The evolution of new functions takes place partially through changes in the way transcription is controlled. Transcriptional control is brought about by the interactions of transcription factors with short target motifs in the DNAs of promoters and enhancers. One way in which changes in gene expression can evolve is through the acquisition of new transcription factor targets in enhancer sequences. Since such target sites are simple, they can be produced rapidly from random DNA by mutation and selection. Here we consider a population of organisms that finds itself in an ecological situation where a particular target gene under the control of a particular transcription factor would be favored by natural selection. What will be the time required for such a process, as a function of the selection for the new target, the mutation rate, and the population size? The starting sequences considered are either real enhancers from the *Drosophila melanogaster* genome, or randomized versions of these. We find that the time required to find binding sites is strongly dependent on the existence of starting sequences of sites that differ from binding sites by single substitutions (pre-sites). The process of converting pre-sites to binding sites is driven by natural selection, and thus the time required typically reduces with the strength of selection. However, if there is a strongly distorted G:C ratio in the starting sequence, pre-sites will typically be absent, and the finding of binding sites will be preceded by a long time period of neutral evolution, however strong is the selection favoring sites. The positions of pre-sites largely determine where binding sites will evolve. One result of this is that any incremental selective benefits that result from the relative positioning of sites have a surprisingly small impact on the final binding-site positions.

**Introduction**

The recent increase in the study of regulatory regions and their evolution is a reflection of their importance to our understanding of the mechanisms of developmental evolution. The key interaction through which transcriptional regulatory sequences in DNA act is the binding of transcription factors to binding sites in enhancers. Enhancers are defined as sequences that can stimulate transcription in either orientation, from upstream or downstream and they can act from distances over 1 kb. Enhancers contain binding sites for multiple transcription factors, usually with multiple sites for each protein that are often of varying binding affinity. For a review of enhancer structure and function, see Arnone and Davidson (1997).

Evolutionary changes in the *cis* control of gene expression can occur through a number of mechanisms recently reviewed by Tautz (2000), Ludwig (2002), and Wray et al. (2003). These reviews give many examples of how enhancers evolve in various species and in a variety of regulatory systems.

Many studies have found within-species variation in enhancer sequences and expression patterns, providing the raw material for natural selection. For example, a large proportion of human *cis*-regulatory polymorphisms have a greater than twofold effect on transcription levels (Rockman and Wray 2002). It has also been shown that enhancer variation is acted upon by selection (Crawford et al. 1999; Segal et al. 1999).

There does not appear to be a single model to fit the evolution of all enhancers, but many systems have features in common. For example, a number of studies have shown how enhancer function can be maintained without maintaining sequence. The best-studied example of this is the *even-skipped* stripe 2 enhancer system, where stabilizing selection is thought to be maintaining the expression pattern but the sequence is changing, possibly due to weak deleterious changes followed by compensatory mutations (Ludwig et al. 1998; Ludwig et al. 2000).

A previous simulation study examined how long it would be expected to take for new binding sites to arise in a regulatory sequence via point mutations (Stone and Wray 2001). Here the authors considered a sequence of DNA evolving in a random neutral walk by point mutation, and observed how long it would take to find a particular binding site for a transcription factor. They showed, for example, for *Drosophila*, that to find two 6-bp transcription factor binding sites in a 200-bp sequence would take around 54,000 years, which seems reasonable given the rate at which *Drosophila* may evolve its gene expression patterns. However, the corresponding time for humans, with their longer generation times, was over 13 million years, which seems too slow a rate to be likely to be useful in evolutionary change. In addition, each of these times was unrealistically low, given the neutral random walk model, in that the effect of the assumed population size of a million (in each species) was incorporated inappropriately. The authors, in assuming an effective population size of a million, converted the expected time to find binding sites for a single lineage to an expected time for a population of a million by dividing the expected time by two million (with the two representing diplodcy). This procedure assumes that the two million copies of the gene in the population are carrying out independent random walks, and the time being calculated is the time until the first of these random walks finds the binding sites that are required. In reality, the two million gene copies will not all be expected to be independently evolving, but will be evolving in concert, since they will all be expected to share a common ancestry (under neutrality) around four million generations ago. Simply dividing the expected time by two million gives a value that is far too small. In addition, the time that is calculated is the expected time until the required sequence arises in one of the two million gene copies existing in a diploid population.
However, unless the selection favoring this evolved sequence is very strong, the sequence will probably be lost by drift soon after its creation. Its probability of fixation is only approximately $2s$, where $s$ is its selective advantage (or approximately $\frac{1}{2N}$, in a diploid of effective size $N$ if this is larger). Once these problems with the incorporation of the effective population size are appreciated, the appropriate conclusion from such a model is that the method of random neutral walk would not be effective at evolving multiple new binding sites in organisms with long generation times and small population sizes (such as metazoans).

