Chapter 4

SEQUENCE CHANGES EXPECTED TO BE ASSOCIATED WITH GAINS AND LOSSES OF TFBSS

4.1 INTRODUCTION

An interesting finding was highlighted at the end of the chapter on the TransFac-based analysis, concerning the DNA surrounding a TFBS. If the TFBS was conserved during evolution, then the sequence of the surrounding DNA also tended to be better conserved than the DNA sequence surrounding diverged TFBSs. The present chapter considers possible explanations for this finding. It will also consider how these explanations might be tested; this means considering any predictions that might be made.

4.2 OUTLINE OF MODELS

The explanation developed at the end of the chapter on the TransFac data was that "the probability of gaining/losing a TFBS is affected by the presence of other TFBSs". This explanation is rather vague; within this general concept, it was possible to develop two more detailed hypotheses, which involved distinctly different concepts.

One was called the "correlated-evolution" hypothesis. If, when a TFBS was lost during evolution, other nearby TFBSs were also lost, then mutations could more easily accumulate in the nearby DNA and this would give an effect that was qualitatively the same as that shown in Table 2.18. Similarly, for gain-of-TFBS, the hypothesis also assumes that it is usual for several TFBSs close together to all be gained at the same time, or at nearly the same time.

The other, developed later, was called the "site-density" hypothesis. Here "site-density" was defined as the number of TFBSs in a region divided by its length. It was assumed that DNA that was not part of a TFBS would be "spacer" DNA that could easily accumulate mutations. Consequently, regions of low site-density would have plenty of "spacer" DNA, so the DNA sequence would not be highly conserved *even if all the TFBSs were conserved*. This is illustrated in figure 4.1, which should be compared with the high site-density case illustrated in fig 4.2. If we also suppose that TFBSs are more likely to be gained/lost if they are in a region of low site-density than if they are in a region of high site-density, this would also give an effect that was qualitatively the same as that shown in Table 2.18.

The TransFac-based results might also be explained by a third hypothesis, the "varying mutation rate" hypothesis. Consider what will happen to regions of the genome that are subject to a higher mutation rate than average (where "mutation rate" means the rate for bases that are not subject to natural selection). This high mutation rate might cause TFBSs to be gained/lost more rapidly, and would also reduce the general level of sequence conservation, thus producing a correlation between the two. However, it is doubtful if

Figure 4.1: A region of low "site-density"
This illustration shows a stretch of DNA (blue line) which contains some TFBSs (each blue block represents a TFBS). It assumes that mutations within each TFBS tend to be rejected by natural selection, whereas mutations elsewhere are tolerated.

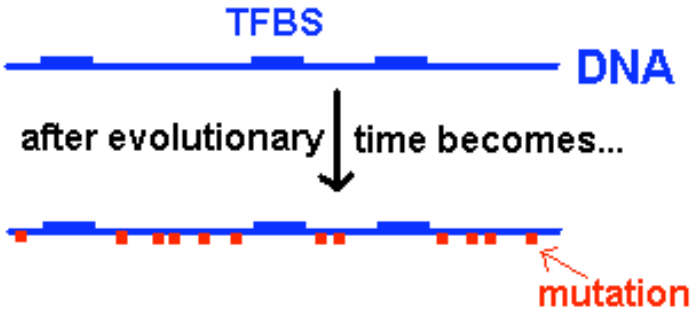
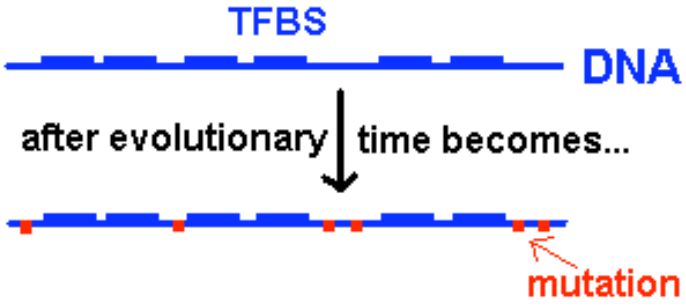


