Appendix A

APPENDIX: MUTATION RATES AT FOURFOLD SYNONYMOUS SITES

A.1 INTRODUCTION

This will be brief, dealing with something that is not a “core” topic; nevertheless the results are needed to justify some of the "input variables" of the simulations described in the earlier chapters.

It was desired to have an estimate of how much a regulatory region would mutate if it was not subject to selection pressure. The mutation rate at fourfold degenerate bases in the protein-coding sequence was used. Although some published estimates were available, for example for human-mouse (Hardison et al., 2003) , it was decided to produce estimates during this project, partly because of all the combinations of species used, but also because the variability of mutation rates was required.

A.2 METHOD

The Perl module SynonymousMutations.pm was written for this purpose. This had to be supplied with information about the two homologous genes to be compared. At least one of these had to have the protein sequence already available.

Where no protein sequence was available for one gene, it would be produced from the transcribed DNA sequence (if available) or from the genomic DNA sequence. Using Blast2 (Tatusova and Madden, 1999) and the StandAloneBlast module of BioPerl (Stajich et al., 2002), the DNA sequence was aligned against the known protein sequence for the other, homologous gene. The coordinates from each High-Scoring Pair (HSP) from the alignment were used to splice out the parts of the DNA sequence that were aligned; these were concatenated to form a protein-coding DNA sequence. This might be smaller than the true protein-coding sequence, since it would not include anything that failed to align. Also, as noted in an earlier chapter, only 10000 bases of sequence were retrieved downstream of the supposed transcription start site, which could leave out later exons. If this process produced <100 residues of protein sequence, no estimate of mutation rates was produced.

Aligning the DNA sequences of the two homologous genes could produce an unsatisfactory alignment, as gaps could be inserted into the middle of a codon. Instead, therefore, the two protein sequences were aligned using ClustalW (Thompson et al., 1994); the DNA sequences then had a triple gap inserted at each position corresponding to a gap in the protein alignment.

The mutation rates dN and dS were at one time estimated using the method of (Yang and Nielsen, 2000). However, as what was required was simply the %id at the fourfold synonymous positions, these were later replaced by an estimate which merely involved counting the number of synonymous bases and how many of them had changed.

The sample of proteins used was the same as for the analysis of TransFac TFBSs. This will have biassed the sample towards proteins from genes that have many TFBSs. For instance, a protein could be counted twice if its gene has two TFBSs recorded in TransFac. This bias is presumably desirable, since the purpose is to produce estimates of mutation rates for use in models of regulatory regions.

A.3 RESULTS

Table A.1 shows the results. The "species with known protein sequence" is a species shown in the TransFac database as having a TFBS associated with the protein sequence; the "comparison species" may not have any TFBSs known for that particular protein.

Table A.1: Mutation frequencies at synonymous sites: Proportion of synonymous bases that were identical

Comparison species	Species with known protein sequence								
	human			mouse			rat		
	Mean	σ	N	Mean	σ	N	Mean	σ	N
human	N/A	N/A	0	0.648	0.081	41	0.622	0.080	36
rhesus	0.894	0.049	67	0.610	0.077	15	0.588	0.049	13
cow	0.613	0.146	60	0.457	0.051	18	0.569	0.105	25
dog	0.673	0.149	91	0.680	0.078	20	0.541	0.069	28
mouse	0.639	0.065	116	N/A	N/A	0	0.814	0.039	36
rat	0.632	0.082	85	0.819	0.043	26	N/A	N/A	0
opossum	0.475	0.086	82	0.486	0.105	23	0.442	0.052	16
chicken	0.433	0.074	48	0.418	0.064	12	0.421	0.097	10

An important question is whether the standard deviations shown in table A.1 were really due to variations in mutation rates. An alternative explanation would be that they were merely statistical variation caused by the limited number of residues in each protein. An analogy would be if one tossed a number of identical coins 100 times each; most coins would not produce *exactly* 50 "heads", thus causing an apparent variation in the percentage of "heads". This was addressed by examining the human-mouse comparisons (where mouse was the comparison species). A chi-squared test for each protein was done, in which the number of identical synonymous bases and the number of non-identical synonymous bases were compared with the value expected from the average proportion of 0.639 given in table A.1. 20 proteins (out of 116) had a chi-squared value > 3.84 (which is the p=0.05 value). If all proteins except these 20 were assumed to have 0.639 of their synonymous bases being identical, the standard deviation was 0.049; this was not a great reduction on the 0.065 given in table A.1. This indicates that most of the variation in mutation rates was "real" in the sense that it could not be explained as mere statistical variation caused by the limited number of residues in each protein.

