Position-specific evolution in transcription factor binding sites, and a fast likelihood calculation for the F81 model

Pavitra Selvakumar[1,2] and Rahul Siddharthan[1,2,*]

[1]The Institute of Mathematical Sciences, Chennai, India
[2]Homi Bhabha National Institute, Mumbai, India
[*]*Corresponding author, rsidd@imsc.res.in*

Transcription factor binding sites (TFBS), like other DNA sequence, evolve, and are subject to selection pressures relating to their function. Various models of nucleotide evolution describe DNA evolution via single-nucleotide mutation. A stationary distribution (stationary vector) of such a model is the long-term distribution of nucleotides, unchanging under such a model. Neutrally evolving sites have stationary vectors reflective of genome-wide nucleotide distributions, but one expects that sites within a TFBS instead have stationary vectors reflective of the fitness of various nucleotides at those positions. We introduce "position-specific stationary vectors" (PSSVs), the collection of stationary vectors at each site in a TFBS locus, analogous to the position weight matrix (PWM) commonly used to describe TFBS. We infer PSSVs for human TFs using two evolutionary models (Felsenstein 1981 and Hasegawa-Kishino-Yano 1985). We find that PSSVs largely reflect the nucleotide distribution from PWMs, but with reduced specificity. We infer ancestral nucleotide distributions at individual positions and calculate "conditional PSSVs" conditioned on specific choices of majority ancestral nucleotide. We find that certain ancestral nucleotides exert a strong evolutionary pressure on neighbouring sequence while others have a negligible effect. Finally, we present a fast likelihood calculation for the F81 model on moderate-sized trees that makes this approach feasible for large-scale studies along these lines, and a Jupyter notebook implementing our methods.

Introduction

Transcription factor binding sites (TFBS) are short regions in DNA where regulatory proteins bind, typically by recognizing a short, conserved "motif". These motifs are not exact strings and are commonly represented by "position weight matrices" (PWMs) [1], whose columns give position-specific nucleotide distributions: for a motif of length L, a PWM $W_{n\alpha}$ is a $L \times 4$ matrix giving the probability, if a sequence of length L is a binding sequence, that the nucleotide at position n is α.

The PWM representation assumes that different positions within a motif (different columns of the matrix) have independent distributions of nucleotides. However, positional dependencies are known to occur in binding sites of many TFs. Extensions to the basic PWM format have been proposed to account for positional dependencies [2, 3, 4, 5]. Nevertheless, PWMs persist as the dominant motif representation format, because of ease of calculation and because they are easily visualized via sequence logos [6].

In this work we ask: in the course of evolution, what are the selective pressures on TFBS? Are these pressures position-independent, or do nucleotide preferences at one position affect evolution at other positions?

Several models of DNA evolution have been developed in the past decades, briefly reviewed in Methods. Neutral models of DNA evolution assume that individual nucleotides mutate at a rate given by a rate matrix, from which can be calculated transition probabilities at finite time, and long-term distributions at individual loci. Symmetric rate matrices, such as in the Jukes-Cantor [7] and Kimura [8] models, yield a uniform stationary distribution of nucleotides under neutral evolution. Commonly, selection is treated as an additional process affecting nucleotide evolution and distribution. In the context of site-specific evolution, the Halpern-Bruno model, developed to describe protein codon evolution [9], was modified for TFBS evolution by Moses *et al.* [10]. The assumptions are that TFBS are under purifying selection (where deleterious mutations are removed from the population), and that time to fixation is small compared to time between fixations ("weak mutation model"). This is further explored in Discussion.

Here we take an alternative approach: we consider evolution via a rate matrix that yields a non-uniform stationary distribution, and seek to infer that stationary distribution (which is indicative of fitness under purifying selection) via an evolutionary model. The earliest such model, and the simplest, was Felsenstein's 1981 model (F81) [11]. (In the same paper, Felsenstein supplied his now-standard recursive "pruning" algorithm for likelihood calculation). We also use the Hasegawa-Kishino-Yano 1985 (HKY85) [12] model, which yields similar results on this problem.

Our approach is as follows: we use a standard tree topology from the literature (for primates, and for mammals); we obtain instances of motifs (with 20bp flanking sequence) in human, and orthologous motifs (without indels) in other species, for various TFs; we estimate the tree's branch lengths using the least informative positions in these sequences, which we assume to be neutrally evolving (stationary vector equal to background probability distribution of nucleotides in human); and then, given this complete tree with topology and branch lengths, we estimate "position-specific stationary vectors" (PSSVs) at all positions within the motif.

These PSSVs are the long-term stationary distribution of sites evolving under our evolutionary model and are obtained via maximum likelihood, as described in Methods. They can be seen as combining mutation and selection information.

Our methodology enables inference of ancestral nucleotide probabilities at each locus. This allows us to study compensatory mutations, by identifying a subset of sites where the ancestral nucleotide at a given position is, say, G with probability > 0.5, and then studying the effect of the PSSV at other positions within this subset of sites.

