Deterministic models for an evolving transcription factor and its binding sites

We consider the case where a transcription factor protein SP has two variants, SP$_A$ and SP$_C$. Protein SP$_A$ originally represents the wild-type allele, while mutant protein SP$_C$ first arises in a single individual at generation $t = 0$. These two transcription factors recognize different binding sequences, denoted as BOX$_A$ and BOX$_C$, respectively. We assume some adaptive benefit for the mutant protein SP$_C$ to bind to BOX$_C$, and our goal is to determine the course of events by which the frequency of allele SP$_C$ as well as the frequencies of BOX$_A$ and BOX$_C$ change within the population over the course of evolution.

We model only sequences for which binding of the SP protein is beneficial. For the wild-type variant SP$_A$, a promoter containing BOX$_A$ has the relative fitness 1. The binding of mutant SP$_C$ to BOX$_C$ has an adaptive advantage, so that promoters containing BOX$_C$ in the presence of SP$_C$ have a relative fitness $1 + s_C$ (where $s_C > 0$). Since we consider only genes for which SP protein binding is beneficial, promoters without BOX$_A$ in the presence of SP$_A$ and without BOX$_C$ in the presence of SP$_C$ have a lower relative fitness, given by $1 - s_0$ ($s_0 > 1$).

We allow both BOX$_A$ and BOX$_C$ to be present in the same promoter, each either present or absent at a given gene. Thus, there are four possible haplotypes for each promoter: that containing no binding sites (h_0), those with only BOX$_A$ (h_A), those with only BOX$_C$ (h_C), and those containing both binding sequences (h_{AC}). We will set the frequencies for these haplotypes to be y_0, y_A, y_C, and y_{AC}, respectively, where $y_0 + y_A + y_C + y_{AC} = 1$. We denote the frequency of SP$_A$ and SP$_C$ within the population to be p and q, respectively, where again $p + q = 1$.

Thus, for a given gene in an individual, there exist several possible phenotypes, which we denote as $H_{i,j}$ for $i, j \in \{0, A, C, AC\}$. The fitness $w_{i,j}$ of each of these phenotypes is given in Table 1. Given the current frequency y_x of haplotype x within the population (where $x \in \{0, A, C, AC\}$), the new frequency y'_x of haplotype x in the next generation is given by

$$y'_x = \frac{y_x \sum_{j \in \{0,A,C,AC\}} y_j w_{x,j}}{\sum_{i,j \in \{0,A,C,AC\}} y_i y_j w_{i,j}} \tag{1}$$

Also of interest is the change in frequency of the SP protein alleles within the population, i.e., the change in p and q over time. The possible phenotypes of the SP protein are AA, AC, and CC. If we let $W_{A,A}$, $W_{A,C}$, and $W_{C,C}$ be the fitnesses of each of these phenotypes, respectively, then we

Phenotypes ($H_{i,j}$)	Fitness ($w_{i,j}$)
$H_{0,0}$	$1 - s_0$
$H_{0,A}$ $H_{A,A}$	$p^2 + 2pq + q^2(1 - s_0)$
$H_{0,C}$ $H_{C,C}$	$p^2(1 - s_0) + (2pq + q^2)(1 + s_C)$
$H_{0,AC}$ $H_{A,C}$ $H_{A,AC}$ $H_{C,AC}$ $H_{AC,AC}$	$p^2 + (2pq + q^2)(1 + s_C)$

Table 1: Relative fitness values for individual promoter phenotypes.

see that

$$p' = \frac{p^2 W_{A,A} + pqW_{A,C}}{p^2 W_{A,A} + 2pqW_{A,C} + q^2 W_{C,C}} \tag{2}$$

and

$$q' = \frac{q^2 W_{C,C} + pqW_{A,C}}{p^2 W_{A,A} + 2pqW_{A,C} + q^2 W_{C,C}} \tag{3}$$