A.4 DISCUSSION

(Hardison et al., 2003) have also estimated mutation rates for a human-mouse comparison and observed an identity of 67.2% at fourfold degenerate sites, and 66.7% at sites within ancestral repeats (another type of site thought to be under no selection pressure). This is fairly similar to the 64%-65% in table A.1.

The mutation rate estimates here were produced for use in estimating how much change there would be amongst bases of regulatory regions that are not subject to selection pressure. However, note should be taken of the limitations of mutation rates which are based on substitutions. The "%id of neighbourhood" statistic used elsewhere in this thesis was calculated by treating gaps as mismatches. Thus, indels would cause this statistic to be less than 100% even if there were no substitutions.

Appendix B

APPENDIX: DETAILS OF COMPUTER CODE

B.1 INTRODUCTION

This appendix records details of computer codes that would not be appropriate to insert in the main text, but could be useful to anyone using the computer codes to repeat any of the research.

B.2 METHOD

B.2.1 PWMs

A Perl module written during the author's time at the University of Exeter, DNAWeightMatrix.pm, was used within the software to store PWMs, score sequences against them, and generated shuffled matrices. When used to search a sequence for "matches" to the PWM, it would return an array of Match.pm objects. Match.pm was also written by the author when at the University of Exeter.

B.2.2 Obtaining a list of TFBSs

This gives details of the procedure outlined on page 83.

For a given species, the module GetSites.pm obtained a list of TFBSs by sending web queries to the TransFac Professional database. Generally, version 7.4 was used for earlier work. More recent work used 10.2/10.3.

It started by searching the TransFac SITE table, using the symbol of the transcription factor (e.g. NF-kappaB) as a query. The resulting list of sites was examined and TFBSs retained only if they referred to the species the program had been asked to use (e.g. human), which will be referred to as the "source" species. (This removes not just sites from other species, but those that are "artificial sequence".) Data obtained included the description of the gene to which the TFBS belongs, the actual DNA sequence of the TFBS (which typically had been extracted from an experimental paper describing the TFBS), the co-ordinates of each end of that sequence, and the quality. The gene symbol was obtained by another query to TransFac; the Human Genome Nomenclature Committee symbol was used if the reply contained it, if not then UniGene (or if that failed, LocusLink) was contacted to obtain a gene symbol (Wheeler et al., 2004). All the information obtained was stored as a TFBS::Site object; TFBS::Site is from the "TFBS" Perl package (Lenhard and Wasserman, 2002). GetSites.pm returned an array of such objects.

The module GeneName.pm checked the acronym of each gene by querying if it was in the "symbol" field of the database at the HUGO human gene nomenclature website (www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl) (Wain et al., 2004). If the HUGO website found one record, from it the official gene symbol and name were extracted.

If the HUGO website did not find a record, the module attempted to find a gene by sending a second query, now examining the "symb_prev_and_alias" field. If this query produced a record for just one gene, that was used. If it produced no records, or records for more than one gene, the analysis of that TFBS was abandoned, with the message "Unclear which gene referred to by

..." going to the output file of results.

B.2.3 Identity of a Gene

This details the procedure outlined on 83.

For human TFBSs, the HUGO symbol was available from TransFac as standard, so the check was merely a precaution. For mouse/rat TFBSs, a standard gene symbol was not necessarily available (this has changed in the recent version of TransFac), with identification often being by non-standard gene names (eg GM-CSF, which the above procedure converted to CSF2), and a considerable number of TFBSs were lost because it was unclear to which gene they referred. Note that the HUGO human gene database was used even for rodent genes, because homologous human and mouse genes often (but not always) have identical symbols. If the wrong gene symbol was used, it would probably have resulted in the analysis of that TFBS being abandoned later if the sequence upstream of the gene did not contain an exact match to the TFBS sequence.

Obtaining the homologous genes in other species

HomologueList.pm was used to obtain homology data from the websites mentioned. The module returns an array of array-references, with the "source" species first (eg if we are looking at a TFBS that was first discovered in experiments on human DNA, the first element will be a reference to an array containing the RefSeq number (etc) of that human gene, and the next element a reference to the RefSeq number of the homologous mouse gene).

Obtaining DNA sequence near the TFBS

The modules that do this are called Gene.pm; GetMRNA.pm; and GetGenomeSeq.pm.

Local storage of data

This was done using DataStore.pm.

B.2.4 Locating the precise TFBS

Figure 2.3 refered to circumstances when a 2nd-best match has a score nearly as good as the best match. "Nearly as good as the best match" was usually defined as having a match score which differs by less than two standard deviations from that of the best match (this quantity being called "$scoreMargin"). If they met this criterion, the 3rd-best and 4th-best matches were also analysed, the 5th, 6th and 7th best may have been (depending on a technical detail) and the 8th best never was. The criterion given in the figure Figure 2.3 flowchart, whilst correct for the main TransFac-based analysis, does not always apply, since for the chip-chip analysis, different values of $scoreMargin were used.

