5e-04	46	3898	4.32	6.29
GO:0030528	**MF**	**transcription regulator activity**	**0.0e+00**	**360**	**3584**	**8.47**	**6.10**
GO:0003700		transcription factor activity	0.0e+00	259	3685	9.33	6.12
GO:0003702		RNA polymerase II transcription factor activity	1.8e-05	84	3860	8.12	6.24

Continue at next page

Table 4.3 (continue)

GO Term Accession#	Term Definition Type	Name	Test p-value	#genes Yes	#genes No	Median Yes	Median No
GO:0016564		transcriptional repressor activity	4.5e-01	19	3925	8.06	6.26
GO:0003712		transcription cofactor activity	4.6e-03	96	3848	7.22	6.25
GO:0003711		transcriptional elongation regulator activity	8.1e-01	5	3939	6.50	6.26
GO:0016563		transcriptional activator activity	6.6e-01	15	3929	5.72	6.27
GO:0045182	**MF**	**translation regulator activity**	**9.9e-01**	**36**	**3908**	**6.64**	**6.26**
GO:0008135		translation factor activity, nucleic acid binding	9.4e-01	35	3909	6.80	6.26
GO:0005215	**MF**	**transporter activity**	**1.8e-01**	**451**	**3493**	**5.99**	**6.30**
GO:0005326		neurotransmitter transporter activity	2.7e-01	6	3938	9.11	6.26
GO:0005319		lipid transporter activity	8.6e-02	12	3932	7.89	6.26
GO:0015457		auxiliary transport protein activity	1.6e-01	11	3933	7.35	6.26
GO:0015267		channel or pore class transporter activity	1.5e-01	107	3837	6.89	6.25
GO:0015075		ion transporter activity	8.0e-01	202	3742	6.10	6.27
GO:0042626		ATPase activity, coupled to transmembrane movement of substances	5.8e-01	33	3911	6.01	6.27
GO:0005275		amine transporter activity	4.7e-01	26	3918	5.65	6.27
GO:0005342		organic acid transporter activity	3.6e-01	33	3911	5.56	6.27
GO:0005489		electron transporter activity	7.3e-02	91	3853	5.57	6.29
GO:0008565		protein transporter activity	3.5e-01	39	3905	5.35	6.27
GO:0005386		carrier activity	1.8e-03	144	3800	5.14	6.32
GO:0015144		carbohydrate transporter activity	2.6e-02	12	3932	4.66	6.27
GO:0005478		intracellular transporter activity	2.9e-01	10	3934	4.34	6.27

Note: The green-colored row indicates the genes in the GO category are significantly more conserved than those not. The red-colored row indicates the genes in the GO category are significantly less conserved than those not. The level 3 GO categories within a level 2 GO category are sorted by the differences of the median alignment scores between the genes in a GO category and those not, so the most conserved GO category is at the top and the least conserved at the bottom.

Table 4.4: Top 10 upstream conserved human-mouse orthologous genes

Official	RefSeq Accession #		Alignment	Conserved Regions	
Gene symbol	human	mouse	Score	Length(*bp*)	Identity(%)
KIAA0319L	NM_024874	NM_133886	275286	4402	81.8
HOXD10	NM_002148	NM_013554	261336	4248	81.4
POU3F3	NM_006236	NM_008900	260227	3874	81.9
BCOR	NM_020926	NM_029510	253203	4222	81.4
HOXB5	NM_002147	NM_008268	233581	3892	79.8
FGF11	NM_004112	NM_010198	231404	4278	77.1
HOXA2	NM_006735	NM_010451	230558	3724	82.1
TOP3A	NM_004618	NM_009410	226746	2988	83.5
UBTF	NM_014233	NM_011551	222929	3347	83.2
PLXNC1	NM_005761	NM_018797	219730	2820	84.8
			Average	3780	81.70

Table 4.5: GO categories of top 30 promoter conserved human-mouse orthologs. The GO category columns indicate whether a gene belongs to one or more of the 9 most significantly overrepresented GO categories in the top 50 list.

Symbol	GO:0007275	GO:0009653	GO:0048513	GO:0030528	GO:0003700	GO:0050789	GO:0050791	GO:0050794	GO:0003676	Gene Name
KIAA0319L	0	0	0	0	0	0	0	0	0	KIAA0319-like
HOXD10	1	0	0	1	1	1	1	1	1	homeo box D10
POU3F3	1	1	1	1	1	1	1	1	1	POU domain, class 3, transcription factor 3
BCOR	0	0	0	1	0	1	1	1	0	BCL6 co-repressor
HOXB5	1	1	0	1	1	1	1	1	1	homeo box B5
FGF11	1	1	1	0	0	0	0	0	0	fibroblast growth factor 11
HOXA2	1	0	0	1	1	1	1	1	1	homeo box A2
TOP3A	0	0	0	0	0	0	0	0	1	topoisomerase (DNA) III alpha
UBTF	0	0	0	1	0	1	1	1	1	upstream binding transcription factor, RNA polymerase I
PLXNC1	1	0	0	0	0	0	0	0	0	plexin C1
HOXA5	1	0	0	1	1	1	1	1	1	homeo box A5
SALL1	1	1	0	1	1	1	1	1	1	sal-like 1 (Drosophila)
HOXC4	1	0	0	1	1	1	1	1	1	homeo box C4
PRSS25	0	0	0	0	0	0	0	0	0	protease, serine, 25
LRRC4	0	0	0	0	0	0	0	0	0	leucine rich repeat containing 4
PCDH7	0	0	0	0	0	0	0	0	0	BH-protocadherin (brain-heart)
GRWD1	0	0	0	0	0	0	0	0	0	glutamate-rich WD repeat containing 1
POU3F4	1	1	1	1	1	1	1	1	1	POU domain, class 3, transcription factor 4
NOG	1	1	1	0	0	1	0	1	0	noggin
PCDH8	0	0	0	0	0	0	0	0	0	protocadherin 8
HMG20A	0	0	0	1	1	1	1	1	1	high-mobility group 20A
STOML2	0	0	0	0	0	0	0	0	0	stomatin (EPB72)-like 2
ZNFN1A2	0	0	0	0	0	1	1	1	1	zinc finger protein, subfamily 1A, 2 (Helios)
ZNF238	0	0	0	1	1	1	1	1	1	zinc finger protein 238
NNAT	1	1	1	0	0	0	0	0	0	neuronatin
TBR1	1	1	1	1	1	1	1	1	1	T-box, brain, 1
FGF10	1	1	1	0	0	1	1	1	0	fibroblast growth factor 10
TLE3	1	1	1	0	0	1	1	1	0	transducin-like enhancer of split 3 (E(sp1) homolog, Drosophila)
MEIS1	0	0	0	1	1	1	1	1	1	Meis1, myeloid ecotropic viral integration site 1 homolog (mouse)

Note: Genes are listed in the descending order of local alignment scores of their upstream regions (upstream 3.5 *kp* and downstream 1k *kp* of TSS).

Figures

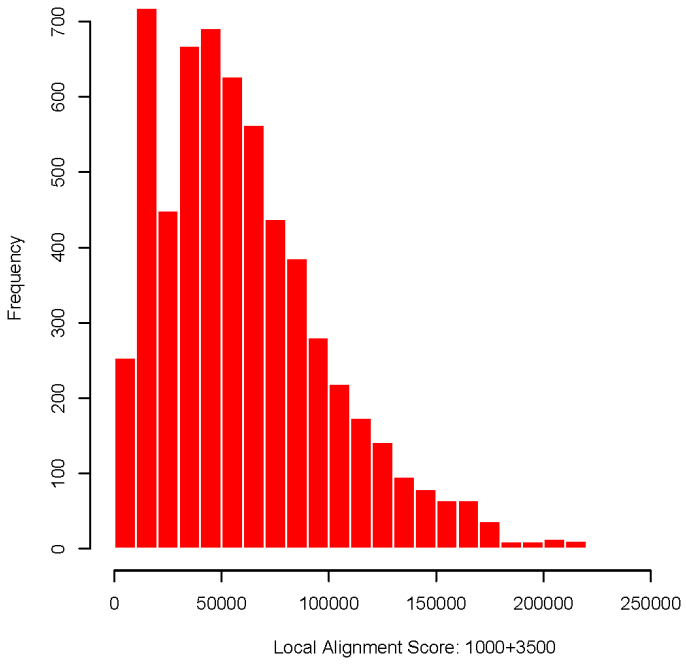


Figure 4.1: The distribution of local alignment scores of the human-mouse 4.5 *kb* regulatory sequences (3.5 *kb* upstream and 1.0 *kp* downstream of TSS)

