Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites

Alan M Moses

Address: Departments of Cell & Systems Biology and Ecology & Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, ON M5S 3B2, Canada
E-mail: alan.moses@utoronto.ca

Abstract

Background: Although cis-regulatory changes play an important role in evolution, it remains difficult to establish the contribution of natural selection to regulatory differences between species. For protein coding regions, powerful tests of natural selection have been developed based on comparisons of synonymous and non-synonymous substitutions, and analogous tests for regulatory regions would be of great utility.

Results: Here, tests for natural selection on regulatory regions are proposed based on nucleotide substitutions that occur in characterized transcription factor binding sites (an important type functional element within regulatory regions). In the absence of selection, these substitutions will tend to reduce the strength of existing binding sites. On the other hand, purifying selection will act to preserve the binding sites in regulatory regions, while positive selection can act to create or destroy binding sites, as well as change their strength. Using standard models of binding site strength and molecular evolution in the absence of selection, this intuition can be used to develop statistical tests for natural selection. Application of these tests to two well-characterized regulatory regions in Drosophila provides evidence for purifying selection.

Conclusion: This demonstrates that it is possible to develop tests for selection on regulatory regions based on the specific functional constrains on these sequences.

Background

The importance of cis-regulatory regions in the evolution of complex organisms is increasingly appreciated (reviewed in [1] and [2]), and general understanding of the molecular evolution of these sequences has grown rapidly [3-13]. An important outstanding question is whether natural selection has driven evolutionary changes in cis-regulatory regions, or whether these result from non-adaptive processes [14].

Many tests for natural selection can be applied to non-coding DNA and several important studies have identified signatures of natural selection in well-characterized regulatory regions (reviewed in [15]). Tests for selection on differences between species often compare the ratio of substitutions in transcription factor binding sites (an important class of functional element within cis-regulatory regions) to the surrounding non-coding DNA [16]. These tests are modelled after tests on coding regions that compare the patterns of amino acid changing differences to synonymous differences, which are amongst the most widely used and most powerful tests to detect the effects of natural selection on individual protein coding genes [17]. However, in applying these tests to binding sites, several
important caveats must be considered [15]. In particular, it must be assumed that all of the functional elements in a regulatory region have been characterized, and that these remain constant in all species considered.

Here I develop a new approach to detect selection on individual cis-regulatory regions that takes advantage of the specificity of transcription factors to assign functional impact to nucleotide changes in binding sites. Recently, evolutionary analyses of large sets of transcription factor binding sites have highlighted the importance of considering the binding affinity or strength of the binding sites for their appropriate transcription factor [10,11,13,18]. Specifically, sequence differences in transcription factor binding sites can increase protein-DNA affinity, decrease it, or have no effect. In the absence of selection, fixation of random mutations will tend to decrease the strength of binding sites [19,20], whereas purifying selection will tend to preserve binding sites, such that the effects of subsequent fixations will cancel out [18]. On the other hand, though the binding sites can arise in regulatory sequences as a result of the action of positive selection [19-21] or through genetic drift alone [22], I show that an increase in binding affinity on average is not expected in the absence of selection. I therefore propose to use the distribution of changes in strength of transcription factor binding sites to develop tests for natural selection on regulatory regions where the binding sites have been identified.

I analyze the fixed differences in two well-characterized regulatory regions in Drosophila (the hb anterior activator and the eve stripe 2 enhancer). These tests reveal statistical evidence for conservation of cis-regulatory information, which is consistent with the known conservation of function of these regulatory sequences.

Results

Quantifying the effects of substitutions in regulatory regions

Motivated by the power of tests for natural selection that exploit the constraints imposed on coding sequences by the genetic code, I sought to develop a test for natural selection on regulatory regions that takes into account the specific constraints on these regions: binding by transcription factors. Using standard matrix models for DNA binding specificity (known as Position Weight Matrices or Position Specific Scoring Matrices [23]), the binding energy of the interaction between a transcription factor and DNA is given by a sum of independent contributions from each residue at each position [23]. An estimate of the relative affinity or ‘strength’ a transcription factor binding site X of length w for its binding protein can be quantified using

\[
S = \sum_{i=1}^{w} \sum_{b \in \{A,C,G,T\}} X_{ib} \log\left(\frac{f_{ib}}{g_b}\right)
\]

(1)

Where \(X_{ib} = 1\) if the sequence X is nucleotide b at position i and 0 otherwise, \(f_{ib}\) is the probability of observing nucleotide b at position i in a binding site for a transcription factor (from the specificity matrix), and \(g_b\) is the probability of observing nucleotide b in the genomic background distribution [23].

Alternatively, the strength of the transcription factor binding sites in a region can be considered the regulatory information in that region, and the formula above can be motivated by information theoretic arguments [23]. Note that the framework and tests for selection presented here can equally be applied to information contained in the cis-regulatory region as to binding affinity. However, because recent work has focused on binding affinity (e.g., [11,13]) this work is presented from that perspective.

In order to quantify the effects of evolutionary changes in binding sites, I consider the effect of a single nucleotide change. In this case I define

\[
\Delta S_{ab} = \log\left(\frac{f_{ib}}{g_b}\right) - \log\left(\frac{f_{ia}}{g_a}\right)
\]

(2)

associated with a change from base a to base b \((a, b \in \{A, C, G, T\})\) where, once again, \(i\) is the position in the motif, \(g\) are background probabilities, and \(f\) are the probabilities in the specificity matrix model. Extending these methods to the general case of arbitrary numbers of substitutions is an area for further research (see Discussion).

The effect of substitutions on binding sites in the absence of selection

Most random mutations will decrease the strength of a transcription factor binding site, and therefore substitutions in the absence of selection will tend to decrease the affinity [19,20]. This follows from the fact that high affinity binding sites represent a small fraction of the possible sequences of a particular length. Since a substitution process that operates independently at each position in the sequence will tend to explore the majority sequence space, sequences that currently represent binding sites are much more likely to move away from these regions of sequence space than to remain in the relatively small regions of sequence space that represent binding sites. This implies that on average \(\Delta S\) should be negative in the absence of selection. To illustrate this, I simulated the evolution of a binding site for Bcd (a developmental transcription factor in Drosophila whose specificity is well-characterized) under an
HKY model (dotted trace in Figure 1a). The strength (S) of the binding site begins high (near the expected value of S for binding sites) and decays as substitutions eventually hit the critical residues. Consistent with this, the distribution of changes in score (ΔS) is concentrated on values less than zero (dotted trace in Figure 1b).