Where the shuffled-matrix search resulted in analysis of a TFBS being abandoned, "shufMat assessment of experiment site raises doubts" was added to the relevant line of the output file.

B.2.5 Alignment of sequences

This was handled by the module AlignSeqs.pm, which returned an array containing a name followed by an array of AlignedSite objects. Each "AlignedSite" object essentially acts as a container for the information about an alignment containing a TFBS. The task of AlignSeqs.pm was limited, and involved only calling the alignment program (eg BlastZ) and processing the sequences. In particular, the PWMs were not used by AlignSeqs.pm and, from that, it will be evident that AlignSeqs.pm could not determine whether or not a TFBS existed in a given sequence.

B.2.6 Examination of alignments

BlastZ aligned the upstream sequence containing the known TFBS with the homologous sequence from one other species. The module ReadLavFormat.pm was written to convert the BlastZ output format into an array of HSPZ BioPerl objects. If there was more than one aligned region, then there would be more than one HSPZ object in the array.

If the TFBS was in one of these aligned regions, and there were no gaps within the TFBS, the following procedure applied. AlignSeqs.pm would store some statistical information about it (including the length, percent identity, and position of the TFBS within the aligned region) in the CollectedData.pm object for this TFBS. The AlignedSite object also had added to it a SimpleAlign object (SimpleAlign is a BioPerl class), containing the known TFBS with 55 bases of sequence on either side, apparently aligned against the corresponding sequences from the other species. This was not a true alignment, since it was not the result of a multiple alignment, but was the results of several pairwise alignments. Consequently the sequence containing the known TFBS could not be shown with gaps since the different pairwise alignments might have inserted gaps in different places, so it was shown complete and without gaps. All other sequences contained dashes to indicate gaps and to indicate sequence beyond the aligned region; consequently, the sequences were not guaranteed to be aligned except at the TFBS (provided the TFBS did not contain gaps).

Early experience showed there were sometimes “near misses” where the known TFBS was not wholly within the aligned region, but would have been if the aligned region was a few bases longer. Where the “miss” was by 5 bases or less, the 5 bases of sequence were added to the sequence of the aligned region, for the comparison species only.

However, if the TFBS fell across a gap in the alignment, or was not in any aligned region, then additional complications were required to deal with these. In these cases, the general strategy was to create an “AwkwardAlign.pm” object containing information summarising the problem, and to store that object within the AlignedSite object. That information would then be available to be acted upon later in the analysis, if possible, or a message printed explaining why no result had been produced. The detailed handling of these awkward alignments will now be described.

If the alignment put a gap in the known TFBS, then the sequence was not stored in the manner described above. Instead, the TFBS sequence with gaps in, plus the 55 bases of sequence on either side (now including gaps), was stored in a separate SimpleAlign object along with the corresponding aligned sequence from the one comparison species. This was then placed in an AwkwardAlign object.

If the known TFBS was aligned against a gap, AlignSeqs.pm did not take any special action.

A special procedure was followed if the known TFBS was between two aligned regions. This was used if the two aligned regions were in the same order in both species, and provided the sequence between the two aligned regions was of roughly similar length in both species (defined as differing by less than a factor of 2). Then the sequence between aligned regions in the comparison species was stored in an AwkwardAlign object, which was labelled as "BetweenAligns". The motivation for this was the assumption that, since the known TFBS was between the aligned regions, then (if it was conserved) the homologous TFBS would be somewhere in the sequence between the two aliged regions in the comparison species. By storing this sequence, it would be possible to search it for the TFBS later.

Apart from this, if the known TFBS was outside the aligned regions then an AwkwardAlign object was created containing a message saying so.

If no alignment was found at all, an AwkwardAlign object was created containing a message saying so.

The return from AlignSeqs.pm was a name followed by an array of AlignedSite objects. The reason for producing more than one object was that any "second-best" site (etc), as described above, would generate another AlignedSite object of its own.

B.2.7 Deciding if a TFBS was conserved or not

For each comparison species, a "ConservedSiteAssessment.pm" object was set up which, amongst other things, acted as container to store comments indicating that a TFBS was conserved or not, and accompanying explanations.

From the SimpleAlign object returned by AlignSeqs.pm, a "splice" would be extracted that contained only the known TFBS aligned against the sequence from the comparison species

When the known TFBS itself contained gaps, AlignSeqs.pm would have returned the gapped alignment sequence within an AwkwardAlign object.

B.2.8 Examination of "neighbourhood" (DNA surrounding a TFBS)

In fig 2.5, one of the decision diamonds refers to a match which has at most one base different from the known TFBS, or two bases different if it was at least 13 bases long. These criteria were set so that a match was very unlikely to occur by chance; for example, a 10 bases long sequence, compared against 100 bases of random sequence, has a 1 in $\approx$ 100 chance of making a match with 9 out of 10 bases identical. Reverse complement matches were counted, though this may have been unnecessary.