En route, we present an algorithm for calculating likelihoods in the F81 model, which, for moderate-sized trees, is much faster than Felsenstein's pruning algorithm. As described in Methods, the idea is to reduce an arbitrary tree into a sum of products of "star trees", where we define a star tree as a tree all of whose leaves are directly connected to one ancestral node. The speedup comes partly from reducing the complexity in alphabet size Σ from $O(\Sigma^2)$ (pruning algorithm) to $O(\Sigma)$, but mainly from removing the recursive traversal of the tree which otherwise needs to be done at every locus.

Materials and Methods

General approach

Given a collection of orthologous sequences from a set of species, and given a phylogenetic tree linking those species, we can calculate, via standard methods described below, the likelihood of those sequences arising via an evolutionary model from a common ancestor. We take the topology of the tree from the literature; these are shown in Figure 1 (A) and (B). We focus on TFBS in human (including flanking sequence), with gapless orthologous sequence in other species. We assume that uninformative positions have a uniform stationary vector (equal distribution of probabilities under evolution) and estimate the branch lengths (the evolutionary time t between neighbouring nodes of the tree) from these positions, via maximum likelihood. We use two evolutionary models, the Felsenstein 1981 (F81) and the Hasegawa-Kishino-Yano 1985 (HKY85) model, described below, that are capable of reproducing arbitrary stationary distributions; having learned the tree branch lengths from uninformative positions, we then learn the site-specific stationary vectors that would be best explain the data at other positions, again via maximum likelihood.

Models of nucleotide evolution, stationary vectors

For a detailed review of models of nucleotide evolution, see, eg, Tavaré [13]. A brief summary follows.

The instantaneous rate of mutation of nucleotides from α to β ($\alpha, \beta \in$ A, C, G, T $\equiv 1, 2, 3, 4$) is given by a 4×4 rate matrix $R_{\beta\alpha}$, whose columns sum to zero: $R_{\alpha\alpha} = -\sum_\beta R_{\beta\alpha}$. The "transition matrix" $T_{\beta\alpha}(t)$ is the probability, given an ancestral nucleotide α at a given locus at time T, of observing β at the same locus at a later time $T + t$, and is equal to $\exp(Rt)$. The equilibrium distribution of nucleotides at a locus is

denoted by $\overrightarrow{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$ and, for a time-reversible model, we have

$$T_{\alpha\beta}(t)\pi_\beta = T_{\beta\alpha}(t)\pi_\alpha.$$

Here time t is geological time, measured in appropriate units.

Transition matrices also satisfy the following multiplicative property

$$\sum_\beta T_{\gamma\beta}(t_2)T_{\beta\alpha}(t_1) = T_{\gamma\alpha}(t_1 + t_2),$$

in other words, an intermediate node (here, the unknown β) with just one child can be removed from the tree and the distance of its child can be added to its distance from its parent. In combination with time-reversibility, this also implies that the branch lengths of the two children of a root node cannot be determined uniquely, only their sums can be, and indeed any node can be chosen to be the root without changing the likelihood (what Felsenstein called the "pulley principle" [11]).

Different models of nucleotide evolution differ in their choice of R. The simplest, the Jukes-Cantor model [7], assumes all off-diagonal elements of R are equal (r) and the diagonal elements are $-3r$. The Kimura model [8] distinguishes between transition rates (A↔G, C↔T, which are typically more frequent) and tranversion rates (all other mutations). Both of these result in a uniform equilibrium frequency $\overrightarrow{\pi} = (0.25, 0.25, 0.25, 0.25)$.

Reproducing more general $\overrightarrow{\pi}$ requires asymmetric matrices R. The simplest of these, the Felsenstein 1981 (F81) model [11], uses

$$R = \mu \begin{pmatrix} \pi_A - 1 & \pi_A & \pi_A & \pi_A \\ \pi_C & \pi_C - 1 & \pi_C & \pi_C \\ \pi_G & \pi_G & \pi_G - 1 & \pi_G \\ \pi_T & \pi_T & \pi_T & \pi_T - 1 \end{pmatrix} \tag{1}$$

where μ is a mutation rate. This yields

$$T_{\alpha\beta}(t) = e^{-\mu t}\delta_{\alpha\beta} + (1 - e^{-\mu t})\pi_\alpha.$$

We choose units of time such that $\mu = 1$, and rather than use the time t, we use a quantity $q \equiv e^{-t}$ (which we previously called the "proximity" [14]):

$$T_{\alpha\beta}(t) = q\delta_{\alpha\beta} + (1 - q)\pi_\alpha.$$

This has a simple interpretation: with a probability q, the nucleotide is unmutated from its ancestor, and with a probability $1 - q$, it has undergone one or more mutation events, and also undergone selection, such that the probability of the new nucleotide α is given by the stationary distribution π_α indicating the relative fitness of α.