The effect of substitutions in binding sites under selection

In contrast, in functionally constrained regulatory regions, purifying selection will preferentially remove nucleotide changes that greatly alter the affinity of the binding sites [6,13]. When these substitutions do become fixed (albeit rarely), positive selection will tend to fix additional nucleotide changes that restore the binding affinity [18]. This process will tend to preserve the binding affinity, and ΔS will therefore tend to be zero if the regulatory region is under functional constraint.

Finally, consider adaptive evolution, which could have arbitrary effects on ΔS. For example, new transcription factor binding sites could be created from background sequence through successive adaptive fixations that increase binding site strength; this would lead to an increase in S, and therefore ΔS would be greater than zero on average. However, because new binding sites can also appear by genetic drift [21,22], it is possible that ΔS can be greater than zero in the absence of selection. To illustrate this, I simulated a background sequence of length equal to the Bcd binding site under the same HKY model as above, and found examples where binding sites arose in the absence of selection (Figure 1a solid trace). I argue that, although arbitrarily strong binding sites (high values of S) can be generated in the absence of selection, the distribution of changes in score (ΔS) is specified by the substitution process. Interestingly, since evolution in the absence of selection is unbiased with respect to the strength of the binding site, the distribution of changes in score is symmetric, with mean equal to zero (Figure 1b solid trace). This indicates that in the absence of selection, in background sequences we expect changes in score to cancel out. Therefore, while the creation of binding sites from background sequence cannot be considered evidence for positive selection, if the distribution of ΔS observed is statistically different from the pattern expected in the absence of natural selection, this can only be consistent with adaptive evolution.

Creation of new binding sites in regulatory regions is an intuitive case of adaptive regulatory evolution. However, depending on the situation, natural selection could also favour mutations that remove functional binding sites within a regulatory region, thus leading to an average ΔS of less than zero. Therefore, although a decrease in S on average is expected in the absence of selection, it could also occur in the presence of selection. Nevertheless, if ΔS is more negative than expected in the absence of selection, we have evidence that natural selection must be acting to remove binding sites.

In summary, for substitutions in a set of characterized binding sites we expect:

- ΔS < 0 in the absence of constraint or adaptive destruction of binding sites
- ΔS = 0 in the presence of functional constraint
- ΔS > 0 during the creation of new binding sites (due to selection or genetic drift)

Statistical tests for natural selection in regulatory regions

An attractive feature of using ΔS for a single substitution (as defined above) in a test for natural selection on
regulatory regions is that its distribution can be computed exactly under standard models of molecular evolution in the absence of selection (see Methods, Figure 1b). I therefore propose to use the distribution of $\Delta S$ to test for the presence of natural selection on regulatory regions. If the distribution of $\Delta S$ is significantly different from that expected in the absence of selection, we can rule out the null hypothesis of evolution in the absence of selection.

Here I consider the tests for selection in the following cases.

1. If the observed $\Delta S$ in real binding sites is greater on average than $\Delta S$ expected for binding sites in the absence of selection, this indicates purifying selection to retain binding sites.
2. If the observed $\Delta S$ in real binding sites is less on average than $\Delta S$ expected for binding sites in the absence of selection, this indicates adaptive destruction of binding sites.
3. If the observed $\Delta S$ in real binding sites is greater on average than the $\Delta S$ expected for binding sites arising from background sequence in the absence of selection, this indicates adaptive creation of new sites.

Case 1: Here the pattern of evolution is consistent with purifying selection to preserve the function of the binding sites in the regulatory region. To rule out the null hypothesis of no constraint, we must compare the observed values of $\Delta S$ to the distribution of $\Delta S$ in sequences we know to be transcription factor binding sites, but in the absence of selection.

In the case of binding sites evolving in the absence of constraint:

$$E[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} \Delta S_{ab} \frac{f_{ia}P_{ab}}{\varphi}$$

where $E[X]$ and $V[X]$ are the mean and variance of the random variable $X$, respectively,

$$\varphi = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} f_{ia}P_{ab}$$

and $P_{ab}$ is the probability of substitution between bases $a$ and $b$ (i.e., $a$, $b$ in \{A, C, G, T\}), computed under a standard model of molecular evolution, such that $P = e^{\varphi t}$ where $R$ are the instantaneous rates of substitution and $t$ is time (see Methods). The dotted trace in Figure 1b shows the distribution of $\Delta S$ for binding sites evolving in the absence of constraint.

In a practical setting, we expect to have observed some relatively modest number ($N$) of substitutions in characterized binding sites. Therefore, in order to test the significance of a set of observed $\Delta S$ values, I propose the statistic:

$$Z = \frac{\frac{1}{N} \sum_{k=1}^{N} \Delta S_{k} - E[\Delta S]}{\sqrt{\frac{V[\Delta S]}{N}}}$$

where $k$ indexes $N$ observed values of $\Delta S$.

Since we can compute the mean and variance of $\Delta S$ under standard models of evolution (see Methods), according to the central limit theorem this statistic should be normally distributed with mean = 0 and variance = 1 (the standard normal) under the hypothesis that the model of evolution is correct. We can therefore perform a one-tailed test that the observed mean is greater than that expected in the absence of selection.

I sought to confirm that the distribution of this statistic was as expected, particularly in the case of small $N$ (few observed substitutions in binding sites) which is typical of real datasets. To simulate the null hypothesis of binding sites evolving in the absence of constraint, I simulated molecular evolution of the 6 real BCD sites in the $hb$ anterior activator under an HKY model with the transition-transversion rate ratio estimated from the alignment of the $hb$ anterior activator (see methods) and evolutionary distance scaled so that we would observe approximately 5 substitutions in the 6 binding sites. I computed $Z$ using $E[\Delta S]$ and $V[\Delta S]$ either under this model, and I observed good agreement with the expected standard normal behavior (Figure 2, ‘exact’).

Case 2: If we wish to test for adaptive destruction of transcription factor binding sites in a regulatory region, the average of $\Delta S$ should be significantly less than expected in the absence of selection. To test for this, we can perform a one-tailed test using the statistic defined above, but in the opposite direction.