B.2.9 Awkward alignments

In "Examination of alignments" above (page 227), it was explained how certain problems could cause an AwkwardAlign object to be added to an AlignedSite object. These were dealt with by adding an "Unclear" comment to the ConservedSiteAssessment object, with a couple of exceptions: those dealing with a gap in the known TFBS, which have already been covered in "Deciding if a TFBS was conserved or not: when a gap is in a TFBS" above; and the "BetweenAligns" cases.

B.2.10 Generating a conclusion

The “comments” will have been stored in a ConservedSiteAssessment object. ConservedSiteAssessment.pm also contained the Perl code to summarise all these comments into a single conclusion, as described in fig 2.7. If a comment did not contain the word “conserved” nor the word “diverged”, the words of that comment would be copied into the “explanatory remark”.

B.2.11 Minor utilities

To control the amount of information output, near the start of divSiteData.pl the user was asked how verbose an output they would like, on a scale of 1 to 5. Choosing 5 would produce a very wordy output. Choosing 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. To control verbosity, the program would print information by calling the module PrintWithControl.pm, which was written during the course of this PhD. Each call to this module would pass the number or text to be printed - typically just a few words - but also a number indicating how important that text was (the higher the number, the less important it was). If that number exceeded the verbosity that had been selected by the user, then the text/number would not be printed. Where large numbers of TFBSs were being analysed at once, as was the case for the main analyses described in this thesis, a verbosity of 1 was normally used.

One practical problem was inconsistencies in the exact format of the name of each species; for instance, the mouse will be called “M.musculus” or “Mus musculus” depending on from which website the data is downloaded. SpeciesName.pm was written to deal with this; it contains several alternative names for each of the species used in this project. Functions provided by this module included sameSpecies, which accepts two names within its input and returns 1 if the names refer to the same species, 0 otherwise; and convertToParticularName, which can convert a name to whichever format is specified in the list of arguments.

CollectedData.pm acted as a container to store data that was generated during the program but needed to be printed out later on; in divSiteData.pl, objects belonging to this class were named $statsForThisSite.

B.2.12 Generating a sample of fictional TFBSs

The “raw” list of fictional TFBSs was filtered into the “filtered” list by the program filterFalseSites.pl.

The program did this placed each fictional TFBS in a “bin”, where the first bin contained fictional TFBSs whose binding score was within one standard deviation of the threshold, the second bin contained fictional TFBSs whose binding score was between one and two standard deviations above the threshold, etc. (Note that this standard deviation was based on the scores of the real TFBSs). For each bin, the program calculated the ratio of the actual number of fictional TFBSs to the desired number of fictional TFBSs (the latter being 5 times the number of real TFBSs in that bin). Each fictional TFBS was then selected if a random number exceeded this ratio. Thus the probability of selection was the same for all the fictional TFBSs in a bin, even thought the scores were not exactly the same. Each type of TF was treated separately.

B.2.13 The chip-chip data

Using a file downloaded from

http://jura.wi.mit.edu/young_public/pancregulators/HNF_data_v2.xls

my routine GetSites::fromChIPchip extracted data and returned it in the form of an array of TFBS::Site objects (Lenhard and Wasserman, 2002), each of which had a "quality" tag set to "ChIP-Chip".

B.2.14 Phylogenetic trees: distinguishing gains and losses of TFBSs

This analysis used a Perl program written for this project, extractTrees2.pl. Each node of the phylogenetic tree was represented by a Bio::Tree::NodeNHX object, which is part of the BioPerl (Stajich et al., 2002) system.

Appendix C

APPENDIX: PARSIMONY, TFBS LOSS AND PHYLOGENETIC TREES

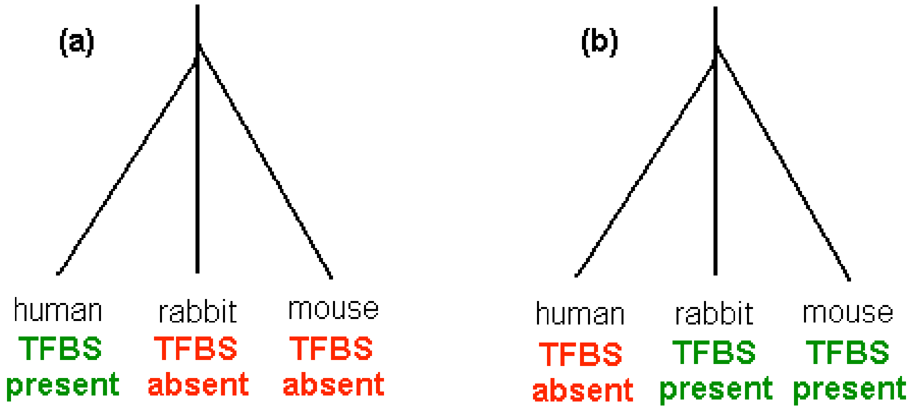
In the chapter on Phylogenetic trees, a question was raised about interpretting TFBSs in phylogenetic trees (section 5.4.8). Suppose a TFBS, which existed in an ancient vertebrate, was lost *twice* - that is, it was lost in two separate lineages. This could be misleading, as it might appear to be a "gain", if for instance it produced the situation in figure 5.2(a). But is this something that is likely to happen in practice, or can we assume that it is so rare that it can be ignored?

