Chapter 5

PHYLOGENETIC TREES: DISTINGUISHING GAINS FROM LOSSES OF TFBSs

5.1 INTRODUCTION

Phylogenetic trees can extend the work described earlier by showing that certain diverged TFBSs are ancient sites that have been lost in some species, whereas certain other diverged TFBSs have been gained recently.

This information may be useful in answering these questions: (i) Are gain-of-TFBS events about as frequent as loss-of-TFBS events, or is one much more frequent than the other? (ii) Do "gains" show different characteristics to "losses"? Discussion of the "correlated evolution" model suggested that the DNA sequence flanking "lost TFBS" would be conserved more highly than DNA sequence flanking "gained TFBSs" (page 163). In contrast, simulations based on the "site density" model showed no such difference. If there were enough data on gained/lost TFBSs, it would help decide which model was correct.

Since a mammalian gene tends to have a more complicated regulatory system

than the gene of simple organisms, with more TFBSs, it seems likely that the number of TFBSs per gene has increased during vertebrate evolution (or before). If so, TFBS gains must have been more frequent than losses. It would be interesting to confirm (or refute) this intuition with data. If gains are indeed more frequent than losses, it would also be useful to have data to show whether gains are *much* more frequent than losses, or whether they are only slightly more frequent than losses. The former would imply that regulatory systems are evolving to be more complicated and that most gains are contributions to this process. The latter would imply that most gains are part of a process of turning over TFBSs without altering the overall complexity of regulation.

"Phylogenetic analysis" often means the development of methods to determine phylogenetic trees as accurately as possible. But in this study, the approach was completely different; it was assumed that the phylogenetic tree of vertebrates was already well known (except for certain details) and not likely to be improved by TFBS analysis. Instead, the task was to take a phylogenetic tree that was already known, and add labels indicating where particular TFBSs were lost or gained.

5.2 METHOD

5.2.1 Use of phylogenetic tree

Earlier chapters have shown how bioinformatic methods were used to examine TFBSs known to exist (eg, from the TransFac database (Matys et al., 2006)), and to determine whether a particular TFBS was conserved in certain other species. This chapter describes how this information was further analysed using phylogenetic trees.

A phylogenetic tree containing eight vertebrates was used. The tree used, figure 5.1, is supported by both morphology (Shoshani and McKenna, 1998) and molecular sequence evidence (Murphy et al., 2001).

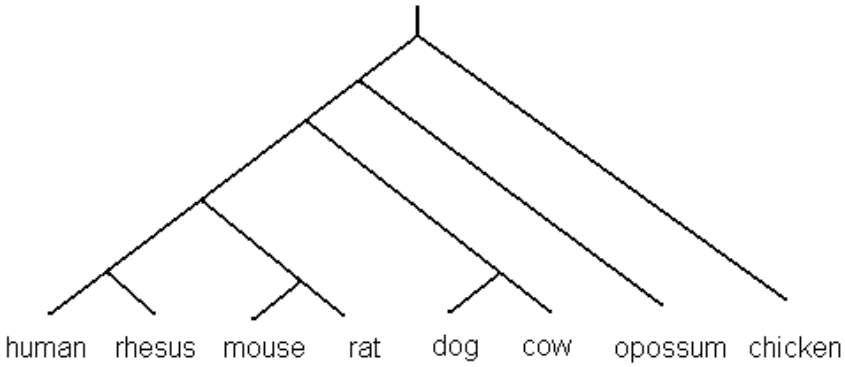
BioPerl (Stajich et al., 2002) modules were used to represent this phylogenetic tree in software.

To estimate where a TFBS was gained or lost, it was assumed that the correct evolutionary history was the one with the smallest number of changes needed to explain the data. This was following a long tradition of basing methods on parsimony - parsimony was the foundation for many of the early methods for constructing phylogenetic trees (Felsenstein, 2004). In this case, the detailed procedure was as follows.

