would like, on a scale of 1 to 5. 5 would produce a wordy output, suitable for someone who was not experienced at using the software and who was analysing a single TFBS; whereas 1 would produce a condensed output largely consisting of a sequence of numbers, which could easily be converted to a table by importing into a spreadsheet.

2.2.7 Generating a sample of fictional TFBSs

This project used data that ultimately came from a vast number of scientific papers by many different groups, which makes it very difficult to be sure that all the data were free from mistakes. Therefore, it was desired to know the effects that would be produced if faulty data was fed into the analysis. This was done by creating a database of "fictional TFBSs".

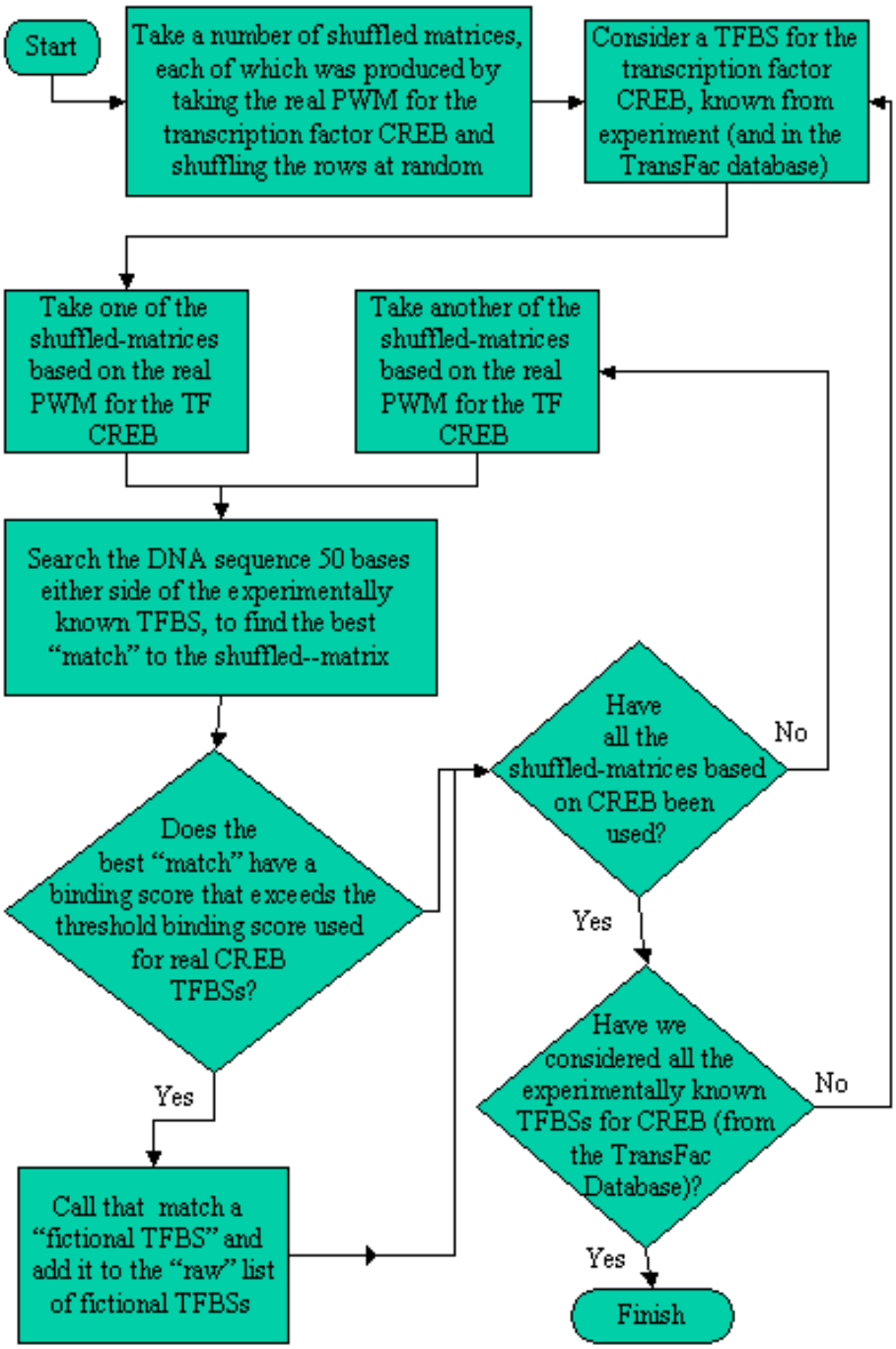
The following example illustrates the method. Starting with a known real TFBS for the transcription factor CREB, a shuffled-matrix based on CREB would be used to search a 100-bp DNA sequence centred on the real TFBS. If a match was found (and the match score exceeded the threshold described earlier on page 80), the matching sequence would be added to the "raw" list of fictional TFBS. Thus, each fictional TFBS was related to a particular transcription factor, and was near a genuine TFBS for that real transcription factor.

This method ensures that fictional TFBSs have a similar location to real ones. For instance, it has been claimed that binding sites for HNF4 are nearly always within 600 bases of the TSS, whereas PPAR sites are sometimes much further upstream (Levitsky et al., 2002). If that were the case for the TFBSs used in this project, then fictional TFBSs based on PPAR would sometimes be more than 600 bases upstream, whereas fictional TFBSs based on HNF-4 would not.

Following this procedure with a single shuffled matrix would produce much fewer fictional TFBSs than was desired. Therefore, a large number of shuffled matrices would be produced from a single real PWM, and used to generate fictional TFBSs in the large numbers required.

Figure 2.8: Flowchart showing how fictional TFBSs were generated

Flowchart showing how fictional TFBSs were generated. As an example, it is fictional TFBSs related to the TF CREB that are generated.



A target was set of producing, on average, five fictional TFBSs for each real TFBS. It was thought desirable to produce more fictional TFBSs than real ones, since otherwise the limited number of fictional TFBSs could become the main limitation on the statistical power of the survey. On the other hand, a ratio of five meant that typically five fictional TFBS were produced from the 100-base sequence around each real TFBS, so the fictional TFBSs would quite often overlap. The 50 bases of sequence flanking one fictional TFBS would be very likely to overlap with the 50 bases flanking another fictional TFBS. It was thought that this overlapping might reduce the statistical power of the sample because the TFBSs were not strictly independent; it was never determined whether this could be a really serious problem or not, but it was thought prudent to avoid a sample so big that most fictional TFBS were overlapping. A ratio of five was the largest that was compatible with this.