To quantify if this is serious, take the estimate that 30% of TFBSs are non-conserved in a human-rodent comparison (Dermitzakis and Clark, 2002), assume that gain/loss events occur with equal frequency on all arms of the lineage, and assume (quite arbitrarily) that losses are 3 times as frequent as losses, thus giving a probability of 0.11 that a loss will occur along a single arm of the phylogenetic tree. Consider a TFBS present in human but not mouse or rabbit. Assume that it is known that mouse diverged a very short time before the human-rabbit divergence, so that almost nothing happened between these two divergences, and so the 0.11 probability applies to all three species-specific lineages. (In fact, the order in which these species diverged is a matter of controversy (Misawa and Janke, 2003)). Thus we have the tree shown in Figure C.1(a), which could be explained by a single gain. But this tree could also be explained as a double loss: if the TFBS was present in ancient vertebrates, then the probability of producing this tree by double loss would be

(1 - 0.11) * 0.11 * 0.11 = 0.011

There would also be the same probability of producing a tree with the TFBS present in the rabbit only. This gives a total probability of 0.022 that a double-loss will produce a tree that looks like a single-gain tree.

Figure C.1: Example involving human, rabbit and mouse
Whilst not completely to scale, the very short time between the two divergences should be noticed.



A tree that looked as if it was produced by a single loss is shown in figure C.1(b). If the TFBS was present in ancient vertebrates, the expected number of losses (counting all 3 branches of the tree) will be 3 * 0.11 = 0.33. Since we are assuming (for this example) that there are 3 losses for every gain, there will be 0.11 gains; but one-third of these gains will produce a tree that is not a clear example of a gain tree (where the gain was in mouse). Thus, for every tree produced by a TFBS present in ancient vertebrates, there will be 0.07 trees that both look like they were produced by a single recent gain, and were, in fact, produced in that way.

This should be compared with the 0.022 trees that look like they were produced by a single recent gain, but were in fact produced by a double loss. Thus in this case, of the cases that appear to be recent gains, 24% (= 100 * 0.022 / (0.07+0.022)) will really be double-losses.

Hence the number of gains will be exaggerated. This will produce a biased estimate of the ratio (frequency of gains : frequency of losses), but in this case the size of this bias does not look that important, in comparison with the actual ratio of 3 upon which this example is based. Clearly this example could be improved; for instance, it was assumed that two losses occurred as independent events (the probability of one is not affected by the fact that the other occurred), which is doubtful.

This is enough to raise concerns about the procedure of always assuming the most parsimonious tree, but not to prove that it is definitely an unacceptable method.

Appendix D

SUMMARY: TFBSs in Vertebrates: Do They Evolve Together, Or Individually?

This Appendix is a shorter description of work described in the thesis.

Background

The evolution of complex organisms is likely to depend, in part, on the evolution of mechanisms that control genes. Transcription factor binding sites (TFBSs) are one part of this. A particular aspect of TFBS evolution will be addressed: the role of flanking DNA on either side of the TFBS.

Some studies have suggested that the gain/loss of a TFBS can be related to events elsewhere in the same regulatory region. A study of a fly enhancer (Ludwig et al., 2000) found that evolutionary changes in the first half of the enhancer were compensated for by evolutionary changes in the second half of the enhancer, so there was no change in the expression pattern caused by the enhancer. It has been suggested that in yeast, the gain of a “RAP1” TFBS has caused the subsequent loss of a “Homol-D” TFBS in the same regulatory region (Tanay et al., 2005); and this has happened, not just once in a particular gene, but rather it has occured many times, each in a gene belonging to the ribosomal protein class. For human TFBSs binding the Estrogen Receptor, in a comparison against mice, it was reported that TFBSs that were conserved, the sequence of the TFBS was conserved more strongly than the flanking sequence; whereas for TFBSs that were not conserved, the sequence of the TFBSs was conserved to about the same extent as the flanking sequence (O'Lone et al., 2004).