For each particular TFBS (upstream of a particular gene), the first step in the analysis was to add a tag to each modern-species node indicating that the TFBS was "Present" or "Absent". Often, however, the earlier bioinformatics analysis would have concluded that it was "Uncertain" whether a TFBS existed in a particular species. In such a case, the species for which the TFBS was "Uncertain" would be removed from the phylogenetic tree, which would then be simplified as much as possible.

In figure 5.2 can be seen two examples of trees of this type; intuitively, it is fairly easy to decide that one of these can most simply be explained as a "gain", but the other can most simply be explained as a "loss"; to make this same decision by computer, the following procedure was used.

Figure 5.1: The eight species used, in an evolutionary tree



(Not to scale.)

Each ancestor node was assigned a TFBS "present" or TFBS "absent" tag. Each such node represents an ancient animal, so the choice of tag will be a little speculative. Therefore, a large number of versions of the labelled phylogenetic tree were produced, containing every possible combination of "present"/"absent" tags for the ancestor nodes. The simplest version was then selected - where "simplest version" was defined as the version with the smallest number of gain-or-loss events. In some cases there would be more than one simplest tree.

Where there was only one simplest tree, it could be be used to decide whether the change of TFBS was a "gain" or a "loss". An extra field of information would be added to an enlarged version of the data file that had been used as input; where the TFBS was "present" for the oldest ancestor node in the tree, then this TFBS would be labelled as "Loss", whereas if the TFBS was "absent" for the oldest ancestor, then the TFBS would be labelled "Gain". Note that these labels apply to a *species-pair* (the species in which the TFBS is already known to exist, and another species which the bioinformatics analysis concluded did not contain the TFBS).

5.2.2 Fictional TFBSs

An analysis of fictional TFBSs is intended to show what results can be expected if incorrect data are fed into the analysis. The method of generating fictional sites has been detailed earlier (see page 108). In outline, it starts with a matrix describing the DNA-binding properties of a real transcription factor, and a list of real TFBSs for that transcription factor. The rows of the matrix are shuffled so as to produce a matrix describing a fictional transcription factor. Genomic DNA near the real TFBSs is then searched to find short sequences of DNA that match the fictional matrix; these are the fictional TFBSs.

For this purpose, a "fictional TFBS" is defined as one that is known to not exist and was deliberately generated for testing purposes. In contrast, a "mistaken TFBS" is defined as one that is believed to exist on the basis of

experimental evidence, but in fact does not exist.

5.2.3 Conservation of DNA surrounding the TFBS

For this, the measure used was the amount of similarity in the DNA when two species are compared, relative to the same measure for conserved TFBSs for the same species-pair. For example, take a TFBS for which there is experimental evidence of its existence in human; which the bioinformatic evidence suggests does not exist in dogs; and where the phylogenetic analysis suggests that this is a loss of TFBS in the dog lineage. The DNA 50 bases either side of the human TFBS (not including the TFBS itself) would be compared with the aligned DNA in the dog to find the percent similarity. This would be compared with the average corresponding percent similarity for all TFBSs that are conserved between human and dog (as identified by the bioinformatic analysis). This difference would then be included in an average that included all the "loss" TFBSs.

Errors were estimated by taking 100 bootstrap samples, where the experimentally determined TFBS was the unit of analysis. Thus, if a particular TFBS was selected to be included into a particular bootstrap sample, *all* the similarity data relating to that TFBS would be included.

5.2.4 TFBSs located by a Chip-Chip Experiment

These data were also used, as an alternative to TransFac data, and were described in more detail earlier (see page 137). In summary, a list of regulatory regions bound by the transcription factor HNF4 were taken from the literature (Odom et al., 2004). The weight-matrix describing the binding of this transcription factor was constructed, based a list of TFBSs in the literature (Ellrott et al., 2002). As the chip-chip method does not provide the exact location of a TFBS, this was determined by searching for the best "match" within 500 bases upstream of the start of the gene. The TFBS would only be used if the "match" was a moderately good one ($p<0.1$).