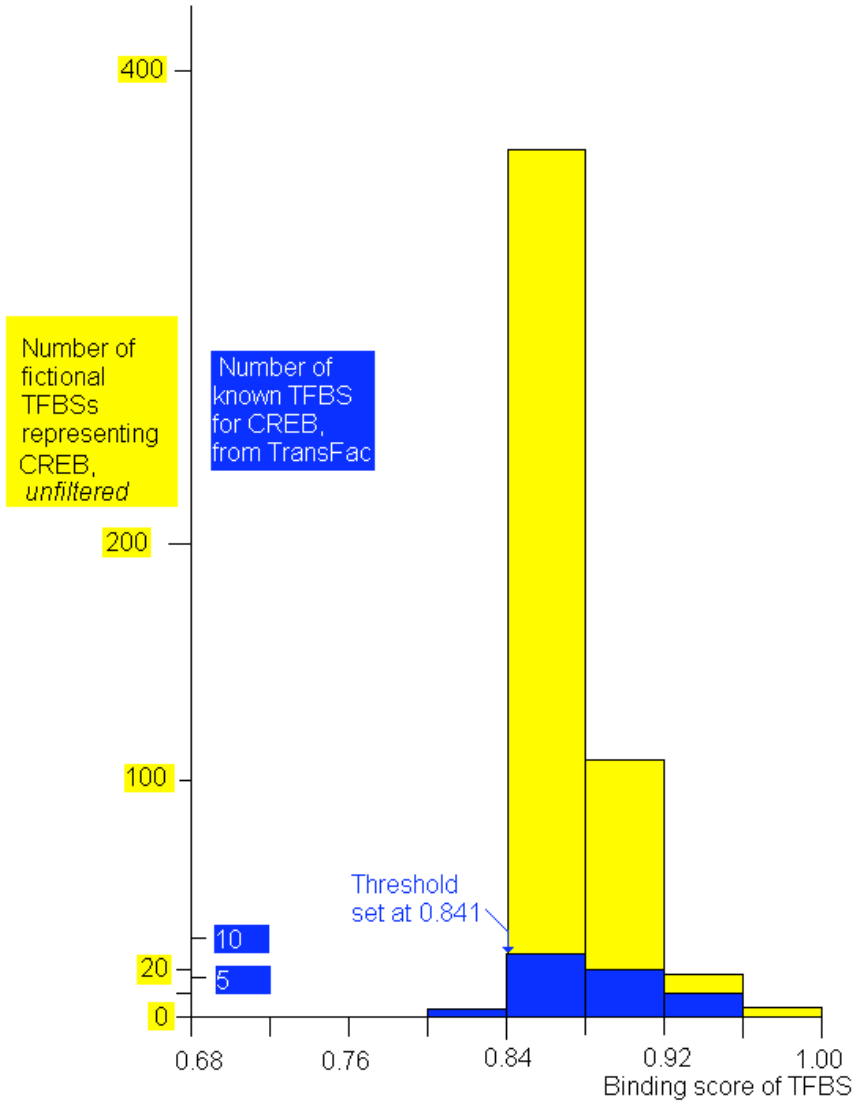
To produce the target number of fictional TFBSs, an initial run was done in which each real PWM would produce - arbitrarily - 50 shuffled matrices. The number of fictional TFBSs produced was noted and, based on this, the number of shuffled matrices was altered before another run took place. This was done individually for each PWM, so the number of shuffled-matrices varied greatly from one TF to another.

The “raw” list of fictional TFBSs produced by this method was considered unsuitable because it contained a high proportion of TFBSs whose binding-score only just exceeded the threshold score, and very few TFBSs whose binding-score was well above the threshold score. In that respect it did not resemble the sample of real TFBSs. An example showing this difference is given in fig 2.9. This undesirable difference was largely eliminated by rejecting many of the "raw" list TFBSs at random, the probability of rejection depending on the score, to produce a "filtered" list of fictional sites. Details of how the software implemented this are on page 231.

Fig 2.10 shows the effect of filtering compared with fig 2.9. Tables 2.10 and 2.11 show the numbers of fictional TFBSs before and after filtering.

A limitation of the filtering was that it was based only on the binding score. For instance, for some PWMs a very high scoring TFBS is much more likely to

Figure 2.9: Fictional CREB TFBSs, unfiltered
This histogram shows unfiltered, fictional TFBSs for CREB (yellow). Notice how TFBSs with relatively low scores (0.84-0.88) are many times more frequent than fictional TFBSs with high scores (0.92-0.96). For real CREB TFBSs (blue), although frequency does vary with score, the *size* of the differences are much



smaller.
Note: because unfiltered fictional TFBSs greatly outnumber real TFBSs, two different scales have been used on the y-axis in order to show both histograms on one figure.

Table 2.10: Numbers of fictional TFBSs - unfiltered
Within each "score bin", this table shows three columns: one containing the number of real TFBSs (R), the next containing the number of *unfiltered* fictional TFBSs (F), and the third column containing the ratio of these numbers (Ra). It can be seen that for TFBSs with a binding score slightly above the threshold score (0 to 1 σ above), fictional TFBSs generally outnumber real TFBSs by an enormous ratio. For very high scoring TFBSs (3 to 4 σ above threshold), the ratio tends to be much smaller.

Binding score →	below thresh -old		0-1 σ above threshold			1-2 σ above threshold			2-3 σ above threshold			3-4 σ above threshold		
	R	F	R	F	Ra	R	F	Ra	R	F	Ra	R	F	Ra
GR	4	0	3	39	13	4	32	8	4	8	2	0	0	NA
NF-Y	0	0	0	101	NA	1	35	35	2	2	1	0	0	NA
NF-AT	0	0	3	89	29.7	3	42	14	5	14	2.8	0	0	NA
HNF-3bet	0	0	1	53	53	2	42	21	1	2	2	0	0	NA
HNF-1	2	0	2	627	314	5	385	77	7	14	2	0	0	NA
AP-2	3	0	3	35	11.7	1	16	16	1	5	5	0	1	NA
STAT	3	0	1	403	403	7	141	20.1	5	13	2.6	0	0	NA
SRF	3	0	1	503	503	5	27	5.4	4	1	0.3	0	0	NA
c-Myb	1	0	2	166	83	1	162	162	2	58	29	2	3	1.5
HNF-3gam	0	0	0	4	NA	1	2	2	0	1	NA	0	0	NA
Crx	0	0	2	146	73	3	77	25.7	1	23	23	1	1	1
PR	0	0	1	13	13	0	7	NA	1	6	6	0	1	NA
SP1	1	0	9	281	31.2	34	401	11.8	37	200	5.4	4	18	4.5
ETS	1	0	1	177	177	3	47	15.7	2	3	1.5	0	0	NA
USF	1	0	0	340	NA	3	112	37.3	1	4	4	1	0	0
Pax	0	0	4	151	37.8	2	162	81	0	5	NA	0	0	NA
p53	1	0	0	350	NA	2	348	174	5	87	17.4	5	3	0.6
YY1	0	0	1	89	89	6	81	13.5	0	5	NA	0	0	NA
CREB	1	0	7	349	49.9	6	110	18.3	4	26	6.5	0	4	NA
C/EBP	1	0	6	394	65.7	17	316	18.6	17	95	5.6	5	7	1.4
POU1F1	1	0	4	31	7.8	2	11	5.5	1	3	3	0	1	NA
HNF-3alph	1	0	3	133	44.3	2	50	25	1	11	11	2	2	1
TTF-1	0	0	2	303	152	6	400	66.7	7	129	18.4	3	7	2.3
NF-1	1	0	1	756	756	5	184	36.8	5	12	2.4	0	0	NA
NF-kappaB	1	0	4	450	113	5	67	13.4	9	9	1	1	5	5

Based on run I

Table 2.11: Numbers of fictional TFBSs - filtered

This Table has the same format as Table 2.10, except that the F columns now show the number of fictional TFBSs that remain *after* the filtering process. The aim was to produce 5 fictional TFBSs for every real TFBS, and this was broadly achieved. However, in some cases there were too few fictional TFBSs, particularly for very high-scoring TFBSs (2 to 4 σ above threshold). The way that random numbers were used to select TFBSs during the filtering means that it is very rare for the ratio to be exactly 5.