Case 3: If the average $\Delta S$ in a regulatory region is greater than 0, we wish to test whether the average $\Delta S$ is greater than we would expect to observe in the absence of selection. Now the null hypothesis is that the average increase in binding affinity we have observed is due binding sites arising in background sequence by genetic
drift. Once again the distribution of $\Delta S$ can be computed exactly, and the mean and variance are:

$$E[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} \Delta S_{ab} \frac{S_a P_{ab}}{\varphi}$$  \hspace{1cm} (7)$$

$$V[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} (\Delta S_{ab} - E[\Delta S])^2 \frac{S_a P_{ab}}{\varphi}$$  \hspace{1cm} (8)$$

with

$$\varphi = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} S_a P_{ab} = w \sum_{a} \sum_{b \neq a} S_a P_{ab}$$  \hspace{1cm} (9)$$

The solid trace in Figure 1b shows this distribution. This distribution is symmetric, and the expectation is zero. This follows from the fact that the substitution processes in the absence of selection is unbiased with respect to the binding site strength, and that the residue frequencies in background genomic sequence are assumed to be drawn from the equilibrium of the substitution process. The means and variances can be used to form a Z-statistic as illustrated above, and simulations again confirm the expected distribution of the statistic (data not shown). If the observed average $\Delta S$ is significantly greater than expected in the absence of selection, we find evidence for adaptive evolution. For example, for the 20 substitutions shown in Figure 1a (solid trace) the average $\Delta S$ is 0.45, which gives $Z = 0.97$ and is not significant. Thus, although there is a large change in $S$, the pattern of changes is consistent with the absence of selection.

**An approximation to the distribution of $\Delta S$**

Under substitution models with no transition-transversion bias [24], the distribution of $\Delta S$ does not depend on evolutionary distance. For example, I can show (see Methods) that for binding sites evolving in the absence of selection,

$$E[\Delta S] = \sum_{i=1}^{w} \sum_{a} (S_a - f_{ia}) \log\left(\frac{f_{ia}}{S_a}\right)$$

$$\sum_{i=1}^{w} \sum_{b \neq a} f_{ib}(1-g_b)$$  \hspace{1cm} (10)$$

A similar, albeit more complicated expression is available for the variance (see Methods). These expressions depend only on the equilibrium probabilities of the four nucleotides and the probabilities in the specificity matrix model for the transcription factor.

In the general case, the distribution of $\Delta S$ depends very weakly on the evolutionary distance (Figure 3a) and only somewhat more strongly on the transition-transversion bias (Figure 3b). It is therefore possible to obtain a good approximation of the distribution of $\Delta S$ using the formulas obtained under the simpler substitution models. I refer to this approximation of the distribution of $\Delta S$ as the ‘long-time limit’ because it becomes exact in the limit of long evolutionary time even in the presence of transition-transversion bias (Figure 3). As expected, using the long-time limit $E[\Delta S]$ and $V[\Delta S]$ when calculating the Z statistic described above also gives the standard normal behaviour (Figure 2, ‘Long time limit’). Thus, this approximation allows application of tests based on the distribution of $\Delta S$ without estimates or assumptions about the evolutionary process in the absence of selection.

**Application to the hb anterior activator**

The $hb$ anterior activator (Figure 4a) responds to the Bcd gradient in the early $D.\ melanogaster$ embryo [25]. It is thought to have been conserved since the divergence of $D.\ melanogaster$ and $D.\ virilis$ [26] and contains well-defined binding sites for Bcd [27]. We therefore expect to see evidence of functional constraint on this regulatory region. Using $D.\ virilis$ and $D.\ pseudoobscura$ as outgroups, I identified 10 substitutions in the alignment of the 6 well-characterized Bcd binding sites (Figure 4b) and used equation 2 above to compute $\Delta S$ for each substitution (Table 1). The average $\Delta S$ for these substitutions was...
I noted that 3 substitutions had occurred in a single binding site on a single lineage (A3, Table 1) and was concerned that this might indicate that the assumption of single substitutions at each site was invalid on this lineage. I therefore performed the test excluding these substitutions and found similar results (Table 2). I also found evidence for constraint when excluding substitutions along the relatively long branch leading to the melanogaster group. Noting that removing the substitutions led to an average $\Delta S$ for the region greater than 0, I tested for evidence for adaptive creation of binding sites in this regulatory region (case 3). However, performing the test above (equations 7-9) yielded $Z = 0.22$, which is not significant, indicating that the observed increase in binding strength could have been observed in the absence of selection.

**More complex regulatory regions**

The hb anterior activator serves as a good test case for this method because it contains multiple binding sites for the same transcription factor. However, in general regulatory regions contain multiple binding sites for multiple different transcription factors. Note that the arguments above regarding the expected $\Delta S$ in regulatory regions apply regardless of whether the binding sites are for a single transcription factor or many different transcription factors.

To extend the statistical test to regulatory regions with multiple binding sites for different factors, two approaches are possible. If enough substitutions in each type of binding site are present, the test above can be performed for each type, and then their results can be combined. However, in the case of few substitutions, it may be preferable to pool the substitutions first. To do so, we must compute the distribution of $\Delta S$ expected from a mixture of transcription factor binding sites. $\Delta S$ is now drawn from a $\nu$ component mixture model,

$$p(\Delta S) = \sum_{j=1}^{\nu} \pi_j p(\Delta S)_j \tag{11}$$

where $\nu$ is the number of types of transcription factor binding sites, $\pi_j$ is the probability that the substitution occurred in the $j$-th type, and $p(\Delta S)_j$ is the distribution of $\Delta S$ for the $j$-th type of binding motif. We can compute these $\pi_j$ for any regulatory region given the numbers of each type of binding site in a characterized regulatory region (see Methods):

$$\pi_j = \frac{n_j \sum_{i=1}^{\nu} \sum_a \sum_{b \neq a} f_{jia} p_{ab}}{\sum_{j=1}^{\nu} n_j \sum_{i=1}^{\nu} \sum_a \sum_{b \neq a} f_{jia} p_{ab}} \tag{12}$$
where \( n_j, f_j \) and \( w_j \) are the number of occurrences, specificity matrix and width of the \( j \)-th type of binding site, and \( P \) is a substitution matrix as above. Since it is possible to compute the means and variances of mixture distributions as a function of the component distributions (see Methods),

\[
E[\Delta S] = \sum_{j=1}^{\nu} p_j E[\Delta S_j] \\
V[\Delta S] = \sum_{j=1}^{\nu} p_j V[\Delta S_j] + \sum_{j=1}^{\nu} \pi_j (E[\Delta S] - E[\Delta S_j])^2
\]

we can apply the test suggested above in the mixture case as well. Once again, we can obtain a good approximation to this distribution using the long-time limit. To confirm the distribution of the test statistic proposed above in the case of mixtures of binding sites, I again performed simulations under the null hypothesis of no constraint, this time using the 6 Kr sites and 5 Bcd sites in the \( \text{eve} \) stripe 2 enhancer, and once again found the expected standard normal behavior (data not shown).