5.3 RESULTS

5.3.1 Based on TFBSs from the TransFac Database

Table 5.1 shows results based on TFBSs extracted from the TransFac database. Only a small proportion of these gave evidence of a gain-of-TFBS event (20 examples) or of a loss-of-TFBS event (21 examples). This was because, for many TFBSs, no evidence was found that the TFBS was diverged and, where there was evidence that the TFBS was diverged, often it was unclear whether this was a gain or a loss. As an example of the latter, if a TFBS is present in mouse and rat, and absent in human and rhesus, then this could be an ancient TFBS that was lost in the primate lineage, but it could also be a TFBS that was gained recently in the rodent lineage.

Table 5.2 shows results based on fictional TFBSs. The most striking result is that "gains" are much more frequent than "losses". This suggests that, if any of the "real TFBS" data contains TFBSs that do not really exist (but are mistakes), then the effect of these mistaken TFBSs will be to bias the results in Table 5.1 towards a result in which gains far exceed losses.

The gain-to-loss ratio of Table 5.1 appears different to that of Table 5.2, but a statistical significance test examined whether this was, in fact, significant; Table 5.3 shows the contingency table for the statistical test used. This showed that the difference was significant ($p = 0.046$). Hence, the real TFBSs are giving different results to the fictional TFBSs; this is evidence that the results for the real TFBSs are not merely the result of them being contaminated by mistaken TFBSs from the database.

Another aspect that was investigated was the question, noted in the Introduction to this chapter, of whether DNA flanking "losses" is conserved more highly than DNA flanking "gains". Table 5.4 shows this for each pair of species, so that an overall answer can be calculated using a "species-stratified" analysis. (The reason for using a "species-stratified" analysis was explained in an earlier chapter, where table 2.12 gave an example of it being needed).

The main result from Table 5.4 is that "losses" are flanked by DNA that is

Figure 5.2: Examples showing apparent gains and losses
(a) has the appearance of a TFBS that has been gained recently in evolutionary history; (b) has the appearance of an ancient TFBS, that has been lost recently in one lineage. (Not to scale).

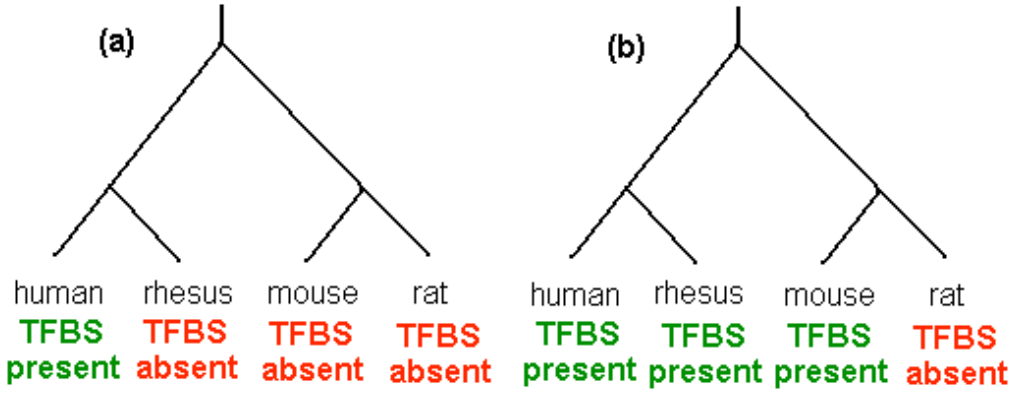


Table 5.1: Gains and losses: TransFac data
Numbers of TFBSs which, when analysed using a phylogenetic tree, suggested that the TFBS had been gained during vertebrate evolution or lost during vertebrate evolution. Based on TFBSs in the TransFac database.