Binding score →	below thresh -old		0-1 σ above threshold			1-2 σ above threshold			2-3 σ above threshold			3-4 σ above threshold		
	R	F	R	F	Ra	R	F	Ra	R	F	Ra	R	F	Ra
GR	4	0	3	18	6	4	17	4.3	4	8	2	0	0	NA
NF-Y	0	0	0	0	NA	1	4	4	2	2	1	0	0	NA
NF-AT	0	0	3	11	3.7	3	19	6.3	5	14	2.8	0	0	NA
HNF-3bet	0	0	1	5	5	2	7	3.5	1	2	2	0	0	NA
HNF-1	2	0	2	12	6	5	30	6	7	14	2	0	0	NA
AP-2	3	0	3	19	6.3	1	1	1	1	5	5	0	0	NA
STAT	3	0	1	5	5	7	37	5.3	5	13	2.6	0	0	NA
SRF	3	0	1	6	6	5	24	4.8	4	1	0.3	0	0	NA
c-Myb	1	0	2	8	4	1	3	3	2	10	5	2	3	1.5
HNF-3gam	0	0	0	0	NA	1	2	2	0	0	NA	0	0	NA
Crx	0	0	2	12	6	3	18	6	1	3	3	1	1	1
PR	0	0	1	7	7	0	0	NA	1	6	6	0	0	NA
SP1	1	0	9	40	4.4	34	168	4.9	37	187	5.1	4	18	4.5
ETS	1	0	1	7	7	3	15	5	2	3	1.5	0	0	NA
USF	1	0	0	0	NA	3	14	4.7	1	4	4	1	0	0
Pax	0	0	4	19	4.8	2	6	3	0	0	NA	0	0	NA
p53	1	0	0	0	NA	2	8	4	5	27	5.4	5	3	0.6
YY1	0	0	1	6	6	6	29	4.8	0	0	NA	0	0	NA
CREB	1	0	7	38	5.4	6	34	5.7	4	20	5	0	0	NA
C/EBP	1	0	6	29	4.8	17	92	5.4	17	88	5.2	5	7	1.4
POU1F1	1	0	4	17	4.3	2	10	5	1	3	3	0	0	NA
HNF-3alph	1	0	3	18	6	2	16	8	1	7	7	2	2	1
TTF-1	0	0	2	7	3.5	6	29	4.8	7	37	5.3	3	7	2.3
NF-1	1	0	1	3	3	5	21	4.2	5	12	2.4	0	0	NA
NF-kappaB	1	0	4	17	4.3	5	25	5	9	9	1	1	5	5

Based on run I

be found in sequence with low GC content, which could lead to an undesirable bias in which fictional TFBSs tend to be located in such sequences, but the filtering system did not correct for this.

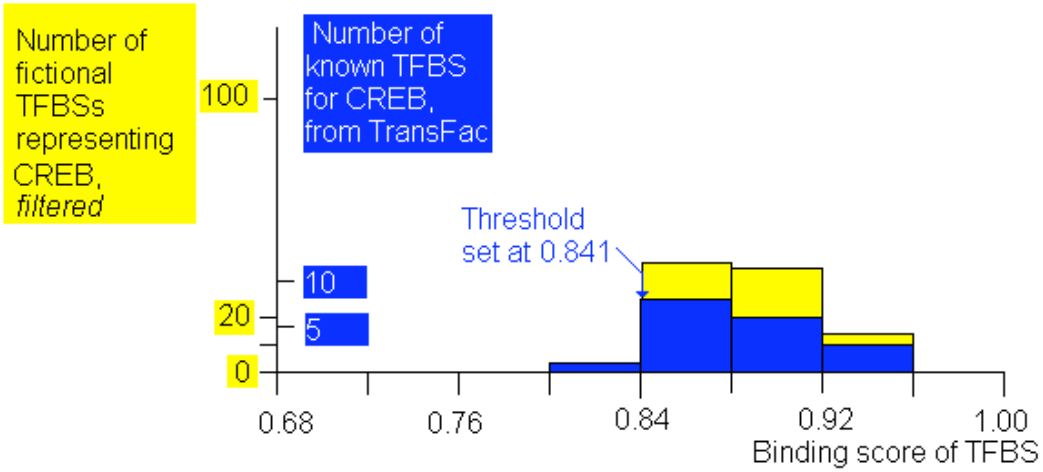
2.2.8 Using the samples of conserved and non-conserved TFBSs

Basic approach

A number of possible measures were considered that might differ between conserved and diverged TFBSs, such as the length of aligned region or the location of the TFBS relative to the TSS. For each measure, the average value for the conserved TFBSs was compared with the average value for the diverged TFBSs.

The errors for such a comparison were estimated by standard statistical mea-

Figure 2.10: Fictional CREB TFBSs, filtered
This histogram shows *filtered*, fictional TFBSs for CREB (yellow), and real CREB TRBSs (blue). Notice how the distribution of fictional TFBSs is much more similar to the distribution of real TFBSs than it was in figure 2.9.



Note: to match fig 2.9, two different scales have been used on the y-axis.

sures, as shown in this box:

Let, for some measure x, the standard deviation of it be σ_c for conserved TFBSs, and the number of such TFBSs in the sample be N_c;
and let the corresponding values for diverged TFBSs be σ_d and N_d;
let x_{diff} be the difference between the average value of x for conserved TFBSs and the average value of x for diverged TFBSs;
then,
The average value of x for conserved TFBSs was subject to an error $\sigma_c/\sqrt{N_c}$
The error of x_{diff} was subject to an error of $\sqrt{\sigma_c^2/N_c + \sigma_d^2/N_d}$
x_{diff} would have to be at least twice as large as its error to be regarded as a "significant difference".

A difficulty was that a single known TFBS might contribute several items to the sample - for example if a known TFBS in human was conserved in rhesus and in mouse it would contribute two items to the sample of conserved TFBSs. This caused two problems; the problems, and methods of addressing them, will be explained in the next two sub-subsections. The two methods are quite distinct conceptually.

"Species-stratified" analysis"

One problem was that the samples of conserved and diverged TFBSs contained different combinations of species-pairs. For instance, human-mouse comparisons produced a "diverged" result far more often than human-rhesus comparisons did, and hence the proportion of human-rhesus pairs within the conserved sample would be different from that in the diverged sample. Whilst, in principle, this could generate misleading results, the analysis of fictional TFBSs would also be affected and thereby draw attention to the problem. Table 2.12 explains the problem using a simple example.

To avoid this problem, a "species-stratified" analysis was devised: conserved TFBSs for a particular pair of species were compared with diverged TFBSs *for the same pair of species*. The resulting differences were then averaged over all possible species-pairs. A simple average was considered unsuitable, since

Table 2.12: Simple example illustrating why species-stratification is desirable

An imaginary example will illustrate why species-stratification is desirable.

Suppose that, in a human-rhesus comparison, DNA surrounding any TFBS was always found to be 90% identical, irrespective of whether the TFBS was conserved or not. Suppose also that 1 in 10 human TFBSs were diverged in rhesus.

Suppose also that, in a human-mouse comparison, DNA surrounding any TFBS was always found to be 60% identical, irrespective of whether the TFBS was conserved or not. Suppose also that 1 in 3 human TFBSs were diverged in mouse.

If we have 100 TFBSs in each of these two types, the data can be summarised as:

A

	Conserved TFBSs		Diverged TFBSs		
Species being compared	Number of TFBSs	%identity of nearby DNA	Number of TFBSs	%identity of nearby DNA	Difference in %id
Human-rhesus	90	90%	10	90%	0%
Human-mouse	67	60%	33	60%	0%

However, if these two data sets are combined, this table becomes:

B

Combined data	157	77%	43	67%	10%

In this last table, notice how the diverged TFBSs are overwhelmingly from the human-mouse comparison, whereas conserved TFBSs are not. From that, it will be evident how the combined data gives a difference of 10% in the right hand column - even though 0% difference was obtained when each species was considered separately.

Clearly the 10% difference, from the combined data, can be misleading. This imaginary example illustrates the dangers of using data from a number of species all mixed together. This danger can be avoided by breaking the data down into individual species-pairs and analysing them separately, as in table *A* above; this is referred to as "species-stratified analysis".

some species-pairs had a much larger sample size than others; a weighted average was preferred, in which the weight was the minimum sample size (that is, the number of conserved TFBSs or the number of diverged TFBSs, whichever was smaller).

In some cases, the real-TFBS difference for a pair of species had substracted from it the fictional-TFBS difference for the same pair of species; when these figures were averaged over a number of different pairs of species, the weight was based on the real-TFBS data only - the grounds for this being that fictional-TFBS data was usually based on a much larger sample than real-TFBS data.