Figure 4.2: A region of high "site-density"
This is similar to fig 4.1, except that the "site-density" is much higher - that is, there are more TFBSs within the same length of DNA. Compare the number of mutations that accumulate, with the number of mutations in fig 4.1.



this can be applied to TFBSs that are being conserved by natural selection, since these will be conserved irrespective of the mutation rate. It seems to be more plausible for TFBSs that are not subject to natural selection, such as the following:

- Suppose a TFBS in the database is a mistake, and does not really exist. Clearly this (apparent) TFBS will not be subject to natural selection. (Admittedly, it may overlap real TFBSs that are being conserved by natural selection, which would reduce the probability of it being disrupted by mutations; but this seems unlikely to eliminate completely disruption by mutations). Therefore, if the database was heavily contaminated with mistaken TFBSs, the "varying mutation rate" hypothesis might be a plausible explanation for trends observed in the data.
- Some TFBSs may not be subject to natural selection. This seems plausible for TFBSs that are in the process of being lost, and perhaps to certain parts of the gain-of-TFBS process. In addition, a *Drosophila* study (Emberly et al., 2003) has suggested that a substantial proportion of TFBSs do not seem to be subject to natural selection.

It will be noted that this has already been (implicitly) considered in previous chapters. The fictional-TFBS analysis, described in previous chapters, gives a prediction of what results would be produced by mistaken TFBSs in the database.

Thus, there are at least three different models for the evolution of regulatory regions, the "correlated evolution", "site-density" and "varying mutation rate" models, each of which gives a possible explanation why the level of neighbourhood conservation is related to whether a TFBS is conserved or not. It can be broadly understood how they do this without going into mathematical details.

It became clear that deciding which hypothesis was correct would require additional information, which seemed to require a larger sample of TFBSs than the TransFac data provided. This was the motivation for analysing a

different source of data that would provide a large sample of TFBSs, the "chip-chip" data described in an earlier chapter; this however failed to shed light on the matter.

4.3 DIFFERENCES BETWEEN MODELS

4.3.1 Conceptual differences

It is important not to confuse the "correlated-evolution" and "site-density" hypotheses. Thus it is useful to consider some particular differences between them.

One difference is this: The "site-density" hypothesis only requires the *presence* of other TFBSs, which do not have to be gained or lost. In contrast, the "correlated-evolution" hypothesis requires that other TFBSs are gained or lost at the same time as the TFBS being studied.

A second difference between the two hypotheses is this: In the "correlated-evolution" model, mutations in the surrounding DNA can accumulate at a high rate only *after* loss of TFBSs. In contrast, in the "site-density" model, regions with low site-density will accumulate mutations at a high rate and are more likely to have gains/losses, but the high mutation rate is independent of whether the gains/losses actually occur.

4.3.2 Possible observable differences

Another difference between the hypotheses involves a distinction between TFBS loss and TFBS gain. The difference concerns the "neighbourhood" (ie, the flanking DNA on either side of a TFBS), and is as follows. The "correlated-evolution" model predicts that the neighbourhood around "losses" will be conserved more highly than the neighbourhood around "gains". The "site-density" model does not predict this difference. This could provide a

method of testing which model is more plausible, and will be explored in a later chapter.