Another previous approach to the evolution of enhancers was that of Carter and Wagner (2002), who explicitly imagined enhancer evolution occurring through pairs of individually deleterious but compensatory mutations. They considered a population genetics model in which, in large populations, evolution of enhancer sequences through compensatory mutations may be quite rapid. In this, they differ from our model described below, in which only neutral or advantageous mutations may spread. Gerland and Hwa (2002) studied the problem of the evolution of the interaction between a transcription factor and a particular binding site in terms of how strongly selection must favor binding in order to preserve a site in the face of inactivating mutations.

### The Model

Any realistic model of enhancer evolution must include selection, as many studies have shown selection to be important in regulatory evolution (Doebley et al. 1997; Schulte et al. 1997). In order to relate selection to our model we assume that each binding site contributes to the “output” of the sequence, where output can be thought of as analogous to transcriptional activation. Any mutations that increase the regulatory output of the element will be selectively advantageous up to a threshold, above which further increases in output provide no selective advantage. A position weight matrix (PWM) is used to describe the binding characteristics of a transcription factor, rather than simply a consensus sequence. The PWM uses all the information gained from DNA-protein interaction studies in the following way:

<table>
<thead>
<tr>
<th>Position Weight Matrix for Bicoid (M00140) Showing Calculated $C_i$ Values, Based on Twelve Bicoid Binding Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>

In this study, we consider a situation where a population of organisms is in a new environment, where the expression of an existing protein in a new tissue or at a new time would be advantageous. The evolution of a new expression pattern for one of a pair of genes following a gene duplication provides another example of a situation in which the evolution of new expression pattern could be favored by selection.

The model we present is one of a population, not of individual alleles. Mutations are introduced into a population and may go to fixation. If a particular base change that is a neutral mutation is occurring at rate $\mu$, then the rate at which the population will be expected to change by incorporating this mutation is still simply $\mu$. However, a particular base change that creates a selective advantage $s$ will also be arising at rate $\mu$ in an individual and thus at a rate $2N\mu$ in a diploid population of effective size $N$. If this advantageous mutation has a probability of fixation of $2s$, the evolutionary rate for this mutation will be $4N\mu s$. Thus, the likelihood of a particular selected mutation arising and spreading in the population relative to the rate for a particular neutral mutation is $4Ns$. Using this approach, in each step of the simulations, a mutation arises and is fixed in the population. The relative likelihoods of all possible base changes in evolution of the sequence are calculated, and used to weight the probabilities of each possible change. The particular evolutionary change occurring is chosen, using a pseudorandom number generator, from these probabilities. The expected time for each of these mutational/evolutionary steps can also be calculated from the proportion of changes that will be selectively driven and their individual $Ns$ values.

We consider the situations of de novo enhancer formation from nonfunctional DNA, and those in which an already existing enhancer is modified to make it respond to a transcription factor that it has previously ignored. Using real enhancer sequences, and randomized sequences, and specific Drosophila transcription factors, we consider the effect of variation in selection and population size on the time taken to reach a specified transcriptional output.

### Model Description

A position weight matrix (PWM) is used to describe the binding characteristics of a transcription factor, rather than simply a consensus sequence. The PWM uses all the information gained from DNA-protein interaction studies in the following way:

$$C_i(i) = \frac{100}{\ln 5} \times \left[ \sum_{b \in \text{A,C,T,G}} P(b, i) \times (\ln P(b, i) + \ln 5) \right]$$

where $P(b, i)$ is the relative frequency of nucleotide $b$ at position $i$. The $C_i$ for the $i$th position is a measure of the
The minimisation of false positives and that of false negatives, whereas the number of false positives tends to increase. The statistic gives a value for base at position that is required to create a functional binding site. Known binding sites can be compared to predicted binding sites at different levels of matrix similarity. The exact level used was calculated following the method of Papatsenko et al. (2002). Here the matrix similarity, used as the cutoff between functional and nonfunctional sequences, is chosen as the level that gives the maximum value from the following equation.