Thus, there are a number of examples where the gain/loss of a TFBS, to be understood fully, must be explained by taking account of the DNA flanking the TFBS. However, this is difficult to model because few details are available. It is not clear if such effects occur for most cases of TFBS gain/loss, or merely for a minority of cases.

To examine this, the study reported here considered a considerable number of cases of TFBS evolution in mammals. The main focus was on the DNA flanking the TFBS - in particular, how strongly it was conserved. Several possible explanations for the results were considered, and further analyses were devised to test these explanations. One of these analyses involved "fictional TFBSs". Another analysis required "gain-of-TFBS" to be distinguished from "loss-of-TFBS". These analyses provided grounds for rejecting some possible explanations and focussing on others.

Results and Discussion

Conservation of the DNA flanking a TFBS

Table 1 is presented as an example of a comparison involving just two species, human and dog. TFBSs for 25 transcription factors (TFs) were considered. It can be seen that conserved TFBSs were flanked by DNA sequence that was moderately well conserved; in contrast, a lower rate of conservation was observed in DNA flanking TFBSs that have diverged (ie, are present in human but not in dog).

Table 1 - Conservation of sequence flanking conserved and diverged TFBSs

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 %	

A much larger sample of diverged TFBSs was obtained by considering eight species rather than two, but it required a more complicated analysis. Table 2 shows the corresponding results for all the pairs of species considered, in the column headed “experimental TFBSs”; here the human-dog results are summarised by showing only the 12.0% difference from table 1. The bottom of the table shows the weighted average over all pairs. Almost every entry in this column is positive; so this again shows that the DNA flanking conserved TFBSs is usually more highly conserved than the DNA flanking diverged TFBSs.

Table 2 - Conservation of flanking sequence, for every combination of species

The human-dog result from the bottom of table 1 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).

Table 2. 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%	

Additional data sets

Additional sets of data were assembled to see if they would confirm results from the main analysis. One was the "reserve list" of 14 TFs, which was similar though smaller than the list of 25 TFs used for the main analysis. The other was more different, as it used ChIP-chip data (Odom et al., 2004), and just one TF (HNF-4). Table 3 summarises results from these data sets as well as the main analysis. The results for experimental TFBSs are broadly consistent.

Table 3 - Flanking sequence conservation, for three different samples of data

The "main analysis" results from the bottom of table 2 is shown here, along with corresponding results from two other samples of transcription factors

Flanking sequence conservation, for three different samples of data			
Sample of experimental TFBS data	%id of 50 bases either side of TFBS: difference between average for conserved TFBSs and average for diverged TFBSs		Average difference between experimental and fictional TFBSs
	Experimental TFBSs	Fictional TFBSs	
Main analysis (TransFac data for 25 TFs)	10.8% ± 1.2%	5.9% ± 0.6%	5.5% ± 1.2%
Reserve list (TransFac data for 14 TFs)	14.6% ± 2.3%	7.8% ± 1.2%	5.3% ± 2.4%
ChIP-chip data for 1 TF (HNF-4)	8.4% ± 2.1%	7.4% ± 1.2%	0.5% ± 2.0%

Possible explanations

To explain why DNA is more highly conserved if it flanks a conserved TFBS, several hypotheses were considered:-

The "correlated-evolution hypothesis"

This supposes that several adjacent TFBSs are all lost simultaneously (or that they are all gained simultaneously). Here "lost" means the time when the TFBSs, which used to contribute to the fitness of the animal, cease to do so. Hence with this hypothesis, if a TFBS is conserved, then the flanking TFBSs are also conserved, and hence the flanking DNA sequence is fairly well conserved (illustrated in fig 1A). Conversely, if a TFBS is lost, then the flanking TFBSs are lost also - consequently, mutations can accumulate in the flanking DNA without being rejected by natural selection - hence the sequence of this DNA is not well conserved (fig 1B).

This hypothesis is particularly plausible in the light of the "enhanceosome" model (Arnosti and Kulkarni, 2005), in which "disruption or displacement of a single binding site ... causes the element to be inactive". Hence, for regulatory regions that conform to this model, once one TFBS is disrupted, all other TFBSs in the regulatory element cease to have any function.

Thus simultaneous "losses" are easy to imagine, but in contrast, the simultaneous "gain" of (say) 12 TFBSs seems very unlikely (Stone and Wray, 2001). However, if a series of "gains" took place in rapid succession (rapid compared with the evolutionary timescale of mammals) it might form an approximation to simultaneous "gains".