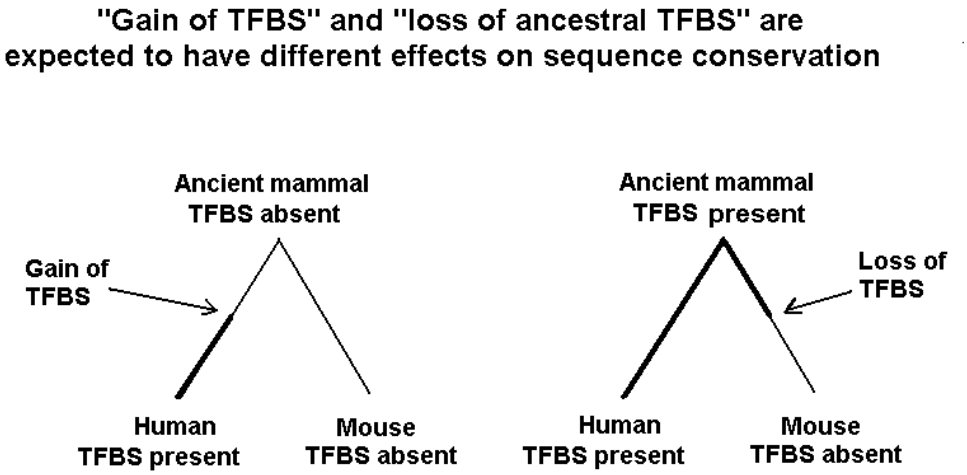
Why does the "correlated-evolution" model give this prediction? For a simple explanation, see Figure 4.3. This explains why, *for the sequence of the TFBS itself*, we expect "losses" to show higher sequence conservation than "gains". However, we want to know about conservation of the sequence *flanking* the TFBS, and on its own, Figure 4.3 does not say anything about this; we need to combine it with one of our evolutionary models to make a prediction. If we use the "correlated-evolution" model, then since the TFBS is lost (or gained) at the same time that other TFBSs in the surrounding sequence are lost (or gained), then the sequence effects outlined in Figure 4.3 will apply to the surrounding sequence as well. Thus, on this basis, the "correlated-evolution" model predicts that the neighbourhood around "losses" will be conserved more highly than the neighbourhood around "gains".

4.3.3 Complications if TFBSs are subject to counterselection when they are lost

One limitation to this prediction (figure 4.3) is that it assumes that evolution has a neutral effect on a TFBS that has just been lost; that is, mutations that disrupt binding are neither selected for, nor selected against. Clearly this assumption will be violated if a TFBS suddenly changes from being "useful" to being "harmful", so that immediately after the change, mutations that disrupt binding will be selected for. If so, sequence conservation may be low even in the circumstances shown in the right hand diagram of figure 4.3.

It is therefore worth asking if TFBSs that lose their purpose tend to be counterselected against. In the Introduction chapter, it was noted that the general rule was that there is no counterselection of unnecessary TFBSs. However, whilst there is some evidence in favour of that rule (Tronche et al., 1997), the evidence relates to potential TFBSs throughout regulatory DNA; it does not specifically relate to TFBSs that performed a useful function in the past

Figure 4.3: Sequence conservation expected in lost and gained TFBSs



but have stopped being useful. The explanation for the rule does not seem to be known, and therefore it is difficult to predict the circumstances (if any) where the rule does not apply. In prokaryotes, counterselection has been observed, although it is weak (it is only strong enough to reduce spurious TFBSs to 80-90% of the expected value) (Hahn et al., 2003).

Looking at how this question is treated in the literature, there are models of TFBS evolution that assume that a TFBS, once it is lost, is subject to neutral evolution only. For example, one paper (Doniger and Fay, 2007) considers several possible models under which a TFBS either (a) evolves neutrally, or (b) is conserved, or (c) is conserved until it is lost and then evolves neutrally. But it does not consider a model under which a TFBS is lost by selection of disruptive mutations, nor is any explanation given why this is not done. Similarly, another paper (Mustonen and Lassig, 2005) examines binding site loss without considering if this might involve disruptive mutations being selected for. (They give equations for the probabilities of sequence changes under "neutral" conditions, and alternatively under "constant selection" conditions; the equation for the probability of binding site loss is built up from these two probabilities, without including any term for sequence changes caused by selection of mutations that disrupt binding). In contrast, another paper (Berg et al., 2004) does indeed recognise that negative selection may eliminate "spurious binding sites"; however, they apply this idea chiefly to new binding sites created by chance mutation, rather than to the loss of existing binding sites; the loss of existing sites they still consider modelling using a neutral model - "If the selection pressure on an existing site ceases, that site will disappear on the larger time scale T_0" (where T_0 is the time scale of neutral evolution).

Thus, in the literature it is common to model TFBS loss as involving neutral selection only (ie, the models do *not* include selection for mutations that disrupt binding). The rest of this thesis will also adopt this assumption.