\[
\ln((H \times H)/(H + FN)) \ln(FN + FP)
\]

where \(H\) is the number of matches between experimentally confirmed and predicted binding sites, \(FP\) is the number of false positive sites, and \(FN\) is the number of false negative sites. As the cutoff level is reduced, so is the number of false negatives, whereas the number of false positives tends to increase. The statistic \(p\) represents a compromise between the minimisation of false positives and that of false negatives, and the cutoff point is chosen where \(p\) is maximised. This matrix similarity level can then be used in the simulation as the cutoff level at which a sequence becomes a functional binding site. For the transcription factors in this study the cutoff values used were as follows: Bicoid, 0.94; Krüppel, 0.84; and Abdominal-B: 0.84. These cutoff values were used as they gave the maximum value when compared with sequences with known binding sites. The sequences used for calculating the bicooid cutoff were Kr730 (Hoch et al. 1991) and S2E (AF042709) (Ludwig et al. 1998). For Krüppel the S2E was again used. For Abdominal B the empty spiracles enhancer (AF029219) was examined (Taylor 1998). However, the low number of binding sites known for Abd-B meant an accurate \(p\) value could not be produced. In this case, a threshold level was selected that gave the lowest number of false positives in sequences that were known not to bind Abd-b (Kr730 and S2E) and that gave the highest number of experimentally confirmed sites in the empty spiracles enhancer.

In order to include selection in the model, we need to calculate the effect on fitness of any given change in the DNA sequence. We achieved this by assigning an “output” value to a sequence based on its similarity to the matrix. If a point mutation increases the total output of the sequence then it can be selected for, as this implies an increase in transcriptional activation and therefore an increase in protein expression.

A maximum output level of two units is set for a sequence that has a matrix similarity score of 1. This is the maximum level of output a single binding site can contribute on its own. Each position in the matrix is assigned a proportion of this output based on its \(C_i\) level. The equation for this is shown as follows.

\[
O(i) = \frac{C_i(i)}{\sum_{j=1}^{n} C_j(j)} \times t
\]

where \(O(i)\) is the output at position \(i\), \(n\) is the length of the matrix, and \(t\) is the maximum output. Using the calculated \(O(i)\) values, each possible base at each position is assigned a proportion of the output for that position. This is calculated as shown below.

\[
Ob(b, i) = \frac{score(b, i)}{max\_score(i)} \times O(i)
\]

where \(Ob(b, i)\) is the output of base \(b\) at position \(i\), score \((b, i)\) is the matrix value of base \(b\) at position \(i\), max\_score \((i)\) is the matrix value of the most frequent base at position \(i\), and \(O(i)\) is the output at position \(i\).

This way of defining output makes it possible to weight base changes based on the importance of each position in the binding site. Positions that are in the core-binding region contribute more to the output of the binding site than do positions that are less important.

From these equations, it is possible to compare any sequence to the matrix, calculate its matrix similarity and therefore, by comparison to the threshold, determine whether it is an active binding site. If it is an active binding site then its output level can be determined.

The output level is calculated for individual binding sites, and the total output of the element is the sum of these individual outputs. However, the model also incorporates position effects. Binding sites which are close together and pairs of sites that are on the same side of the DNA stimulate...
transcription more than do those which are further apart and on opposite sides of the DNA (Mao et al. 1994; Liu and Little 1998). The model simulates this effect by adding what is called an "output bonus," shown as follows.

This graph is created using the formula below, using a range, \( r \), of 100 bp.

\[
ob = 1 - \left( \frac{d}{r} \right) \times \left( 1 + \cos(d \times 36) \right) + \frac{1 - \left( \frac{d+5}{r} \right) \times \left( 1 + \cos((d+5) \times 36) \right)}{3}
\]

where \( d \) is the distance in base pairs between the centers of the two binding sites being considered, \( r \) is the range over which the bonus is active, and \( ob \) is the output bonus.

As figure 1 shows, the further a site is from an existing binding site, the lower the bonus awarded. In addition, if a binding site is on the opposite side of the DNA strand, i.e., not a multiple of 10 bp away, then it also receives a lower bonus. The output bonus is added to the output that is calculated as the sum of all sites’ matrix similarities, multiplied by 2. Each binding site receives a bonus for every binding site within range, so output is increased by the sum of the output bonuses from all pairwise interactions. In this way, the effect of sites being clustered together can be to increase the total output of the sequence greatly, as is seen in real enhancers.