**Application to the \( \text{eve} \) stripe 2 enhancer**

The individual binding sites in the \( \text{eve} \) stripe 2 enhancer are well-characterized [28,29] and this enhancer exemplifies complex \( \text{cis} \)-regulatory sequences in that it contains multiple binding sites for multiple transcription factors [30]. Here I consider the 5 Bcd binding sites and 6 Kr binding sites to illustrate the application of the test for selection to a complex regulatory region (see Additional file 1 for alignments of these binding sites). Two of the Kr and Bcd binding sites the \( \text{eve} \) stripe 2 enhancer overlap,
and I excluded two substitutions that affect the strength of more than one binding site (Table 3). This left 16 substitutions (9 in bcd binding sites and 7 in Kr binding sites) for which I used equation 2 to compute associated ΔS values (Table 3). The average ΔS for all the substitutions was -0.23, and I performed the test described above with evolutionary distance and transition-transversion rate ratio estimated from the alignment of the eve stripe 2 enhancer (see methods). Using equations 11-14, I computed the distribution of ΔS for 6 Kr sites and 5 Bcd sites evolving in the absence of constraint. This yields E[ΔS] = -1.56 and V[ΔS] = 2.53, and provides evidence for constraint (case 1) on the regulatory sequence with Z = 2.53 and P-value = 0.0004 (Table 4).

Although its function has been conserved over evolution [31], the eve stripe 2 enhancer has undergone some lineage specific evolution [32], as well as gained and lost individual binding sites; its evolution is characterized by rapid sequence divergence [31,33]. Consistent with this, the alignments of D. pseudoobscura for BC-3 were not possible, as this site seems to have appeared recently [32]. Within the closely related species in the melanogaster subgroup, BC-3 contains seven inferred substitutions, four of which are inferred to occur along the lineage leading to D. yakuba. In addition to the rapid divergence of BC-3, I again found cases where more than one substitution had occurred along the D. ananassae lineage in a single binding site. In addition, I therefore performed the tests excluding lineages with multiple substitutions, or excluding BC-3 entirely. In all cases there is still sufficient power to provide statistical evidence against the null hypothesis of no constraint (table 4). In no case could I find evidence for adaptive evolution (case 2 or case 3, data not shown).

Discussion and Conclusion

A new test for natural selection on regulatory regions

One of the difficulties in many current evolutionary analyses of cis-regulatory regions is that it is difficult to choose an appropriate set of unconstrained sites to which to compare the functional regulatory sites. In general, studies either choose the rate of substitution in surrounding

---

**Table 2: Tests for selection on the hb anterior activator**

| N  | E[ΔS] | V[ΔS] | Observed average ΔS | Case | Z   | P-value | Notes                                    |
|----|-------|-------|----------------------|------|-----|---------|-----------------------------------------|
| 10 | -1.61 | 2.77  | -0.31                | 1    | 2.48| 0.007   | Phylogenetic model estimated from hb anterior activator alignment |
| 10 | -1.66 | 2.87  | -0.31                | 1    | 2.53| 0.005   | Long time limit distribution of ΔS       |
| 7  | -1.61 | 2.77  | 0.07                 | 1    | 2.67| 0.004   | Excluding A3, phylogenetic model estimated from hb anterior activator alignment |
| 5  | -1.61 | 2.77  | 0.20                 | 1    | 2.43| 0.007   | Excluding A3 and substitutions on the lineage leading to the melanogaster group, phylogenetic model estimated from hb anterior activator alignment |
| 5  | 0     | 4.26  | 0.20                 | 2    | 0.22| 0.415   | As above, test for adaptive evolution    |

**Table 3: Substitutions in Bcd and Kr sites in the eve stripe 2 enhancer**

| From To | Pos | ΔS | site | lineage | Coordinate |
|---------|-----|----|------|---------|------------|
| G       | C   | N.A.| N.A. | BC-5, KR-5 | ana 2R:5489670 |
| G       | T   | 0.62| KR-5 | ana     | 2R:5489674  |
| A       | G   | 0   | -0.44| KR-4     | mel 2R:5489850 |
| C       | G   | 3   | -0.51| KR-4     | ana 2R:5489853 |
| G       | C   | N.A.| N.A. | KR-3, BC-1 | ere/yak 2R:5490048 |
| C       | T   | 3   | -0.92| KR-2     | mel Subgroup 2R:5490098 |
| A       | G   | 0   | -0.44| KR-2     | ana 2R:5490094 |
| T       | C   | 9   | -1.39| KR-1     | ana 2R:5490140 |
| A       | G   | 7   | -2.73| KR-1     | ana 2R:5490142 |
| A       | G   | 0   | 0.12 | BC-3     | yak 2R:5489835 |
| A       | C   | 0   | 1.22 | BC-3     | sim/sec 2R:5489835 |
| T       | C   | 1   | -1.67| BC-3     | yak 2R:5489836 |
| A       | G   | 2   | -2.30| BC-3     | ere 2R:5489837 |
| T       | C   | 3   | 0.41 | BC-3     | yak 2R:5489838 |
| T       | A   | 3   | 3.85 | BC-3     | mel/sim/sec 2R:5489838 |
| C       | G   | 7   | 0.48 | BC-3     | yak 2R:5489842 |
| C       | G   | 7   | -0.48| BC-4     | mel/sim/sec 2R:5489683 |
| G       | C   | 7   | 0.48 | BC-4     | sim 2R:5489683 |

N.A. - not applicable, other abbreviations are as in Figure 3. Naming of binding sites is as in [31]. Coordinates are based on mapping of sites to the D. melanoster genome [52].

and I excluded two substitutions that affect the strength of more than one binding site (Table 3).