Bootstrapping: the "unit of selection"

The second problem was that standard statistical methods assume each item in the sample is independent, but that is unlikely to be the case if more than one item derives from a single known TFBS. For example, if a conserved TFBS is in a genomic region which has an unusually low mutation rate, the human-rhesus comparision and the human-mouse comparison will each contribute to the sample of conserved TFBSs, and each will have a low mutation rate associated with it, but they are not two independent estimates. The effect of this will be to make the estimated error smaller than the true error, unless special methods are used to allow for this.

This problem was dealt with by bootstrapping. The bootstrapping technique, introduced by Efron, is described by him as a technique "for assessing the accuracy of almost any statistical estimate. It is particularly useful in complicated nonparametric estimatation problems, where analytical methods are impractical" (Efron et al., 1996). He notes that one example of its use is in assessing the reliability of phylogenetic trees, an application of it that is particularly well-known to bioinformaticians; however, that particular application will not be used in this thesis.

The bootstrap method can be summarised as follows. A main step is to produce a "bootstrap sample" of items, which is similar to an ordinary sample

of items. Each item in a bootstrap sample is selected from a real sample at random; thus a particular item might appear in the bootstrap sample once, or more than once, or not at all. For example, the real sample of TFBSs included a CREB-binding site that was 70 bases upstream of the ADRB2 gene; in the real sample of TFBSs, that particular TFBS was, of course, included once; but in a boostrap sample of TFBSs, that particular TFBS might be included twice, or it might be included once, or it might not be included at all. A number of bootstrap samples are taken, each sample is used to produce an average, and the standard deviation of these is an estimate of the sampling error. For example, suppose a real sample of TFBSs had a GC content of 41.8% on average - a bootstrap sample of those TFBSs might have an average GC content of 40.5%, but another bootstrap sample of those TFBSs might have an average GC content of 42.9%, and another bootstrap sample might have an average GC content of 41.2%, etc; this variation from one bootstrap sample to the next suggests the amount of error likely to be present.

An important point for this project was that the unit of selection was the known TFBS, rather than a comparison of that TFBS in two species. For example, if a known human TFBS was selected to go in a bootstrap sample, then *all* the comparisons involving that TFBS would be included (human-rhesus, human-mouse, etc), potentially increasing the number of conserved TFBSs by several additional members. This could result in a larger estimate of error than if the unit of selection was each species-pair comparison (an example of the latter would be if the human-rhesus comparison for a TFBS was included in a bootstrap sample when the human-mouse comparison for the *same* TFBS was not included).

The species-stratified analysis and the bootstrapping were not used in the preliminary analysis, but used for a more detailed examination of the most interesting results from the preliminary analysis.

2.3 RESULTS

2.3.1 Problems using TransFac co-ordinates

Following the procedure described above would produce an estimate of the location of a TFBS. This estimate would not always be the same as the location given in the TransFac database. These two estimates are compared in the scatterplot shown in fig 2.11. (This figure omits TFBSs more than 250 bases upstream of their gene, in order to keep the scale sensible).

From the scatterplot, it can been seen that there are a considerable number of TFBSs where the two estimates differ by more than 50 bases. Such a discrepancy is not necessarily because the TransFac estimate is incorrect. For instance, if a gene has more than one TSS, that could cause discrepancies in locations measured relative to the TSS.

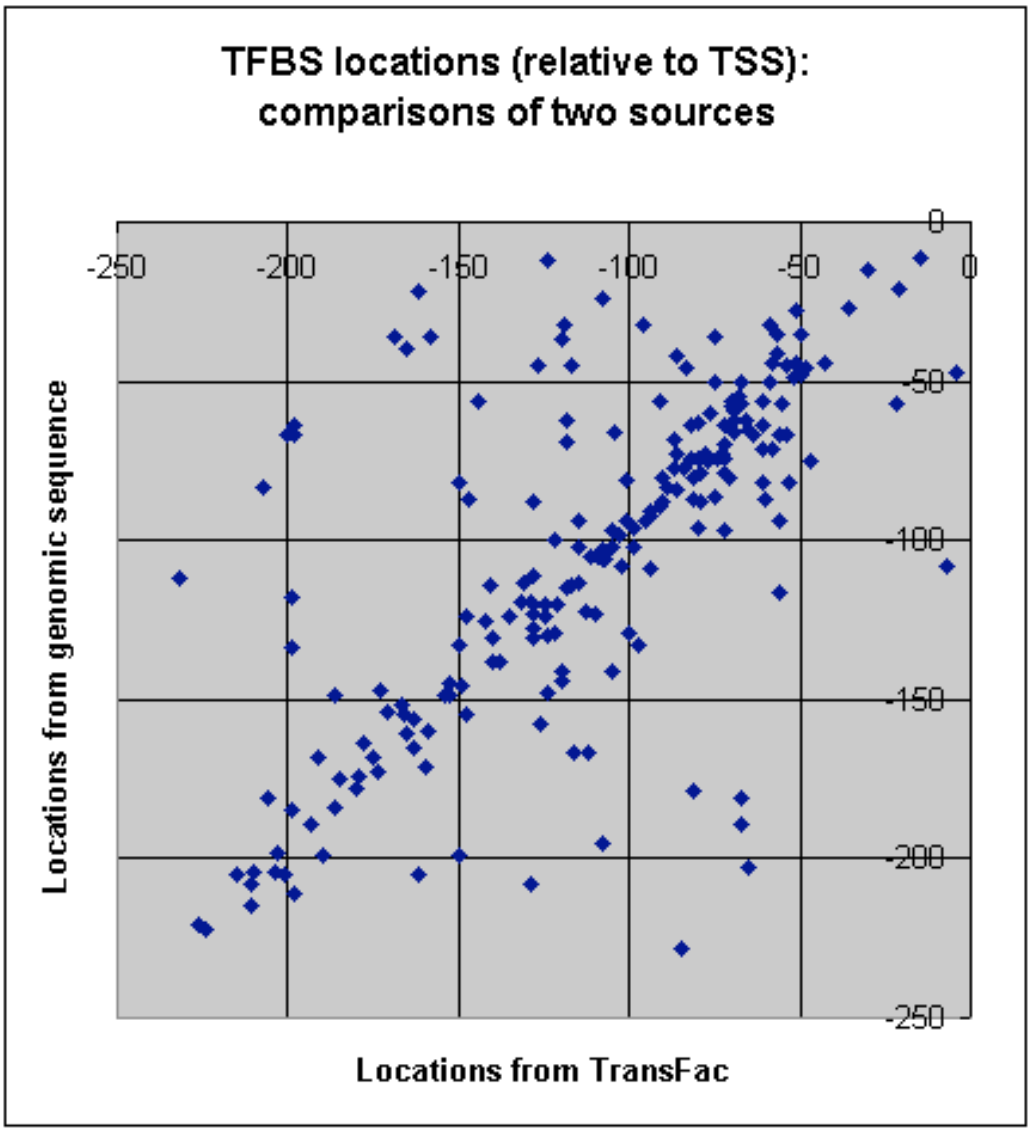
The scatterplot does emphasise that, if the TransFac coordinates alone were used to locate TFBSs in genomic sequence, then the locations would frequently be in error - and the size of this error would often exceed the size of the TFBS. Hence it is important to use the TFBS *sequence* given in TransFac when determining the TFBS's exact location in genomic sequence.

2.3.2 Counting the proportion of TFBSs that are not conserved

Figure 2.12 shows the proportion of human TFBSs that are conserved, or not, in comparisons with each of the other seven species. They are classified into conserved/uncertain/diverged according to the conclusion reached at the end of the flowchart in figure 2.7.

A striking feature is the large proportion of cases that are "uncertain". This leads to much ambiguity in interpretting the graph. It is possible that the vast majority of "uncertain" cases are, in fact, diverged TFBSs, which should be added to the diverged TFBSs actually shown. Alternatively, it is possible

Figure 2.11: Locations of TFBS: two estimates compared
The x-axis shows where the upstream end of a TFBS is located relative to the gene it controls, as shown in the TransFac database. The y-axis shows the same, except that it is estimated in a different manner, using the locations determined from the procedure described earlier (see fig 2.3), relative to the start of the gene (determined using RefSeq sequence).



that the vast majority of "uncertain" cases are, in fact, conserved TFBSs (in which case the percentages of diverged TFBSs in the figure might be fairly close to the true figure). Consequently, for a human-mouse comparison, the true percentage of diverged TFBSs could be between 10% and 57%.

That is a large range of uncertainty. One might wonder whether it could be reduced; some improvement is presumably possible, but eliminating the uncertainty would be extremely difficult. For instance, one group found that most regulatory regions could not be aligned in a comparsion of humans with marsupials or more distant vertebrates (King et al., 2007). That problem alone means that more than 50% of TFBS comparisons with chicken will be placed in the "uncertain" category (since, using the method described in this thesis, unaligned TFBSs are put in the "uncertain" category). Because alignment programs sometimes fail to detect orthologous sequences (Margulies et al., 2007), it is very difficult to determine the true orthology of an unaligned TFBS.

These results could be compared with a somewhat similar survey, which was also an *in silico* examination of TransFac TFBSs (Sauer et al., 2006). This found 72% of TFBSs conserved in a human-rodent comparison. They did not have an "uncertain" category, and so did not obtain a range-of-uncertainty estimate which could be compared to the one given above (although they did consider alternative ways of treating unaligned TFBSs, and thereby obtained an alternative estimate of 60% of TFBSs being conserved).

Other estimates, for TFBSs in a human-mouse comparison, were covered in more detail in the literature review earlier, and can be summarised as: 32-40% of TFBSs non-conserved (Dermitzakis and Clark, 2002); 60-80% of Estrogen Receptor TFBSs non-conserved (O'Lone et al., 2004); 41-89% of regulatory links not conserved, plus the TFBS itself being non-conserved in two-thirds of cases where the regulatory link *was* conserved (implying that much more than 41-89% of TFBSs are non-conserved) (Odom et al., 2007). So the estimates in the literature cover a large range. Even the wide 10%-57% estimate given above does not completely cover all the estimates in the literature. However, the minimum estimate of 10% of TFBSs being non-

conserved does look over-cautious, since it is so much lower than all the estimates from the literature just quoted.

Figure 2.13 shows similar information, but subdivided by the molecular structure of the TF. To get sufficient data, comparisons relating to all combinations of species have been combined, and TFBSs known experimentally to exist in human or mouse or rat were used. The data on molecular structures came from the classification system used by TransFac. This classification system could be used to divide the data into a very large number of detailed categories (for example, category 1 "basic domain" could be subdivided into "1.1 - leucine zippers", "1.2 - basic HLH", etc), but given the amount of data available, for the present purpose it was thought unwise to break the data down into a large number of categories.

Examining figure 2.13, it is difficult to prove that there is any difference between the categories, because of the large number of "uncertain" cases. For example, if it was suggested that exactly 60% of TFBSs comparisons are "conserved" (independent of TF structure), then figure 2.13 does not contain evidence to disprove this, because 60% is always between the minimum estimate and the maximum estimate of the percentage of conserved cases (since the maximum estimate is based on the numbers of "conserved" cases plus "uncertain" cases).

Thus, because of the large range of uncertainty implied in figure 2.12, it was doubted that it would be a fruitful line of research to persue this further by breaking the data down into other categories (although figure 2.13 shows an attempt at this). Thus, at a fairly early stage in the project, attention was turned to a different line of research. This was to compare the diverged TFBSs against the conserved TFBSs (ignoring the "uncertain" cases), looking for any characteristics that differed.

2.3.3 Initial examination of several measures

Tables in this section show some properties that were examined to see if they differed between conserved and diverged TFBSs. The results shown are

Figure 2.12: Proportion of TFBSs that are not conserved
This shows, for TFBSs known to exist in humans, what percentage were conserved, what percentage were diverged, and what percentage were placed in the "uncertain" category because the evidence was not clear enough. The seven comparison species are shown in order of mutational distance from humans (estimated by divergences at fourfold synonymous sites, as detailed in Appendix A). This chart only includes cases where the human TFBS was successfully located in the human genome, and orthologous genome sequence from the comparison species was successfully retrieved. (Therefore, the "uncertain" percentages do *not* include cases where the human TFBS could not be located in the human genome). For run Q0.

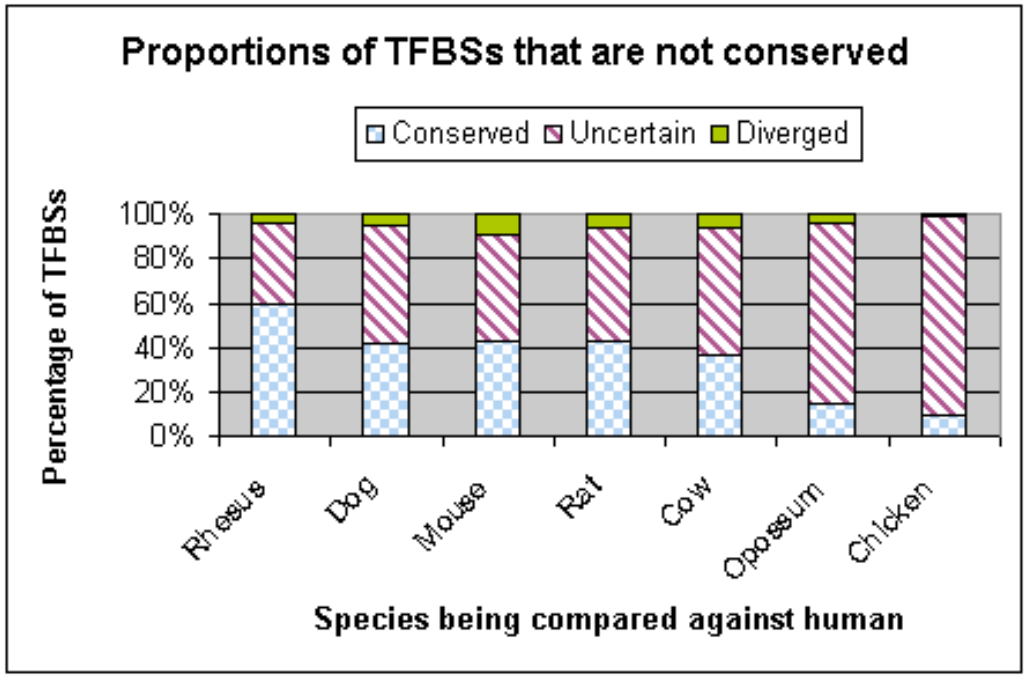
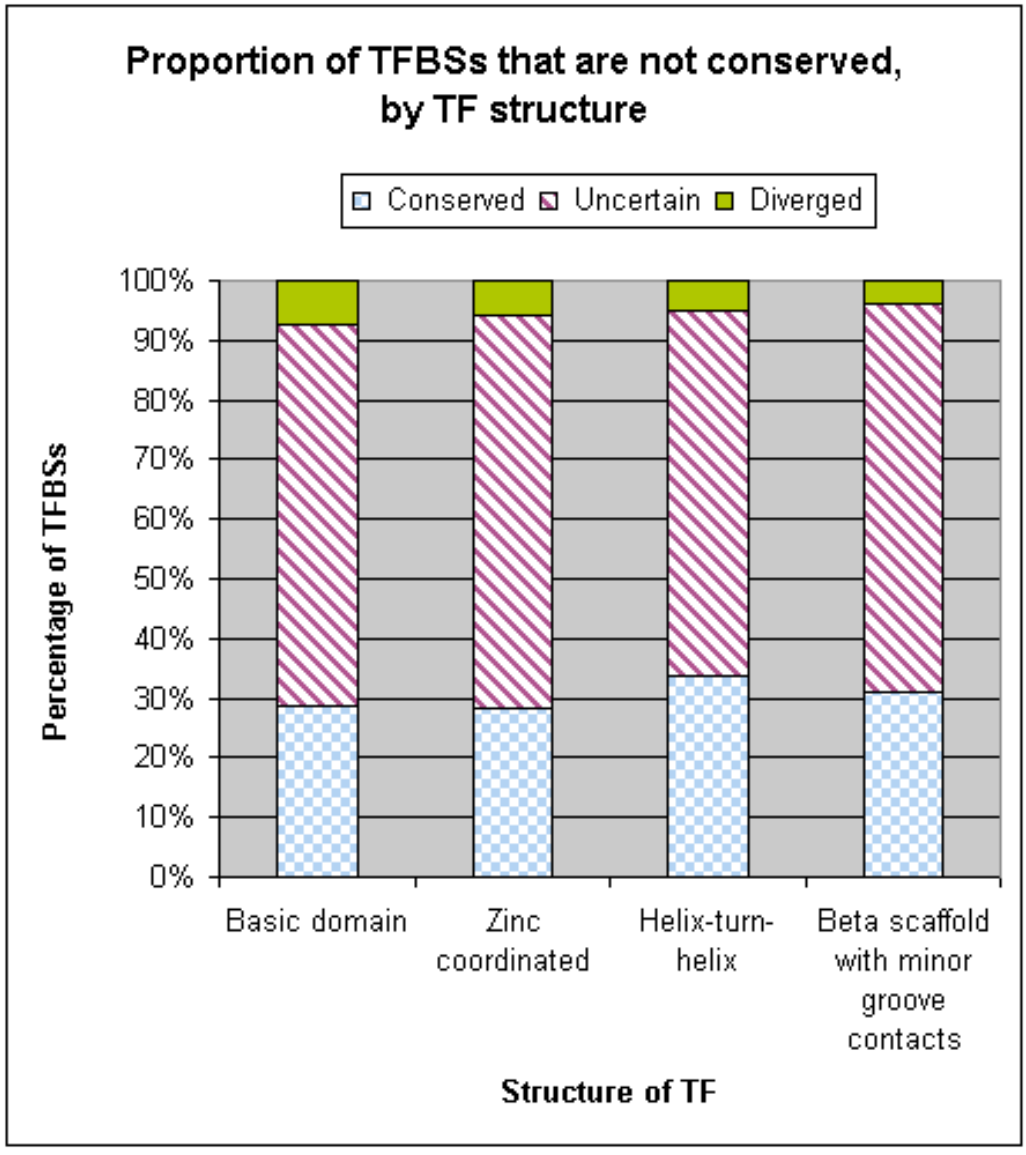


Figure 2.13: Proportion of TFBSs that are not conserved, by structure This shows, for TFBSs known to exist one species, what percentage were conserved, what percentage were diverged, and what percentage were placed in the "uncertain" category because the evidence was not clear enough. Each TFBSs has placed in into one of four categories, depending on the molecular structure of the TF that binds it, using classification data from the TransFac database. For run Q0.



from "Run F". This was an early analysis which only included human, mouse, rat and dog, and contained imperfections that were addressed later in the project.

For 802 experimentally-known TFBSs, the analysis was abandoned before the TFBS could be compared with any other species. Table 2.13 show the most frequent reasons why this happened.

Consequently, there were only 279 experimentally-known TFBSs for which a comparison was carried out with at least one other species. The results of these comparisons are shown in table 2.14. Often a single known TFBSs could be compared with more than one other species, which is why the numbers shown in table 2.14 add up to more than 279. For instance, a known human TFBS might produce a human-dog comparison in the "diverged" category, a human-mouse comparison in the "conserved" category, and a human-rat comparison also in the "conserved" category; consequently, that one TFBS would add 3 to the total of table 2.14.

Table 2.13: Most common reasons for abandoning analysis of a TFBS before it was compared with any other species (Run F)

Problem that caused analysis to be abandoned	Number of known TFBSs
"Experimental site seq not found in genomic sequence"	270
No good match to the PWM found	163
"Failed to identify homologous genes"	135
"Experimental binding site ... is too small"	89
"Unclear which gene referred to by ..."	62
"Failed to retrieve genome sequence ..."	40

Table 2.14: Number of TFBS comparisons (in Run F)

Conclusion of comparison	Number of comparisons
Conserved	244
Diverged	43
An "Unclear" or "Mixed evidence" conclusion	357

Thus, 244 examples of TFBS conservation had been identified; and so had

43 examples of TFBS divergence (where a TFBS in one species did not have an orthologue in another). These were used in the next stage of research, a number of different measures were considered - for instance, GC content of the sequence around the TFBS, and location of the TFBS relative to the TSS. The average value of these for TFBSs that had diverged was calculated, and so was the average value of these for TFBSs that were conserved. The hope was to identify some measures where these two values differed. The results of this exercise for Run F are summarised in tables 2.15 to 2.17.

For some properties, such as GC content, no significant difference was observed between conserved and non-conserved TFBSs. These are listed in table 2.15. Here "no significant difference" was determined using the two-standard-errors rule mentioned earlier; at this stage, no use was made of species-stratified analysis or bootstrap errors.

Table 2.15: Similarities: Measures for which no significant difference was observed between features associated with conserved TFBSs and those associated with diverged TFBSs

Measure	Notes
Distance of TFBS from TSS	Using the value from the TransFac entry
GC content	Based on the known TFBS plus 50 bases either side
CpG content (relative to GC content)	Based on the known TFBS plus 50 bases either side
Length of protein encoded by regulated gene	
dN (non-synonymous mutation rate) for regulated gene	Not always based on full protein (see Appendix A). Estimated by yn00 (Yang and Nielsen, 2000)
dS (synonymous mutation rate) for regulated gene	As with dN

For a couple of properties, there was a difference between conserved and diverged TFBSs, yet a very similar difference was observed for fictional TFBSs (table 2.16). Clearly, when the difference within fictional TFBSs is almost the same as the difference within real TFBSs, it does reduce confidence that

the differences represent something meaningful. So the measures in table 2.16 were not considered for further work.

Table 2.16: Measures for which a significant difference was observed between conserved and diverged TFBSs, but a similar difference was observed for fictional TFBSs

Measure	Experimental TFBSs			Fictional TFBSs		
	Conserved TFBSs	Diverged TFBSs	Difference	Conserved TFBSs	Diverged TFBSs	Difference
Length of aligned region	2221 ± 117 bases	1724 ± 207 bases	497 bases	2458 ± 93 bases	2027 ± 111 bases	431 bases
%id for aligned region	63.5% ± 0.6%	58.7% ± 1.0%	4.8%	63.6% ± 0.5%	59.6% ± 0.6%	4.0%

There were a few properties where not only was there a difference between conserved and diverged TFBSs, but also, that difference was much larger than anything observed for fictional TFBSs. Table 2.17 shows these.

CpG content was higher around conserved TFBSs, but the difference was only just large enough to be included in this Table.

"Position of TFBS within aligned region" requires some explanation: it was the distance of the TFBS from the start of the aligned region, divided by the length of the aligned region. Thus, for example, it would be 0.5 for a TFBS exactly in the middle of the aligned region; it would be 0 for a TFBS that was as far upstream as it was possible to be whilst still remaining entirely within the aligned region; and it would be 1 for a TFBS that was as far downstream as it was possible to be whilst still remaining entirely within the aligned region. From Table 2.17, it is evident that most TFBSs were towards the downstream end of the aligned region. The key result is that conserved TFBSs tended to be more downstream than diverged TFBSs.

The distance from the edge of the aligned region is a similar concept, except that a TFBS that was nearer the downstream end of the aligned region would have its distance measured from the downstream end; consequently its value could never exceed 0.5. The numbers in table 2.17, therefore, suggest that conserved TFBSs tend to be nearer the edge of the aligned region than

diverged TFBSs do. (In interpretting this, it should be remembered that sequence downstream of the TSS was removed before alignment, which will sometimes force the downstream end of the aligned region to be at the TSS).