We note that the fitnesses $W_{A,A}$, $W_{A,C}$, and $W_{C,C}$ are determined by the frequencies of binding sites BOX_A and BOX_C across genes within the population. We let G represent the set of L genes considered, where $G = \{g_1, g_2, ..., g_L\}$. Each gene g_i then has a corresponding frequency of binding site alleles, $y_0(g_i)$, $y_A(g_i)$, $y_C(g_i)$, and $y_{AC}(g_i)$. Assuming Hardy-Weinberg equilibrium, the expected numbers L_A and L_C of genes containing BOX_A and BOX_C, respectively, are

$$L_A = \sum_{i=1}^{L} [y_A(g_i) + y_{AC}(g_i)]^2 + 2[y_A(g_i) + y_{AC}(g_i)][y_0(g_i) + y_C(g_i)] \tag{4}$$

$$L_C = \sum_{i=1}^{L} [y_C(g_i) + y_{AC}(g_i)]^2 + 2[y_C(g_i) + y_{AC}(g_i)][y_0(g_i) + y_A(g_i)] \tag{5}$$

We then calculate $W_{A,A}$, $W_{A,C}$, and $W_{C,C}$ assuming multiplicative fitnesses across loci:

$$W_{A,A} = (1)^{L_A}(1 - s_0)^{L - L_A} \tag{6}$$

$$W_{A,C} = (1 + s_C)^{L_C}(1)^{L_A - L_{AC}}(1 - s_0)^{L - L_A - L_C + L_{AC}} \tag{7}$$

$$W_{C,C} = (1 - s_0)^{L - L_C}(1 + s_0)^{L_C} \tag{8}$$

Here, L_{AC} is the expected number of genes containing both BOX_A and BOX_C, which is estimated to be $L_{AC} = (L_A \cdot L_C)/L$.

The above model considers changes in *trans*- and *cis*-regulatory element frequencies according only to natural selection acting upon phenotypes existing in the initial population. However, the process of regulatory element evolution also involves mutations, including gains and losses of regulatory elements as well as transitions between different binding sequences. Thus, we must incorporate such processes into the model.

For the mutational process, we consider a birth-death-transition model, where sequence elements can be gained, lost, or converted to the alternate binding sequence (i.e., $\mathrm{BOX}_A \rightarrow \mathrm{BOX}_C$ or vice versa). The mutation process, then, alters the frequencies of the haplotypes within the population individually for each gene. Thus, if y_0, y_A, y_C, y_{AC} represent the frequencies of the haplotypes in a given generation, the mutation process creates new population frequencies $y'_0, y'_A, y'_C, y'_{AC}$ according to mutation rate parameters and the previous haplotype frequencies.

The mutation process is defined by three parameters, the birth parameter (θ_α), the death parameter (θ_β), and the conversion parameter (θ_T). The birth parameter θ_α represents the rate (per generation) at which a new binding site appears in a previously empty promoter, while the death parameter θ_β represents the rate (per generation) at which a promoter with an existing binding site looses its binding site and becomes empty. The conversion parameter θ_T represents the rate at which a BOX$_A$ binding site converts to a BOX$_C$ binding site, or vice versa. For consistency, we assume that these parameters are identical for BOX$_A$ and BOX$_C$ sequence elements.

We define our parameters θ_α and θ_β according to the birth/death parameters α and β estimated in our birth-death model. We recall that α represents the probability that a binding site appears at a given unoccupied nucleotide site in one year, and that β represents the probability that an existing binding site is lost in one year. If the number of years between generations is R and the width of the binding site target region is D, then approximate the birth and death parameters to be $\theta_\alpha = RD\alpha$ and $\theta_\beta = R\beta$. Choosing a value for the conversion parameter θ_T is more arbitrary, so we conducted several simulation analyses assuming different values for this parameter.

Determining the population frequencies following this mutation process is straightforward. Given the initial haplotype frequency vector $[y_0, y_A, y_C, y_{AC}]^T$, we can determine the haplotype frequency vector $[y'_0, y'_A, y'_C, y'_{AC}]^T$ following the mutation process through matrix algebra:

$$\begin{bmatrix} 1-2\theta_\alpha & \theta_\beta & \theta_\beta & 0 \\ \theta_\alpha & 1-\theta_\beta-\theta_T-\theta_\alpha & \theta_T & \theta_\beta \\ \theta_\alpha & \theta_T & 1-\theta_\beta-\theta_T-\theta_\alpha & \theta_\beta \\ 0 & \theta_\alpha & \theta_\alpha & 1-2\theta_\beta \end{bmatrix} \begin{bmatrix} y_0 \\ y_A \\ y_C \\ y_{AC} \end{bmatrix} = \begin{bmatrix} y'_0 \\ y'_A \\ y'_C \\ y'_{AC} \end{bmatrix} \tag{9}$$