Using these formulae, the simulation assesses the output of a sequence, based on the number, positions and output values of the binding sites. It then assesses, for each possible base change at each position, how this change would affect the output of the sequence. If a base change would increase the output of the sequence, it gains a selective advantage. If it does not affect the output then it receives a neutral value and if it is deleterious it receives a value of zero. Each neutral base change is given a value, \( S \), of \( \frac{1}{3} \) (so that the summed value for three possible changes at a neutral base is 1). The value for a base change that is advantageous is \( S = 4Ns/3 \), where \( N \) is the effective population size, and \( s \) is the product of the selection coefficient that is created by an increase in output of one unit, which we define as \( \sigma \), and the increase in output. Deleterious base changes are given \( S = 0 \). Following this procedure each possible base change at each position receives a value for \( S \). The higher the \( S \) value, the higher the chance that this base change will be the one chosen to be the next evolutionary change in the sequence. The simulation adds one change at a time to the sequence. The expected time for any change will depend on the degree to which changes are possible that will be selectively favored. The more adaptive evolution that is happening, then the more rapid will be the rate of evolution and thus the lower the expected time required to make a change to the sequence. If \( Si,j \) represents the \( S \) value for the \( i \)th (of three) possible changes at the \( j \)th base in the sequence, then the total expected rate of change in the sequence will be \( \mu \sum_{i,j} Si,j \) changes per generation. The expected time in generations required to make a change will thus be \( \frac{1}{\mu \sum_{i,j} Si,j} \) generations. The mutation rate, \( \mu \), is set at \( 10^{-9} \) per base per generation throughout all simulations, and the effective population size, \( N_e \), is always \( 10^4 \). Unless otherwise specified, \( \sigma \) is always 0.0025. However, for any step in the evolutionary process, the actual time taken will not be the same as the expected time—rather the actual times are chosen from an exponential distribution with a mean given by the expected time. The value of \( Si,j \) will change as evolution proceeds, so, in every step of the simulations, an expected time for the step is calculated and an actual time selected from the exponential distribution with this expectation.

Simulations were started from real enhancer sequences and continued until a required level of output was reached. Unless otherwise stated, this output level is 30 units, with a single site perfectly matching the matrix providing an output of 2 units, and the distance bonus given by the formula earlier. The time in generations taken to reach this output level was recorded. Simulations to examine the effects of the distance model used a target of twenty binding sites rather than a particular output level. This was to allow the ratio of the binding sites’ individual outputs and the distance output bonus to be examined. In some simulations, the sequence was allowed to evolve further once the output has been reached as a way of modeling the turnover in binding sites possible for a sequence with constant function.

Randomized sequences of length equal to the original sequence were constructed by picking bases based on the nucleotide composition of the original sequence with replacement. Where random sequences are used, a different random sequence is used for each individual run.

Sequences and binding site matrices were chosen in order to cover a range of lengths and types of sequence. Transcription factor matrixes used were those for Bicoid (M00140), Krüppel (M00021), and Abd-B (M00090). Accession numbers are for the TFMATRIX databank from the TRANSFAC database (Wingender et al. 2001).

Sequences used were the Krüppel Kr730 enhancer (Hoch et al. 1991; Hoch et al. 1992), the even-skipped stripe 2 element (S2E) (AF042709), and the alcohol dehydrogenase intron 1 (AF201423). The computer program was written in ActivePerl 5.6.1.633 and run on a Compaq Tru64 UNIX system. Twenty replicates of each simulation were carried out unless otherwise stated.

Harmonic means are used to express the averages over runs unless otherwise stated. The use of this form of mean was chosen due to the nature of the results, where some runs may take very long times to reach a target relative to the
would have the same effects as are seen with changes to the effective population size and selection through the value of \( \sigma \). Selection only acts on the frequency of the selection parameter, not on the selection itself. As selection only acts on the frequency of the selection parameter, not on the selection itself. By varying the selection parameter and combining it with the same sequence, which was used to give an expected value. The level of simplicity was then calculated by the ratio of observed to expected values. Hereafter, this measure is referred to as a nucleotide simplicity score.

Analysis of the nucleotide simplicity of transcription factor binding sites was carried out by comparing every possible sequence of the same length as a matrix to the factor’s matrix. Every sequence that matched the matrix with a similarity score equal to or greater than that of the cutoff value was stored. The nucleotide simplicity score and base compositions of these sequences could then be analyzed.

**Results**

**Selection**

An important aspect of this study was the incorporation of selection into the evolution of binding sites. By varying the selection parameter \( \sigma \), we could examine the effect of selection on this process. As selection only acts through the value of \( S \), which is a linear function of the product of the effective population size and \( \sigma \), changes to \( N \) would have the same effects as are seen with changes to \( \sigma \). \( \sigma \) values ranging from \( 2.5 \times 10^{-5} \) to \( 2.5 \times 10^{-7} \) were used and the results can been seen in figure 2.