	Number	%
Number of cases where the presence or absence of the TFBS has been determined for at least two species	301	100 %
"Gains" - the number of TFBSs for which the simplest explanation is a single gain of the TFBS	20	6.6 %
"Losses" - the number of TFBSs for which the simplest explanation is a single loss of the TFBS	21	7.0 %
The number of TFBSs for which the simplest explanation involved more than one gain/loss event	3	1.0 %

Table 5.2: Gains and losses: fictional data

Numbers of TFBSs which, when analysed using a phylogenetic tree, suggested that the TFBS had been gained during vertebrate evolution or lost during vertebrate evolution. Based on fictional TFBSs that were generated to test the methodology.

	Number	%
Number of cases where the presence or absence of the TFBS has been determined for at least two species	1061	100 %
"Gains" - the number of TFBSs for which the simplest explanation is a single gain of the TFBS	233	22.0 %
"Losses" - the number of TFBSs for which the simplest explanation is a single loss of the TFBS	117	11.0 %
The number of TFBSs for which the simplest explanation involved more than one gain/loss event	65	6.1 %

Table 5.3: Gain/loss significance test

Contingency table for testing if the ratio of gains to losses for the real data is different from the corresponding ratio in the fictional data

	Real data	Fictional data
Number of trees showing a gain of TFBS	20	233
Number of trees showing a loss of TFBS	21	117

Statistical significance estimated by chi-squared test: $p = 0.046$ (2-sided, with Yates' correction)

more highly conserved than the DNA flanking “gains” is. The difference is 4.3% (but subject to an error of 2.8%). The 4.3% average is on the borderline of being significantly different from zero (p=0.05) (based on a negative average being obtained in 52 out of 1000 bootstrap samples of the data in Table 5.4). Thus, this is tentative evidence for “losses” being flanked by DNA that is more highly conserved than the DNA flanking “gains”.

A “species-stratified” analysis of the fictional TFBSs (not shown in the table) gave an average difference between “losses” and “gains” of only 0.3% (error: 1.4%). This suggests that any mistaken TFBSs in the real data will not cause any spurious difference between "gains" and "losses", so far as the conservation of DNA surrounding the TFBS is concerned.

5.3.2 Based on TFBSs located by a Chip-Chip Experiment

Table 5.5 shows the ratio of "gains" to "losses" for real chip-chip data. Whereas the TransFac-based data (Table 5.1) showed "gains" and "losses" occurring in roughly equal numbers, instead we now have far more "gains" than "losses". In that respect, it resembles the fictional TFBSs (Table 5.6), which also show far more "gains" than "losses". Moreover, the ratio of “gains” to “losses” is very similar in both cases; both tables show “gains” being about four times as frequent as “losses”.

No tables are given showing conservation of flanking DNA, as with only three “losses” in the sample, there is almost no chance of detecting any difference between “gains” and “losses”.

Thus, the “gain” to “loss” ratio for real data resembles that for fictional data. Therefore, we should consider the possibility that the real chip-chip data contains mistaken TFBSs in enough numbers that they produce most of the "gains" from the real data.

Table 5.4: Neighbourhood conservation, TransFac data, species-stratified analysis
Conservation of DNA surrounding TFBSs: differences between "gains" and "losses", stratified by species.