This left 16 substitutions (9 in bcd binding sites and 7 in Kr binding sites) for which I used equation 2 to compute associated ΔS values (Table 3). The average ΔS for all the substitutions was -0.23, and I performed the test described above with evolutionary distance and transition-transversion rate ratio estimated from the alignment of the eve stripe 2 enhancer (see methods). Using equations 11-14, I computed the distribution of ΔS for 6 Kr sites and 5 Bcd sites evolving in the absence of constraint. This yields E[ΔS] = -1.56 and V[ΔS] = 2.53, and provides evidence for constraint (case 1) on the regulatory sequence with Z = 2.53 and P-value = 0.0004 (Table 4).
non-coding sequence [16] or in synonymous sites in adjacent protein coding regions [34]. Both assumptions may be problematic. The former assumes that the surrounding DNA is under no functional constraint (as opposed to some unknown constraints). In the latter case, because non-coding sequences show larger numbers of insertions and deletions than coding regions, it is not always clear that rate estimates based on alignments of coding and non-coding regions can be directly compared.

Tests based on the distribution of $\Delta S$, such as those proposed here, do not rely on these assumptions, as they consider only substitutions that occur in binding sites. Practically, this is an attractive feature of these tests, as they only require accurate alignments of the binding sites, which are generally more reliable than alignments of unconstrained non-coding DNA [35].

Another attractive feature of tests based on the distribution of $\Delta S$ in the absence of selection is that they make few assumptions about the nature of selection on binding sites. For example, it is not assumed that binding sites are all under the same strength of selection, or that they all have the same binding affinity - only the changes in strength of binding are important. Further, even under a stabilizing selection model, where binding sites for a given transcription factor are gained and lost over evolution [33], $\Delta S$ will be zero on average if the total output of the regulatory sequence is preserved: the negative $\Delta S$ associated with the binding site loss will be compensated by positive $\Delta S$ associated with the binding site gain. However, if binding sites for one transcription factor are replaced by binding sites for another, $\Delta S$ may no longer be zero on average and testing for selection in this case is an area for further research.

**Practical considerations, limitations and future improvements**

Application of these tests to two well-characterized regions demonstrates that they have enough power to detect constraint on individual regulatory regions with ~10 substitutions in binding sites, and perhaps even as few as 5 or 6 substitutions (tables 2 and 4). However, application to the *eve* stripe 2 enhancer illustrates several practical difficulties: First, I didn’t include the Hb binding sites in this enhancer [36] because these binding sites contain homopolymeric runs, and it is difficult to assign a ‘position’ to a substitution; $\Delta S$ cannot be reliably computed for each substitution in this case. Second, although the *eve* stripe 2 enhancer has characterized sites for Gt, I did not include these because the sequence specificity of this transcription factor is poorly characterized. Third, the *eve* stripe 2 enhancer contains substitutions in overlapping binding sites, for which it is not clear how to calculate $\Delta S$; these were therefore excluded from the analysis. Finally, the distribution of $\Delta S$ is sensitive to the estimation of the frequency parameters in the specificity matrix. For example, I excluded the Bcd binding sites in the *eve* stripe 2 enhancer and reconstructed the Bcd matrix for analysis of that regulatory region. If the binding sites in the regulatory region of interest are included in the estimation of the specificity matrix, there is a potential for circularity in the analysis. Thus, the tests require (i) well-characterized transcription factor binding specificity and (ii) confident alignment of a binding site to a single specificity matrix. None of these constraints are present for tests that compare binding sites to surrounding regions or synonymous sites [16,34] or for tests of natural selection based on spacing between conserved blocks [35-37]. However, rapid advances in methods to characterize DNA-protein interactions are making specificity information available for large numbers of transcription factors [38-40]. Among these are methods that yield information about binding to each sequence, such that the assumption of independent contributions to binding of each DNA base in the binding site could in principle be relaxed [38,41].

In addition, the tests I have proposed assume that only a single substitution has occurred at any position in binding sites. Although for most of the data analyzed

| N  | E($\Delta S$) | V($\Delta S$) | Observed average $\Delta S$ | Case | Z    | P-value | Notes                                                                 |
|----|--------------|--------------|----------------------------|------|------|---------|-----------------------------------------------------------------------|
| 16 | -1.56        | 2.53         | -0.23                      | I    | 3.35 | 0.0004  | Phylogenetic model estimated from *eve* stripe 2 alignment            |
| 16 | -1.60        | 2.59         | -0.23                      | I    | 3.41 | 0.0003  | Long time limit distribution of $\Delta S$                            |
| 9  | -1.56        | 2.53         | 0.05                       | I    | 3.04 | 0.001   | Phylogenetic model estimated from *eve* stripe 2 alignment, excluding substitutions on lineages with multiple substitutions in individual binding sites |
| 6  | -1.56        | 2.51         | -0.39                      | I    | 1.81 | 0.035   | Phylogenetic model estimated from *eve* stripe 2 alignment, excluding BC-3 and substitutions on lineages with multiple substitutions in individual binding sites |
here this assumption seems valid, I noted several cases were multiple substitutions occurred on a single lineage, suggesting the possibility of ‘multiple hits’ at a single site. Furthermore, there is clear evidence of insertions and deletions occurring near or within the binding sites considered here. These are likely to affect their binding affinity, but are not included in the null model of molecular evolution in the absence of selection. More sophisticated models of molecular evolution [42] may be able to account for these effects, and these could be applied in this framework. Similarly, the evolutionary models here do not account for di-nucleotide substitution bias, particularly the elevated rate of CpG toTpG found in mammals; these could be included using an improved null model [43,44].

Finally, I note that I have suggested one simple statistical test based on the observed average ΔS, however many tests based on distribution of ΔS are possible. For example, purifying selection might also be expected to reduce the variance of ΔS. Indeed, in the case of the Bcd sites in the hb anterior enhancer, the observed variance of ΔS is less than expected, though this difference is not significant (e.g., 1.15 vs. 2.86, n = 10, chi-square test P = 0.089). Determining what tests have the most power to detect various types of selection in regulatory regions is an area for further research. In general, however, it seems very likely that tests that consider the effects of substitutions on transcription factor binding site affinity will facilitate the detection of adaptive evolution in regulatory regions.

Methods
Construction of motif matrices
I used publicly available compilations of characterized binding sites for Bcd and Kr [33,45] to construct specificity matrices using a pseudocount of 1 per position. Throughout this study, I use as the background distribution (gA, gC, gG, gT) = (0.3, 0.2, 0.2, 0.3) which is close to the observed nucleotide probabilities in drosophila non-coding DNA. In order to avoid the possibility of circularity, for analysis of each regulatory region I excluded the characterized sites from that region and reconstructed the matrix, such that (for example) Bcd sites from the hb anterior activator were not included in the matrix used for analysis of the hb anterior activator. These matrices were used to compute ΔS for each substitution (Tables 1 and 3) and E[ΔS] and V[ΔS] (Tables 2 and 4).