However, the papers just cited do not include explicit discussion of this assumption, nor present any evidence to demonstrate that it is true; presumably it was adopted because of simplicity or plausibility. Indeed one paper

suggests counterselection many occur, even though their models do not take it into account (Doniger and Fay, 2007). It is thus worth asking if this assumption is reliable.

In principle, it is easy to imagine circumstances in which a TFBS changes suddenly from "useful" to "harmful". As an imaginary example, suppose a species of bird has in its genome a TFBS that helps it to develop large wings. For this species, the main advantage of having wings is that it enables the bird to escape from ground-dwelling predators. On a particular day, a hurricane blows a flock of these birds far out to sea, where some of them become stranded on an island. The island provides adequate food for them, and contains no large ground-dwelling predators. It is now a disadvantage to have large wings, as they are "expensive" (in terms of the food consumption required to develop and power them), and they no longer provide any commensurate advantage to compensate for this. The birds will survive better if they become flightless. Therefore, the TFBS that causes large wings is now harmful; any mutation disrupting it will be selected for. Note that this TFBS changed from being "useful" to being "harmful" in a few hours, without going through any intermediate "neutral" stage.

The example just given is an illustration of a more general principle. Any animal will have a number of "expensive features" that require a considerable amount of food consumption to maintain (obvious examples are wings, tails, brains, etc; but one could consider less obvious examples, such as molecular pathways). The fact that they have these features shows that they aid survival, indicating that their usefulness outweighs their cost. But, if a change of circumstances means that a feature is no longer useful, then (as it is expensive) it changes directly from "useful" to "harmful". Consequently, the TFBSs that cause its development will also change from "useful" to "harmful".

Thus, in principle, it is plausible to imagine that TFBS loss will sometimes involve selection for mutations that disrupt binding.

It is also possible that counterselection occurs for some but not all TFBSs that are lost. Indeed, that idea is particularly plausible if one adopts the "enhanceosome" model (Arnosti and Kulkarni, 2005) already mentioned in

the Introduction chapter. Suppose that, during the course of evolution, an enhancer conforming to this model changes from being "useful" to being "harmful". Clearly, mutations that disrupt this enhancer will be positively selected. However, the "enhanceosome" model implies that the enhancer will completely cease to function if a single one of its TFBSs is lost. Therefore, once selection of mutations has destroyed the binding ability of *one* of the TFBSs, the enhanceosome will cease to function, and so further losses will not affect the fitness of the animal either positively or negatively; consequently the enhanceosome will no longer be subject to selection. In other words the enhanceosome model seems to predict that, even if there is counterselection against TFBSs, one TFBS in an enhanceosome will be disrupted by counterselection, after which the other TFBSs in that enhanceosome will be disrupted by the slow accumulation of mutations which are not selected.

Evidence of counterselection might be obtained by examining cases of TFBS loss. If disruptive mutations were selected for, then for a period the number of substitutions will have exceeded those predicted by a neutral model. In principle, this might be observed and thus provide evidence to address the question. Indeed, when the TFBS data collected for the present project were examined, a number of cases were noticed where TFBS loss appeared to have been caused by more substitutions than would be expected under neutral selection. However, it is not clear that the number of such cases exceeds the number expected to occur by chance, so these data are not considered worth presenting in detail. But if a procedure were developed to predict the number of such cases that would be expected to occur by chance (which would be complicated), then this approach might yield evidence that would address whether TFBS loss is accompanied by counterselection.

4.3.4 Time taken to disrupt a TFBS that is not under natural selection

The reader may have noticed that, in figure 4.3, "loss of TFBS" was used in a way inconsistent with the rest of this thesis. In figure 4.3, "loss of TFBS"

referred to the moment when a TFBS ceased to be useful for the survival of the animal. In the rest of the thesis, "loss of TFBS" refers to when a TFBS loses its ability to bind (see page 66). This was necessary, but as the two events may not take place at the same time, it is worth asking if the time-difference is trivial, or if it amounts to a significant amount of evolutionary time. The question is essentially one of how long it takes to disrupt a TFBS that is not under natural selection.