The Felsenstein 1984 (F84) [15] and the Hasegawa-Kishino-Yano 1985 (HKY85) [12] models generalize F81 to account for differing transition and transversion rates. We discuss HKY85 below, and its results on this problem are similar to F81. However, we

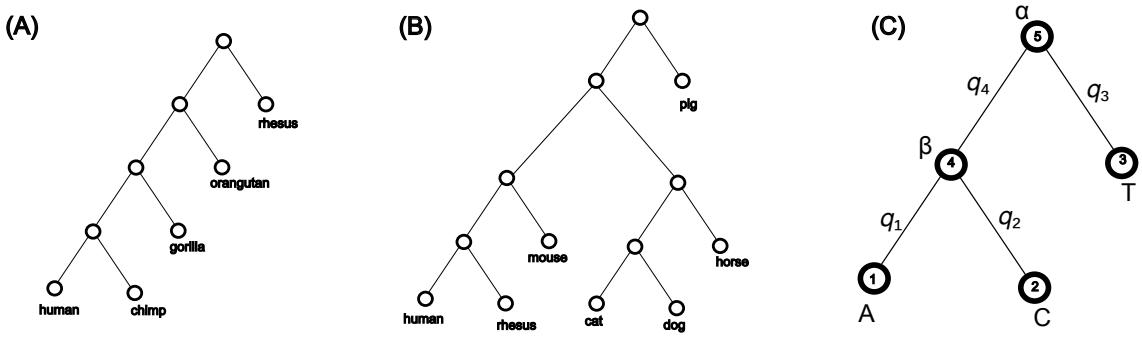

Figure 1: (a) Primate tree topology used in this work (b) Mammal tree topology used in this work (branch lengths are calculated and not shown in this figure) (c) Sample tree with proximity labels on branches

demonstrate a fast likelihood algorithm for the F81 model (developing on an approach in a previous work by one of us [14]) which makes it more useful for this task. We argue that the similar results are because the evolution of TFBS are dominated by selection, so the differing rates of transition and transversion are not important: at most positions the rate matrix is dominated by the value of the equilibrium frequency $\vec{\pi}$ which indicates selection pressures at that position.

Likelihood calculation on a phylogenetic tree; Felsenstein's algorithm

Consider a large collection of transcription factor binding sites in a species of interest, as well as orthologous sequence in other species. These species are leaves on a phylogenetic tree, assumed to be binary (each internal node has exactly two children). Two examples, featuring primates and mammals, are in figure 1. For a tree with n species (leaves), we label the leaves with numbers from 1 to n and the ancestral nodes with numbers from $n+1$ to $2n-1$.

The branch lengths indicate the evolutionary distance. Rather than an additive distance in time units (t) we label the branches with multiplicative proximities $q = e^{-t}$: in the F81 model, if node/leaf i has descended from its parent for a time t_i with a mutation rate μ (which we take to be 1), then $q_i = e^{-\mu t_i}$, that is, q_i the probability that a given nucleotide under i is conserved (not mutated) from its parent. The proximities are therefore multiplicative along branches. For the root node, the children's proximities are not independent and only their product can be determined ("pulley principle" mentioned above).

Each position within the binding site is a separate locus, with a separate collection of leaves, on the same tree. The likelihood, at a particular locus x, of seeing the nucleotides $\{S^k\}$ at the leaves (k denotes the leaf or species label) is then a product over the tree of all transition matrices along edges, summed over all ancestors. unknown ancestors. For

the tree in fig. 1 (c), this is

$$P(\{S^k\}|\{q\}) = \sum_\alpha \pi_\alpha T_{T\alpha}(q_3) \sum_\beta T_{\beta\alpha}(q_4) T_{A\beta}(q_1) T_{C\beta}(q_2) \tag{2}$$

where, for the F81 model, as above, $T_{\beta\alpha}(q) = q\delta_{\beta\alpha} + (1-q)\pi_\beta$. For the HKY85 model too, explicit formulas for $T_{\beta\alpha}$ exist, discussed below. This sum depends only on the product $q_3 q_4$ and not on their individual values. In our code we assign the right child of the root node a proximity $q = 1$, without loss of generality, and the root node also has a proximity 1 (it has no parent), so that for a tree of n leaves, and $2n-1$ nodes we need to calculate only $2n-3$ proximities.

Felsenstein, in the same paper where he introduced the F81 model [11], introduced a recursive algorithm to calculate this likelihood for any tree (and for any transition matrix), which can be summarised (for binary trees, but generalizable) as

- Define $L_i(\alpha)$ = likelihood of leaves below node i given that the nucleotide at node i is α
- If i is a leaf node with nucleotide x, set $L_i(\alpha) = \delta_{\alpha x}$
- Otherwise, let the two children of i be j and k with proximities q_j and q_k. Then $L_i(\alpha) = \sum_\beta T_{\beta\alpha}(q_j) L_j(\beta) \sum_\gamma T_{\gamma\alpha}(q_k) L_k(\gamma)$. (For non-binary trees this step can be generalized to a product over all children, with the appropriate number of sums.)
- Termination: let the root node be $2n-1$, then the likelihood of the leaves is $\sum_\alpha L_{2n-1}(\alpha)\pi_\alpha$.