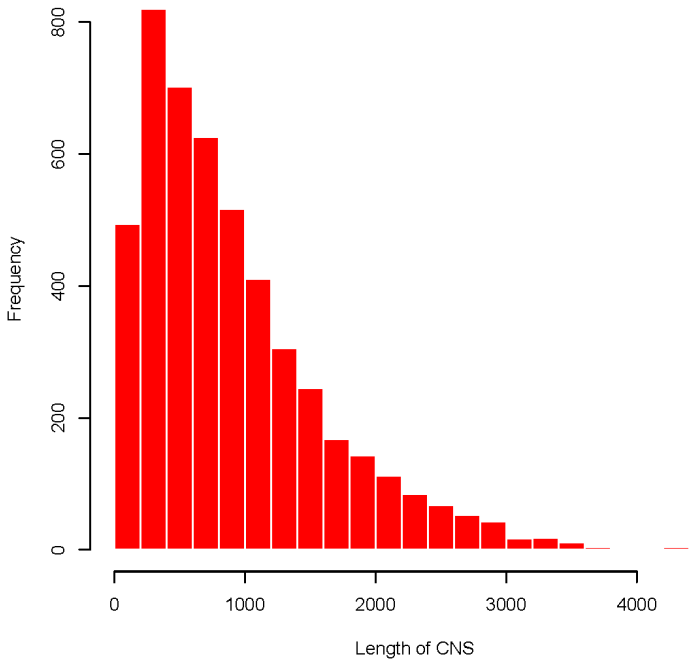


Figure 4.2: The length distribution of the conserved regions (identity $\geq 75\%$) in the 4.5 *kb* regulatory sequences (3.5 *kb* upstream and 1.0 *kp* downstream of TSS) of human-mouse orthologs.

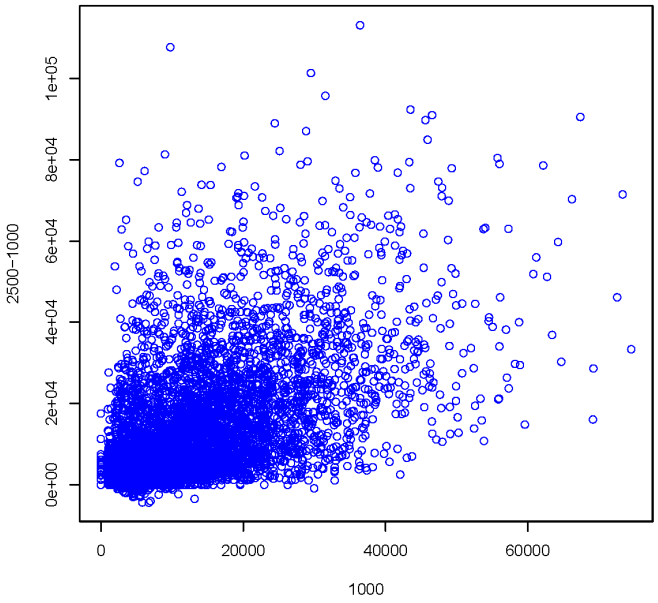


Figure 4.3: The plot of the differences of alignment scores of 2500 *bp* and those of 1000 *bp* upstream regions of TSS of human-mouse orthologs *vs* those of 1000 *bp* upstream regions

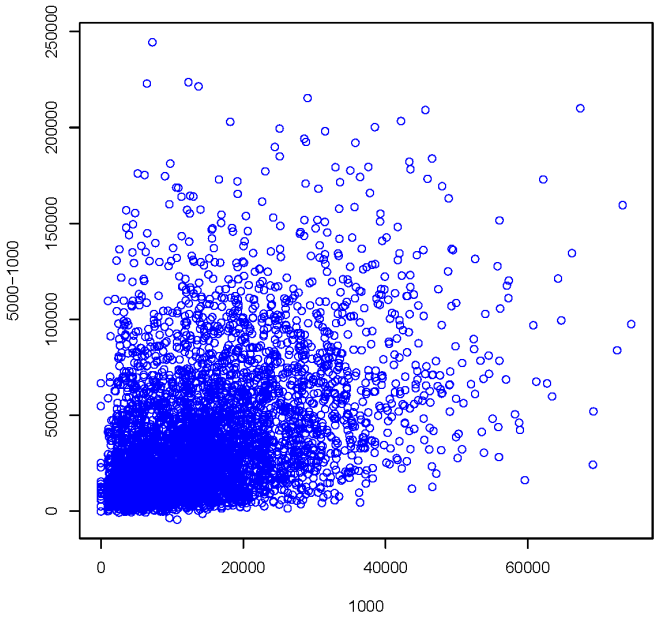


Figure 4.4: The plot of the differences of alignment scores of 5000 *bp* and those of 1000 *bp* upstream regions of TSS of human-mouse orthologs *vs* those of 1000 *bp* upstream regions

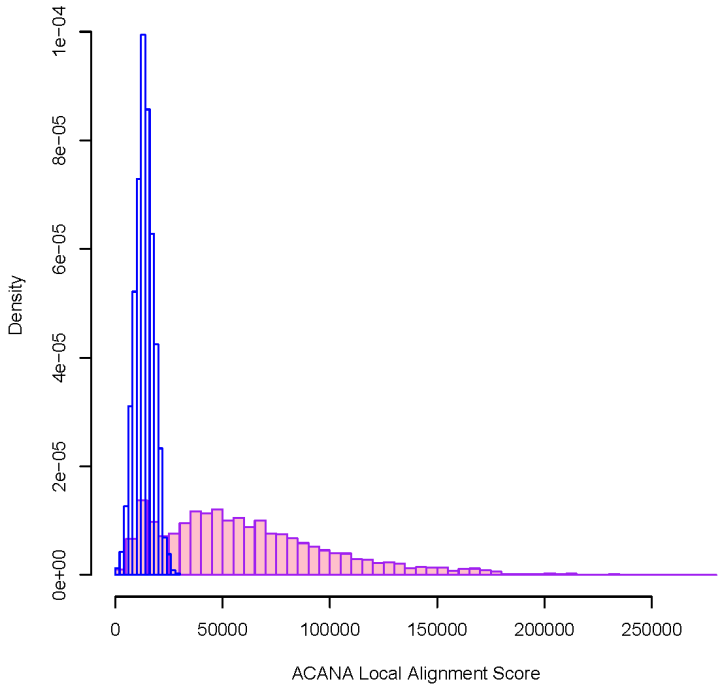


Figure 4.5: Comparison of the ACANA alignment score distribution (red) of upstream sequences (3.5 *kb* upstream and 1.0 *kp* downstream of TSS) of the human-mouse orthologs and that (blue) of randomly permuted pairs.

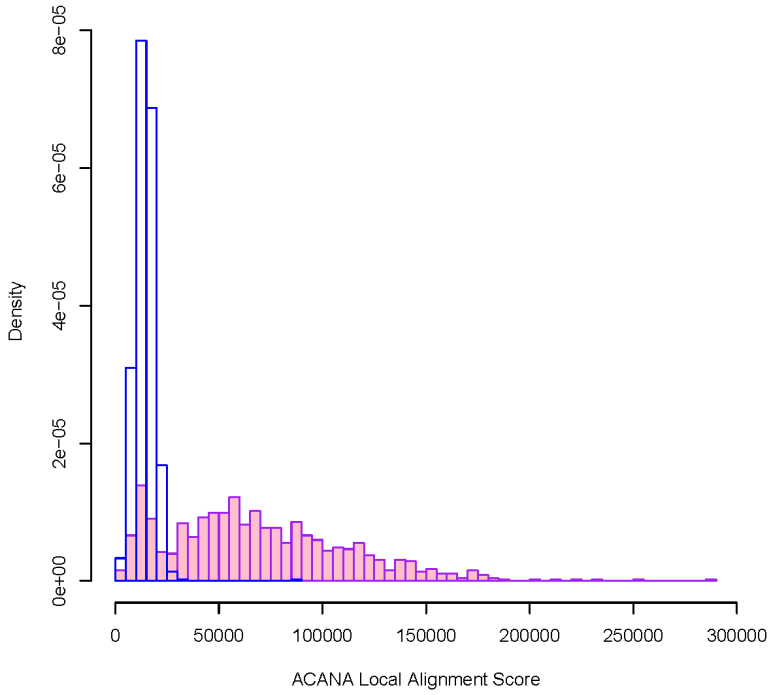
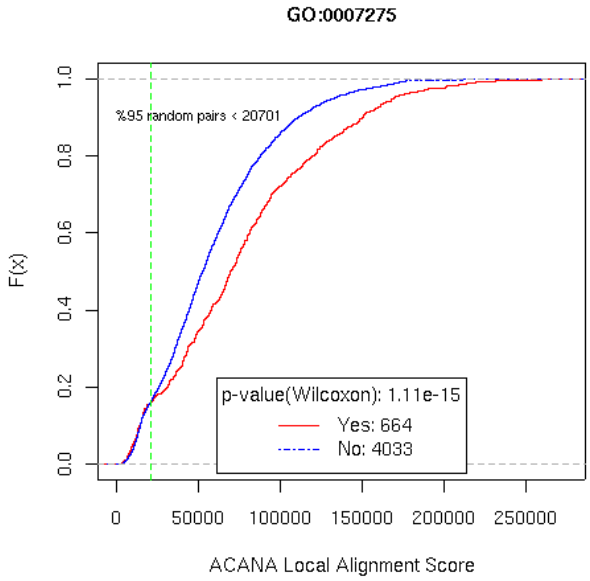
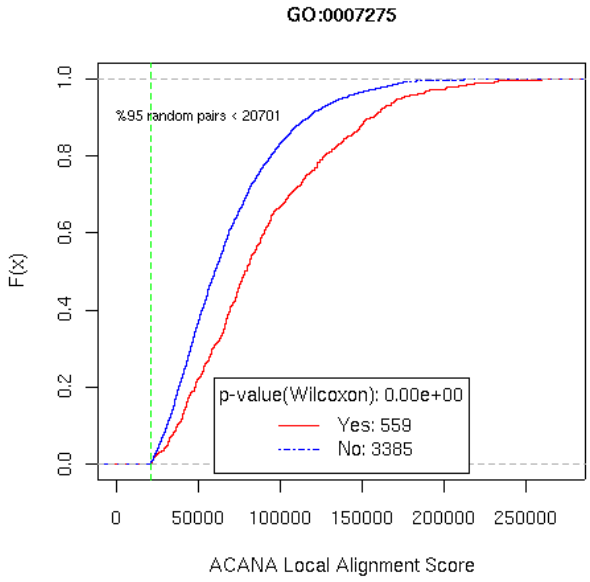


Figure 4.6: Comparison of the ACANA alignment score distribution (red) of upstream sequences (3.5 *kb* upstream and 1.0 *kp* downstream of TSS) of the human-rat orthologs and that (blue) of randomly permuted pairs.

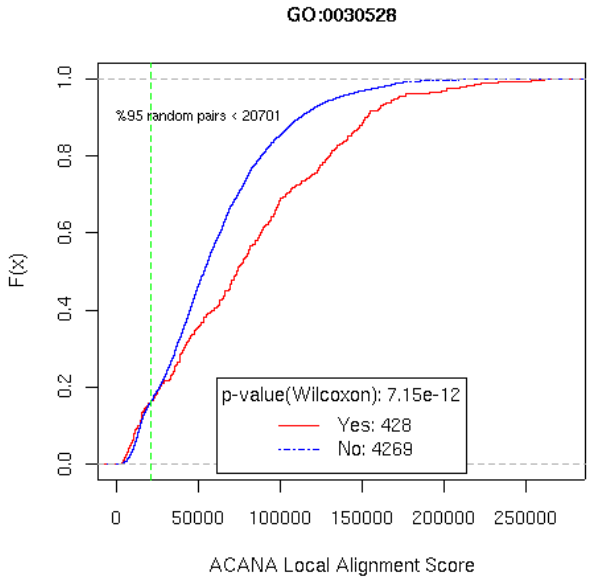


(a) Using all putative human-mouse orthologs

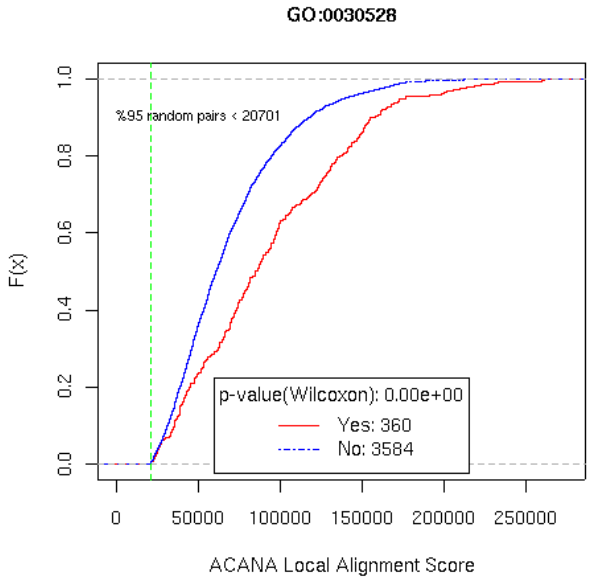


(b) Using only putative orthologs of alignment scores $\geq E_{.95}$

Figure 4.7: Comparison of empirical CDF plots of ACANA local alignment scores of two functional classes of human-mouse orthologs: development-related (GO:0007275) and non-development.



(a) Using all putative human-mouse orthologs



(b) Using only putative orthologs of alignment scores $\geq E_{.95}$

Figure 4.8: Empirical CDF of upstream region alignment scores for human-mouse orthologs with and without transcription regulator activity (GO:0030528), respectively.

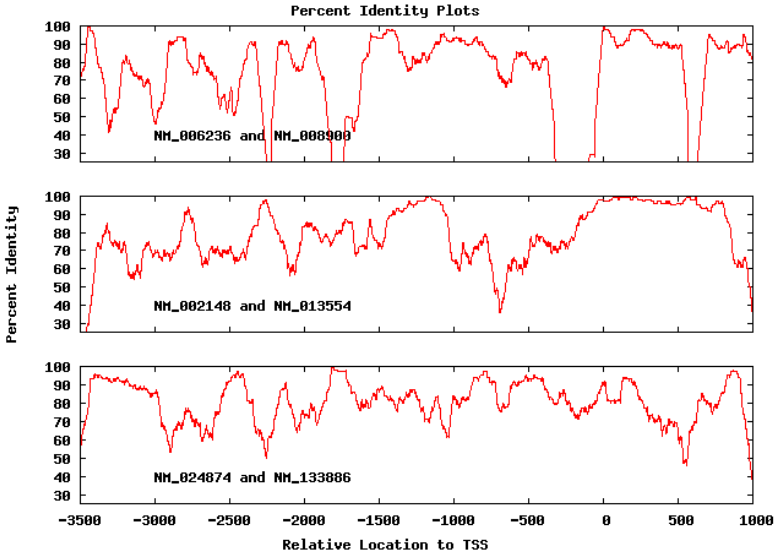


Figure 4.9: Percent identity plots for the top 3 upstream conserved human-mouse orthologs.

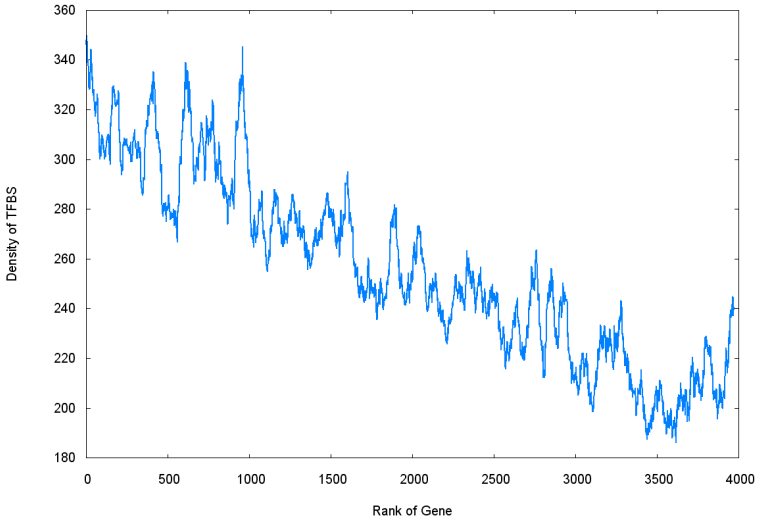
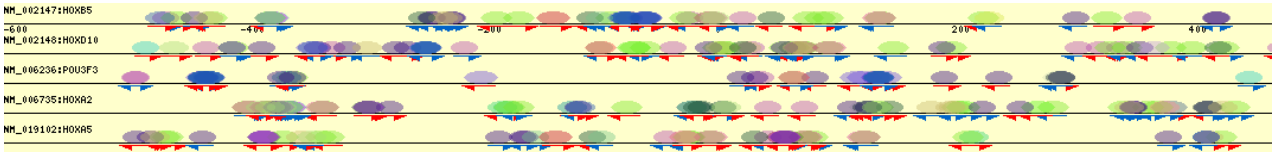
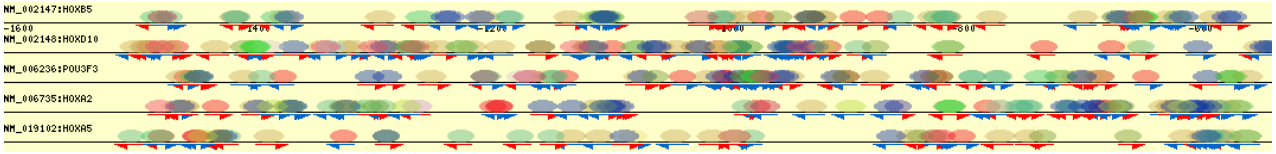


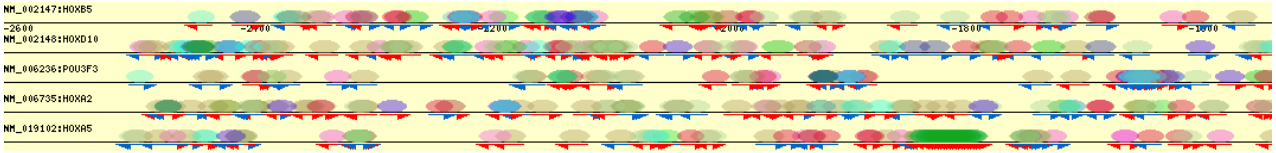
Figure 4.10: The plot shows that TFBS density in the upstream 4.5 *kb* region generally decreases as gene rank increases (the degree of conservation decreases)



(a) Region: -500 to 500



(b) Region: -1500 to -500



(c) Region: -2500 to -1500

Figure 4.11: The *cis*-element map in the promoter regions, -2500 to 500 relative to TSS, of the top listed transcription factor encoding genes.

Appendix: supplementary materials

Table 4.6: Statistical tests of the association between gene function and the degree of upstream region conservation of human-mouse orthologs. The Wilcoxon rank sum test p-values are for the two tails tests of hypothesis that there is no difference in the distribution of alignment scores between the genes in a GO category and those not. In the table, the rows in bold font are level 2 GO terms, and the remaining rows are level 3 GO terms. The last column is the median of ACANA alignment scores of genes in a GO category (Yes) and those not (No).

GO Term	Term Definition		Wilcoxon Test p-value			
Accession#	Type	Name	3k5+1k	1k	2k5	5k
GO:0007610	**BP**	**behavior**	**1.4e-02**	**7.4e-01**	**2.8e-01**	**7.9e-03**
GO:0007631		feeding behavior	1.4e-01	1.1e-01	6.4e-02	3.1e-01
GO:0007611		learning and/or memory	7.3e-01	4.2e-01		
GO:0007626		locomotory behavior	5.2e-02	9.0e-01	4.7e-01	2.8e-02
GO:0009987	**BP**	**cellular process**	**3.7e-02**	**8.9e-02**	**1.7e-01**	**2.2e-01**
GO:0030154		cell differentiation	2.1e-08	1.4e-05	1.7e-08	1.1e-06
GO:0007154		cell communication	0.0e+00	4.8e-08	7.5e-05	4.5e-04
GO:0050794		regulation of cellular process	6.2e-15	4.1e-13	1.4e-09	1.8e-08
GO:0050875		cellular physiological process	6.0e-02	3.6e-01	4.7e-01	3.3e-01
GO:0007275	**BP**	**development**	**0.0e+00**	**1.1e-13**	**3.2e-13**	**6.7e-16**
GO:0007389		pattern specification	2.5e-02	2.0e-02	1.5e-01	3.7e-02
GO:0009790		embryonic development	2.6e-03	5.9e-03	1.3e-02	4.7e-02
GO:0007498		mesoderm development	7.3e-02	5.2e-01	6.5e-02	7.5e-02
GO:0030154		cell differentiation	2.1e-08	1.4e-05	1.7e-08	1.1e-06
GO:0048513		organ development	0.0e+00	6.3e-12	1.7e-12	1.5e-13
GO:0009653		morphogenesis	0.0e+00	2.8e-11	1.6e-11	2.1e-12
GO:0040029		regulation of gene expression, epigenetic	1.1e-01	4.1e-01	4.1e-01	5.4e-01
GO:0040007		growth	4.4e-02	2.6e-01	2.8e-01	1.0e-01
GO:0050793		regulation of development	2.5e-01	1.0e-01	5.6e-01	2.3e-01
GO:0048468		cell development	1.0e-00	9.5e-01	6.5e-01	5.3e-01
GO:0007548		sex differentiation	6.7e-01	1.9e-01	9.2e-01	5.6e-01
GO:0000003		reproduction	7.5e-02	6.6e-02	1.1e-01	6.4e-02
GO:0007582	**BP**	**physiological process**	**1.4e-03**	**2.3e-01**	**1.3e-01**	**1.3e-01**
GO:0044419		interaction between organisms	1.0e-01	3.1e-01	2.3e-01	
GO:0048511		rhythmic process	8.1e-02	1.4e-01	9.3e-02	1.8e-01
GO:0050791		regulation of physiological process	3.9e-13	2.0e-10	5.0e-08	1.6e-07
GO:0043062		extracellular structure organization and biogenesis	4.6e-01	6.4e-01	8.9e-01	4.1e-01
GO:0050874		organismal physiological process	8.2e-01	5.1e-01	6.2e-01	4.8e-01
GO:0050896		response to stimulus	4.7e-01	4.7e-01	3.2e-01	1.4e-01
GO:0050875		cellular physiological process	6.0e-02	3.6e-01	4.7e-01	3.3e-01
GO:0008152		metabolism	2.5e-03	2.0e-01	8.5e-01	4.1e-01
GO:0051179		localization	2.5e-01	8.5e-01	2.2e-01	4.5e-02
GO:0016265		death	3.3e-01	4.1e-01	1.7e-01	7.1e-01
GO:0046903		secretion	2.2e-01	1.1e-01	7.0e-02	1.2e-01
GO:0050817		coagulation	1.1e-01	4.5e-01	3.3e-01	1.8e-01
GO:0042592		homeostasis	5.3e-02	8.6e-01	3.0e-01	4.6e-01
GO:0050789	**BP**	**regulation of biological process**	**3.3e-14**	**1.8e-11**	**1.2e-08**	**6.5e-07**
GO:0040029		regulation of gene expression, epigenetic	1.1e-01	4.1e-01	4.1e-01	5.4e-01
GO:0050794		regulation of cellular process	6.2e-15	4.1e-13	1.4e-09	1.8e-08
GO:0050791		regulation of physiological process	3.9e-13	2.0e-10	5.0e-08	1.6e-07

Continue at next page

Table 4.6 (continue)

GO Term Accession#	Term Definition Type	Name	Wilcoxon Test p-value 3k5+1k	1k	2k5	5k
GO:0050793		regulation of development	2.5e-01	1.0e-01	5.6e-01	2.3e-01
GO:0050790		regulation of enzyme activity	4.8e-01	4.5e-01	5.8e-01	4.7e-01
GO:0048518		positive regulation of biological process	4.4e-01	9.5e-01	3.2e-01	1.1e-01
GO:0048519		negative regulation of biological process	1.8e-01	7.2e-01	2.6e-01	7.1e-03
GO:0016032	**BP**	**viral life cycle**	**6.4e-01**	**9.5e-01**	**9.9e-01**	**4.1e-01**
GO:0019058		viral infectious cycle	9.6e-01	6.3e-01	9.6e-01	
GO:0005623	**CC**	**cell**	**2.0e-03**	**2.8e-03**	**2.6e-02**	**8.9e-02**
GO:0009986		cell surface	3.4e-01	3.0e-01	4.5e-01	9.4e-01
GO:0016020		membrane	2.8e-02	5.3e-01	6.3e-01	8.9e-01
GO:0042995		cell projection	3.2e-01	7.5e-01	6.7e-01	6.6e-01
GO:0000267		cell fraction	4.7e-01	8.1e-01	5.7e-01	9.6e-01
GO:0005622		intracellular	2.5e-01	8.7e-01	6.5e-01	5.2e-01
GO:0031012	**CC**	**extracellular matrix**	**1.1e-01**	**2.2e-02**	**9.4e-02**	**2.9e-01**
GO:0005578		extracellular matrix (sensu Metazoa)	1.1e-01	2.2e-02	9.4e-02	2.9e-01
GO:0005576	**CC**	**extracellular region**	**2.2e-01**	**5.4e-03**	**1.7e-01**	**4.7e-01**
GO:0005578		extracellular matrix (sensu Metazoa)	1.1e-01	2.2e-02	9.4e-02	2.9e-01
GO:0005615		extracellular space	1.2e-01	4.1e-01	5.4e-01	6.9e-01
GO:0043226	**CC**	**organelle**	**9.1e-01**	**8.8e-01**	**2.5e-01**	**4.2e-02**
GO:0043233		organelle lumen	4.1e-01	5.1e-01	2.2e-01	6.3e-01
GO:0043228		non-membrane-bound organelle	6.7e-01	5.5e-02	7.9e-01	4.3e-01
GO:0043229		intracellular organelle	9.1e-01	8.8e-01	2.5e-01	4.2e-02
GO:0043227		membrane-bound organelle	9.8e-01	5.1e-01	2.1e-01	7.0e-02
GO:0043234	**CC**	**protein complex**	**4.0e-03**	**1.4e-01**	**2.8e-02**	**5.4e-02**
GO:0008076		voltage-gated potassium channel complex	5.3e-06	2.2e-03	3.9e-04	6.6e-03
GO:0030880		RNA polymerase complex	2.1e-02	6.8e-03	7.1e-03	1.8e-02
GO:0016585		chromatin remodeling complex	5.8e-01	2.4e-01	4.0e-01	2.4e-01
GO:0016591		DNA-directed RNA polymerase II, holoenzyme	8.8e-01	1.5e-01	4.2e-01	6.5e-01
GO:0005667		transcription factor complex	5.7e-01	9.4e-01	8.2e-01	1.8e-01
GO:0005834		heterotrimeric G-protein complex	4.5e-01	2.0e-01	1.6e-01	2.9e-01
GO:0000786		nucleosome	5.6e-01	7.4e-01	8.5e-01	3.8e-01
GO:0005875		microtubule associated complex	5.6e-01	9.3e-01	4.1e-01	4.8e-01
GO:0043235		receptor complex	4.4e-01	9.2e-01	8.7e-01	1.4e-01
GO:0008180		signalosome complex	6.9e-01	6.8e-01	2.9e-01	
GO:0016010		dystrophin-associated glycoprotein complex	8.5e-01	7.0e-01	7.7e-01	
GO:0016282		eukaryotic 43S preinitiation complex	4.4e-01	8.1e-01	7.7e-02	3.3e-01
GO:0005941		unlocalized protein complex	5.5e-01	7.6e-01	2.6e-01	8.5e-02
GO:0016469		proton-transporting two-sector ATPase complex	7.4e-01	3.3e-01	4.0e-01	3.0e-01
GO:0000151		ubiquitin ligase complex	2.7e-01	8.8e-01	1.3e-01	1.0e-01
GO:0045259		proton-transporting ATP synthase complex	5.7e-01	6.5e-01	6.3e-01	2.6e-01
GO:0030894		replisome	3.5e-01	3.0e-02	5.5e-01	5.7e-01
GO:0000502		proteasome complex (sensu Eukaryota)	1.7e-02	9.5e-04	2.6e-02	1.4e-01
GO:0016012		sarcoglycan complex	7.9e-01	7.4e-01		
GO:0016011		dystroglycan complex	7.9e-01	7.4e-01		
GO:0030529		ribonucleoprotein complex	2.5e-05	2.2e-03	6.9e-03	4.2e-02
GO:0016209	**MF**	**antioxidant activity**	**9.9e-01**	**3.4e-01**	**4.5e-01**	**7.8e-01**

Continue at next page

Table 4.6 (continue)

GO Term	Term Definition		Wilcoxon Test p-value			
Accession#	Type	Name	3k5+1k	1k	2k5	5k
GO:0004601		peroxidase activity	6.0e-01	5.4e-01	9.8e-01	9.9e-01
GO:0005488	**MF**	**binding**	**1.1e-06**	**1.2e-06**	**1.8e-05**	**2.6e-04**
GO:0042165		neurotransmitter binding	6.9e-02	3.1e-01	2.7e-01	9.0e-01
GO:0043021		ribonucleoprotein binding	5.0e-01	3.0e-01	9.8e-01	9.7e-01
GO:0001871		pattern binding	6.3e-02	1.4e-01	4.6e-02	3.3e-01
GO:0005102		receptor binding	7.7e-04	1.1e-03	2.5e-02	5.8e-02
GO:0003676		nucleic acid binding	1.4e-06	2.4e-06	9.8e-07	4.2e-05
GO:0005515		protein binding	6.1e-05	4.3e-02	1.3e-02	3.4e-02
GO:0043176		amine binding	6.6e-01	9.1e-01	4.3e-01	9.5e-01
GO:0043167		ion binding	1.0e-00	3.2e-01	6.0e-01	3.0e-01
GO:0005496		steroid binding	9.1e-01	3.0e-01	6.1e-01	2.0e-01
GO:0000166		nucleotide binding	5.4e-02	3.0e-01	4.4e-02	3.4e-02
GO:0030246		carbohydrate binding	9.9e-01	5.3e-01	2.8e-01	7.4e-01
GO:0008289		lipid binding	1.4e-01	8.5e-01	1.4e-01	5.1e-01
GO:0042277		peptide binding	2.4e-01	6.6e-02	5.0e-01	6.4e-01
GO:0046906		tetrapyrrole binding	2.9e-01	9.1e-01	6.1e-01	6.1e-01
GO:0019840		isoprenoid binding	3.8e-01	9.4e-01	5.0e-01	6.9e-01
GO:0048037		cofactor binding	6.6e-03	5.6e-01	4.5e-01	9.7e-01
GO:0019842		vitamin binding	1.3e-03	2.7e-01	1.2e-01	6.3e-01
GO:0008144		drug binding	2.7e-02			
GO:0003824	**MF**	**catalytic activity**	**4.0e-14**	**6.8e-11**	**2.5e-07**	**3.1e-07**
GO:0008639		small protein conjugating enzyme activity	1.5e-01	3.7e-02	5.5e-02	4.7e-01
GO:0016740		transferase activity	4.1e-02	2.4e-03	1.2e-01	4.1e-02
GO:0016829		lyase activity	2.6e-02	9.2e-02	7.3e-01	4.9e-01
GO:0016787		hydrolase activity	1.7e-06	8.8e-05	2.0e-04	1.9e-04
GO:0009975		cyclase activity	5.1e-01	8.0e-01	5.8e-01	1.7e-01
GO:0016874		ligase activity	6.4e-02	7.7e-01	1.6e-01	3.5e-02
GO:0016491		oxidoreductase activity	3.5e-05	9.5e-03	6.4e-02	4.1e-01
GO:0016853		isomerase activity	2.5e-02	2.6e-02	3.5e-02	1.1e-01
GO:0004386		helicase activity	7.8e-02	2.1e-01	2.5e-02	1.1e-02
GO:0019239		deaminase activity	5.5e-02	8.6e-01	4.1e-01	
GO:0030234	**MF**	**enzyme regulator activity**	**7.6e-01**	**5.1e-01**	**8.6e-01**	**5.0e-01**
GO:0019207		kinase regulator activity	1.2e-01	3.3e-02	3.2e-02	1.5e-01
GO:0019208		phosphatase regulator activity	1.4e-01	6.4e-02	3.7e-01	2.2e-01
GO:0030695		GTPase regulator activity	6.0e-01	4.1e-01	7.3e-01	7.2e-01
GO:0004857		enzyme inhibitor activity	6.4e-01	8.8e-01	9.1e-01	8.2e-01
GO:0008047		enzyme activator activity	4.4e-01	9.2e-01	8.5e-01	9.6e-01
GO:0003774	**MF**	**motor activity**	**7.6e-01**	**8.2e-01**	**9.3e-01**	**5.6e-01**
GO:0003777		microtubule motor activity	6.1e-01	8.3e-01	8.9e-01	8.1e-01
GO:0004871	**MF**	**signal transducer activity**	**4.0e-06**	**2.7e-04**	**9.4e-03**	**3.7e-01**
GO:0005057		receptor signaling protein activity	1.6e-01	2.9e-01	2.9e-01	9.8e-01
GO:0005102		receptor binding	7.7e-04	1.1e-03	2.5e-02	5.8e-02
GO:0004872		receptor activity	3.1e-02	1.9e-02	6.5e-01	6.1e-01
GO:0005070		SH3/SH2 adaptor activity	2.7e-01	2.8e-01	7.2e-01	4.3e-01
GO:0005198	**MF**	**structural molecule activity**	**2.4e-01**	**5.3e-01**	**3.7e-02**	**1.4e-02**
GO:0008307		structural constituent of muscle	1.8e-02	9.8e-02	5.6e-02	5.0e-03
GO:0005200		structural constituent of cytoskeleton	3.1e-02	7.2e-01	9.1e-01	1.4e-01
GO:0005201		extracellular matrix structural constituent	4.1e-02	2.4e-02	1.1e-01	3.7e-01
GO:0005212		structural constituent of eye lens	5.5e-01	7.8e-01	2.9e-01	1.3e-01
GO:0003735		structural constituent of ribosome	6.5e-04	5.0e-02	1.4e-01	3.5e-01
GO:0030528	**MF**	**transcription regulator activity**	**0.0e+00**	**8.7e-13**	**1.1e-15**	**1.5e-11**
GO:0003700		transcription factor activity	0.0e+00	5.1e-13	0.0e+00	5.5e-14

Continue at next page

Table 4.6 (continue)

GO Term	Term Definition		Wilcoxon Test p-value			
Accession#	Type	Name	3k5+1k	1k	2k5	5k
GO:0003702		RNA polymerase II transcription factor activity	1.8e-05	1.4e-03	5.9e-04	2.8e-02
GO:0016564		transcriptional repressor activity	4.5e-01	1.3e-01	1.2e-01	7.1e-02
GO:0003712		transcription cofactor activity	4.6e-03	3.8e-02	2.3e-02	1.8e-01
GO:0003711		transcriptional elongation regulator activity	8.1e-01	8.8e-01	4.5e-01	
GO:0016563		transcriptional activator activity	6.6e-01	6.4e-01	8.6e-01	6.0e-01
GO:0045182	**MF**	**translation regulator activity**	**9.9e-01**	**7.7e-01**	**5.1e-01**	**7.4e-01**
GO:0008135		translation factor activity, nucleic acid binding	9.4e-01	7.7e-01	4.6e-01	5.9e-01
GO:0005215	**MF**	**transporter activity**	**1.8e-01**	**7.9e-01**	**1.7e-01**	**3.8e-01**
GO:0005326		neurotransmitter transporter activity	2.7e-01	3.3e-01	3.4e-01	1.7e-01
GO:0005319		lipid transporter activity	8.6e-02	6.8e-02	4.0e-01	1.1e-01
GO:0015457		auxiliary transport protein activity	1.6e-01	7.0e-01	2.0e-02	9.4e-02
GO:0015267		channel or pore class transporter activity	1.5e-01	1.3e-01	1.2e-01	6.9e-01
GO:0015075		ion transporter activity	8.0e-01	8.9e-01	5.5e-01	8.6e-01
GO:0042626		ATPase activity, coupled to transmembrane movement of substances	5.8e-01	5.4e-01	4.1e-01	2.1e-01
GO:0005275		amine transporter activity	4.7e-01	8.5e-01	6.2e-01	3.7e-01
GO:0005342		organic acid transporter activity	3.6e-01	8.1e-01	6.4e-01	1.7e-01
GO:0005489		electron transporter activity	7.3e-02	6.3e-01	3.3e-01	3.7e-01
GO:0008565		protein transporter activity	3.5e-01	5.8e-01	5.0e-01	5.3e-01
GO:0005386		carrier activity	1.8e-03	2.7e-01	1.3e-02	1.8e-01
GO:0015144		carbohydrate transporter activity	2.6e-02	3.5e-01	8.3e-02	2.5e-01
GO:0005478		intracellular transporter activity	2.9e-01	4.2e-01	2.0e-01	2.6e-01

Note: The green-colored row indicates the genes in the GO category are significantly more conserved than those not. The red-colored row indicates the genes in the GO category are significantly less conserved than those not. The level 3 GO categories within a level 2 GO category are sorted by the differences of the median alignment scores between the genes in a GO category and those not, so the most conserved GO category is at the top and the least conserved at the bottom.

Table 4.7: Statistical tests of the association between gene function and the degree of upstream region conservation of human-rat orthologs. The table showed the Wilcoxon rank sum test results for 4.5 *kb* upstream region (3.5k *bp* upstream and 1k *bp* downstream of TSS). In the table, the rows in bold font are level 2 GO terms, and the remaining rows are level 3 GO terms. The last column is the median of ACANA alignment scores of genes in a GO category (Yes) and those not (No).

GO Term	Term Definition		Test	#genes		Median	
Accession#	Type	Name	p-value	Yes	No	Yes	No
GO:0007610	**BP**	**behavior**	**9.2e-01**	**24**	**735**	**6.80**	**7.17**
GO:0007631		feeding behavior	1.0e+00	7	752	7.27	7.12
GO:0007626		locomotory behavior	9.3e-01	13	746	6.56	7.20
GO:0009987	**BP**	**cellular process**	**2.8e-01**	**682**	**77**	**7.23**	**6.71**
GO:0030154		cell differentiation	1.4e-02	31	728	9.78	7.07
GO:0050794		regulation of cellular process	8.2e-04	164	595	7.98	6.82
GO:0007154		cell communication	1.3e-04	279	480	7.85	6.70
GO:0050875		cellular physiological process	5.5e-03	587	172	6.86	8.13
GO:0007275	**BP**	**development**	**7.7e-07**	**142**	**617**	**8.90**	**6.75**
GO:0030154		cell differentiation	1.4e-02	31	728	9.78	7.07
GO:0048513		organ development	2.3e-06	90	669	9.46	6.85
GO:0009653		morphogenesis	4.1e-07	116	643	9.05	6.75
GO:0040007		growth	7.2e-01	25	734	8.05	7.12
GO:0050793		regulation of development	6.5e-01	20	739	7.55	7.14
GO:0000003		reproduction	1.2e-01	10	749	4.92	7.20
GO:0007582	**BP**	**physiological process**	**2.5e-03**	**637**	**122**	**6.87**	**8.33**
GO:0042592		homeostasis	1.5e-01	12	747	10.60	7.10
GO:0050791		regulation of physiological process	2.2e-03	159	600	7.95	6.86
GO:0050874		organismal physiological process	5.1e-01	133	626	7.58	7.06
GO:0051179		localization	2.3e-01	193	566	6.88	7.27
GO:0016265		death	9.5e-01	32	727	6.54	7.17
GO:0050896		response to stimulus	1.5e-01	106	653	6.50	7.34
GO:0046903		secretion	1.9e-01	23	736	6.27	7.19
GO:0050817		coagulation	2.1e-01	7	752	6.18	7.16
GO:0050875		cellular physiological process	5.5e-03	587	172	6.86	8.13
GO:0008152		metabolism	1.1e-04	399	360	6.58	7.88
GO:0050789	**BP**	**regulation of biological process**	**2.2e-03**	**183**	**576**	**7.76**	**6.83**
GO:0050794		regulation of cellular process	8.2e-04	164	595	7.98	6.82
GO:0050791		regulation of physiological process	2.2e-03	159	600	7.95	6.86
GO:0050793		regulation of development	6.5e-01	20	739	7.55	7.14
GO:0048519		negative regulation of biological process	6.4e-01	39	720	7.40	7.12
GO:0048518		positive regulation of biological process	1.2e-01	33	726	6.41	7.28
GO:0050790		regulation of enzyme activity	4.3e-01	16	743	6.17	7.20
GO:0005623	**CC**	**cell**	**2.5e-01**	**586**	**173**	**7.27**	**6.75**
GO:0016020		membrane	1.7e-01	318	441	7.37	6.86
GO:0000267		cell fraction	9.9e-01	85	674	7.34	7.11
GO:0005622		intracellular	5.8e-02	353	406	6.79	7.59
GO:0031012	**CC**	**extracellular matrix**	**5.9e-01**	**28**	**731**	**7.28**	**7.11**
GO:0005578		extracellular matrix (sensu Metazoa)	5.9e-01	28	731	7.28	7.11
GO:0005576	**CC**	**extracellular region**	**4.0e-01**	**67**	**692**	**7.58**	**7.08**
GO:0005615		extracellular space	2.6e-01	17	742	7.87	7.12
GO:0005578		extracellular matrix (sensu Metazoa)	5.9e-01	28	731	7.28	7.11

Continue at next page

Table 4.7 (continue)