Alignments and phylogenetic analysis of regulatory sequences
I obtained homologous regions for each regulatory region from the UCSC genome-browser alignments [46]. The sequences were then realigned using mLAGAN [47]. Using these alignments and the known species relationships for these species [48], I estimated the evolutionary distance and transition-transversion rate ratio bias under an HKY model [24] using paml [49]. The parameters estimated using paml for each regulatory region were then used to compute the exact E[ΔS] and V[ΔS] shown in Tables 2 and 4.

Simulations of molecular evolution
To confirm that the test statistic had a standard normal distribution under the null hypothesis, I simulated the evolution of the 6 known binding sites in the hb anterior activator. To do so, I inferred the ancestral sequences using maximum parsimony [50], and then simulated their evolution by introducing substitutions using an HKY model with kappa estimated from the alignment, 60% AT content for the equilibrium distribution of nucleotides, and evolutionary distance scaled to observe an average of 5 substitutions over the 6 binding sites. I then computed the average ΔS for the substitutions we observed, and calculated the Z statistic using E[ΔS] and V[ΔS] computed exactly using the evolutionary model or using the long-time limit approximation. I repeated this simulation until I observed 1000 cases with at least 3 substitutions in total. Simulations for the eve stripe 2 enhancer were similar, except I used the actual D. melanogaster binding sites (because reliable inference of the ancestral sites was difficult) and that the evolutionary distance was scaled so that the 5 substitutions were distributed over the 5 Bcd and 6 Kr sites.

Distribution of ΔS
I sought to compute the distribution of ΔS in the absence of selection. Because the number of observed evolutionary differences in any particular binding site is typically small, I make the assumption that each DNA difference in a transcription factor binding site occurs independently, and presence of a single change has no effect on the probability of other changes. Under this assumption, the probability of observing the particular change from base a to base b (a, b in {A, C, G, T}) at position i is

$$p(a \rightarrow b \text{ at } i | \text{ one subs.}) = \frac{p(a \rightarrow b \text{ at } i)}{p(\text{one subs.})} = \frac{\sum_{w=1}^{w} f_{i} p_{i}^{w} a b}{p(\text{one subs.})}$$

(15)

where $p(\text{one subs.}) \equiv \varphi$, and $\varphi = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} p(a \text{ at } i) p(b \mid a)$, so that

$$\varphi = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} f_{i} p_{i}^{w} a b$$

(16)
and \( P = e^{\mu t} \) is a substitution probability matrix. The
expected value of \( \Delta S \) for binding sites that evolve in the
absence of selection is \( \Sigma \Delta S_{ab} \rho(a \rightarrow b \text{ at } i|\text{one subs.}) \) or

\[
E[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} \Delta S_{ab} \frac{f_{ia}P_{ab}}{\phi} \quad (17)
\]

Similarly, for the variance, we have

\[
V[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} (\Delta S_{ab} - E[\Delta S]) \frac{f_{ia}P_{ab}}{\phi} \quad (18)
\]

As computed here, the distribution of \( \Delta S \) is exact only for
the first substitution at each site in a particular sequence. Therefore it is important to apply the tests described here
to cases where only small numbers of substitutions have occurred on each lineage. For the regulatory sequences
considered here, this assumption seems appropriate. However, if enough substitutions have occurred such
that multiple subsequent substitutions occur at the same position, the distribution of \( \Delta S \) computed based on
the sequence of a reference species or inferred ancestral
sequence will no longer be exact. Computing the distribution of \( \Delta S \) under more relaxed assumptions is
area for further research.

Since I am considering the conditional probability that
one particular substitution occurs out of all the possible
substitutions that could have occurred, under some
substitution models such as F81 [51], or in the limit of
long evolutionary time, this probability does not depend
on time and mutation rate (evolutionary distance). I
refer to this time independent approximation as the
‘long time limit’ distribution, and derive formulas under
this assumption. Under the F81 [51] substitution model
\( P_{ab} = \delta_b(1 - e^{-u}) \), where \( u \) is the mutation rate and \( t \) is time. We have

\[
\varphi = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} f_{ia}\delta_b(1 - e^{-u}) = (1 - e^{-u}) \sum_{i=1}^{w} \sum_{a} f_{ia}(1 - \delta_a)
\]

and therefore \( \rho(a \rightarrow b \text{ at } i|\text{one subs.}) = \frac{f_{ia}P_{ab}}{\varphi} \)

\[
= \frac{f_{ia}\delta_b(1 - e^{-u})}{\varphi} = \frac{f_{ia}\delta_b}{\Sigma_{i=1}^{w} \Sigma c f_{ic}(1 - \delta_c)} \quad (20)
\]

which depends only on the frequencies in the matrix, \( f \),
and the background distribution of nucleotides \( \delta \), where
now \( a \), \( b \) and \( c \) index the bases \{A, C, G, T\}. Therefore
under this model, the long time limit is exact. Substitution
into the general formulas for the expectation gives

\[
E[\Delta S] = \sum_{i=1}^{w} \sum_{a} \sum_{b \neq a} \Delta S_{ab} \frac{f_{ia}\delta_b}{\Sigma_{i=1}^{w} \Sigma c f_{ic}(1 - \delta_c)} \quad (21)
\]

for the case of binding sites evolving in the absence of
selection (case 1). This formula can be simplified using
the fact that \( \Delta S = 0 \) if \( a = b \):

\[
\sum_{a} \sum_{b \neq a} \Delta S_{ab} f_{ia}\delta_b = \sum_{a} \sum_{b} \left[ \log \left( \frac{f_{ib}}{\delta_b} \right) - \log \left( \frac{f_{ia}}{\delta_a} \right) \right] f_{ia}\delta_b = \sum_{a} (\delta_a - f_a) \log \left( \frac{f_{ia}}{\delta_a} \right)
\]

Therefore, we have for case 1,

\[
E[\Delta S] = \frac{\Sigma_{i=1}^{w} \Sigma a (\delta_a - f_{ia}) \log \left( \frac{f_{ia}}{\delta_a} \right)}{\Sigma_{i=1}^{w} \Sigma b f_{ib}(1 - \delta_b)} \quad (22)
\]

