

1 Introduction

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 BOX_C , and our goal is to determine the course of events by which the frequency of allele SP_C within the population changes over the course of evolution. We set the frequency of the wild-type SP_A phenotype at time t to be p_t and the mutant SP_C to be q_t .

2 Relative fitness of individuals

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 $1 - s_0$ in both cases ($s_0 > 1$). We allow both BOX_A and BOX_C to be present in the same promoter, each with potentially multiple instances in the same gene. Moreover, the number of each binding element within a single gene may vary within the population, as does the SP protein phenotype.

Let H represent the SP protein phenotype, which in haploids can be either A or C , depending on whether SP_A or SP_C is present in an individual. In diploids, H has three possible values, AA , AC , or CC . Within an individual, we can set L_A be the number of genes containing at least one copy of BOX_A , and we set L_C be the number of genes containing at least one copy of BOX_C . The total number of genes is denoted as L . For haploids, we denote the total fitness F_H be the fitness of that individual with SP protein phenotype H . For an individual carrying SP_A , the total fitness is

$$F_A = (1)^{L_A}(1 - s_0)^{L-L_A} \quad (1)$$

while the fitness of an individual carrying SP_C is

$$F_C = (1 + s_C)^{L_C}(1 - s_0)^{L-L_C} \quad (2)$$

The diploid case is more complex, because both the SP protein allele as well as the number of binding sites BOX_A and BOX_C in each gene can be heterozygous. We assume here that the existence of BOX_A and/or BOX_C in a promoter is a dominant trait, such that heterozygosity produces a fitness identical to that of a homozygous individual carrying a particular binding site. In this framework, we let L_A and L_C be the number of genes containing BOX_A or BOX_C in at least one of the chromosome copies.

Fitnesses for homozygous SP alleles AA and CC are similar to those of haploids, and are given by

$$F_{AA} = (1)^{L_A}(1 - s_0)^{L-L_A} \quad (3)$$

$$F_{CC} = (1 + s_C)^{L_C}(1 - s_0)^{L-L_C} \quad (4)$$

In addition, in cases where the phenotype for the SP proteins are heterozygous, we consider binding of SP_C to be dominant to the binding of SP_A . Thus, we see a corresponding increase of fitness upon

binding of SP_C to BOX_C , regardless of whether BOX_A is present or not. In the heterozygous SP phenotype AC , we need to define an extra value L_{AC} , which represents the number of promoters in which both BOX_A and BOX_C are present. Under the simplifying assumption that BOX_A and BOX_C elements occur independently, we can estimate L_{AC} to be $L_{AC} = L_A L_C / L$. The total fitness of these individuals are

$$F_{AC} = (1 + s_C)^{L_C} (1)^{L_A - L_{AC}} (1 - s_0)^{L - L_A - L_C + L_{AC}} \quad (5)$$

3 SP protein phenotypes within the population

The fitness of an individual, given by both the SP protein phenotype and the existence of binding sites in each promoter, stochastically determines the probability of passing along the total phenotype to the next generation. We remember that, in generation t , the frequency of the SP_A allele is given by p_t , while the frequency of allele SP_C is q_t . Note that $p_t + q_t = 1$ for all t .

First, let us consider the haploid case. Given frequencies p_t and q_t at generation t , we can determine the frequencies p_{t+1} and q_{t+1} in the next generation according to the relative fitnesses F for each phenotype. However, we must remember that individuals within the population may carry varying binding site phenotypes. We let $p_t(L_A)$ represent the fraction of the population carrying allele SP_A and L_A promoters containing BOX_A elements, so that $\sum_k p_t(k) = p_t$. Similarly, we let $q_t(L_C)$ be the fraction of the population carrying the SP_C allele and L_C promoters containing BOX_C elements. Then,

$$p_{t+1}(L_A) = \frac{p_t(L_A) F_A(L_A)}{\sum_k p_t(k) F_A(k) + \sum_k q_t(k) F_C(k)} \quad (6)$$

where $F(k)$ represents the relative fitness given k promoters containing the corresponding binding site elements.

For diploids, we introduce heterozygosity for both SP protein phenotypes as well as the number of binding elements. For each generation t , we have frequencies p_t and q_t , which represent the frequencies of alleles SP_A and SP_C , respectively. If we let g

Here, we will denote values involving p and q to be similar as for the haploid case, assuming homozygous alleles AA and CC for the SP protein, respectively. We also introduce values r , which represent frequencies for heterozygous individuals carrying both SP_A and SP_C . Given this notation, we see that

$$p_{t+1}(L_A) = \frac{p_t(L_A)^2 F_{AA}(L_A) + p_t(L_A) \sum_k q_t(k) F_{AC}(L_A, k)}{P + Q + R} \quad (7)$$

$$q_{t+1}(L_C) = \frac{q_t(L_C)^2 F_{CC}(L_C) + q_t(L_C) \sum_k p_t(k) F_{AC}(k, L_C)}{P + Q + R} \quad (8)$$

where

$$P = \sum_k p_t(k)^2 F_{AA}(k) \quad (9)$$

$$Q = \sum_k q_t(k)^2 F_{CC}(k) \quad (10)$$

$$R = 2 \sum_{k,m} p_t(k) q_t(m) F_{AC}(k, m) \quad (11)$$

where $F_{AC}(L_A, L_C, L_{AC})$ represents the relative fitness of an individual heterozygous for the SP protein with L_A promoters containing BOX_A , L_C promoters containing BOX_C , and L_{AC} promoters containing both. Total frequencies of each SP protein phenotype can be obtained through the summation of $p(k)$, $q(k)$, and $r(k_A, k_C, k_{AC})$ across all values of $k = 0, 1, 2, \dots, L$ (or k_A , k_C , and k_{AC} for heterozygous SP protein individuals), i.e., all possible numbers of promoters containing the corresponding binding sites.

4 Cis-regulatory mutations

So far, we have considered cases without any gains, losses, or conversions between binding elements BOX_A and BOX_C . However, mutations clearly arise over the course of evolution, and thus must be incorporated into the model.

Considering the diploid case, we remember that L_A and L_C represent the number of genes with $\text{BOX}_A/\text{BOX}_C$ on at least one chromosome within a single individual. Assuming Hardy-Weinberg equilibrium and linkage equilibrium, we can then calculate the overall frequency (b_A and b_C) of promoters with $\text{BOX}_A/\text{BOX}_C$. That is,

$$L_A/L = b_A^2 + 2b_A(1 - b_A) \quad (12)$$

$$L_C/L = b_C^2 + 2b_C(1 - b_C) \quad (13)$$

where L is the total number of genes. This gives $b_A = 1 - \sqrt{1 - L_A/L}$ and $b_C = 1 - \sqrt{1 - L_C/L}$. The total number of promoters L'_A containing BOX_A (2 per gene in a diploid) is then $L'_A = 2b_AL$, and similarly for BOX_C .

Allowing for mutations stochastically changes these allele frequencies at some probability during each generation. In some cases, binding sites can be gained *de novo* or lost completely. In other cases, we may observe binding site conversions between the two binding elements, where BOX_A converts to BOX_C , or vice versa. Let us suppose that the probability of gaining a copy of BOX_A or BOX_C *de novo* in the promoter of a single individual is μ_A and μ_C , respectively. Then, the number of BOX_A binding sites gained in an individual in one generation (U_A) follows the binomial distribution, with probability of success μ_A and number of trials $2L(1 - b_A)$. Similar rules follow for BOX_C .

In addition to the birth of new binding sites, binding sites can also be lost over the course of evolution. We let ν_A and ν_C represent the probability that a promoter with BOX_A or BOX_C loses that binding site in one generation. The number of deaths (V) then is also given by the binomial distribution with probability ν and number of trials $2Lb$.

Binding site conversions, where $\text{BOX}_A \rightarrow \text{BOX}_C$ or $\text{BOX}_C \rightarrow \text{BOX}_A$, can also be modeled. For simplicity, we assume that the frequency of conversion is the same in either direction, and occur at

probability κ per promoter in any individual per generation. If we first consider the case for which BOX_A is present and BOX_C is absent from the same promoter, the number of $\text{BOX}_A \rightarrow \text{BOX}_C$ conversions (K) should also be given by the binomial distribution, with probability κ and number of trials $2Lb_A(1 - b_C)$. The opposite $\text{BOX}_C \rightarrow \text{BOX}_A$ conversion should also follow the binomial distribution, with probability κ and number of trials $2Lb_C(1 - b_A)$.

We can estimate the values for μ , ν , and κ empirically. Under neutral processes, the rate of mutation per individual locus is equal to the fixation rate within the population. Thus, a reasonable estimation for μ and ν can be determined according to the observed birth and death rates in the mammal phylogeny. More specifically, we can consider birth and death rates for (the neutrally evolving) BOX_C sites in lineages known to prefer BOX_A and vice versa.