The "%id of neighbourhood" shows a difference that is 7 times larger than its error, which is a much greater ratio than for any other property in the table. On the other hand, the fictional TFBSs also show a difference that is similar in direction, though of half the magnitude. Evidently there was the potential for an effect of this sort to be generated spuriously by faulty data. However, as the effect seemed too large to be explained this way, it also seemed likely that the difference represented something real and was worth further investigation.

Table 2.17: Measures for which a significant difference was observed between conserved and diverged TFBSs, and where the difference was much larger than any observed for fictional TFBSs

Measure	Experimental TFBSs			Fictional TFBSs		
	Conserved TFBSs	Diverged TFBSs	Difference	Conserved TFBSs	Diverged TFBSs	Difference
%id of neighbourhood	75.7% ± 0.9%	62.8% ± 1.6% %	12.9%	75.8% ± 0.7%	69.3% ± 1.0%	6.5%
Position of TFBS within aligned region	0.807 ± 0.014	0.68 ± 0.04	0.127	0.890 ± 0.011	0.770 ± 0.017	0.039
Distance of TFBS from edge of aligned region	0.153 ± 0.009	0.225 ± 0.022	-0.072	0.139 ± 0.007	0.167 ± 0.010	-0.028
CpG content of neighbourhood	0.034 ± 0.0024	0.024 ± 0.004	0.010	0.030 ± 0.002	0.031 ± 0.002	-0.001

2.3.4 Further examination of sequence conservation near each TFBS

The "%id of neighbourhood" effect was examined further. Table 2.18 shows a result from a later analysis (run Q0). Unlike the other Tables, this shows only

human-dog comparisons, thus ignoring the other species. Because there are only two species, the analysis is simpler; however, it also means the number of TFBSs is small and so the sampling error is larger.

In Table 2.18, real TFBSs show a difference (12.0%) which implies the same thing as the "%id of neighbourhood" line of Table 2.17. That is, it implies conserved TFBSs are flanked by DNA that tends to be more highly conserved than the DNA flanking diverged TFBSs.

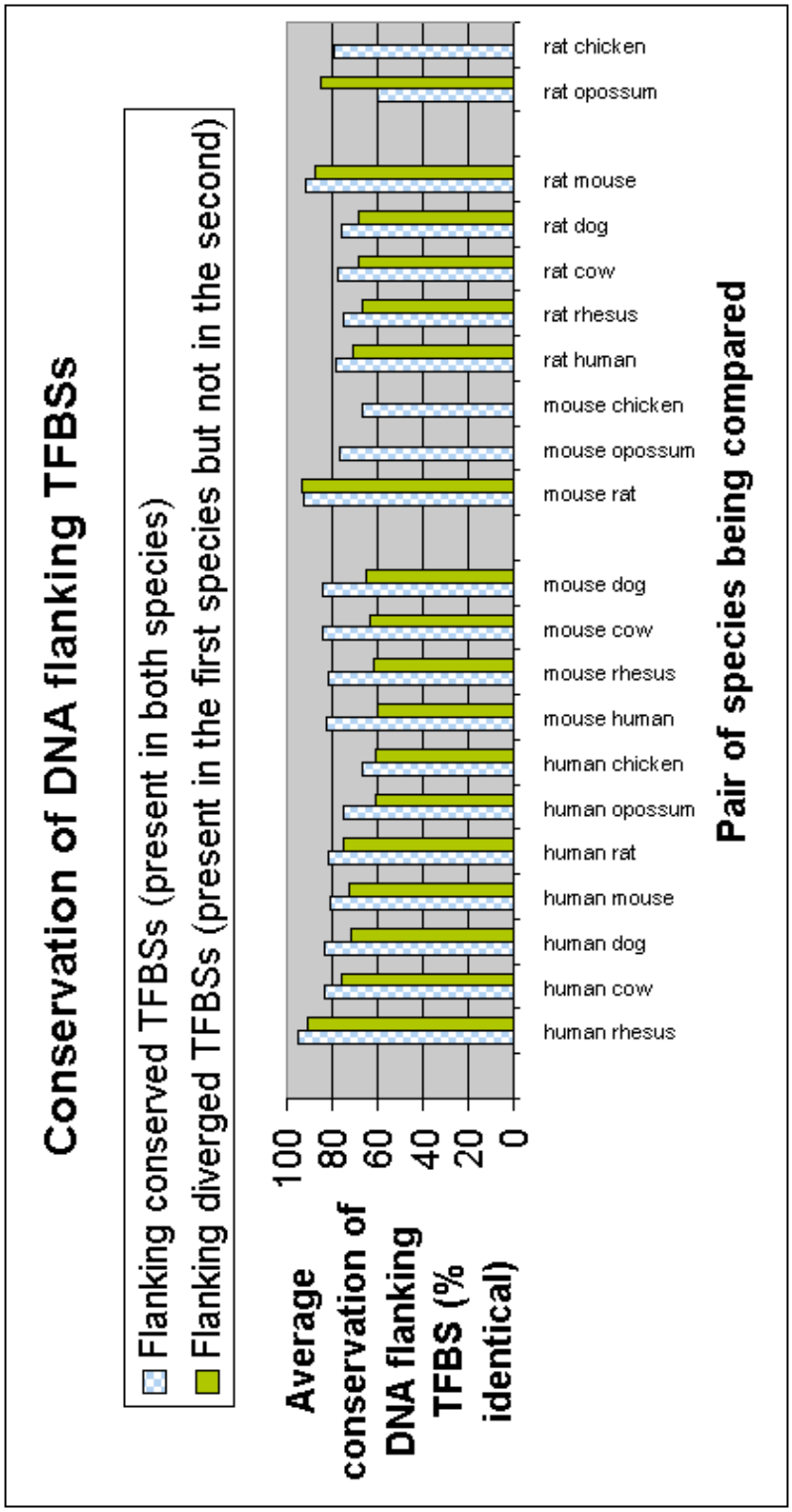
Could the difference shown in Table 2.18 be caused by faulty TFBS data? To address this, Table 2.19 shows a similar analysis of fictional TFBSs, which produces a difference (6.9%) that is similar to the difference (12.0%) shown in Table 2.18. This shows that faulty TFBS data could produce an effect which is *qualitatively* the same as that shown by the real data in Table 2.18; but would the *size* of the effect be the same? It is difficult to answer this using the data in Tables 2.18 and 2.19, because of the large error estimates. To examine this more thoroughly, it is desirable to reduce the error by taking a larger sample. This was done by using data from eight species, although this results in a more complicated analysis.

Figure 2.14 shows that is is fairly consistently found that DNA flanking conserved TFBSs is better conserved than DNA flanking diverged TFBSs. However, it is desirable to know if the differences shown in this figure are larger than would be produced by faulty TFBS data, which requires comparison with fictional TFBS data; such a comparison is shown in Table 2.20.

Results based on eight species are shown in Table 2.20. In this table, the results from Tables 2.18 and 2.19 are included but condensed into the single line marked "human dog". Tables 2.18 and 2.19 might be considered as "worked examples" of how the numbers in Table 2.20 are calculated. Most of the numbers in Table 2.20 are, in effect, intermediate numbers during the calculation of the key result shown at the bottom of the table, but they are given here to help make the details of this sort of "species-stratified" calculation clear.

The main result from Table 2.20 is that the effect for real TFBSs exceeds the

Figure 2.14: For each pair of species, this shows how well conserved is the DNA for 50 bases either side of known TFBSs.



effect from fictional TFBSs, by 5.5% ± 1.4%. This shows that faulty TFBS data is not able to explain the effect observed for real TFBSs, since the faulty data would produce an effect which is smaller than the effect actually observed from real data.

Thus the conclusion from this section is that, for real TFBSs, conserved TFBSs are surrounded by DNA whose sequence is more conserved than the sequence surrounding diverged TFBSs, by 10.8%. This is too large to be explained as being caused by faulty data.