To compute the variance, I use \( V[\Delta S] = E[\Delta S^2] - E[\Delta S]^2 \), where

\[
E[\Delta S^2] = \frac{\Sigma_{i=1}^{w} \Sigma a (f_{ia} + \delta_a) \log \left( \frac{f_{ia}}{\delta_a} \right)^2}{\Sigma_{i=1}^{w} \Sigma c f_{ic}(1 - \delta_c)}
\]

\[
2 \Sigma_{i=1}^{w} \Sigma a f_{ia} \log \left( \frac{f_{ia}}{\delta_a} \right) \Sigma b \delta_b \log \left( \frac{f_{ib}}{\delta_b} \right) - \frac{2 \Sigma_{i=1}^{w} \Sigma \delta_a \log \left( \frac{f_{ia}}{\delta_a} \right)^2}{w(1 - \Sigma b \delta_b^2)} \quad (23)
\]

Similarly, for the case of background sequences evolving
into binding sites in the absence of selection (the null
hypothesis for case 3), the same calculations give \( E[\Delta S] = 0 \), and

\[
V[\Delta S] = E[\Delta S^2] = \frac{2 \Sigma_{i=1}^{w} \Sigma a \delta_a \log \left( \frac{f_{ia}}{\delta_a} \right)^2}{w(1 - \Sigma b \delta_b^2)} - \frac{2 \Sigma_{i=1}^{w} \Sigma a \delta_a \log \left( \frac{f_{ia}}{\delta_a} \right)^2}{w(1 - \Sigma b \delta_b^2)} \quad (24)
\]

While these formulas are complicated, they depend only
on the residue probabilities in the matrix \( f \) and the
Mixtures of binding sites

In the case of transcription factors binding a regulatory region, ΔS is drawn from a mixture distribution,

\[ p(ΔS) = \sum_{j=1}^{v} \pi_j p(ΔS)_j \]

where \( v \) is the number of types of transcription factor binding sites, \( \pi_j \) is the probability that the substitution occurred in the \( j \)-th type, and \( p(ΔS)_j \) is the distribution of ΔS for the \( j \)-th type of binding motif. To compute this we need \( \pi_j = p(\text{subs. in type } j | \text{ one } \text{ subs.}) \), so

\[ \pi_j = \frac{p(\text{subs. in type } j)}{p(\text{one } \text{ subs.})} \]  

(25)

This can be computed using

\[ p(\text{subs. in type } j) = φ_j = n_j \sum_{i=1}^{w_j} \sum_{a} \sum_{b \neq a} f_{iab}P_{ab} \]  

(26)

where \( w_i \) is the length of the \( j \)-th motif and \( n_i \) is the number of times that motif occurs in the regulatory region. In this case

\[ φ = \sum_{j=1}^{v} φ_j \]  

(27)

and therefore

\[ \pi_j = \frac{φ_j}{φ_j} = \frac{n_j \sum_{i=1}^{w_j} \sum_{a} \sum_{b \neq a} f_{iab}P_{ab}}{\sum_{j=1}^{v} n_j \sum_{i=1}^{w_j} \sum_{a} \sum_{b \neq a} f_{iab}P_{ab}} \]  

(28)

for case 1. To compute the mean and variance of arbitrary mixture models we proceed as follows. To simplify the notation, I will indicate sums over \( i, a, b, a \) as sums over ΔS. In this notation,

\[ E[ΔS] = \sum_{ΔS} ΔS p(ΔS) \]

= \[ \sum_{ΔS} \sum_{j=1}^{v} \pi_j p(ΔS)_j = \sum_{j=1}^{v} \pi_j \sum_{ΔS} ΔS p(ΔS)_j \]  

(29)

= \[ \sum_{j=1}^{v} \pi_j E[ΔS]_j \]

using the linearity of the expectation. For the variance,

\[ V[ΔS] = \sum_{ΔS} \frac{(ΔS - E[ΔS])^2 p(ΔS)}{ΔS} \]

= \[ \sum_{ΔS} \frac{(ΔS - E[ΔS])^2 \sum_{j=1}^{v} \pi_j p(ΔS)_j}{ΔS} \]  

(30)

= \[ \sum_{j=1}^{v} \pi_j \sum_{ΔS} p(ΔS)_j (ΔS^2 - 2E[ΔS]_j ΔS + E[ΔS]_j^2) \]  

(31)

We now add and subtract the square of \( E[ΔS] \) for the \( j \)-th motif.

\[ V[ΔS] = \sum_{j=1}^{v} \pi_j \sum_{ΔS} p(ΔS)_j (ΔS^2 - E[ΔS]_j^2) \]

+ \[ E[ΔS]_j^2 - 2E[ΔS]_j ΔS + E[ΔS]_j^2 \]  

(32)

We now reorder the terms and take the expectations out of the summations,

\[ V[ΔS] = \sum_{j=1}^{v} \pi_j \left[ E[ΔS]^2 + E[ΔS]_j^2 \right. \]

+ \[ \left. \sum_{ΔS} p(ΔS)_j (ΔS^2 - E[ΔS]_j^2) \right] \]  

(33)

\[ -2E[ΔS] \sum_{ΔS} p(ΔS)_j ΔS \]  

\[ V[ΔS] = \sum_{j=1}^{v} \pi_j \left[ E[ΔS]^2 + E[ΔS]_j^2 \right. \]

+ \[ \left. V[ΔS]_j - 2E[ΔS] E[ΔS]_j \right] \]  

(34)

And finally

\[ V[ΔS] = \sum_{j=1}^{v} \pi_j V[ΔS]_j + \sum_{j=1}^{v} \pi_j (E[ΔS] - E[ΔS]_j)^2 \]  

(35)

Scripts to compute \( E[ΔS] \) and \( V[ΔS] \) will be provided from the author’s website.

Additional material

Additional file 1
eve stripe 2 enhancer binding sites. Alignments of binding sites in the eve stripe 2 enhancer

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-286-S1.PDF]
References

1. Prud’homme B, Gompel N and Carroll SB: Emerging principles of regulatory evolution. Proceedings of the National Academy of Sciences of the United States of America 2007, 104(Suppl 1): 8605–8612.

2. Wray GA: The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 2000, 2(6):225–228.

3. Wasserman WW, Palumbo M, Thompson W, Fickett JW and Lawrence CE: Human-mouse genome comparisons to locate regulatory sites. Nature Genetics 2000, 26(3):206–216.

4. Dermitzakis ET and Clark AG: Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Molecular Biology and Evolution 2002, 19(7):1141–21.

5. Dermitzakis ET, Bergman CM and Clark AG: Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Molecular Biology and Evolution 2003, 20(5):703–714.