For efficiency, the values of $L_i(\alpha)$ need to be stored, which we do within the tree structure.

This likelihood applies to nucleotides at one locus. It is assumed that nucleotides at different positions are independent, so the likelihood of a collection of sequences is the product of the likelihoods at individual loci. It is preferable to use log likelihoods to avoid underflows. There have been efforts to optimize the algorithm by sorting columns of the aligned sequence and storing and looking up identical calculations [16].

Fast likelihood calculation for F81 model

The F81 transition matrix has a peculiar interpretation that allows us to greatly simplify likelihood calculations. First, we generalize trees beyond binary trees, so that nodes can have more than two children. We define a "star topology" as a tree topology where all nodes, except the root, are leaves (i.e. every leaf is directly attached to the root). It turns out that, under the F81 model, *every tree is equivalent to a sum of products of star-topology trees.*

We illustrate with fig. 1 (c), where we had

$$P(\{S^k\}|\{q\}) = \sum_\alpha \pi_\alpha T_{T\alpha}(q_3) \sum_\beta T_{\beta\alpha}(q_4) T_{A\beta}(q_1) T_{C\beta}(q_2).$$

Consider the sum over β. The node corresponding to that sum is what we call a "star node", i.e. a node all of whose children are leaves. We plug in the explicit expression $T_{\beta\alpha} = q_4\delta_{\alpha\beta} + (1-q_4)\pi_\beta$, to obtain

$$\begin{aligned} P(\{S^k\}|\{q\}) &= \sum_\alpha \pi_\alpha T_{T\alpha}(q_3) \sum_\beta (q_4\delta_{\alpha\beta} + (1-q_4)\pi_\beta) T_{A\beta}(q_1) T_{C\beta}(q_2) && (3) \\ &= q_4 \sum_\alpha \pi_\alpha T_{T\alpha}(q_3) T_{A\alpha}(q_1) T_{C\alpha}(q_2) \\ &\quad +(1-q_4)\left(\sum_\alpha \pi_\alpha T_{T\alpha}(q_3)\right)\left(\sum_\beta \pi_\beta T_{A\beta}(q_1) T_{C\beta}(q_2)\right) && (4) \end{aligned}$$

. Basically, with probability q_4 the node β is unmutated from its parent, so its leaves can be directly joined to its parent (the first term in eq. 4). And with probability $1-q_4$ it is mutated but is now completely independent of its parent α and its likelihood now merely multiplies the rest of the tree.

For this simple tree, this is all, but for more complicated trees, this can be done recursively:

1. Initialize a list of lists of trees, $[L_1, L_2, L_3 \ldots]$, initially with a single tree list [L] with a single element $L = [T]$ (the original tree)
2. For each list of trees $L = [T, T', T'' \ldots]$
   a) The first element T is the "main" tree, the rest are multipliers
   b) While T has non-star nodes
      i. identify a star node in T, whose proximity to its parent we call q
      ii. create three new trees from T: T_m (where star node's leaves are merged into parent), T_p (where star node and its leaves are removed from parent), T_s (the star node that has been removed).
      iii. Replace original T with T_m with a weight q; and append a new element to the list consisting of $[T_p, T_s, \ldots]$ with weight $1-q$, where $\ldots$ indicates that all previous multipliers are retained
      iv. All multipliers are star nodes, so the only possible non-star trees are the first elements in the tree lists
   c) Repeat until no non-star trees remain in this list

Under this process, the list of lists of trees may evolve as follows:

$$[[T]] \to [q_1[T_m], (1-q_1)[T_p, T_s]] \to [q_1[Tm], (1-q_1)q_2[T_{pm}, T_m], (1-q_1)(1-q_2)[T_{pp}, T_s, T_{ps}]] \ldots$$

where T_{pm} indicates a previously pruned tree that has been merged, and so on. This need be done only once for each tree, and then the "sum-of-stars" tree list can be used for fast likelihood calculations on that tree.

HKY85 model

The F81 assumes equal rates of transition and transversion. The HKY85 [12] model (and also the similar but non-equivalent F84 model [15]) are modifications of the F81 model to take account of a non-unity transition-transversion ratio κ. The rate matrix of the HKY model is

$$R = \mu \begin{pmatrix} * & \pi_A & \kappa\pi_A & \pi_A \\ \pi_C & * & \pi_C & \kappa\pi_C \\ \kappa\pi_G & \pi_G & * & \pi_G \\ \pi_T & \kappa\pi_T & \pi_T & * \end{pmatrix} \tag{5}$$

where the diagonal elements are such that columns sum to zero. We choose time units such that $\mu = 1$ and, as before, define proximities as $q = e^{-t}$ (however, these no longer have the simple interpretation of probability of conservation as in the F81 model).