Table 2.18: How strongly had evolution conserved the DNA sequence flanking experimentally found TFBSs, in a human-dog comparison? Divided by whether the TFBS were conserved or not. The conservation is of 50 bases either side of the TFBS (not including the TFBS itself), measured as the percentage of bases identical in a 2-species comparison, ignoring columns with gaps. The error was estimated from 100 bootstrap samples.

Conservation of sequence flanking conserved and diverged TFBSs			
	Mean	Standard deviation	Number in sample
TFBSs that were conserved	83.5 %	8.9 %	83
TFBSs that were diverged	71.5 %	12.1 %	10
Difference	12.0 %	Error of this difference: 4.2 %	

Table 2.19: How strongly had evolution conserved the DNA sequence flanking fictional TFBSs, which were generated to test the methodology? Human-dog comparison; for other details, see table 2.18.

Conservation of sequence flanking fictional TFBSs			
	Mean	Standard deviation	Number in sample
Fictional TFBSs that were conserved	85.3 %	9.0 %	194
Fictional TFBSs that were diverged	78.3 %	10.3 %	145
Difference	6.9 %	Error of this difference: 1.1 %	

Table 2.20: The human-dog result from the bottom of table 2.18 is shown in this table, as well as the corresponding result for each combination of species analysed. The weighted average is shown at the bottom (this was calculated using the minimum sample size as the weight).

Conservation of flanking sequence, for every combination of species				
Pair of species being compared (the first species has the experimentally determined TFBS; the second is the comparison species)	%id of 50 bases either side of TFBS: difference between average for conserved TFBSs and average for diverged TFBSs		The difference between the two columns to the left	Minimum sample size for experimental TFBSs (number of conserved TFBSs or number of diverged TFBSs, whichever is smallest)
	Experimental TFBSs	Fictional TFBSs		
human rhesus	4.1 %	0.5 %	3.6 %	9
human cow	7.7 %	7.0 %	0.7 %	13
human dog	12.0 %	6.9 %	5.1 %	10
human mouse	8.7 %	6.8 %	1.9 %	16
human rat	6.5 %	9.6 %	-3.1 %	5
human opossum	14.8 %	6.5 %	8.4 %	7
human chicken	6.3 %	5.5 %	0.8 %	2
mouse human	22.7 %	6.3 %	16.4 %	6
mouse rhesus	19.7 %	4.6 %	15.1 %	5
mouse cow	21.0 %	4.8 %	16.3 %	8
mouse dog	19.2 %	9.0 %	10.2 %	9
mouse rat	-0.1 %	0.9 %	-1.0 %	2
rat human	7.0 %	3.6 %	3.4 %	9
rat rhesus	8.8 %	3.3 %	5.4 %	7
rat cow	9.3 %	3.5 %	5.8 %	9
rat dog	7.3 %	1.7 %	5.6 %	5
rat mouse	3.3 %	0.9 %	2.4 %	2
rat opossum	-25.3 %	7.2 %	-32.4 %	1
Weighted average	10.8 %. Error (bootstrap): 1.2 %	5.9 %. Error (bootstrap): 0.6 %	5.5%. Error (bootstrap): 1.4%	

2.3.5 Reserve List

The Reserve List of transcription factors was intended to be used to verify the key results from the analysis of the main list of transcription factors. It was not to be used for any exploratory work, and therefore used only rarely.

Table 2.21 shows the results from the Reserve List that correspond to Table 2.20. Comparing these two Tables, they are qualitatively in agreement that the real TFBSs show a difference from the fictional TFBSs; for Table 2.20 the real TFBSs give an effect $5.5\% \pm 1.4\%$ in excess of that for fictional TFBSs, whilst for the Reserve List (Table 2.21) the real TFBSs give an effect $5.3\% \pm 2.7\%$ in excess of the fictional TFBSs. Thus the Reserve List helps confirm the conclusion relating to flanking DNA that was originally obtained using the Main List of transcription factors.

2.4 DISCUSSION

Table 2.17 identified some differences that might be of interest.

One was that conserved TFBSs tended to be more downstream than diverged TFBSs within the aligned region ("Position of TFBS within aligned region"). The idea behind this analysis was that, upstream of a gene, the regulatory regions may be more highly conserved than other upstream DNA, so perhaps each regulatory region would form one aligned region, whilst the non-regulatory DNA would be unaligned. If so, measuring the position of a TFBS relative to the ends of the aligned region will be equivalent to measuring its position relative to the ends of the regulatory region it is in. On this interpretation, the result suggests that, within a regulatory region, the downstream TFBSs rarely change and it is the upstream TFBSs that get altered by evolution. This interpretation is rather speculative since there is no guarantee that the ends of the aligned regions really do correspond to the ends of regulatory regions. Nevertheless, it is striking that this measure produced an interesting result when the apparently similar measure of "Distance

Table 2.21: BASED ON RESERVE LIST: How strongly had evolution conserved the DNA sequence around transcription factor binding sites: difference between conserved and diverged TFBSs, for a variety of species

Conservation of flanking sequence, for every combination of species				
Pair of species being compared (the first species has the experimentally determined TFBS; the second is the comparison species)	%id of 50 bases either side of TFBS: difference between average for conserved TFBSs and average for diverged TFBSs		The difference between the two columns to the left	Minimum sample size for experimental TFBSs (number of conserved TFBSs or number of diverged TFBSs, whichever is smallest)
	Experimental TFBSs	Fictional TFBSs		
human cow	8.9	4.2	4.7	3
human dog	15.4	9.0	6.4	9
human mouse	13.4	15.3	-1.9	7
human rat	20.0	2.1	17.9	2
human opossum	13.1	12.9	0.2	2
mouse human	13.7	8.6	5.1	3
mouse rhesus	24.5	9.9	14.6	3
mouse cow	13.2	12.6	0.6	4
mouse dog	21.0	13.8	7.2	8
mouse rat	6.5	3.5	3.0	3
rat human	8.8	-6.0	14.8	1
rat rhesus	25.8	-2.4	28.2	1
rat cow	11.9	4.7	7.2	1
rat dog	-5.2	1.7	-7.0	2
Weighted average	14.6 %. Error (bootstrap): 2.3 %	7.8 %. Error (bootstrap): 1.2 %	5.3%. Error (bootstrap): 2.7%	

of TFBS from TSS" (Table 2.15) produced a difference far too small to meet the criterion for significance.

"The distance from the edge of the aligned region" measure is similar in many ways and does not appear to lead to any new interpretation.

The CpG result suggests TFBSs are more likely to be conserved if they are in a region with higher-than-usual CpG content. A worry related to this result occurs if a TF has a C next to a G in its consensus sequence, and consequently the probability of obtaining a match varies considerably depending on the CpG content of the sequence. When a shuffled matrix is produced, this property will be destroyed and so this is a circumstance when a shuffled matrix can be an inaccurate model; the practical importance of this has been pointed out (Wingender et al., 2002). Whether this problem would particularly affect the CpG result in Table 2.17 has not been ascertained.

The "%id of neighbourhood" result might have an interesting explanation. A plausible explanation why some regulatory regions are highly conserved is that they contain many TFBSs, so that mutations that disrupt these TFBSs will be rejected by natural selection. But, if sequence conservation depends on TFBSs, and if the probability of gaining or losing a TFBS is correlated with the sequence conservation (as Table 2.17 suggests), then it is plausible to suggest that the probability of gaining/losing a TFBS is affected by the presence of other TFBSs in nearby DNA. This idea seemed particularly interesting, and might shed light on one of the mechanisms causing gain/loss of TFBSs. Therefore, work later in the project was focussed on measuring this effect more accurately, and investigating possible explanations. Some of the work examining the effect has already been shown in Table 2.20. As for the explanations, the somewhat vague idea that "the probability of gaining/losing a TFBS is affected by the presence of other TFBSs" was developed into two more specific hypotheses; these ideas will be described in a later chapter.

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)