In simulations starting with the Kr730 enhancer, we can see that, on log scales, \( \sigma \) and the mean number of generations required to reach the target output are in a linear relationship, with a slope of approximately 1. Since a linear relationship with a slope of 1 would be expected if all the changes to the sequence were being driven by natural selection, this shows that most of the evolutionary changes in these cases are selective. The reason for this is that there exist, in the sequence, motifs that already act as binding sites, or, more often, that can be converted to binding sites by single nucleotide mutations (which may thus be fixed by selection). We call the latter type of sequence “presites,” and the presence of these will allow the rapid accumulation of binding sites. The presence of such sites that can be converted to sites subject to selection was also noted by Berg, Lässig, and Radic (2003). However, in the case of the highly skewed AT:GC ratio sequences (10:90 AT:GC), which evolve more slowly, there is no clear relationship between \( \sigma \) and generations. This results from these highly skewed sequences having fewer presites and most of the time taken by the evolutionary process is occurring independently of selection via a random walk of neutral changes.

The effect of the GC ratio on the time taken to evolve new binding sites can be seen clearly in figure 3. Here sequences with AT:GC ratios that vary from 10:90 to 90:10 were used to evolve binding sites for three transcription factors.

Sequences with a high GC bias were slower at evolving sites than those with an equivalent high AT bias. This is probably due to the transcription factors recognizing sequences that are AT rich. (The mean AT:GC ratios for these transcription factors’ matrices are Bicoid: 64:36, Krippel: 53:47, and Abdominal-B: 56:44.) Sequences with a high AT bias will possess more presites than will high GC sequences and will contain more sequences that are only two or three bases changes from active binding sites, which can thus walk to active sites more rapidly. In general, binding sites for transcription factors in *D. melanogaster* have a GC content that approximately matches that in the genome. The analysis of 338 *Drosophila* transcription factor binding sites (summed across transcription factor types) from the TRANSFAC database showed 41% GC,
pairs of binding sites, where the evolution of sites. If so, the binding sites' evolution to the transcriptional output. It could be that some appropriate distance between sites is the addition of a bonus, are recruited. In this model, the effect of an more distant presites, which do not yield a large distance, shows that distance and spacing play an important role as level as that seen when no distance bonus is applied. This measure is calculated as 1 \(- (\text{distance between sites} / r)\) summed over all pairs of binding sites, where \(r\) is 100 bp.

whereas the genome itself is 43% GC (Genome Release 3). A transcription factor whose required target has a GC content that matches the genomic average will be expected to find target sites more rapidly. This will make the adaptive evolution of new binding sites more rapid, but will also have the effect that binding sites will also arise more often by mutation in genomic locations where their presence is deleterious.

Distance Model

An interesting question to ask about the evolution of binding sites is what causes the clustering of binding sites seen in real enhancer sequences. To explore this, various ratios of binding site output to distance bonus were explored. The results of these simulations are shown in figure 4.

We can see that, in the initial stages of evolution, sites are more clustered when a distance bonus is applied. As the collection of presites that happen to be tightly clustered is used up, the extra clustering created by further sites is reduced, and even with a very high distance bonus, where a binding site’s output can be increased by up to 540 times by each adjacent site, clustering ultimately reaches the same level as that seen when no distance bonus is applied. This shows that distance and spacing play an important role as the first few sites are being formed from a number of presites, later, when fewer presites are available, spacing becomes less important and sites begin to spread out as more distant presites, which do not yield a large distance bonus, are recruited. In this model, the effect of an appropriate distance between sites is the addition of a bonus to the transcriptional output. It could be that some transcription factors absolutely require their binding sites to be in specific positions relative to each other (Berg, Lässig, and Radic 2003). If so, the binding sites' evolution would be described by the uppermost line in figure 4, indicating strong clustering, which is seen when all of the transcriptional output results from the distance bonus alone.

Enhancer Versus Intron Sequences

We can examine how real sequences evolve compared to randomized versions of themselves. One such study is the evolution of Abdominal B binding sites in the Adh intron, the Kr730 enhancer, and the eve stripe 2 enhancer (S2E). None of these sequences is thought to have any binding sites for Abd-B, and so all sites that arise in the sequence must be newly evolved sites.

Of the three sequences, Kr730 reaches the target fastest and the Adh intron the slowest. Our analysis of the nucleotide simplicity score of the sequences show that Kr730 has a very high level of simplicity, and that the Abd-B binding site is itself simple (table 3).

If a sequence has a high level of nucleotide simplicity then it can rapidly evolve binding sites for transcription factors that are themselves looking for simple sequences, provided that there is some agreement between the bases that are repeated in the genome and in the transcription factor target. Thus we would expect the rate at which sites are found to be increased both by similarity in GC content between the genome and the matrix for the protein considered, and by the presence of repeats (quantified by nucleotide simplicity) if the matrix of target sites itself includes such direct nucleotide repeats. The stripe 2 element sequence has a lower nucleotide simplicity than Kr730, and thus, since the transcription factor binding site itself has high nucleotide simplicity, one would expect the stripe 2 element sequence to have fewer presites than Kr730 and thus would be expected to take longer to evolve. More random walking needs to occur to reach sites, which is a slow process.