We further define $\pi_R = \pi_A + \pi_G$; $\pi_Y = \pi_C + \pi_T$; $e_2 = e^{-t} \equiv q$; $e_3 = q^{\kappa\pi_R + \pi_Y}$; and $e_4 = q^{\kappa\pi_Y + \pi_R}$.

In terms of these, the transition probabilities are: [12]
Identity for purines ($\beta = \alpha$ for $\alpha = A, G$):

$$P_{\beta\alpha} = \pi_\alpha + \pi_\alpha \frac{\pi_Y}{\pi_R} e_2 + \left(1 - \frac{\pi_\alpha}{\pi_R}\right) e_3$$

Identity for pyrimidines ($\beta = \alpha$ for $\alpha = C, T$):

$$P_{\beta\alpha} = \pi_\alpha + \pi_\alpha \frac{\pi_R}{\pi_Y} e_2 + \left(1 - \frac{\pi_\alpha}{\pi_Y}\right) e_4$$

Transition, purine ($\beta = A, \alpha = G$ or vice versa):

$$P_{\beta\alpha} = \pi_\beta \left(1 + \frac{\pi_Y}{\pi_R} e_2 - \frac{1}{\pi_R} e_3\right)$$

Transition, pyrimidine ($\beta = C, \alpha = T$ or vice versa)::

$$P_{\beta\alpha} = \pi_\beta \left(1 + \frac{\pi_R}{\pi_Y} e_2 - \frac{1}{\pi_Y} e_4\right)$$

Transversion (all other cases):

$$P_{\beta\alpha} = \pi_\beta (1 - e_2)$$

These can be derived by, for example, diagonalizing R, or via an argument involving likelihoods of different mutations analogous to the F81 picture [17].

Position-specific stationary vectors for TFBS

Our approach is to calculate the q's first (and, for HKY85, κ), which we assume are not site-specific, using unconserved/uninformative sites (which we assume are evolving more-or-less neutrally) and assuming a stationary vector $\pi_A = \pi_T = 0.3$, $\pi_C = \pi_G = 0.2$

(the approximate nucleotide distribution in mammals), by maximising the likelihood of these neutral positions. Having fixed the q's and κ, we calculate the site-specific $\vec{\pi}$'s within TFBS by by, again, numerically maximising the likelihood of the observed nucleotides at those positions.

The calculation of q's and κ in this work was done, using a specific uninformative position in the neighbourhood of one motif (combining all sites for that motif); the same q's were used for all further calculations in that species. In principle, the q's can also be calculated from other neutral loci, such as synonymous sites in coding sequence (correcting for codon bias).

Ancestral weight vectors

It is of interest to infer the likely ancestral nucleotide at every locus.

At every locus, we have a collection of leaves L. Let the ancestral node be denoted by X. At a locus i, the probability of the ancestor X having nucleotide α can be written as

$$p_{iX}(\alpha|L) \propto p_i(L|\alpha; X)p(\alpha; X)$$

We take $p(\alpha; X)$ (the prior for α at X) to be b_α (the background probability for α, = 0.3 for A, T, 0.2 for G, C) and $p_i(L|\alpha; X)$ is the likelihood of the leaves at that locus given α at the root. This likelihood can be calculated by the fast "sum of stars" method for F81, or by the standard Felsenstein algorithm for other models.

Data used

ChIP-seq data was obtained from Gene Transcription Regulation Database [18, 19] (GTRD). For each TF, TFBS within chip-seq peaks were predicted using FIMO [20], with motif models (PWMs) from HOCOMOCO [21]. 20bp flanking sequence was added to each motif, in order to examine possible selection effects on flanking sequence. UCSC LiftOver (command line tool from `https://genome.ucsc.edu`)was used to find orthologous regions in other organisms. Only gapless liftOver matches were considered. The organisms and assemblies used were:

Primates: human (hg38), chimpanzee (PanTro6), gorilla (GorGor6), orangutan (PonAbe3), rhesus monkey (RheMac10)

Mammals: human (hg38), rhesus monkey (RheMac10), mouse (Mm39), cat (FelCat9), dog (CanFam6), pig (SusScr11), horse (EquCab3)

The human and orthologous TFBS sequence were stored in a custom file format based on the FASTA format. Processed data for the factors presented here, and other factors, are available in the github link cited in "Data availability".

Figure 2: The Felsenstein pruning algorithm's running time, compared to our "fast-star" algorithm. As a function of tree size, we achieve a speedup of $> 50\times$ over the pruning algorithm for trees with 5 leaves, and $> 4\times$ for trees with 10 leaves, over a range of sequence lengths. For larger trees our performance deteriorates.

Results

The sum-of-stars F81 likelihood algorithm shows significant speedup for moderate-sized trees

Figure 2 shows the speed of our implementation of the Felsenstein pruning algorithm, compared with our "fast-star" algorithm, for the F81 model on trees of varying sizes and with sequences of varying length. For trees with $\lesssim 10$ leaves we obtain a significant speedup.