Conservation of flanking sequence, when gain-of-TFBS distinguished from loss-of-TFBS						
Pair of species being compared (the first species has the experimentally determined TFBS; the second is the comparison species without the TFBS)	%id of 50 bases either side of TFBS		Sample size		Difference (value for losses minus value for gains)	Minimum sample size (number of lost TFBSs or number of gained TFBS, whichever is smallest)
	TFBSs lost	TFBSs gained	TFBSs lost	TFBSs gained		
human rhesus	90.7 %	90.9 %	4	4	-0.2 %	4
human cow	81.8 %	76.0 %	2	6	5.8 %	2
human dog	73.6 %	72.3 %	1	3	1.3 %	1
human mouse	76.4 %	68.4 %	7	6	8.0 %	6
human rat	81.9 %	73.2 %	1	4	8.7 %	1
human opossum		63.2 %	0	3		0
human chicken		55.1 %	0	1		0
mouse human		63.6 %	0	4		0
mouse rhesus		69.6 %	0	3		0
mouse cow		59.6 %	0	3		0
mouse dog	65.5 %	65.4 %	1	4	0.1 %	1
mouse rat		92.0 %	0	1		0
rat human	83.8 %	73.3 %	1	4	10.5 %	1
rat rhesus		69.7 %	0	3		0
rat cow	64.4 %	65.4 %	3	2	-1.1 %	2
rat dog		71.5 %	0	4		0
rat mouse		88.0 %	0	2		0
Weighted average					4.3%. Error (bootstrap): 2.8%	

The minimum sample size was used as the weight, when calculating the weighted average

Table 5.5: Gains and losses: chip-chip data
Numbers of TFBSs which, when analysed using a phylogenetic tree, suggested that the TFBS had been gained during vertebrate evolution or lost during vertebrate evolution. Based on target genes for the transcription factor HNF4, as determined by a chip-chip experiment.

	Number	%
Number of cases where the presence or absence of the TFBS has been determined for at least two species	124	100 %
"Gains" - the number of TFBSs for which the simplest explanation is a single gain of the TFBS	12	9.7 %
"Losses" - the number of TFBSs for which the simplest explanation is a single loss of the TFBS	3	2.4 %
The number of TFBSs for which the simplest explanation involved more than one gain/loss event	1	0.8 %

Table 5.6: Gains and losses: fictional chip-chip data
Numbers of TFBSs which, when analysed using a phylogenetic tree, suggested that the TFBS had been gained during vertebrate evolution or lost during vertebrate evolution. Based on fictional TFBSs that were generated to test the methodology and designed to resemble chip-chip data.

	Number	%
Number of cases where the presence or absence of the TFBS has been determined for at least two species	536	100 %
"Gains" - the number of TFBSs for which the simplest explanation is a single gain of the TFBS	106	19.8 %
"Losses" - the number of TFBSs for which the simplest explanation is a single loss of the TFBS	27	5.0 %
The number of TFBSs for which the simplest explanation involved more than one gain/loss event	15	2.8 %

5.4 DISCUSSION

5.4.1 Ratio of gains to losses

The only reasonably firm result of interest is the ratio of "lost TFBSs" to "gained TFBSs", based on the TransFac sites. The proportion of "lost TFBSs" for the real data (table 5.1) was significantly larger than it was for the fictional-TFBS analysis (table 5.2). This suggests that "loss" events make up a substantial proportion of "gain/loss" events. It is evidence against any idea that "loss of TFBS" is extremely rare compared to "gain of TFBS".

What is the smallest proportion that "loss of TFBSs" could have, consistent with this evidence? The fictional-TFBS analysis gave the ratio (frequency of "loss of TFBS" : frequency of "gain of TFBS") as being 0.50. Although the ratio for real-TFBS results was significantly difference from this, the level of significance was not very high, indicating that 0.50 is close to the limits of confidence. It is therefore reasonable to regard ˜0.50 as being the lower limit of the ratio for real TFBSs. (This does not allow for any bias in the analysis, notably the "discovery bias" mentioned below).

Thus the conclusion is that, for every "gain of TFBS" event, there are at least 0.50 "loss of TFBS" events, and possibly a much larger number.

It is more difficult to put a lower limit on the frequency of "gain of TFBS" events. In order to do so, we need to know that at least some of the "gain of TFBS" events identified are genuine and not cause by mistaken TFBSs, but tables 5.1 and 5.2 provide no help in demonstrating this.