6. Moses AM, Chiang DY, Kellis M, Lander ES and Eisen MB: Position specific determination in the rate of evolution in transcription factor binding sites. BMC Evolutionary Biology 2003, 3:19.

7. Emberly E, Rajewsky N and Sigga ED: Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics 2003, 4:57.

8. Sinha S and Sigga ED: Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. Molecular Biology and Evolution 2005, 22(4):874–885.

9. Cameron RA, Chow SM, Berney K, Chiou T, Yuan Q, Krämer A, Helguero A, Ransick A, Yun M and Davidson EH: An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(33):11769–11774.

10. Mustonen V and Lassig M: Evolutionary population genetics of promoters: predicting binding sites and functional phylogeny. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(44):15936–15941.

11. Mustonen V, Kinney J, Callan CG and Lassig M: Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences of the United States of America 2008, 105(34):12376–12381.

12. Gaffney DJ, Blekhman R and Majewski J: Selective constraints in experimentally defined primate regulatory regions. PLoS Genetics 2008, 4(6):e1000157.

13. Kim J, He X and Sinha S: Evolution of regulatory sequences in 12 Drosophila species. PLoS Genetics 2009, 5(1):e1000330.

14. Lynch M: The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences of the United States of America 2007, 104(Suppl 1): 8597–8604.

15. Hahn MW: Detecting natural selection on cis-regulatory DNA. Genetics 2007, 179(1):7–18.

16. Jenkins DL, Ortorti CA and Brookfield JF: A test for adaptive change in DNA sequences controlling transcription. Proc Biol Sci 1995, 261(1361):203–7.

17. Fay JC and Wu C: Sequence divergence, functional constraint, and selection in protein evolution. Annual review of genomics and human genetics 2003, 4:213–35.

18. Mustonen V and Lassig M: From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends in Genetics: TIG 2009, 25(3):111–119.

19. Schneider TD: Evolution of biological information. Nucleic Acids Research 2000, 28(4):2794–2799.

20. Berg J, Willmann S and Lassig M: Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology 2004, 4(1):42.

21. MacArthur S and Brookfield JF: Expected Rates and Modes of Evolution of Enhancer Sequences. Molecular Biology and Evolution 2004, 21(6):1064–1073.

22. Stone JR and Wray GA: Rapid evolution of cis-regulatory sequences via local point mutations. Molecular Biology and Evolution 2001, 18(9):1767–1770.

23. Storomo GD: DNA binding sites: representation and discovery. Bioinformatics 2000, 16(1):16–23.

24. Yang Z, Goldman N and Fitch WM A: Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Molecular biology and evolution 1994, 11(2): 316–24.

25. Drierer W and Nüsslein-Volhard C: The bcd protein is a positive regulator of hunchback transcription in the early Drosophila embryo. Nature 1989, 337(6263):138–143.

26. Lukowitz W, Schröder C, Glaser G, Hulskamp M and Tautz D: Regulatory and coding regions of the segmentation gene hunchback are functionally conserved between Drosophila virilis and Drosophila melanogaster. Mechanisms of Development 1994, 45(2):105–115.

27. Drierer W, Thoma G and Nüsslein-Volhard C: Determination of spatial domains using pairwise gene expression in the Drosophila embryo by the affinity of binding sites for the bicoid morphogen. Nature 1989, 340(6232):363–367.

28. Small S, Kraut R, Hoey T, Warrior R and Levine M: Transcriptional regulation of a pair-rule stripe in Drosophila. Genes & Development 1991, 5(8):827–839.

29. Stanojevic D, Small S and Levine M: Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. Science 1991, 254(5036):1385–1387.

30. Howard ML and Davidson EH: cis-Regulatory controls of development. Developmental Biology 2004, 271(1):109–118.

31. Ludwig MZ, Patel NH and Kreitman M: Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 1998, 125(5):949–958.

32. Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J and Kreitman M: Functional evolution of a cis-regulatory module. PLoS Biology 2005, 3(4):e93.

33. Ludwig MZ, Bergman C, Patel NH and Kreitman M: Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 2000, 403(6769):564–567.

34. Wong WSW and Nielsen R: Detecting selection in noncoding regions of nucleotide sequences. Genetics 2004, 167(2):949–958.

35. Pollard DA, Moses AM, Iyer VN and Eisen MB: Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics 2006, 7:376.

36. Stanojevic D, Hoey T and Levine M: Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Krüppel in Drosophila. Nature 1989, 341(6240):331–335.

37. Kim J: Macro-evolution of the hairy enhancer in Drosophila species. The Journal of Experimental Zoology 2001, 291(2):175–185.

38. Lavery R: Recognizing DNA. Quarterly Reviews of Biophysics 2005, 38(04):339–344.

39. Sadowski D and Kodadek T: Genomic studies of transcription factor-DNA interactions. Current Opinion in Chemical Biology 2005, 9(1):38–45.

40. Bryk LM: Protein binding microarrays for the characterization of DNA-protein interactions. Advances in Biochemical Engineering/Biotechnology 2007, 104:65–85.

41. Allely TM, Peña-Castillo L, Badis G, Talukder S, Berger MF, Gehrke AR, Philippakis AA, Bryk ML, Morris QD and Hughes TR: Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics 2009, 25(8):1012–1018.

42. Thorne JL, Kishino H and Felsenstein J: Inching towards reality: an improved likelihood model of sequence evolution. Journal of Molecular Evolution 1997, 45(4):412–426.

43. Hwang DG and Green P: Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America 2004, 101(39):13994–14001.

44. Siepel A and Haussler D: Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Molecular Biology and Evolution 2004, 21(3):402–416.

45. Bergman BP, Nibu Y, Pfeiffer BD, Tomanic P, Celinker SE and Levine M, et al: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proceedings of the National Academy of Sciences of the United States of America 2002, 99(2):575–62.
46. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H and Diekhans M, et al: The UCSC Genome Browser Database: 2008 update. Nucl Acids Res 2008, 36(suppl_1):D773–779.
47. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E and Green ED, et al: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Research 2003, 13(4):721–731.
48. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B and Markow TA, et al: Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007, 450(7167):203–218.
49. Yang Z: PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 2007, 24(8):1586–1591.
50. Durbin R, Eddy SR, Krogh and Mitchison G: Biological sequence analysis. Cambridge University Press: 1998.
51. Felsenstein J: Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 1981, 17(6):368–76.
52. Bergman CM, Carlson JW and Celniker SE: Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 2005, 21(8):1747–1749.

Publish with BioMed Central and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp