Quantifying the strength of natural selection of a motif sequence

Chen-Hsiang Yeang[1]

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, chyeang@stat.sinica.edu.tw.

Abstract. Quantification of selective pressures on regulatory sequences is a central question in studying the evolution of gene regulatory networks. Previous methods focus primarily on single sites rather than motif sequences. We propose a method of evaluating the strength of natural selection of a motif from a family of aligned sequences. The method is based on a Poisson process model of neutral sequence substitutions and derives a birth-death process of the motif occurrence frequencies. The selection coefficient is treated as a penalty for the motif death rate. We demonstrate that the birth-death model closely approximates statistics generated from simulated data and the Poisson process assumption holds in mammalian promoter sequences. Furthermore, we show that a considerably higher portion of known transcription factor binding motifs possess high selection coefficients compared to negative controls with high occurrence frequencies on promoters. Comparison of SP1 and TP53 binding motifs indicates that a higher portion of the SP1 motifs are conserved between the orthologous promoters of human and other mammals. Preliminary analysis supports the potential applications of the model to identify regulatory sequences under selection.

Summary

Motivation

Many sequence motifs are present in multiple locations of the genomes and are utilized in various contexts. Prominent examples include protein-binding sites on DNAs or RNAs and domain sequences on proteins. Due to their functional constraints, selective pressures are often exerted on the evolution of sequence motifs. Quantification of selective pressures on motifs is a powerful tool to study the evolution of biological systems (such as gene regulatory networks) and to identify functionally important motifs (such as transcription factor binding motifs) from sequence data. However, most existing methods either target specific sites (rather than motif sequences) or apply only to two species. Consequently, a quantitative model and algorithm to evaluate the selective strength of motif sequences from multiple species need to be developed.

Main results

We propose a simple model of neutral (i.e., selection-free) evolution of motif sequences of fixed lengths. The model hypothesizes that each position undergoes an independent random sequence substitution. In a long stretch of sequence (e.g., a 5Kb promoter of a gene), the occurrence of a motif sequence results from stochastic additions (birth) and removals (death) based on sequence substitutions. We derive the differential equations of the neutral evolutionary model on motifs and confirm the consistency of the model on simulated data. Furthermore, we augment the neutral model with a parameter of selective strength and develop an algorithm to evaluate the selective strength of a motif from aligned sequences. To demonstrate its utility, we calculate the selective strengths of known transcription factor binding motifs from a curated database (TRANSFAC) and random motifs on aligned 5Kb promoters of 27667 genes across 34 mammalian genomes. A higher proportion of TRANSFAC motifs have strong selective strengths compared to random controls (1/4 versus 1/20). In contrast, a conservation score fails to separate TRANSFAC motifs and the short random motifs that occur frequently on promoters.

Significance

Initial analysis indicates that the birth-death model is adequate for the neutral evolution of motifs. Furthermore, selective coefficients outperform conservation scores in separating functional motifs from random sequences. The results justify the use of our model and algorithm in studying the evolution of functional motifs and identifying de novo functional motifs.

1 Introduction

High sequence similarity of many genes between distant species raises a fundamental question in molecular evolution: how do diverse phenotypes arise from the highly similar protein families? Many recent studies have shifted focus to the evolution of non protein-coding regions, as the sequence substitutions and recombinations of cis-regulatory elements may drive the evolution of gene regulatory networks. One central issue in the study of non protein-coding regions is to gauge the selective pressure of a sequence motif. Cis-regulatory elements or regulatory RNAs may possess strong sequence specificity and resist random drifts. It is therefore possible to identify these elements from the sequences of multiple genes and organisms. One can align the promoter sequences of orthologous genes and apply motif-finding algorithms to identify the conserved motifs [1]. Conservation alone, however, may not confer natural selection since it also depends on the rate of neutral evolution, sequence length and complexity, population structure, and other factors. A variety of methods have been proposed to detect/quantify natural selection from sequences, including the ratios of non-synonymous to synonymous substitution rates $\frac{K_a}{K_s}$, [2], likelihood scores from a background sequence substitution model [3], comparison of intra-specific variation versus inter-specific divergence [4], deviation between heterozygosity and number of segregation sites [5], and comparison of SNP frequencies in distinct haplotype groups [6]. Despite the rich literature in detecting natural selection from sequences, the majority of the studies consider the evolution of single sites instead of motifs. Furthermore, most of these models require intra-specific polymorphism data which may not be available, and the $\frac{K_a}{K_s}$ test applies only to protein-coding regions.

To overcome these drawbacks, we propose a method of evaluating the strength of natural selection of a motif from aligned sequences. The method is based on a simple neutral model of sequence substitution: what is the distribution of motif occurrences in a sequence of fixed length if each position undergoes an independent sequence substitution? The rate of sequence substitution, the entire sequence length, evolutionary distances of sampled species and sequence complexity of the motif determine the rates of addition (birth) and deletion (death) of motifs in neutral evolution. In contrast, a motif under purifying selection such as a transcription factor binding site often populates on promoters and resists deletions. We quantify natural selection by a coefficient penalizing the rate of motif deletions, and develop an algorithm to estimate the maximum-likelihood selection coefficient.

Our model resembles the probabilistic model of promoter evolution in [7] as both models define a motif as a collection of fixed-length sequences and employ continuous-time Markov processes on sequence substitutions. However, our model differs from [7] by discarding sequence-specific substitution rates, considering the evolution of the motif occurrence frequencies, and being applied to the aligned sequences of more than two species.