5.4.2 Ratio of gains to losses - comparison from the literature

Examining the literature, a study has been published which distinguishes gains from losses and estimates the rate of each (Moses et al., 2006). For "excess" (that is, functional) TFBSs, they estimate 41.7 ± 37.6 gains in the

regions of the *Drosophilia* genome they studied. They also estimated that 24.8 ± 16.2 losses occurred, giving a net gain of 16.9 ± 34.2. (The $\pm$ estimates, copied from the paper, represent twice the standard error, unlike the $\pm$ estimates presented elsewhere in this thesis, which represent one standard error). Thus nominally, gains are more frequent than losses, but (noticing the size of the error estimates), it is not certain which really occurs more frequently.

That study uses some methods that are similar to those described in the current chapter of this thesis, but there are notable differences.

The similarities will be listed first. Both studies distinguish gains from losses by using a phylogenetic tree. Both rely on parsimony to argue that a particular result can be regarded as evidence of a gain (or of a loss, as the case may be). In both studies, a proportion of TFBSs cannot be distinguished as being a gain or a loss, and so these TFBSs have to be left out of the analysis. Both studies use experimental evidence to locate TFBSs in one species, but use bioinformatic methods to obtain TFBS information for the other species. Both examine how false (or non-functional) TFBSs will affect the analysis, and produce numerical estimates of results that could be produced by false/non-functional TFBSs.

However, there are considerable differences in the use made of estimates from the false/non-functional TFBS analysis. The paper cited (Moses et al., 2006) used the estimated rates of false/non-functional TFBSs as corrections, used to adjust the rates based on the experimental TFBSs. In contrast, the methodology described in the current chapter of this thesis did not do this - instead, the fictional TFBS analysis was used as evidence whether or not the experimental TFBSs were heavily contaminated by incorrect data. For the TransFac-based analysis, the results based on experimental data were different from the results based on fictional data (as detailed in Table 5.3) - which is evidence that the experimental data was not dominated by any incorrect data within it. However, for the analysis of chip-chip data, the experimental data gave a ratio of gains to losses that was almost the same as the ratio obtained from fictional data, so (in this chapter) it was thought that no clear

conclusion could be drawn from the chip-chip data.

In contrast, in the paper (Moses et al., 2006), although their analysis of chip-chip data was also affected by many false/non-functional TFBSs, they adopted a different procedure, and corrected their estimates by "subtracting out" the estimated effects of non-functional TFBSs. In principle this is valid, though it does require a very accurate estimate of the rate at which false/non-functional TFBSs are encountered.

To illustrate the need for accuracy, suppose their estimate of false/non-functional gains is a slight underestimate, and that the true figure is 634 gains per 425000bp, rather than the 602 gains per 425000bp they gave in their paper. Redoing their calculations, their estimate of 41.7 gains would be reduced to 24.7 gains. Thus in this example, if their non-functional rate was underestimated by just 5%, the number of functional gains will have been overestimated by 69%. It is not suggested that such an error has actually occurred. It merely illustrates that, in order to use their method, the false-match rate must be estimated with very great accuracy - even a 5% underestimate will be serious. Producing rates of such high accuracy is not an easy task.

To sum up this comparison, both the paper (Moses et al., 2006) and this chapter try to estimate the ratio of gains to losses, using methods that have much in common. However, the paper is bolder in trying to extract estimates from difficult chip-chip data (by correcting for non-functional sites), rather than abandoning any attempt to draw a conclusion from chip-chip data as this chapter has done. In both cases, any estimate of the ratio of gains to losses is subject to considerable error, so only order-of-magnitude conclusions can be drawn. In both cases, the estimate of losses is reliable enough to be evidence against the idea that "loss of TFBS is extremely rare compared to gain of TFBS".