Further speedup will be possible: our current implementation of fast-star still uses a tree data structure, but star trees can be represented using matrices, which will greatly improve cache performance.

PSSVs resemble PWMs, but more weakly

Figure 3 shows PWMs, PSSVs and conditional PSSVs (described below) for five representative TFs: CTCF, GABPA, REST, TCF12, USF2. More examples are available via notebooks on github (see Data availability). The PWMs are nucleotide counts within one species (human). PSSVs, position-specific stationary vectors on individual positions derived from an evolutionary model, closely reflect the pattern of the PWM, but are weaker (only slightly weaker for GABPA, but significantly weaker for TCF12, perhaps indicating relative turnover rates of these binding sites). Figure 4 shows further weakening for PSSVs when more distant mammals are included.

Multiple TFBS exhibit positional correlations based on ancestral nucleotide selections

Figure 3 shows the effect of conditioning certain nucleotides ancestrally on the resulting PSSV, that is, it shows PSSVs for subsets of sites where the inferred ancestral nucleotide

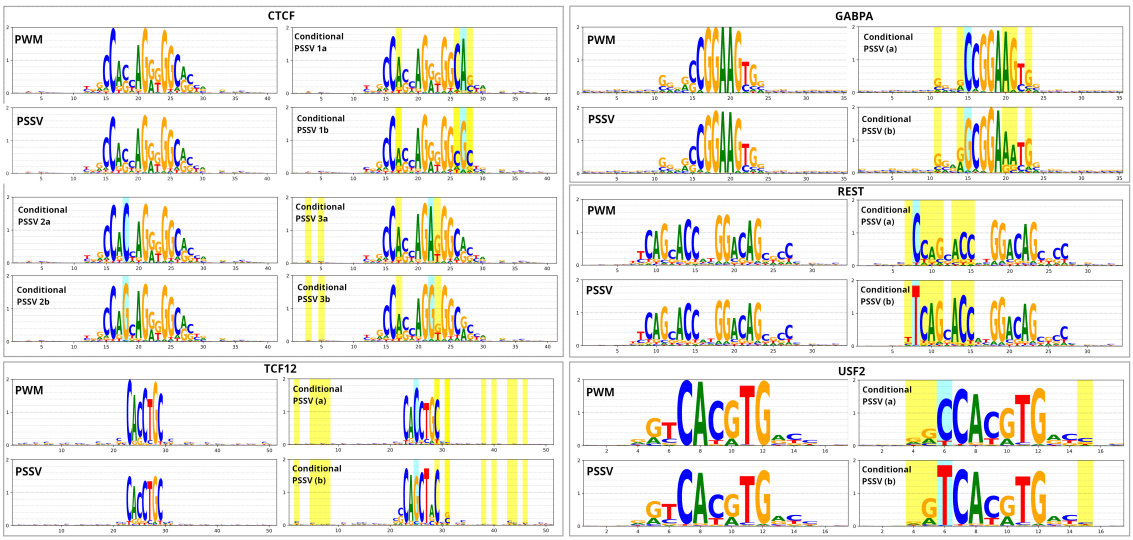

Figure 3: PWMs, PSSVs and conditional PSSVs for five transcription factors, calculated from primate data. In conditional PSSVs, the blue highlight is the ancestral nucleotide upon which the PSSV is conditioned, and the yellow highlights show positions that show significant dependency on this conditioning.

is a particular nucleotide α with a probability ≥ 0.5. Blue highlights indicate the position of ancestral conditioning, and the yellow highlights indicate positions that vary significantly on the ancestor.

CTCF is a widely-studied DNA-binding protein that plays roles in transcriptional regulation, insulator function, and chromatin remodelling [22]. Its binding sequence has long been recognized to have significant positional correlations [23]. We observe here that an ancestral preference for A at position 27 causes a preference of C at position 26 and G at position 28, while an ancestral preference of G at 27 causes a weaker preference of C at 26 but a stronger preference of C at 28. In contrast, an ancestral preference of C or G at position 18 appears to show no significant change in the PSSV at other positions. An ancestral preference of A or G at position 22 shows weak effects at positions 17 and 26. But also, positions 3 and 5 suggest signs of a motif, previously called the M2 motif [24], that has previously been noted to occur in some CTCF instances; it appears an A at 22 is associated with a slightly stronger M2 motif. We will investigate CTCF further in future work.

Other factors shown in figure 3 show similar positional correlations. Of interest: REST has a dimeric motif, but a choice of C in position 8 seems to suppress the first half of the dimer. And TCF12's binding sites have CG-rich sequence signatures extending well beyond the core motif, which also reflect in the PSSV, indicating selective pressure in an extended region around the core site. An ancestral choice of G at position 25 appears to strengthen the PSSV overall, including these peripheral CG signatures, but greatly

weaken position 28.

The F81 and HKY85 models give similar PSSVs

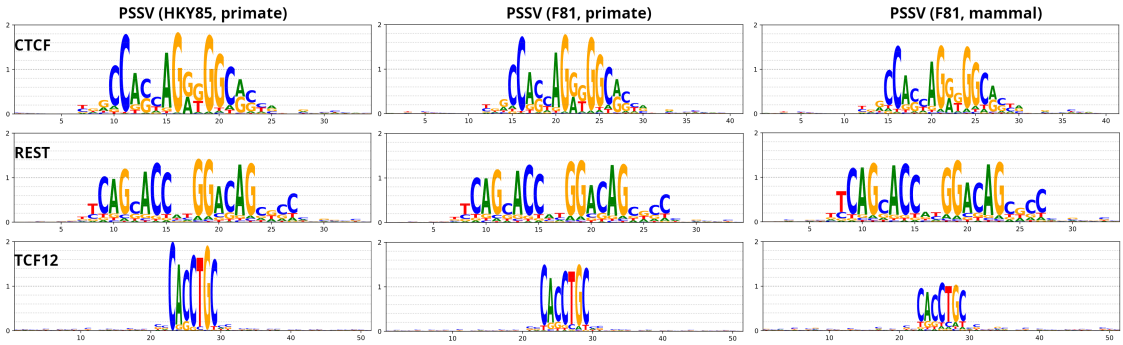

Figure 4: Comparison of PSSVs obtained from the HKY85 model on five primates, the F81 model on primates, and the F81 model on seven mammal species.

Figure 4 compares PSSVs obtained from the HKY85 model on five primates, the F81 model on the same 5 primates, and the F81 model on 7 mammals. All are similar, though the F81 mammal PSSVs are weaker.

Discussion

TFBS evolution has previously been studied from various points of view. Gerland and Hwa [25] presented a model of molecular evolution that includes genotype, phenotype and fitness, focussing on the fitness advantag to maintain a binding sequence and the evolution of multiple sites in a regulatory region. Similarly, Sengupta *et al.* [26] considered, theoretically, the tradeoff between TFBS specificity and robustness to mutation. Bais *et al.* [27] considered TFBS evolution in the context of annotated genome alignments; they consider both the F81 model and the Halpern-Bruno model. Mustonen *et al.* [28] considered a biophysical model for yeast TFBS evolution, where positional correlations emerge as a consequence of maintaining energy-dependent fitness (via compensatory mutations).

Most similarly to our work, Moses *et al.* [10] discuss position-specific evolution in TFBS via the Halpern-Bruno model, originally developed for protein-coding sequence [9]. This model incorporates selection on top of a mutational model. Essentially, the rate of evolution is modelled as a product of two terms: a site-independent rate of mutation and a site-dependent rate of fixation ($R_{\alpha\beta p} \propto r_{\alpha\beta} F_{\alpha\beta p}$ where $r_{\alpha\beta}$ is the rate of mutation from β to α and $F_{\alpha\beta p}$ is the probability of fixation from β to α at position p. They use Kimura's weak-mutation model (time of fixation $\ll$ time between fixations) to estimate F in terms of selection coefficients, which in turn is estimated from the equilibrium frequency (stationary distribution) at position p. A key assumption is that this stationary distribution

(which indicates the long-term distribution of nucleotides at a given locus over the course of evolution) is given by the corresponding column of the position weight matrix (which gives the distribution of nucleotides at that position across TFBS instances in a single species).

We make the contrary assumption, that the stationary distribution at a position is *not necessarily* given by the nucleotide counts for TFBS instances within a species. We therefore opt for the simpler F81 and HKY85 models to ascertain stationary distributions. We find that position-specific stationary vectors (PSSVs) at TFBS loci reflect position weight matrices (PWMs), under both F81 and HKY85, but are weaker than the PWMs in terms of information content, and weaken further when looking at a larger group of mammals compared to using primates alone. Since we have no data on binding affinity in other primates or mammals, and are going purely by gapless alignments, this observation may arise from evolutionary divergence or turnover of TFBS. This varies by TF, with GABPA being strongly conserved among mammals, TCF12 being significantly weakened.

TFBS turnover has been reported widely earlier, in human/mouse [29] and fruitfly [30, 31]. Recently Krieger al. [32] studied the effect of TFBS sequence variation and its effect on TFBS binding in two closely-related yeast species, considering both motif variation and chromatin accessibility, and find that imprecise motifs are bound to a high level by TFs and, in many cases, binding localization in the two species is conserved despite sequence divergence, suggesting "fast and flexible evolution" of TFBS. The same is likely true in other organisms.

It is widely recognized that PWMs are not adequate representations of TFBS complexity, which contain significant positional correlations [22] and multiple efforts have been made to go beyond the PWM model [2, 3, 4, 5]. As a complementary approach, we show that positional dependency effects can be examined by conditioning PSSVs on specific ancestral nucleotides.

We consider the F81 and HKY85 models. The primary difference is that HKY85 accounts for differing transition and transversion rates, which is an important factor in neutral evolution and in general phylogenetic problems. However, both models can reproduce any arbitrary stationary vector, and we observe that both models result in similar stationary vectors (which arises from a combination of mutation and purifying selection, not from mutation alone). The F81 model is conceptually simpler and we demonstrate an algorithm that greatly improves calculation time on moderate-sized trees ($\lesssim 10$ leaves), enabling rapid notebook-style analyses of sets of aligned TFBS. We plan to perform detailed evolutionary analysis of TFBS, and of CTCF in particular, in future studies, and hope our tool and methods will be useful to others.

Author contributions

Conceptualization: RS
Data Curation: PS
Formal Analysis: PS, RS

Funding Acquisition: RS
Investigation: PS, RS
Methodology: RS
Project Administration: RS
Resources: RS
Software: PS, RS
Supervision: RS
Validation: PS, RS
Visualization: RS
Writing – Original Draft Preparation: PS, RS
Writing – Review & Editing: PS, RS

Data and code availability

Processed data for primate and mammal sites, and code in the form of Jupyter notebooks, are available at `https://github.com/rsidd120/TFBS-PSSV`.

Funding

PS was funded by a fellowship at the Institute of Mathematical Sciences (IMSc) via the Department of Atomic Energy (DAE), Government of India. The work was supported by the computational biology project, an internal apex project at IMSc funded by DAE.

References

[1] Stormo GD, Hartzell 3rd G. Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences. 1989;86(4):1183–1187.

[2] Sharon E, Lubliner S, Segal E. A feature-based approach to modeling protein–DNA interactions. PLoS computational biology. 2008;4(8):e1000154.

[3] Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PloS one. 2010;5(3):e9722.

[4] Kulakovskiy IV, Levitsky VG, Oschepkov DG, Vorontsov IE, Makeev VJ. Learning Advanced TFBS Models from Chip-Seq Data-diChIPMunk: Effective Construction of Dinucleotide Positional Weight Matrices. In: International Conference on Bioinformatics Models, Methods and Algorithms. vol. 2. SCITEPRESS; 2013. p. 146–150.

[5] Omidi S, Zavolan M, Pachkov M, Breda J, Berger S, van Nimwegen E. Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. PLoS computational biology. 2017;13(7):e1005176.

[6] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004;14(6):1188–1190.

[7] Jukes TH, Cantor CR, et al. Evolution of protein molecules. Mammalian protein metabolism. 1969;3:21–132.

[8] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution. 1980;16:111–120.

[9] Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Molecular biology and evolution. 1998;15(7):910–917.

[10] Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. Position specific variation in the rate of evolution in transcription factor binding sites. BMC evolutionary biology. 2003;3(1):1–13.

[11] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981;17:368–376.

[12] Hasegawa M, Kishino H, Yano Ta. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of molecular evolution. 1985;22:160–174.

[13] Tavaré S. Some probabilistic and statistical problems on the analysis of DNA sequence. Lecture of Mathematics for Life Science. 1986;17:57.

[14] Siddharthan R, Siggia ED, Van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS computational biology. 2005;1(7):e67.

[15] Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of molecular evolution. 1989;29:170–179.

[16] Pond SLK, Muse SV. Column sorting: Rapid calculation of the phylogenetic likelihood function. Systematic biology. 2004;53(5):685–692.

[17] Xia X. Deriving transition probabilities and evolutionary distances from substitution rate matrix by probability reasoning. J Genet Genome Res. 2017;3(031).

[18] Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation—2019 update. Nucleic acids research. 2019;47(D1):D100–D105.

[19] Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, et al. GTRD: an integrated view of transcription regulation. Nucleic acids research. 2021;49(D1):D104–D111.

[20] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–1018.

[21] Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic acids research. 2013;41(D1):D195–D202.

[22] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194–1211.

[23] Eggeling R, Gohr A, Keilwagen J, Mohr M, Posch S, Smith AD, et al. On the value of intra-motif dependencies of human insulator protein CTCF. PLoS One. 2014;9(1):e85629.

[24] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148(1):335–348.

[25] Gerland U, Hwa T. On the selection and evolution of regulatory DNA motifs. Journal of molecular evolution. 2002;55:386–400.

[26] Sengupta AM, Djordjevic M, Shraiman BI. Specificity and robustness in transcription control networks. Proceedings of the National Academy of Sciences. 2002;99(4):2072–2077.

[27] Bais AS, Grossmann S, Vingron M. Incorporating evolution of transcription factor binding sites into annotated alignments. Journal of biosciences. 2007;32:841–850.

[28] Mustonen V, Kinney J, Callan Jr CG, Lässig M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences. 2008;105(34):12376–12381.

[29] Dermitzakis ET, Clark AG. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Molecular biology and evolution. 2002;19(7):1114–1121.

[30] Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS computational biology. 2006;2(10):e130.

[31] Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS biology. 2010;8(3):e1000343.

[32] Krieger G, Lupo O, Wittkopp P, Barkai N. Evolution of transcription factor binding through sequence variations and turnover of binding sites. Genome Research. 2022;32(6):1099–1111.