The Adh intron has a low level of simplicity and, since randomized sequences have an expected simplicity of zero, the Adh intron takes an equally long period of time to reach the target output. Although the Adh intron does not take a significantly longer number of generations to reach the target than does the random sequence, the difference seen in the harmonic means may be accounted for by the fact that a lack of presites means it has to randomly walk to reach binding sites. However, some of the random sequences will, by chance, contain sites or presites for Abd-B, which will reduce their time to reaching the output required. The randomized sequences show an average of 0.65 sites to start with, compared with 0 for the real Adh.

In order to determine how existing binding sites in a sequence affect their evolution, simulations were run in which known binding sites were conserved (table 2). Here, known binding sites for any transcription factor could not be mutated. However, the positions within existing sites may be used as part of binding sites in an overlapping fashion. This was designed to determine if the presites in enhancer sequences were the result of existing binding sites with similar motifs.

The extreme increase in the length of time taken to reach the target with Kr730 reflects the fact that the presites for Abd-B are within a region of existing binding sites,
which cannot now be modified to make Abd-B sites. However, in S2E, many presites are in previously non-binding regions. So it appears that the large increase in the time required starting from Kr730 is due to the particular positions of presites in the sequence and is not entirely due to changes in length, GC ratio or simplicity.

Length

As mentioned earlier, all three binding sites in this study bind to sequences of similar AT:GC ratios (Bicoid 64:36, Krüppel 53:47, and Abdominal-B 56:44). However, in many cases, such as shown in figure 3, the Krüppel matrix finds binding sites faster than do the other two matrices. This is despite the fact that the Bicoid matrix is only 8 bases long, and any particular 8-base sequence should theoretically occur more frequently in a sequence than will a 14-bp sequence, such as that of the Abd-B binding site. However, due to the composition of the different matrices, and their cutoff values, a much smaller proportion of sequences of the appropriate length match the Bicoid matrix than match the Krüppel and Abdominal-B matrices.

Figure 5 shows that it is feature of matrices that, on average, longer matrices have a lower mean $C_i$ value per position than do shorter matrices. As matrices are calculated from empirical data, we can say that transcription factors with longer binding sites are less specific and will bind to a larger proportion of sequences of their target length than would be predicted by extrapolation from shorter binding sites. This result shows that transcription factors with shorter binding sites will not necessarily evolve targets faster than will those with longer binding sites and that presites for longer binding sites will not necessarily be rare in a random sequence.

Protoenhancers

In cases where binding sites are created in a previously nonfunctional sequence multiple binding sites may be required to have any effect on transcription and therefore to be able to come under the influence of selection. In order to see how this would affect the number of generations needed to reach a certain output level, we imposed a lower selection threshold, below which selection does not operate and only neutral changes occur. Once enough binding sites have been created by chance that a further change would take the output level over the lower threshold, selection can operate.

![Image](image_url)

**Fig. 5.—**Relationship between the lengths of 309 different transcription factor matrices and the mean $C_i$ values for each matrix. Regression analysis gives a $p$ value of $6.32 \times 10^{-13}$ for the slope.

Various levels of this lower threshold were used; the results can be seen in table 4.

The results show that, as expected, the more output that is required before selection can operate, the longer it takes to reach the final output. Binding sites have to be formed by random mutations, with no selection to hold binding sites once formed. As each run of the simulation is carried out on a separate randomly generated sequence some sequences initially have one or more binding sites present by chance, whereas others have no binding sites and require a large number of changes to produce any sites. This produces the large difference in the minimum and maximum numbers of generations taken to reach the target. In the case of the lower threshold of four output units the maximum seen in 20 replicates of the simulation is almost 5,000 times longer than is the minimum time seen.

**Discussion**

The model for mutation that is used here is the simplest possible, with all base changes, whether they are transitions or transversions, occurring with equal frequency. In reality, there would normally be an excess of transitions, and there would also be variation between sites in their rates of mutation. Clearly, particular sites would arise more rapidly if the particular mutations required to create them were occurring at elevated rates, relative to the average.

The results imply that it is possible for there to be rapid evolution of enhancer sequences in their capacity to bind new transcription factors. The rapidity and ease with which evolution can perform this task arise because binding sites show a low complexity and because sites which are not perfect versions of the binding site consensus can nevertheless produce some activational output, and thus can be subject to selection. The rapidity of adaptation, however, will depend on the particular $\sigma$ value chosen, and $\sigma$ is unknown. However, the reciprocal relationship between the time taken and $\sigma$, in examples such as those on figure 2, when the process is driven by selective changes, means that, were we to doubt the $\sigma$ value chosen, we can infer directly the impact on the time taken of differing $\sigma$ values.