Other papers are less suitable for comparison. Whereas the work cited above identifies “gains” and “losses” in a single analysis, another paper uses a different approach with two separate analyses (Doniger and Fay, 2007). To find “losses”, they obtained a sample of 19264 TFBSs using bioinformatic meth-

ods, and identified examples of “loss” within that sample. To find “gains”, they obtained a sample of 654 TFBSs that were documented in a database, and identified examples of “gain” within that sample. Thus, the “losses” came from a sample of TFBSs that was quite different to the sample the “gains” came from; and these two samples were obtained by very different methods. This causes extra difficulties in trying to compare the frequency of “gains” with the frequency of “losses”. The authors of the paper made no attempt to produce an empirical estimate of whether “gains” are more frequent than “losses” or not. Thus, for comparing with this chapter of the thesis, their work is not as suitable as the Moses paper was.

5.4.3 Fictional TFBSs

Table 5.2 showed that, when fictional TFBSs were analysed, "gain trees" were found more frequently than "loss trees".

Intuitively, this is not surprising; for example, a TFBS with the tree shown in Figure 5.2(a) will be classed as a gain of TFBS - whereas a TFBS with the tree shown in Figure 5.2(b) will be classed as a loss of TFBS. Which of these two trees is a fictional TFBS more likely to produce? The "loss" tree is relatively difficult to produce by chance, since it requires a pattern of mutations that disrupts the TFBS in rats yet also conserves the TFBS in a human-mouse comparison. This suggests the "gain" tree is more likely to arise. Table 5.2 confirms this intuition, and quantifies the size of the effect.

In the literature, a similar result exists for *Drosophilia* (Moses et al., 2006). This paper included results for DNA sites that match the PWM for the transcription factor “Zeste”, yet were not bound by “Zeste” in a chip-chip experiment, and therefore may be spurious sites rather than functional TFBSs. 78 “losses” and 602 “gains” were observed. So this again suggests that spurious TFBSs are more likely to be classified as “gains” than as “losses”. It will be noted that their method of generating spurious “Zeste” sites is considerably different from the method of generating fictional-TFBSs used in this thesis, yet both types of data produce more “gains” than “losses”.

Thus the intuitive explanation just given, the results in this chapter (table 5.2), and a result from the literature (Moses et al., 2006) all agree in suggesting that mistaken TFBS data is more likely to appear to be a "gain" than to appear to be a "loss".

This suggests that studies based on thoroughly unreliable TFBS data are likely to give the conclusion that "gains" are more frequent than "losses", even if the truth is quite different.

5.4.4 Conservation of neighbourhood DNA

There was tentative evidence for "losses" being flanked by DNA that is more highly conserved than the DNA flanking "gains" (table 5.4). That is a potentially interesting result, but given the weak significance level (p=0.05), a more reliable result is desirable, which might be obtained when more data becomes available.

As discussed in the previous chapter, such a difference was predicted using the "correlated-evolution" model, but no such difference was predicted using the "site-density" model. Thus, the result obtained is a reason for thinking tentatively that the "correlated-evolution" model is more realistic than the "site-density" model.

5.4.5 Discovery bias

The TransFac-based data may be biased towards containing "lost" TFBSs rather than "gained" ones, simply because a TFBS is more likely to be discovered if it is present in many species. For instance, consider a TFBS, upstream of a particular gene, and that TFBS is present in mouse, but not in human nor in rat - which suggests it is a recently gained TFBS. If the researchers who examine how this gene is regulated choose to experiment with a human cell line, this TFBS will not be discovered, and so will not appear in the TransFac database; it will only be discovered if they choose to experiment with mouse cells. In contrast, if the TFBS is present in human, mouse but

not rat - which suggests a recently lost TFBS - then it will be discovered if the researchers experiment with human cells, or if they use mouse cells. This suggests that "lost" TFBSs are more likely to be discovered than "gained" TFBSs.

How large would this discovery bias be, given that we only used TFBSs that had been discovered in human, mouse or rat? Assuming that: any research group was equally likely to pick one of these three species: and never experimented with more than one species; that "gain" sites were only present in one species; that "lost" sites were present in all three; then, clearly, a "lost" site would be 3 times more likely to be discovered than a "gain" site. But all these assumptions are unrealistically simple and replacing them with more realistic assumptions would reduce the factor of 3 (except possibly the assumption about "gains" being present in one species). Therefore, it seems sensible to regard the factor of 3 as an upper limit, and the true size of the bias is likely to be much smaller.

There could also be bias caused by the combination of comparison species used. As an extreme example, if known TFBSs were only available for human and the only comparison species were mouse and rat, then no gains could be detected, only losses.

5.4.6 Gains or losses of TFBSs not caused by mutations

The methodology used here was designed only to detect gains and losses of TFBSs caused by point mutations, or small indels. The statistics presented here should be understood as referring to these kinds of events only. For instance, the duplication of an entire regulatory region will have increased the number of TFBSs in the genome, but those "gains" will not have been included in this study.

5.4.7 Mistaken TFBSs from chip-chip data

It is plausible to suppose that the "gains" from the real chip-chip data are heavily contaminated with mistaken TFBSs, as this data resembles the fictional data in the gain:loss ratio.

An aspect of mistaken TFBSs is the particular vulnerability of "gains" to this problem. Intuitively, a mistaken TFBS is less likely to appear to be conserved than a true TFBS, since the former will not be preserved by natural selection. But, in addition to this, tables 5.2 and 5.6 show that a mistaken TFBS is much more likely to appear to be a "gained" TFBS than a "lost" TFBS. Hence, if a collection of real TFBS contains (say) 20% that are mistakes, a collection of "gained" TFBSs produced from that sample will contain far *more* than 20% of TFBSs that are mistakes. It was noted in the chapter on chip-chip data that, within the chip-chip data, about 20-30% of TFBSs will have been incorrectly located.

The TransFac-based data appears to be more reliable; the evidence for this is that the real-TFBS results differ from the fictional-TFBS results (table 5.3).

5.4.8 A methodological issue: should losses and gains have equal weight?

The program did contain a provision for giving different weights to "gain" events than to "loss" events, but this was not used in the work reported here. To understand why different weights might be desirable, consider the tree in Figure 5.2(a). This could be produced by a single recent gain of TFBS in the human lineage; but it could also be an ancient, pre-vertebrate TFBS that was lost twice, once in the rodent lineage and once in the rhesus lineage. The "single gain" explanation is the most parsimonious one, and would be the one used in the work reported here. However, if losses were very much more frequent than gains, then it could be argued that the “double-loss” explanation is more plausible than the “single-gain” explanation; and the program could be made to select the double-loss explanation rather than the

single-gain explanation, by increasing the weight given to each gain event. This implies that the most parsimonious explanation will not always be the correct one.

As a preliminary examination of whether this is likely to be a serious problem in practice, a simple though quantitative analysis is presented in Appendix C.

5.4.9 Other problems with parsimony

Another problem occurs because there may be a long delay between the time when a TFBS loses its usefulness, and the time when the TFBS loses its ability to bind. This delay might be tens of millions of years (see page 167). Imagine for example a TFBS that performed a useful function in ancient mammals, but that just before the splitting of the human and rodent lineages, this TFBS ceased to have any effect on fitness (either positive or negative). Suppose that the TFBS survives into modern humans, merely because it did not happen to accumulate enough mutations to disrupt binding; whilst the TFBS is destroyed by mutations in both the rhesus and rodent lineages. On examining the modern species human, rhesus, and mouse, it would *appear* (using parsimony) that the TFBS was a "gain" in the human lineage, when in fact it was a "double-loss" in the rhesus and rodent lineages. Moreover, these two "losses" would not really be independent events, since both were caused by a single event - which was the TFBS losing its usefulness before the human-rodent split.

In a wider context, in the field of constructing phylogenies it has also been noted that parsimony, though a popular method, can sometimes lead to the wrong conclusions ((Felsenstein, 2004), especially chapter 9).

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)