The birth-death model considerably approximates the empirical distributions derived from simulated data. Moreover, analysis on the 5kb upstream promoters of 34 mammalian genomes indicates that the underlying hypothesis of our model – sequence substitution follows a Poisson process – holds. We then calculate the selection coefficients of 388 known transcription factor binding motifs and 500 random sequences of 5 and 10 bases. The selection coefficient distribution of known motifs is significantly tilted to high values compared to random controls, suggesting the tendency of positive selection of many transcription factor binding motifs. In contrast, the magnitudes of conservation (fraction of species containing the motif) on transcription factor binding motifs are not higher than random controls. Detailed analysis of SP1 and TP53 binding motifs further elucidates the selective constraints on motifs. While the TP53 binding motif appears in far more human genes than the SP1 binding motif, a smaller fraction of the former are retained in the orthologous promoters beyond primates.

2 Methods

2.1 Overview

Our method is based on a simple neutral model of independent sequence substitution in each position. The distribution of motif counts depends on (1)the rate of sequence substitution, (2)the time interval of interest, (3)the promoter sequence length, (4)the degeneracy and complexity of the motif in the sequence space.

In the neutral model a Poisson process is employed to the sequence substitution of each position. The instantaneous rates of additions (birth) and deletions (death) of the motif can be derived from the sequence substitution model. In contrast, if purifying selection occurs to the motif then the death rate is penalized by a constant. The evolution of motif occurrence frequency distributions can thus be expressed as a system of differential-difference equations parameterized by the penalty constant and the four factors described above. We can calculate the motif count distributions by simulating the differential-difference equations. Furthermore, according to the simulated distributions we apply binary search to find the penalty constant that maximizes the likelihood score of aligned sequences. The penalty constant characterizes the strength of natural selection.

Our model can be viewed as a generalization of motif conservation and can handle the cases where conservation alone fails to capture. For instance, presence of a motif on the orthologous promoters of multiple species may not confer selection if (1)sequence substitution is slow relative to their divergence time or (2)the motif is degenerate and contains many sequences. In contrast, these spurious cases will yield low selection coefficients of the birth-death model.

2.2 A Poisson process model of sequence substitution

A Poisson process is probably the simplest model of sequence substitution [8]. In an infinitesimal time interval dt the nucleotide sequence of a position transitions to another base with probability λdt. Denote $n_P(t)$ the cumulative number of sequence changes at time t. The transitions from t to $t + dt$ follow

$$\begin{array}{ll} P(n_P(t+dt) = N+1 | n_P(t) = N) = \lambda dt. \\ P(n_P(t+dt) = N | n_P(t) = N) \quad\quad = 1 - \lambda dt. \end{array} \tag{1}$$

and the conditional probability at a finite time interval t is

$$P(n_P(t) = N | n_P(0) = 0) = \frac{(\lambda t)^N}{N!} e^{-\lambda t}. \tag{2}$$

Poisson processes are Markovian as conditional probabilities are invariant with time shifts.

Suppose $n_P(t)$ is observed at time points $t_i's$ with interval $T_i's$:

$$t_{i+1} = t_i + T_i. \tag{3}$$

Denote $m_i \equiv n_P(t_{i+1}) - n_P(t_i)$ as the number of sequence changes in time interval $(t_i, t_{i+1}]$. The log likelihood of the data is

$$\begin{array}{rl} L(\lambda) &= \sum_i \log P(n_P(t_{i+1}) - n_P(t_i) = m_i) \\ &= \sum_i m_i \log(\lambda T_i) - \lambda T_i + C. \end{array} \tag{4}$$

By taking the derivative of $L(\lambda)$ with respect to λ the maximum likelihood rate is

$$\hat{\lambda} = \frac{N}{T}. \tag{5}$$

where N is the total number of changes along each time interval and T is the sum of all time intervals.

In this work we assume the Poisson process rate λ is identical in all positions and across all lineages and estimate λ from a family of aligned sequences and their phylogenetic tree. The parsimonious sequences of the internal nodes of the tree are inferred by a dynamic programming algorithm [9]. We then count the total number of sequence changes along all branches of the phylogeny for all positions (N) and the total length of the intervals (T), and estimate λ by equation (5).

2.3 A birth-death model for the neutral evolution of motif occurrences

The major contribution of this study is a neutral model of motif evolution. Motif occurrences can be modeled as a birth-death process [10]. The birth and death rates are determined by the sequence substitution rate and the degeneracy of the motif sequences.

A motif $\mathcal{M} \subset \mathcal{B}^l$ ($\mathcal{B} = \{A, C, G, T\}$) is defined as a collection of nucleotide sequences of length l. Degenerate symbols in IUPAC format are allowed in $\mathcal{M}$. For instance, R denotes purines (A or G) and

Y denotes pyrimidines (C or T). For mathematical convenience differential nucleotide frequencies in distinct positions are ignored. Given a promoter sequence $\mathcal{S}$ of length l_s and the sequence substitution rate λ at each position, we want to model the distributions of $n(t)$, the frequency of motif occurrences in sequence $\mathcal{S}$ at time t.

We first consider the sequence evolution in a window of length l. There are 4^l possible sequences, and each sequence $s \in \mathcal{B}^l$ can be labeled as either a member of the motif ($s \in \mathcal{M}$) or not ($s \notin \mathcal{M}$). These sequences comprise an undirected graph $G = (V, E)$, where a node $v \in V$ denotes a sequence and an edge $e = (v_1, v_2)$ denotes the two sequences v_1 and v_2 differing at one position. $\mathcal{M}$ can be viewed as a subset of nodes in G, and the evolution of the sequences in an l-mer window can be viewed as a Markov random walk on G. In an infinitesimal time interval a sequence is allowed only to transition to neighboring nodes in G. The overall rate of transitions is the sequence substitution rate of the entire window λl. With an independent and identically distributed (iid) assumption this rate is equally divided among all the neighboring nodes. The left diagram of Figure 1 illustrates the transitions of sequence states in a window.

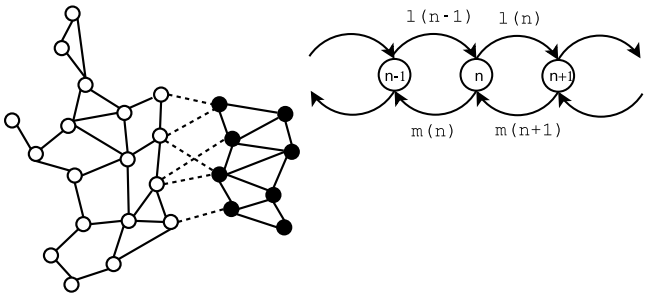

Fig. 1. Left: A sequence space of fixed length as a graph. A node denotes a sequence, and an edge denotes two sequences differing at one position. Black nodes are members of a motif and white nodes are non-motifs. Dotted edges denote transitions between motifs and non-motifs. Solid edges denote transitions within motifs and non-motifs. Right: The state transition diagram of a birth-death model. State n denotes the count of motif occurrence on a promoter. $l(n)$ and $m(n)$ denote the birth and death rates emanating from state n.

We are interested in the transition rate from a non-motif sequence to a motif sequence or vice versa. In principle this rate depends on the initial and final states of each transition and is quite complicated. To simplify the model we use two numbers to characterize the average fraction of motif $\rightarrow$ non-motif transitions and vice versa.

$$\begin{aligned} r_{01} &= \frac{|\{(v_1,v_2)\in E: v_1\notin\mathcal{M}, v_2\in\mathcal{M}\}|}{|\{(v_1,v_2)\in E: v_1\notin\mathcal{M}\}|}. \\ r_{10} &= \frac{|\{(v_1,v_2)\in E: v_1\in\mathcal{M}, v_2\notin\mathcal{M}\}|}{|\{(v_1,v_2)\in E: v_1\in\mathcal{M}\}|}. \end{aligned} \tag{6}$$

r_{01} is the fraction of all non-motif $\rightarrow$ motif transitions among all transitions from non-motifs. For simplicity we expect the non-motif $\rightarrow$ motif transitions and non-motif $\rightarrow$ non-motif transitions are distributed by a ratio $\frac{r_{01}}{1-r_{01}}$. A reciprocal argument applies to the motif $\rightarrow$ non-motif transitions for r_{10}.

An equal transition rate to each sequence may not be an adequate assumption as the distribution of vertebrate genes has a strong bias in the CpG islands [11]. Consequently, we calibrate the ratios r_{01} and r_{10} by the background frequencies of nucleotides (P_A, P_C, P_G, P_T):

$$\begin{aligned} r_{01} &= \frac{\sum_{\{(v_1,v_2)\in E: v_1\notin\mathcal{M}\}} w(v_1,v_2)\delta(v_2\in\mathcal{M})}{\sum_{\{(v_1,v_2)\in E: v_1\notin\mathcal{M}\}} w(v_1,v_2)}. \\ r_{10} &= \frac{\sum_{\{(v_1,v_2)\in E: v_1\in\mathcal{M}\}} w(v_1,v_2)\delta(v_2\notin\mathcal{M})}{\sum_{\{(v_1,v_2)\in E: v_1\in\mathcal{M}\}} w(v_1,v_2)}. \end{aligned} \tag{7}$$

where $w(v_1, v_2)$ is the nucleotide background probability of v_2 at the position where v_1 and v_2 differ. For instance, $w(AGGC, AGTC) = P_T$. $\delta(.)$ is an indicator function. r_{01} and r_{10} are weighted by the background nucleotide frequencies such that more transitions are allocated to GC-rich sequences.

Summarizing the discussions above the transitions of motif occurrence of an l-mer window in an infinitesimal time interval conform with the following equations:

$$\begin{array}{l} P(n(t+dt)=1|n(t)=0)=\lambda l r_{01} dt. \\ P(n(t+dt)=0|n(t)=1)=\lambda l r_{10} dt. \end{array} \tag{8}$$

We then extend the analysis to the entire promoter sequence of length l_s. In an infinitesimal time interval dt, $n(t)=n$ can only increase/decrease by 1 or remain intact. Assuming the motif instances on the promoter do not overlap, there are ln positions occupied by existing motifs and $l_s - ln$ free positions. The ln occupied positions are divided into n independent windows, and the motif $\rightarrow$ non-motif transitions of each window follow equation (8.2). Thus the "death rate" of motif occurrence on the entire sequence is multiplied by n:

$$P(n(t+dt)=n-1|n(t)=n)=\lambda l r_{10} n dt. \tag{9}$$

The "birth rate" of motif occurrence is more difficult to analyze because the number of windows depends on the actual positions of existing motifs and these windows are not independent. For simplicity we approximate the number of independent l-mer windows among free positions by $l_s - ln + l + 1$, the number of l-mer windows in $l_s - ln$ consecutive positions. Thus the "birth rate" of motif occurrence on the entire sequence is multiplied by $l_s - ln + l + 1$:

$$P(n(t+dt)=n+1|n(t)=n)=\lambda l r_{01}(l_s - ln + l + 1) dt. \tag{10}$$

Equations (10) and (9) specify the birth and death rates of motif occurrences in an infinitesimal time interval. The distribution $P_n(t) \equiv P(n(t)=n)$ of motif occurrences over time can be expressed as a system of differential-difference equations:

$$\begin{array}{ll} \frac{dP_0(t)}{dt} & = \mu(1)P_1(t) - \lambda(0)P_0(t). \\ \frac{dP_n(t)}{dt} & = \lambda(n-1)P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - (\lambda(n)+\mu(n))P_n(t). \\ \lambda(n) & = \lambda l r_{01}(l_s - ln + l + 1). \\ \mu(n) & = \lambda l r_{10} n. \end{array} \tag{11}$$

The system can be illustrated by the right diagram of Figure 1. The system is a single server M/M/1 queueing model where the birth and death of motifs correspond to the arrival and completion of the jobs. Here the arrival and completion rates depend on the system state (n).

2.4 A birth-death model of the selective evolution of motif occurrences

The purpose of constructing a neutral model of motif evolution is to identify the motif sequences that undergo purifying selection. Intuitively, purifying selection penalizes decrements of a functional motif on the promoter. Thus we divide the death rates by a selection coefficient:

$$\mu'(n) = \frac{\mu(n)}{s}. \tag{12}$$

When $s > 1$, the process of motif deletion slows down and more motifs are accumulated. This phenomenon is consistent with purifying selection. Notice s is different from the conventional definition of selection coefficients in population genetics, which denotes the deviation of genotype frequencies from the neutral model.

2.5 Evaluating the selection coefficient of the birth-death model

To evaluate the strength of purifying selection we apply both neutral and selective models to aligned sequences. The algorithm in Figure 2 shows the outline of the algorithm to evaluate the selection coefficient.

The inputs of the algorithm are the phylogenetic tree $\mathcal{T} = (V_T, E_T)$ of k species, n orthologous families of aligned sequences, and a sequence motif $\mathcal{M}$. We first apply dynamic programming to reconstruct sequences of internal nodes of $\mathcal{T}$ and infer the Poisson rate λ using equation (5). For simplicity we assume the sequence substitution rates of all families are identical.

Fig. 2. Evaluating the selection coefficient of a motif

Inputs: Motif $\mathcal{M}$, phylogenetic tree $\mathcal{T} = (V_T, E_T)$ of k species, n orthologous families of aligned promoter sequences, s_{ij} denotes the aligned sequence of gene i in species j.
Outputs: Selection coefficient of $\mathcal{M}$ on the aligned sequences.

1. Infer the Poisson rate λ of sequence substitution using dynamic programming and equation (5).
2. Split a promoter sequence into multiple segments of fixed length $l_s = 30l$.
3. Count the motif occurrence in each segment.
4. Reconstruct the motif occurrence for each segment and internal node using dynamic programming.
5. Count the empirical conditional frequency along each branch.
6. Apply binary search to find the selection coefficient that maximizes the log likelihood score.

The birth-death models in equations (10) and (9) apply to sequences of any lengths as long as $l_s \gg l$. In practice, longer sequences are computationally challenging for the following reasons. First, due to frequent recombinations, insertions and deletions, more gaps will appear in a long stretch of aligned sequences. Gaps add complexities in evaluating likelihood scores hence are undesirable. Second, longer sequences accommodate more motif instances by random sequence substitution. Hence more terms in equation (11) need to be considered. We divide the promoter sequence into segments of length $l_s = 30l$. A segment with more than 10% gaps in a species is treated as a missing data.

Motif occurrences in a segment are counted by sliding a window of length l along the segment. Denote s_{ijk} the aligned sequence of the kth segment of gene i in species j, and n_{ijk} the motif count of the corresponding segment. The motif occurrences of internal nodes can be inferred from their reconstructed sequences.

The joint log likelihood of the observed and reconstructed motif counts is

$$\mathcal{L} = \sum_i \sum_k \sum_{(v,w)\in E_T} \log P(n(t_{(v,w)}) = n_{iwk} | n(0) = n_{ivk}) + C. \tag{13}$$

where summation is over indices of gene i, segment k and edge (v, w) in $\mathcal{T}$. $t_{(v,w)}$ denotes the branch length of edge (v, w). Resembling EM, our method fills the missing data of internal nodes with reconstructed motif counts and avoids the cumbersome evaluation of the marginal likelihood.

The log likelihood can also be expressed as

$$\mathcal{L} = \sum_t \sum_{n_0} \sum_{n_1} f(t, n_0, n_1) \log P(n(t) = n_1 | n(0) = n_0) + C. \tag{14}$$

where $f(t, n_0, n_1)$ denotes the frequency of the instances where the motif counts in the parent and child nodes are n_0 and n_1 and the branch length is t. These empirical frequencies can be directly obtained from the observed and reconstructed data. $P(n(t) = n_1 | n(0) = n_0)$ is the conditional probability derived from the birth-death model of the neutral or selective evolution (equations (11) and (12)). In this work we solve the transient responses $P_n(t)$ numerically by simulating the differential-difference equations. Given the relatively short segments ($30l$) only the first few equations in equation (11) are needed.

The only free parameter of the log likelihood is the selection coefficient s. We want to find the s that maximizes equation (14). Because s is integrated in equation (14) in a complex form and $P(n(t) = n_1 | n(0) = n_0)$ has no analytic solutions, we apply a binary search to find the optimum value of s over the interval $[0, 20]$.

3 Results

3.1 The birth-death model agrees with simulation data

We first verified that the birth-death model approximated the motif count distribution derived from a Poisson sequence substitution process with simulation data. A random 100-base initial sequence and 20 random 4-base motifs were constructed. 1000 instances with the identical initial sequence underwent independent Poisson sequence substitutions with $\lambda = 0.2$ and $T = 4.0$. The empirical data were compared to the conditional probabilities predicted by equations (11) and (12). The birth-death model strongly agreed with the simulated data. Figure 3 shows the time evolution of motif count distributions of 3 motifs with 0, 1 and 2 instances in the initial sequence respectively. The predicted

models (dashed lines) closely follow the empirical distributions (solid lines) in each case. The results indicate the birth-death model accurately describes motif count distributions in a Poisson process of sequence substitution.

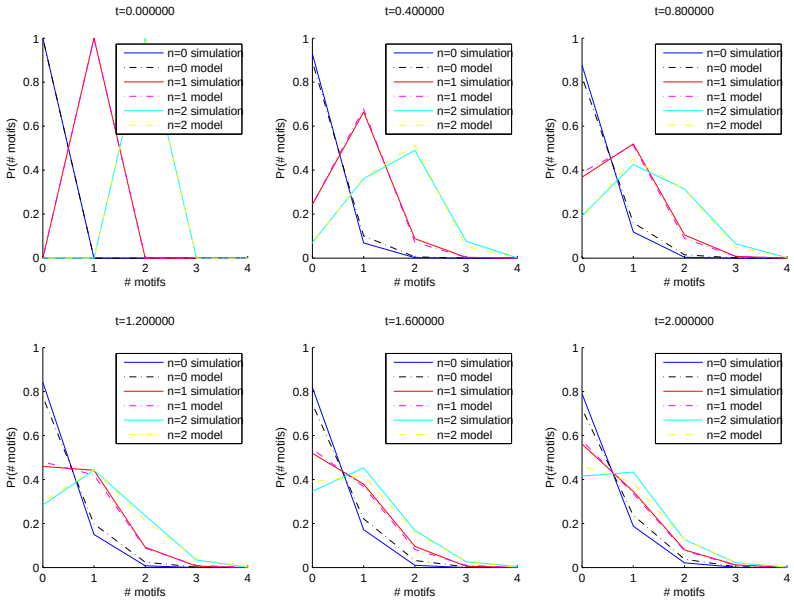

Fig. 3. Comparison of empirical and predicted distributions of motif counts in simulated data. Conditional probabilities of motif counts at 6 time points are shown. Solid lines indicate empirical distributions of 3 motifs with 0 (blue), 1 (red) and 2 (cyan) instances in the initial sequence respectively. Dashed lines are the distributions derived from equation (11) (initial counts: 0:black, 1:magenta, 2:yellow).

3.2 Sequence substitutions on mammalian promoters follow a Poisson process

Aligned 5kb upstream sequences of 27667 orthologous gene families from 44 vertebrate species were extracted from the UCSC Genome Browser [12]. 34 of the 44 species were mammals. Figure 4 shows the average sequence similarities between human and other species versus their distances on the phylogenetic tree. Sequence similarity was negatively correlated with the phylogenetic distance in mammals and was stablized around 0.6 as other non-mammal vertebrates were included. We suspected the saturated sequence similarity was due to a selection bias for alignable sequences. Dissimilar promoter sequences may not be alignable with mammalian sequences thus were treated as gaps. To correct this error we only included the data of 34 mammals (phylogenetic distance below 1.0 in Figure 4). Figure 5 compares the distributions of sequence substitutions from the data and the Poisson process with the maximum-likelihood rate $\lambda = 0.8937$. For each position, the empirical number of sequence substitutions between two species is the number of sequence changes along the path connecting the two species in the phylogenetic tree. The predicted Poisson distributions closely resembled the empirical distributions at various time intervals, suggesting that promoter sequence substitutions in mammals follow a Poisson process.

3.3 Known transcription factor binding motifs have higher selection coefficients than random sequences with high occurrence frequencies

388 transcription factor binding motifs were extracted from the TRANSFAC database [13]. Motif lengths ranged from 5 to 15 nucleotides and the mean length was 10.66 nucleotides. We applied the

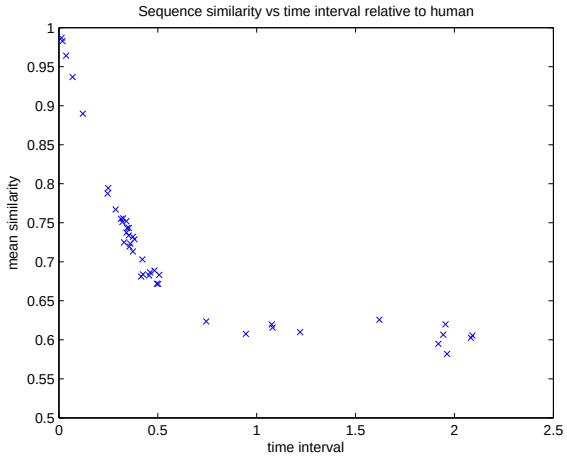

Fig. 4. Promoter sequence similarity (Y-axis) versus phylogenetic distance (time interval, X-axis) relative to human. Sequence similarity between human and one species is the average of the ratio $\frac{\#\text{different sequences}}{\#\text{non-gapped sequences}}$ over the 27667 families. A phylogenetic distance between human and one species is obtained from the phylogenetic tree provided by UCSC Genome Browser. All 34 mammals are at the left half of the diagram (time interval < 1.0).

algorithm in Figure 2 to evaluate the selection coefficient of each motif on the 5kb promoters of 27667 orthologous families in 34 mammals. In addition to selection coefficients, we also evaluated the magnitudes of conservation by counting the fractions of species containing the motifs among 34 mammals and averaging the scores over all segments in all the gene families. As a negative control we generated 10000 random motif sequences (in IUPAC format) of 5 and 10 nucleotides and selected the top 500 sequences according to their occurrence frequencies on mammalian promoters. The left diagram of Figure 6 shows the distributions of selection coefficients in TRANSFAC motifs and two negative control sets. Intriguingly, there are many more high-scoring TRANSFAC motifs than the negative controls. About one quarter (96 of 388) of known motifs have selection coefficients ≥ 4.0. In contrast, only 11 and 24 of 500 5-mer and 10-mer control motifs pass the same threshold. The fraction of high-scoring motifs may be over-estimated as some transcription factors possess multiple similar motifs. By grouping motifs by their transcription factors, the same conclusion was reached (results not shown).

The right diagram of Figure 6 shows the distributions of conservation magnitudes in TRANSFAC motifs and two negative control sets. The conservation mangitude of a motif on a promoter is the fraction of the species containing the motif. We report the average of the conservation magnitudes over the genes where the motif appears at least in one species. Clearly, natural selection is not revealed by conservation alone, as the conservation magnitudes of most TRANSFAC motifs are smaller than those of the control motifs. Moreover, unlike selection coefficients conservation magnitudes of control motifs are sensitive to their lengths. The results are sensible in two aspects. First, the "random motifs" in Figure 6 are the sequences with high occurrence frequencies. They often contain multiple degenerate sequences and are thus expected to appear in more species by chance. The birth-death model can eliminate these spurious motifs as the volumes of motif sequences are taken into account. Second, conservation of motifs is sensitive to sequence length as short sequences are likely to appear in more species by chance. The birth-death model also takes sequence length into account. Therefore, the selection coefficient distributions of 5-mer and 10-mer control motifs are similar.

3.4 SP1 and TP53 binding motifs elucidate motif evolution

Table 1 lists the top-ranking motifs with selection coefficients ≥ 10.0. These high-scoring motifs share two common features. First, they are long stretches of sequences with low degeneracy. They occupy a small volume of the sequence space, hence random drifts into the motif are much less likely than those out of the motif. Second, despite the uniqueness they appear on the orthologous promoters

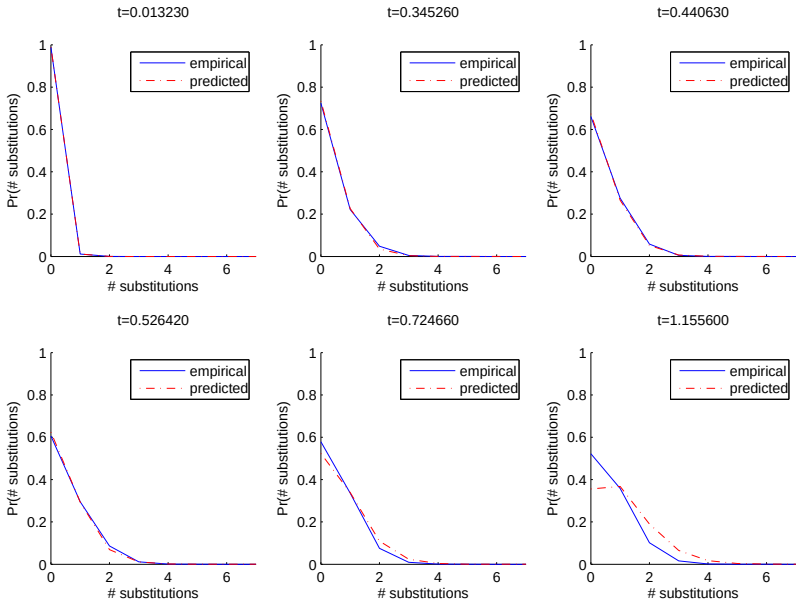

Fig. 5. Sequence substitutions on 5kb promoters of mammalian genes. Empirical distributions of sequence substitutions (solid blue lines) are obtained by the numbers of sequence changes along the paths connecting each pair of species in each position. Predicted distributions (dashed red lines) are calculated by a Poisson process with $\lambda = 0.8937$. The distributions at 6 time intervals are shown.

of distant species with a disproportional rate compared to the neutral model. Twelve TRANSFAC motifs have selection coefficients $s = \infty$. In these cases the death of a motif is not observed in the data, hence the death rate is penalized by an infinite factor according to equation (12).

To elucidate the relation of selection coefficients and motif evolution, we examined the frequency distributions of two transcription factor binding motifs. A SP1-binding motif has a 10-base unique sequence (GGGGCGGGGC) and a high selection coefficient (13.5143). A TP53-binding motif yields multiple 10-base degenerate sequences (NGRCWTGYCY) and a low selection coefficient (1.4350). Figures 7 and 8 show the snapshots at 6 time points of the distributions of SP1 and TP53 motif counts on the promoters of 33 non-human mammals. Here motif counts on human promoters are treated as $n(0)$, the phylogenetic distance from human as t, and the motif counts on the promoters of the target species as $n(t)$. The 6 times intervals are selected by visually inspecting the dynamics of $P(n(t)|n(0))$ and finding the time points that yield distinct shapes of $P(n(t)|n(0))$. For both sequences, the probability mass in the stationary distribution is concentrated at $n = 0$. Thus both empirical and predicted distributions of $P(n(t)|n(0) = 0)$ (blue curves) remain invariant with t. In contrast, $P(n(t)|n(0) = 1)$ quickly moves away from the initial distribution toward stationarity. The empirical distributions of $P(n(t)|n(0) = 1)$ (red solid lines) generally follow the predicted distributions from the neutral model (red dashed lines) in the TP53 motif, confirming its small selection coefficient. However, the empirical distributions of $P(n(t)|n(0) = 1)$ in the SP1 motif assign substantially more probability on $n(t) = 1$ compared to the predicted distributions. This deviation implies a smaller motif death rate than in the neutral model and a higher selection coefficient.

2225 of 27667 5kb upstream human promoters contain at least one TP53-binding motif, and 160 human promoters contain at least one SP1-binding motif. On these 2225 and 160 promoters we examined how many retained TP53 and SP1 binding motifs in other species. Figure 9 shows the conservation of the two motifs between human and other mammals. The fraction of genes retaining either motif drops below 20% beyond primates (species index > 5). However, the SP1-binding motif is retained in much higher fraction of promoters than the TP53-binding motif among non-primate mammals. This observation again confirms the stronger purifying selection of the SP1-binding motif.

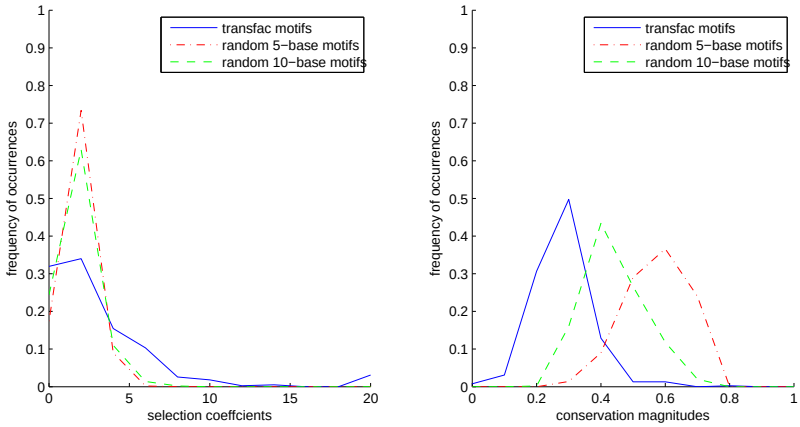

Fig. 6. Left: Distributions of selection coefficients of 388 TRANSFAC motifs (solid blue), 500 frequent 5-mer random motifs (dashed red), and 500 frequent 10-mer random motifs (broken green). Right: Distributions of conservation magnitudes of 388 TRANSFAC motifs (solid blue), 500 frequent 5-mer random motifs (dashed red), and 500 frequent 10-mer random motifs (broken green).

4 Discussions

In this work we propose a model and an algorithm to evaluate the strength of natural selection of a sequence motif from aligned sequences across gene families and species. The neutral model of motif occurrence distributions is based on a simple assumption that each position undergoes an independent Poisson process of sequence substitution. We consequently derive a birth-death model of motif occurrences according to the sequence substitution rate, motif sequence degeneracy and total sequence length. The selection coefficient is the penalty on the rate of motif deletion in the birth-death model. Predictions derived from the neutral model fit both simulated data and the statistics of random motifs on the aligned promoters of 34 mammals. In addition, many more known transcription factor binding motifs have high selection coefficients relative to negative controls, suggesting many of them are under purifying selection.

Despite the success in the preliminary study the current model and algorithm have several limitations. First, the model (and many other models of natural selection) focuses on sequence substitution and does not take other types of mutations – insertions/deletions, recombinations – into account. Second, the model considers a ubiquitous selection along each branch of phylogeny and discards lineage specific selection. Third, the model also discards the gene-specific selection on promoters and only considers the overall effects on all gene families. Fourth, numerical simulations and binary search of the algorithm are time-consuming. Analytic approximations to the transient responses of the birth-death model should be developed. Fifth, the inverse problem of this work – identify the motif sequences with high selection coefficients – is yet to be tackled. In spite of these limitations our model serves as a reasonable tool to validate the computationally or experimentally discovered candidate motifs.