One interesting result is that sequences with high simplicity are more likely to find binding sites for new transcription factors than are sequences with lower
simplicity. Cryptic simplicity (defined slightly differently from our simplicity measure) has been reported in a number of species, and it thought to be the result of DNA slippage (Tautz et al. 1986; Schlotterer and Tautz 1992). Cryptic simplicity may be high in enhancers and, given that binding sites often include repeated bases; this would tend to make the finding of binding sites more rapid. It may be that enhancers have, in the past, contained binding sites for numerous transcription factors that they no longer can bind. The similarities between these binding sites and those forming the targets of the new transcription factor may mean that sites can be reactivated by a few base substitutions.

The ability to find new binding sites increases, as expected, with the length of the sequence examined. Also as expected, the time depends strongly on the relationship between the CG content of the binding site and that of the enhancer sequence that is evolving. However, the time taken to find binding sites does not always increase with the length of the binding site— if the matrix is such that selective rewards are not produced until there is a close match with the consensus, a short binding site could be harder for evolution to find than a longer one that does not have this property.

The simulations take much longer when there is a lower output threshold, with selection only operating above this. Table 4 shows the big increase in the harmonic mean time required when an output of eight is required (corresponding to more than four binding sites) to cross this lower threshold, compared to situations when outputs of zero or four are required. It is possible to calculate the probabilities of there being any given number of binding sites in a random sequence of length, as here, of 730 bp. This depends on the proportion of possible sequences of the appropriate length that would act as binding sites. Thus, 0.0458% of all possible eight-base pair sequences qualify as bicoid sites. This means that, considering both strands of the 730-bp sequence, the expected number of bicoid-binding sites will be 0.6623, and the actual number will be approximately Poisson distributed with this expectation, leading to a probability that a random sequence would contain two sites of 11.31%, but a probability that a random sequence would contain four sites of only 0.41%. (Corresponding figures for Abdominal B and Krüppel sites are 15.02% and 0.86% and 22.29% and 2.87%, respectively.) Since a single mutation in a presite would typically be capable of moving a sequence with two sites above a lower output threshold of four units, or moving a sequence with four sites above a lower output threshold of eight units, we should not be surprised that simulations with a lower output threshold of eight units are typically much slower than those with a lower output threshold of zero or four units.

Some simulations (data not shown) were continued after the output was reached. Output was prevented from dropping below the target level by selection, but was free to change above this target. Once above the target output, sites can be lost and gained, creating a kind of “turnover” of binding sites, although base changes in the binding sites still occur at a slower rate than the rate of neutral change in the sequence. However, such a process of stochastic loss and gain of sites through site number having climbed above the number necessary, may not be the true cause of the turnover sometimes seen in binding sites (Ludwig et al. 2000). Weakly deleterious changes that cause the loss of binding sites (which are disallowed in our model), may create alleles that persist at low frequency in populations and may then be compensated by the appearance of new binding sites in the same haplotype (Carter and Wagner 2002). Apart from their influence on this turnover process, we do not know whether the inclusion of weakly deleterious changes would greatly affect the outcome of the model. It would be possible to allow weakly deleterious changes, which would have an $S$ value, predicting the relative probability of spread to fixation, which is nonzero but is less than the $\frac{1}{3}$ that is used for a neutral change. The $S$ value corresponding to a weakly deleterious mutation would be (following Kimura (1962)) $2N(e^{2s} - 1)/(3(e^{4Ns} - 1))$, where $s$ is now the selective disadvantage of the mutation. Such weakly deleterious mutations might allow access to more advantageous mutations that strictly neutral changes might not.

We have assumed that the population can be considered as a single genotype at all times. New sites appear in this genotype, with selectively advantageous changes having a $4Ns$-fold higher chance of being fixed than have neutral changes. (In more realistic diploid situation, $s$ should be seen as the selective advantage of the changed allele when it is in the heterozygous state, since this is what will determine the probability of its eventual fixation.) It is thus assumed that mutations that will be fixed will spread rapidly, relative to the separation in time between successive fixed mutations. This assumption will be inaccurate if there are many opportunities of selective change in the sequence and if the product of the population size and the mutation rate is not many orders of magnitude less than 1. With the rapid evolution seen in some of our simulations, it would seem likely that a mutation may not have been fixed in the population before the next adaptive change in the sequence arrives. This issue is considered by