As a simple example, consider comparing modern humans with the last common ancestor of humans and mice. For bases that are not subject to natural selection, it is reasonable to assume that substitutions occured in 14% of bases (Taylor et al., 2006). That gives an 86% chance that a base will *not* undergo substitution. Thus, taking (for example) a TFBS 10 bases long, the probability that none of the bases will undergo substitution is $0.86^{10} = 0.22$; that is, a 22% probability that the TFBS sequence will be completely unchanged.

As a second example, a more complicated exercise was done. The method was to take an actual TFBS (from TransFac, as used in the earlier chapter). This would be subject to one/two/three random substitutions. The modified TFBS would then be compared against the original TFBS, to see if it met the criteria for a "diverged" TFBS - the criteria used was the same as that used in the analysis of TransFac data (as described in detail on page 94). Doing this for a large sample of TFBSs, it was found that:

- for TFBSs subject to 1 random substitution, 15% met the criteria to be classified as a diverged TFBS
- for TFBSs subject to 2 random substitutions, 36% met the criteria to be classified as a diverged TFBS
- for TFBSs subject to 3 random substitutions, 56% met the criteria to be classified as a diverged TFBS

As a third example, imagine a TFBS which existed (and was useful) in the last common ancestor of humans and mice, but which stopped being useful

within the mouse lineage. Suppose the time when it stopped being useful was about halfway between that last common ancestor and the modern mouse. From that "halfway time" to the modern mouse, the TFBS would be subject to a substitution rate of, say, 0.22 (Taylor et al., 2006). To determine what would happen to that TFBS, the "second example" exercise (as described just above) was repeated, except that the number of substitutions was chosen to given a substitution rate of 0.22. 49% of cases met the criteria to be classified as a diverged TFBS. Thus, only half these TFBSs would be detected as being non-diverged; the other half would not accumulate enough substitutions to be detected as diverged according to the criteria used in this thesis.

These three examples can be summarised as implying that, when a TFBS loses its usefulness, there will be a delay of perhaps tens of millions of years before the TFBS is disrupted enough to be detected as a loss-of-TFBS by the methods used in this thesis. Indeed, a substantial minority of TFBSs that lost their usefulness around the human-mouse split may have survived into modern humans without accumulating *any* substitutions; this makes the loss of usefulness very difficult to detect by any bioinformatic or *in-vitro* method.

Thus, if we consider "loss of TFBS" cases like that illustrated in figure 4.3, the methods used in this thesis will fail to detect "loss of TFBS" in a high proportion of cases. Moreover, chance alone will cause some TFBSs to accumulate more substitutions than expected from the average substitution rate, whereas other TFBSs will accumulate fewer. Clearly those which accumulate more substitutions are more likely to be detected, so the sample of known "loss of TFBS"s will be biassed towards these.

All this section has been written on the assumption that TFBSs are not under natural selection, once they have lost their usefulness. If, in contrast, a TFBS changes suddenly from being useful to being harmful, then mutations that disrupt the TFBS will be positively selected for, causing a rapid change in sequence of the former TFBS, almost eliminating the delays.

4.4 SUMMARY

To explain the results in an earlier chapter (such as table 2.18), three hypotheses were proposed: the "correlated evolution", "site-density" and "varying mutation rate" models. Each gives a possible explanation why the level of neighbourhood conservation is related to whether a TFBS is conserved or not.

A test was proposed that might demonstrate that the "correlated evolution" hypothesis is more plausible than the "site-density" hypothesis. This required distinguishing cases where a TFBS was gained from cases where a TFBS was lost. The "correlated evolution" hypothesis predicted that the DNA sequence surrounding the TFBS should be more highly conserved for "losses" than for "gains".

However, the proposed test was subject to two limitations. One was that it assumed that a TFBS, once it ceased to be useful, was no longer subject to natural selection. This assumption would be violated if natural selection acted to destroy the TFBS, by positively selecting mutations that disrupted it. The other limitation was the difficulty in detecting TFBSs that had ceased to be useful for survival, since some might not accumulate enough mutations, even after some tens of millions of years, to be detected as a "loss" by the methods used in this thesis.

Declaration: no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning, except for the reuse of some Perl modules (altered as required), originally written during the candidate's Master in BioInformatics course (University of Exeter)