

## 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 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.

## 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  $1 - s_0$  ( $s_0 > 1$ ). We allow both  $BOX_A$  and  $BOX_C$  to be present in the same promoter.

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  $\text{BOX}_A$  and  $\text{BOX}_C$  are present. The total fitnesses of these individuals are

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

For any gene  $g$ , we can imagine that there is some frequency  $y_A$  at which  $\text{BOX}_A$  exists within the population. Similarly, we can set  $y_C$  to be the frequency of  $\text{BOX}_C$  at this gene. Assuming Hardy-Weinberg equilibrium, the probability  $b_A$  that at least one chromosome of an individual contains  $\text{BOX}_A$  at this gene is  $b_A = y_A^2 + 2y_A(1 - y_A)$ ; we can define the probability  $b_C$  similarly using frequency  $y_C$ .

The total fitness  $F_{AA}$  of a population homozygous for the  $\text{SP}_A$  allele can be determined in the following way. Suppose we have a set  $G$  of  $L$  genes, where  $G = \{g_1, g_2, \dots, g_L\}$ , where the frequency of  $\text{BOX}_A$  at gene  $g_i$  is denoted as  $y_A(g_i)$ . For gene  $g_i$  in any individual, we define a random variable  $U_A(g_i)$  representing the presence or absence of  $\text{BOX}_A$ , where  $U_A(g_i) = 1$  if  $\text{BOX}_A$  is present, or  $U_A(g_i) = 0$  if it is absent. We note that  $U_A(g_i)$  simply represents a Bernoulli random variable, with probability of success  $b_A(g_i) = y_A(g_i)^2 + 2y_A(g_i)(1 - y_A(g_i))$ . The expected value, then, for the total number of genes with at least one copy of  $\text{BOX}_A$  ( $U_A$ ) for any individual is then

$$\mathbb{E}[U_A] = \sum_{i=1}^L \mathbb{E}[U_A(g_i)] = \sum_{i=1}^L b_A(g_i) = \sum_{i=1}^L [y_A(g_i)^2 + 2y_A(g_i)(1 - y_A(g_i))] \quad (6)$$

We can determine the expected number of genes with at least one copy of  $\text{BOX}_C$  for any individual ( $\mathbb{E}[U_C]$ ) in the same manner by replacing the values for  $y_A(g_i)$  with those for  $y_C(g_i)$ . These expected values allow us to estimate the average fitness of the population according to their SP protein phenotype using Equations 3-5, replacing  $L_A$ ,  $L_C$ , and  $L_{AC}$  with  $\mathbb{E}[U_A]$ ,  $\mathbb{E}[U_C]$ , and  $(\mathbb{E}[U_A] \cdot \mathbb{E}[U_C])/L$ , respectively.

### 3 Phenotype fluctuations within the population

Using the values for the relative fitness in each population, we can determine the frequency of each SP allele within the entire population. For frequency  $p_t$  of the  $\text{SP}_A$  allele at generation  $t$  and frequency  $q_t$  of the  $\text{SP}_C$  allele at time  $t$ , we see that

$$p_{t+1} = \frac{p_t^2 f_{AA} + p_t q_t f_{AC}}{p_t^2 f_{AA} + 2p_t q_t f_{AC} + q_t^2 f_{CC}} \quad (7)$$

$$q_{t+1} = \frac{q_t^2 f_{CC} + p_t q_t f_{AC}}{p_t^2 f_{AA} + 2p_t q_t f_{AC} + q_t^2 f_{CC}} \quad (8)$$

Also of interest is the change in frequency of binding site occurrences at each gene. We assume that  $p$  and  $q$  are the current frequencies of  $\text{SP}_A$  and  $\text{SP}_C$  at this time, with a total population size of  $N$ . We denote the fitness of a gene carrying neither  $\text{BOX}_A$  or  $\text{BOX}_C$  as  $F_{**}$ , the fitness of a gene carrying only  $\text{BOX}_A$  as  $F_{A*}$ , the fitness of a gene carrying only  $\text{BOX}_C$  as  $F_{*C}$ , and that carrying

both as  $F_{AC}$ . These are given by

$$F_{**} = (1 - s_0)^{p^2+2pq+q^2} \quad (9)$$

$$F_{A*} = (1)^{p^2+2pq}(1 - s_0)^{q^2} \quad (10)$$

$$F_{*C} = (1 + s_C)^{2pq+q^2}(1 - s_0)^{p^2} \quad (11)$$

$$F_{AC} = (1 + s_C)^{2pq+q^2}(1)^{p^2} \quad (12)$$

Given  $b_A$  and  $b_C$ , which again represent the frequency at which we observe at least one copy of  $\text{BOX}_A$  or  $\text{BOX}_C$  in a promoter, we let  $b'_A$  and  $b'_C$  be the new frequencies in the next generation. If we set

$$X_{**} = (1 - b_A)(1 - b_C) \quad (13)$$

$$X_{A*} = b_A(1 - b_C) \quad (14)$$

$$X_{*C} = b_C(1 - b_A) \quad (15)$$

$$X_{AC} = b_A b_C \quad (16)$$

we see that

$$b'_A = \frac{X_{A*}F_{A*} + X_{AC}F_{AC}}{X_{**}F_{**} + X_{A*}F_{A*} + X_{*C}F_{*C} + X_{AC}F_{AC}} \quad (17)$$

$$b'_C = \frac{X_{*C}F_{*C} + X_{AC}F_{AC}}{X_{**}F_{**} + X_{A*}F_{A*} + X_{*C}F_{*C} + X_{AC}F_{AC}} \quad (18)$$