### Table 3

<table>
<thead>
<tr>
<th>Enhancer</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adh</td>
<td>0.14</td>
<td>0.54</td>
<td>0.98</td>
<td>1.11</td>
<td>3.12</td>
</tr>
<tr>
<td>Kr730</td>
<td>1.13</td>
<td>3.63</td>
<td>13.65</td>
<td>44.65</td>
<td>124.00</td>
</tr>
<tr>
<td>S2E</td>
<td>0.90</td>
<td>1.94</td>
<td>2.19</td>
<td>8.09</td>
<td>26.27</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Harmonic Mean</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>653535</td>
<td>21104964</td>
<td>206972</td>
</tr>
<tr>
<td>4</td>
<td>834480</td>
<td>514316909</td>
<td>179196</td>
</tr>
<tr>
<td>8</td>
<td>1732862</td>
<td>1107380833</td>
<td>222707</td>
</tr>
</tbody>
</table>

*NOTE.*—The lower thresholds are zero, four, or eight output units. The harmonic means of twenty runs are shown, along with the maxima and minima seen in these runs.
Gerrish and Lenski (1998) in the context of bacterial evolution. Here, the effect of clonality is a requirement for the spread of one advantageous mutation before the next can be incorporated in the same clone, with the effect that there is an effective speed limit imposed on evolution. In our short sequence, recombination during the spread of an advantageous allele will be so rare that the evolutionary process can also be seen as the successive replacement of one nonrecombining haplotype by another. The first mutation might thus be at a low frequency, $p$, when the next one arrives in the population, in which case the second mutation would have to arise in an allele that had the first mutation. If we imagine a mutation that creates an increase in fitness of $s$, the expected time taken for this mutation to spread, until it has a frequency in the population, $p$, of a half, is almost exactly $(\ln(4N)s)/s$. However, at the population level, the rate of production of mutations into the population that subsequently spread to fixation by selection will be $4\mu_1 Ns$, where $\mu_1$ is the rate of mutation in the sequence as a whole to advantageous mutations. Thus, the expected time between successive advantageous fixations is $1/(4\mu_1 N)$ generations. Thus, for mutations to spread to fixation rapidly, and before the next mutation would be expected, $1/(4N\mu_1)$ must be much greater than $\ln(4N)s/s$. This implies that $1/(4\mu_1) \gg \ln(4N)s/s$. If $s$ is chosen to be equal to the most commonly used value of $\sigma$, at $2.5 \times 10^{-3}$, and $N$ is $10^9$, this means that $\mu_1 \ll 2.71 \times 10^{-8}$. This means that, since we have a mutation rate per base of $10^{-9}$, there cannot be very many possible selectively advantageous mutations in the sequence if mutations are to be able to spread to fixation before the next mutation arises. (It should be remembered that, since, typically, at most only one of the three possible base substitutions at a given base will be advantageous, the mutation rate for any one of these three will be $10^{-9}/3$. This means that $\frac{\mu}{s} \ll 2.71 \times 10^{-8}$ implies that the number of bases in the sequence that can selectively change is much less than around 80, which seems probable given the length of the sequences being considered.) Increasing the selective advantage, $s$, will reduce the time between successive fixations. However, such increases will also reduce the time taken for individual mutations to spread. The result is that, overall, $s$ does not have a strong influence on whether successive substitutions are effectively independent.

If mutations are arising before the last one has fixed this will tend to slow down the rate of evolution, since mutations will have to arise in the subset of the population that received the previous mutation (and assuming that the sequence is so small that recombination is not significant on this time scale). The situation is not simple, however. A mutation of advantage $s$ that arises in a subset of the population that is already increasing as a result of an earlier mutation will have a probability higher than $2s$ of spreading to fixation. In general, our simulations correspond to situations where heterozygosity at the base pair level is low, which would exist when $N\mu \ll 1$, which would typically be seen in multicellular organisms, but not in some viruses, such as RNA viruses, for example (Berg, Lässig, and Radic 2003).

Another factor involving the population genetics of the situation is that we have not included in our simulations the reduction in the expected time required to find the first of the advantageous mutations in the process. This first advantageous mutation will be found more quickly since the starting population will be genetically heterogeneous and many mutations will exist at low frequency in the population. The quantitative effect of this genetic variation would depend on the frequency distribution of genetic variants in the population, and so would be impossible to model without the assumption of a neutral equilibrium. However, since the population genetic variation will affect only the time taken to find the first of what might be very many selectively driven changes as the sequence evolves, its effect on the overall time for the evolutionary process will be comparatively small.

Acknowledgments

We thank the Biotechnology and Biological Sciences Research Council for a studentship to S.M.

Literature Cited


Diethard Tautz, Associate Editor

Accepted January 15, 2004