References

1. Kellis M, Patterson N, Endrizzi M, Birren B and Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory motifs. Nature 423:241-254.
2. Yang Z and Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends of Ecology and Evolution 15:496-503.
3. Siepel A. and Haussler D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. Journal of Computational Biology 11(2-3):413-428.
4. McDonald JH and Kreitman M. 1991. Adaptive evolution at the Adh locus in Drosophila. Nature 351: 652-654.

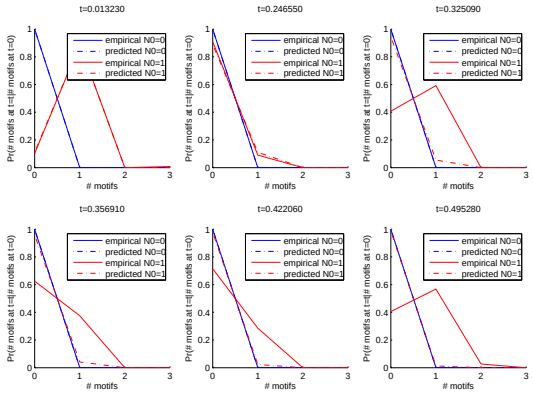

Fig. 7. Empirical and predicted conditional probabilities $P(n(t)|n(0))$ of SP1 binding motif at 6 time points from left-top to bottom-right diagrams. Solid lines: empirical distributions. Dashed lines: predicted distributions from the neutral model. Blue lines: $n(0) = 0$. Red lines: $n(0) = 1$.

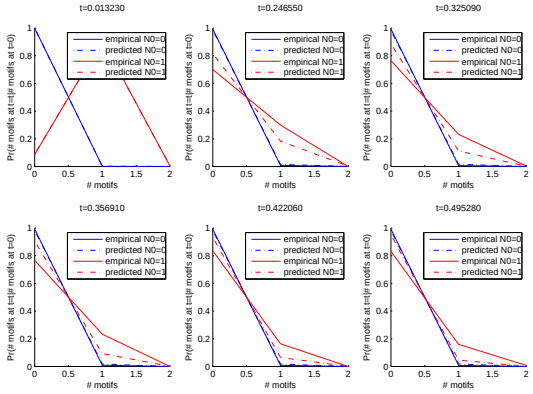

Fig. 8. Empirical and predicted conditional probabilities $P(n(t)|n(0))$ of TP53 binding motif. Legends follow Figure 7.

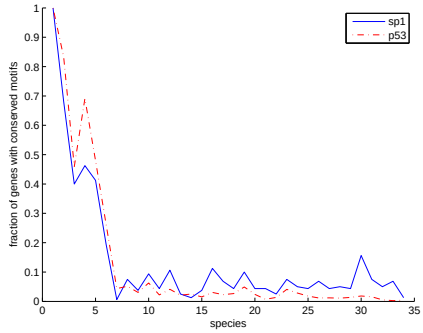

Fig. 9. Conservation of the SP1 and TP53 motifs on the promoters of mammals. X-axis indicates species indices sorted by phylogenetic distances with respect to human. 5:rhesus, 15:sloth, 25:rock hyrax, 34:platypus. Y-axis indicates the fraction of genes containing the motif in both human and the target species among the genes containing the motif in human. Solid blue line: SP1-binding motif. Dashed red line: TP53-binding motif.

Table 1. Top-scoring transcription factor binding motifs

FACTOR	SEQUENCE	s
Sp1	NGGGGGCGGGGYN	∞
C/EBPbeta	CTBATTTCARAAW	∞
POU3F1	CTNATTTGCATAY	∞
HNF-4alpha2	TGAMCTTTGMMCYT	∞
HNF-4alpha2	RGGNCAAAGGTCA	∞
HSF1	TTCCMGARGYTTC	∞
Egr-1	CCCGCCCCCRCCCC	∞
CDC5L	GATTTAACATAA	∞
FXR-alpha	GGGTBAATRACCY	∞
MTF-1	TBTGCACHCGGCCC	∞
ZNF219	CRCCCCCCNCCC	∞
DMRT7	TTGTTACAWTKTKG	∞
Sp1	CCCCGCCCCN	14.4169
Sp1	GGGGCGGGGC	13.5143
POU3F1	TTATGYTAAT	11.1595
Sp1	NNGGGGCGGGGNN	10.6129
HNF-4alpha2	TGACCTTTGNCCY	10.413
c-Krox	SCCCTCCCC	10.25

5. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.
6. Atwal GS, Bond GL, Metsuyanim S et al. 2007. Haplotype structure and selection of the MDM2 oncogene in humans. Proceedings of National Academy of Science USA 104:4525-4529.
7. Raijman D, Shamir R and Tanay A. 2008. Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. PLoS Computational Biology 4:77-87, 2008.
8. Zuckerandl E and Pauling L. 1965. Evolutionary divergence and convergence in proteins, pp. 97-166. In Evolving genes and proteins, edited by Bryson V and Vogel HJ. Academic Press, New York.
9. Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evoution. 17:368-376.
10. Kendall DG. 1948. On the generalized birth-death process. The Annals of Mathematical Statistics 19(1):1-15.
11. Bird AP. 1987. CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics 3:342-347.
12. Kuhn RM, Karolchik D, Zweig AS et al. 2009. The UCSC Genome Browser Database: update 2009. Nucleic Acids Research D755-D761.
13. Matys V., Fricke E., Geffers R., Gossling E., Haubrock M. et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research 31(1):374-378.