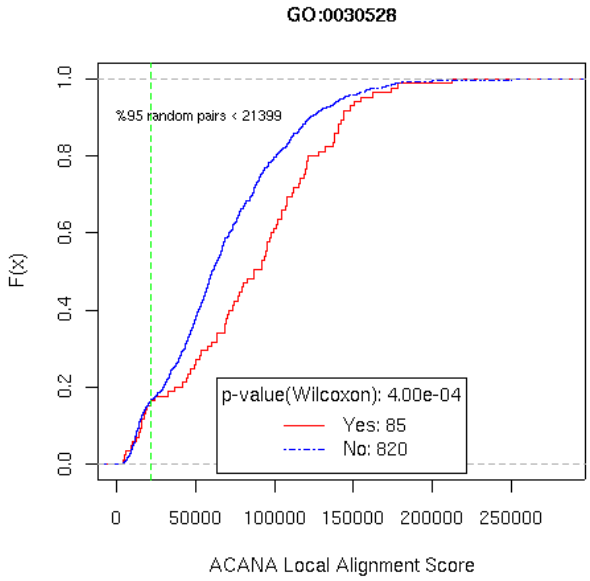
GO Term	Term Definition		Test	#genes		Median	
Accession#	Type	Name	p-value	Yes	No	Yes	No
GO:0043226	**CC**	**organelle**	**1.9e-01**	**278**	**481**	**6.85**	**7.39**
GO:0043228		non-membrane-bound organelle	4.8e-01	46	713	7.27	7.14
GO:0043229		intracellular organelle	1.9e-01	278	481	6.85	7.39
GO:0043227		membrane-bound organelle	4.3e-02	252	507	6.68	7.42
GO:0043234	**CC**	**protein complex**	**5.6e-02**	**90**	**669**	**6.27**	**7.34**
GO:0008076		voltage-gated potassium channel complex	1.1e-01	10	749	9.66	7.10
GO:0005891		voltage-gated calcium channel complex	1.3e-01	6	753	9.54	7.10
GO:0005667		transcription factor complex	7.8e-01	6	753	6.67	7.14
GO:0005941		unlocalized protein complex	5.9e-01	7	752	6.58	7.16
GO:0030529		ribonucleoprotein complex	6.9e-01	5	754	6.39	7.19
GO:0000502		proteasome complex (sensu Eukaryota)	4.7e-02	7	752	5.88	7.23
GO:0043235		receptor complex	6.4e-02	8	751	5.68	7.26
GO:0005875		microtubule associated complex	5.6e-01	5	754	5.31	7.16
GO:0000151		ubiquitin ligase complex	3.2e-01	9	750	5.28	7.16
GO:0005488	**MF**	**binding**	**2.8e-02**	**480**	**279**	**7.37**	**6.72**
GO:0003676		nucleic acid binding	3.8e-03	98	661	8.89	6.89
GO:0005102		receptor binding	3.9e-03	35	724	8.74	7.05
GO:0030246		carbohydrate binding	2.6e-01	6	753	7.98	7.11
GO:0005515		protein binding	7.6e-03	166	593	7.72	7.05
GO:0042277		peptide binding	3.9e-01	18	741	7.77	7.10
GO:0043167		ion binding	6.0e-01	102	657	6.98	7.17
GO:0042165		neurotransmitter binding	2.8e-01	25	734	6.64	7.19
GO:0000166		nucleotide binding	9.7e-02	127	632	6.40	7.31
GO:0005496		steroid binding	1.9e-01	6	753	6.17	7.14
GO:0008289		lipid binding	2.5e-01	20	739	5.91	7.20
GO:0043176		amine binding	1.7e-01	9	750	5.71	7.19
GO:0003824	**MF**	**catalytic activity**	**5.1e-07**	**297**	**462**	**6.10**	**7.75**
GO:0009975		cyclase activity	9.6e-01	6	753	6.64	7.17
GO:0016740		transferase activity	2.6e-01	126	633	6.66	7.28
GO:0016787		hydrolase activity	2.9e-02	109	650	6.35	7.34
GO:0016829		lyase activity	3.0e-01	20	739	6.15	7.20
GO:0016874		ligase activity	1.2e-01	18	741	5.57	7.20
GO:0016853		isomerase activity	8.0e-02	6	753	4.90	7.20
GO:0016491		oxidoreductase activity	6.1e-06	37	722	5.00	7.35
GO:0030234	**MF**	**enzyme regulator activity**	**8.2e-01**	**39**	**720**	**6.57**	**7.20**
GO:0019207		kinase regulator activity	3.6e-03	8	751	12.17	7.10
GO:0030695		GTPase regulator activity	6.9e-01	10	749	7.23	7.14
GO:0004857		enzyme inhibitor activity	9.3e-01	17	742	6.85	7.16
GO:0008047		enzyme activator activity	9.9e-01	19	740	6.57	7.19
GO:0003774	**MF**	**motor activity**	**5.4e-01**	**9**	**750**	**8.23**	**7.11**
GO:0004871	**MF**	**signal transducer activity**	**4.9e-05**	**188**	**571**	**8.03**	**6.72**
GO:0005102		receptor binding	3.9e-03	35	724	8.74	7.05
GO:0005057		receptor signaling protein activity	2.4e-01	16	743	8.28	7.10
GO:0004872		receptor activity	2.5e-02	123	636	7.81	6.88
GO:0005198	**MF**	**structural molecule activity**	**4.1e-03**	**30**	**729**	**9.30**	**7.06**
GO:0005201		extracellular matrix structural constituent	3.8e-02	6	753	9.44	7.10
GO:0030528	**MF**	**transcription regulator activity**	**1.1e-05**	**72**	**687**	**9.63**	**6.85**
GO:0003700		transcription factor activity	9.2e-06	53	706	9.98	6.87
GO:0003702		RNA polymerase II transcription factor activity	6.2e-02	18	741	9.83	7.09
GO:0016563		transcriptional activator activity	3.0e-01	5	754	9.55	7.12

Continue at next page

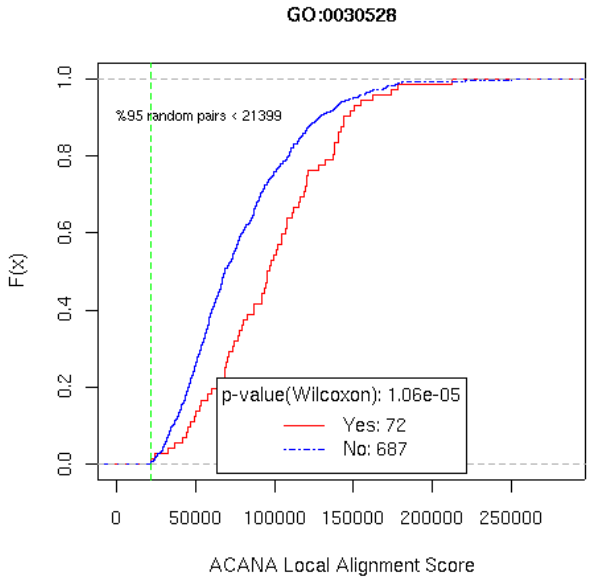
Table 4.7 (continue)

GO Term	Term Definition		Test	#genes		Median	
Accession#	Type	Name	p-value	Yes	No	Yes	No
GO:0003712		transcription cofactor activity	6.3e-01	15	744	8.30	7.10
GO:0005215	**MF**	**transporter activity**	**1.1e-01**	**160**	**599**	**6.70**	**7.33**
GO:0015267		channel or pore class transporter activity	3.9e-01	64	695	7.91	7.09
GO:0015075		ion transporter activity	6.2e-01	90	669	7.62	7.10
GO:0005386		carrier activity	7.5e-01	34	725	7.15	7.14
GO:0005275		amine transporter activity	6.9e-01	15	744	6.59	7.16
GO:0005326		neurotransmitter transporter activity	2.9e-01	7	752	6.01	7.16
GO:0005342		organic acid transporter activity	6.9e-01	11	748	6.01	7.16
GO:0042626		ATPase activity, coupled to transmembrane movement of substances	7.8e-01	11	748	5.68	7.16
GO:0005489		electron transporter activity	4.0e-03	20	739	5.31	7.27
GO:0005478		intracellular transporter activity	2.7e-03	5	754	3.87	7.20

Note: The green-colored row indicates the genes in the GO category are significantly more conserved than those not. The red-colored row indicates the genes in the GO category are significantly less conserved than those not. The level 3 GO categories within a level 2 GO category are sorted by the differences of the median alignment scores between the genes in a GO category and those not, so the most conserved GO category is at the top and the least conserved at the bottom.

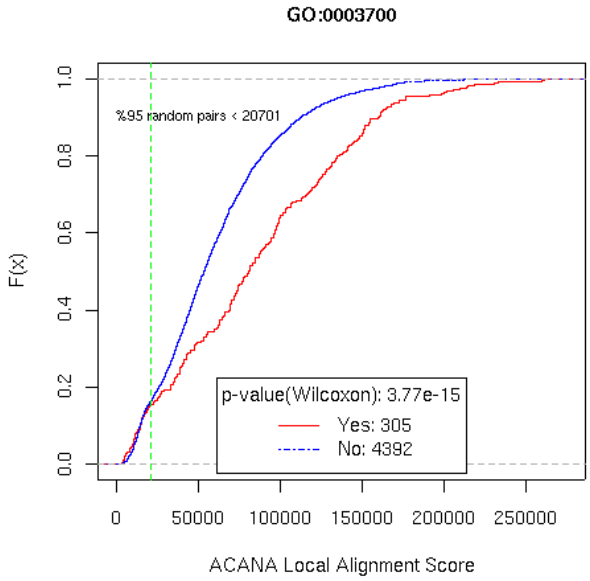


(a) Using all putative human-mouse orthologs

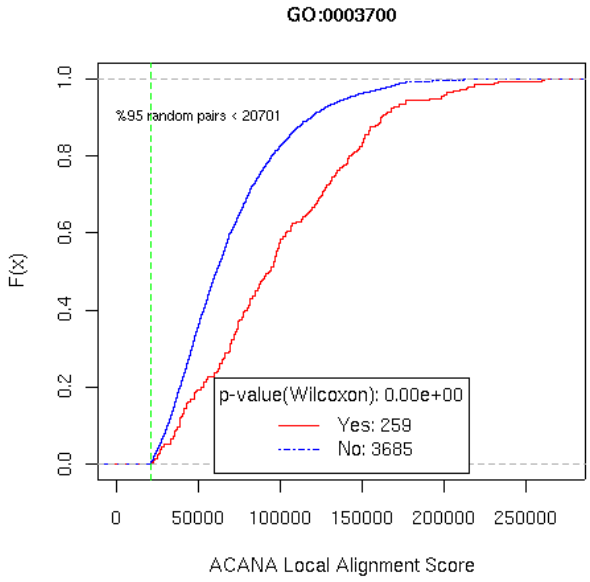


(b) Using only putative orthologs of alignment scores $\geq E_{.95}$

Figure 4.12: Empirical CDF of upstream region alignment scores for human-rat orthologs with and without transcription regulator activity (GO:0030528), respectively.

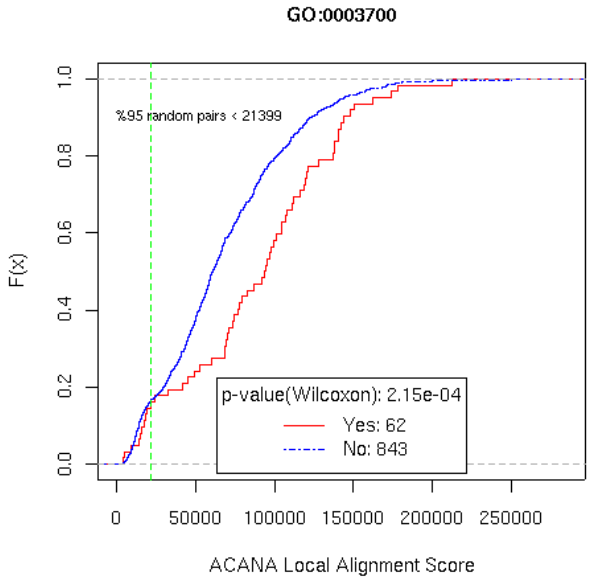


(a) Using all putative human-mouse orthologs

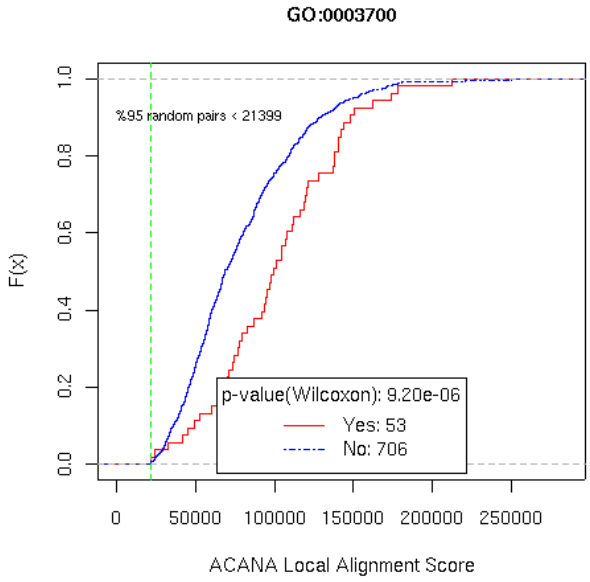


(b) Using only putative orthologs of alignment scores $\geq E_{.95}$

Figure 4.13: Empirical CDF of upstream region alignment scores for human-mouse orthologs with and without transcription factor activity (GO:0003700), respectively.



(a) Using all putative human-mouse orthologs



(b) Using only putative orthologs of alignment scores $\geq E_{.95}$

Figure 4.14: Empirical CDF of upstream region alignment scores for human-rat orthologs with and without transcription factor activity (GO:0003700), respectively.

Bibliography

Altschul, S. F. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol.*, 266:460–480.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.

Bergman, C. M. and Kreitman, M. (2001). Analysis of Conserved Noncoding DNA in Drosophila Reveals Similar Constraints in Intergenic and Intronic Sequences. *Genome Res.*, 11(8):1335–1345.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *PNAS*, 99(2):757–762.

Doniger, S. W., Huh, J., and Fay, J. C. (2005). Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Res.*, 15(5):701–709.

Ferrier, D. E. K. and Minguillón, C. (2003). Evolution of the Hox/ParaHox gene clusters. *Int. J. Dev. Biol.*, 47(7-8):605–611.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, 32(Web Server issue):W273–W279.

Garcia-Fernàndez, J. (2005). Hox, ParaHox, ProtoHox: facts and guesses. *Heredity*, 94(2):145–152.

Halligan, D. L., Eyre-Walker, A., Andolfatto, P., and Keightley, P. D. (2004). Patterns of

Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila. *Genome Res.*, 14(2):273–279.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., and Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32(Database issue):D258–D261.

Huang, W., Umbach, D. M., and Li, L. (2005). Accurate anchoring alignment of divergent sequences. accepted by Bioinformatics.

Ilia, M. (2004). Oct-6 transcription factor. *Int. Rev. Neurobiol.*, 59:471–489.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, 16(9):418–420.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254.

Kmita, M. and Duboule, D. (2003). Organizing axes in time and space; 25 years of colinear tinkering. *Science*, 301(5631):331–333.

Kmita, M., Tarchini, B., Zàny, J., Logan, M., Tabin, C. J., and Duboule, D. (2005). Early developmental arrest of mammalian limbs lacking HoxA/HoxD gene function. *Nature*, 435(7045):1113–1116.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., 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, J., 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, J., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R., Wilson, R., Hillier, L., McPherson, J., Marra, M., Mardis, E., Fulton, L., Chinwalla, A., Pepin, K., Gish, W., Chissoe, S., Wendl, M., Delehaunty, K., Miner, T., Delehaunty, A., Kramer, J., Cook, L., Fulton, R., Johnson, D., Minx, P., Clifton, S., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R., Muzny, D., Scherer, S., Bouck, J., Sodergren, E., Worley, K., Rives, C., Gorrell, J., Metzker, M., Naylor, S., Kucherlapati, R., Nelson, D., Weinstock, G., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R., Federspiel, N., Abola, A., Proctor, M., Myers, R., Schmutz, J., Dickson, M., Grimwood, J., Cox, D., Olson, M., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G., Athanasiou, M., Schultz, R., Roe, B., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W., de la Bastide, M., Dedhia, N., Blör, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D., Burge, C., Cerutti, L., Chen, H., Church, D., Clamp, M., Copley,

R., Doerks, T., Eddy, S., Eichler, E., Furey, T., Galagan, J., Gilbert, J., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L., Jones, T., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W., Kitts, P., Koonin, E., Korf, I., Kulp, D., Lancet, D., Lowe, T., McLysaght, A., Mikkelsen, T., Moran, J., Mulder, N., Pollara, V., Ponting, C., Schuler, G., Schultz, J., Slater, G., Smit, A., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y., Wolfe, K., Yang, S., Yeh, R., Collins, F., Guyer, M., Peterson, J., Felsenfeld, A., Wetterstrand, K., Patrinos, A., Morgan, M., de Jong, P., Catanese, J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y., Szustakowki, J., and Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Levine, M. and Davidson, E. H. (2005). From the Cover. Gene Regulatory Networks Special Feature: Gene regulatory networks for development. *PNAS*, 102(14):4936–4942.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I. (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047.

Negre, B., Casillas, S., Suzanne, M., Sanchez-Herrero, E., Akam, M., Nefedov, M., Barbadilla, A., de Jong, P., and Ruiz, A. (2005). Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. *Genome Res.*, 15(5):692–700.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.*, 15(1):137–145.

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29(2):153–159.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, 23(23):4878–4884.

Santini, S., Boore, J. L., and Meyer, A. (2003). Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res.*, 13(6A):1111–1122.

Schonemann, M. D., Ryan, A. K., Erkman, L., McEvilly, R. J., Bermingham, J., and Rosenfeld, M. G. (1998). POU domain factors in neural development. *Adv. Exp. Med. Biol.*, 449:39–53.

Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., Gocayne, J., Amanatides, P., Ballew, R., Huson, D., Wortman, J., Zhang, Q., Kodira, C., Zheng, X., Chen, L., Skupski, M., Subramanian, G., Thomas, P., Zhang, J., Miklos, G. G., Nelson, C., Broder, S., Clark, A., Nadeau, J., McKusick, V., Zinder, N., Levine, A., Roberts, R., 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, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T., Higgins, M., Ji, R., Ke, Z., Ketchum, K., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G., Milshina, N., Moore, H., Naik, A., Narayan, V., Neelam, B., Nusskern, D., Rusch, D., 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, M., 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, Y., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J., Campbell, R. G. M., Sjolander, K., 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, Y., 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., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vilo, J., Brazma, A., Jonassen, I., Robinson, A., and Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:384–394.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla,

A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, A. C. V., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-

P., Zdobnov, E. M., Zody, M. C., Lander, E. S., and Consortium, M. G. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Wegner, M., Drolet, D. W., and Rosenfeld, M. G. (1993). POU-domain proteins: structure and function of developmental regulators. *Curr. Opin. Cell Biol.*, 5(3):488–498.

Zhang, T.-Y., Kang, Zhang, Z.-F., and Xu, W.-H. (2004). Identification of a POU factor involved in regulating the neuron-specific expression of the gene encoding diapause hormone and pheromone biosynthesis-activating neuropeptide in Bombyx mori. *Biochem. J.*, 380(Pt 1):255–263.