Figure 1 - The “Correlated-evolution” hypothesis

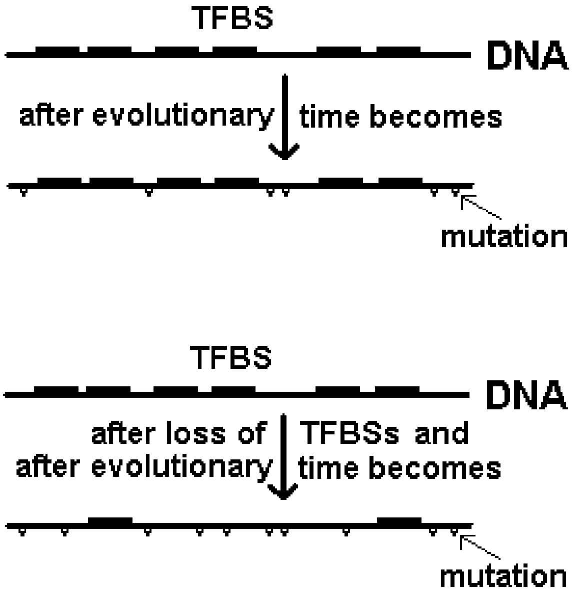

TFBS
DNA
after evolutionary time becomes
mutation
TFBS
DNA
after loss of TFBSs and
after evolutionary time becomes
mutation

Fig 1A

Fig 1B

Figure 1 legend: Each illustration shows a stretch of DNA which contains some TFBSs (each thick block represents a TFBS). It assumes that mutations within each TFBS tend to be rejected by natural selection, whereas mutations elsewhere are tolerated. In Fig 1A, all the TFBSs are conserved. In Fig 1B, during the passage of evolutionary time, an event causes several TFBSs to be lost (no longer be selectively advantagous) simultaneously, after which the DNA is able to accumulate more mutations than in Fig 1A.

The "site-density hypothesis"

Here the "site-density" of a region is defined as the number of TFBSs it contains, divided by its length in base-pairs. Assume that DNA that is not part of a TFBS is "spacer" DNA that can easily accumulate mutations. Consequently, regions of low site-density will have plenty of "spacer" DNA, so the DNA sequence will not be highly conserved *even if all the TFBSs are conserved*. This is illustrated in figure 2A, which should be compared with the high site-density case illustrated in figure 2B. 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 1.

It is not known why the probability of gain/loss should be affected by the site-density. However, for "gains", a simple explanation can be suggested. This is that, in regions of low site-density, mutations within the "spacer" DNA can easily create a new TFBS; whereas in regions of high site-density, this is less likely to happen simply because there is less "spacer" DNA.

The "faulty-data hypothesis"

Suppose a TFBS in the database did not exist in reality, but was only present due to a mistake. If it was in a region where the substitution rate was high, it would be more likely to appear to be diverged than if it was in a region where the substitution rate was low. Hence if mistaken TFBSs were sufficiently common in the database, this would give an effect that was qualititively the same as in table 1.

The "neutral-selection hypothesis"

Here it is supposed that many TFBSs have no effect on the fitness of the animal, even if each is a genuine TFBSs which does alter the expression level of an adjacent gene. It has been suggested that changes in gene expression often have a neutral effect (Khaitovich et al., 2004), or that it is common for TFBSs to cause little or no selection pressure against disruptive mutations (Keighley et al., 2005). Thus, in this hypothesis, these TFBSs will not be subject to natural selection. Consequently the probability that one of these is lost will depend on the substitution rate of the region of the genome it is in. Thus - in a similar way to the previous hypothesis - this would give an effect that was qualititively the same as in table 1.

Figure 2 - Substitution rates may depend on "Site-Density"

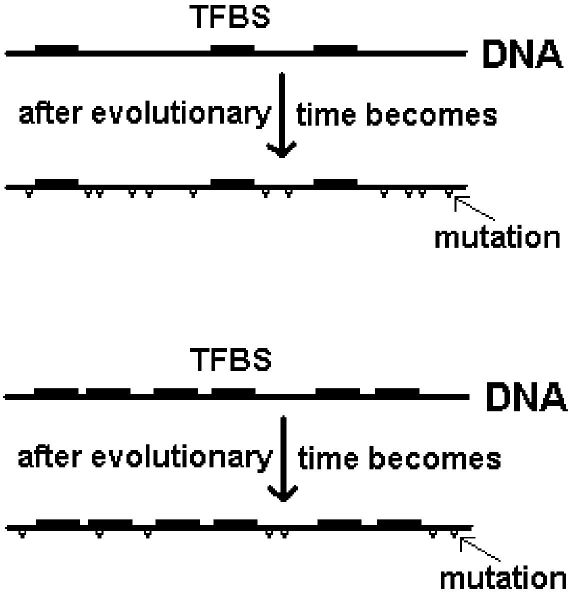


Fig 2A

Fig 2B

Figure 2 legend: Fig 2A shows a regulatory region with "low site-density", whereas Fig 2B shows a regulatory region with "high site-density" (that is, there are more TFBSs within the same length of DNA). As with fig 1, it is assumed that mutations with each TFBS tend to be rejected by natural selection, whereas mutations elsewhere are tolerated. Hence the "low site-density" region can accumulate more mutations that the "high site-density" region. Unlike figure 1, all the TFBSs shown are conserved, showing that the difference in substitution rates can occur even without any evolutionary change in the TFBSs.

Means of deciding which hypothesis was most plausible

Fictional TFBSs

To investigate the “faulty-data hypothesis”, a database of fictional TFBSs was created (see methods). Each fictional TFBS is a DNA sequence that actually occurs at a particular place in a real genome; the fiction is to pretend that it is a TFBS. The fictional TFBS were analysed in the same way as the experimental ones had been. Thus, the fictional-TFBS analysis shows us what would happen in the extreme case of a database that consists entirely of mistaken TFBSs.

Table 4 shows the results of a fictional-TFBS analysis, corresponding to the analysis of experimental TFBSs shown in table 1. Table 4 shows a difference of 6.9%. This is of the same sign as the difference shown in table 1, though the size of the difference is smaller, by 5.1%. Similar results for a larger selection of species are shown in table 2; it can be seen that experimental TFBS result exceeds the fictional TFBS result for all species-pairs except three. The bottom of table 2 shows that the the experimental TFBS result exceeds the fictional TFBS result by, on average 5.5 ± 1.2 %. This is a statistically significant difference (based on 100 bootstrap samples all giving a positive value). This shows that the experimental TFBS data are not like the fictional TFBS data.

Table 4 - Conservation of sequence flanking fictional TFBSs

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 1.

Conservation of sequence flanking fictional TFBSs			
	Mean	Std 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 %	

The analysis of "reserve-list" data also showed that fictional-TFBS give a different result than experimental TFBSs do, thus confirming the result from the main study (table 3). In contrast, for the study of chip-chip data, fictional-TFBSs produced the same result as experimental TFBS (table 3). It is not clear why the chip-chip data gave this result, when the TRANSFAC-based data did not. One possible cause is that the chip-chip data consisted entirely of evidence of TFBSs in humans, whereas the TRANSFAC data provided evidence of TFBSs in humans, mice or rats. Another possible cause is that chip-chip data only provides a location accurate to a few hundred bases, so the exact location had to be found by searching for a good match to the PWM, a procedure that is not very reliable and was estimated to find the incorrect location in about 25% of cases here.

Thus, (with the exception of the chip-chip data), the results from the fictional TFBSs are quantitatively different from the results from the TRANSFAC-based experimental TFBSs. Since the fictional TFBSs are our model for the “faulty-data hypothesis”, this suggest that the “faulty-data hypothesis” is not an adequate explanation, as it predicts an effect that is too small compared with the effect actually observed. (However, for the chip-chip data, the “faulty-data hypothesis” is more plausible).

The fictional-data analysis might also be used as a model for the “neutral-selection hypothesis”. The grounds for this are that the fictional TFBSs are not subject to natural selection (except if they overlap parts of the genome that are under natural selection), and thus are a model for the behaviour of “neutral” TFBSs (ie, ones that are not under natural selection). A difficulty is that some fictional TFBSs will overlap real TFBSs; this was not a problem when the purpose was to model mistaken TFBSs (since some mistaken TFBSs will overlap real TFBSs); however, if “neutral” TFBSs do not overlap real TFBSs (or only very rarely), then the fictional TFBSs will be an imperfect model of them.

As analysis of fictional-TFBSs gave different results to analysis of experimental TFBSs, this is evidence that the “neutral-selection hypothesis” is not a good explanation of the data, subject to the reservation noted in the previous paragraph.

Distinctions between the “correlated-evolution” and “site-density” hypotheses

Having rejected the “faulty-data” and “neutral-selection” hypotheses, it was therefore desired to find a method of analysing data that would indicate which of these two remaining hypotheses was more plausible. Before considering this, however, it is worth examining some conceptual differences between these two hypotheses. Both depend on other TFBSs that are assumed to be in the flanking DNA, so it might at first appear that there is no real distinction between the two hypotheses. But in fact they are distinct, for example:-

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 distinction 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 occurs irrespective of whether the gains/losses actually happen.

Another distinction between the hypotheses, which might actually be tested, is as follows. The "correlated-evolution" model predicts that the DNA sequence flanking "losses" will be conserved more highly than the DNA sequence flanking "gains". The "site-density" model does not predict this difference. This could provide a method of testing which model is more plausible, which will be explored below.

To understand why this prediction is made, see figure 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 3 does not say anything about this; it must be combined with one of the evolutionary models to make a prediction. Using the "correlated-evolution" model, because the TFBS being studied 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 3 will apply to the flanking DNA sequence as well. In contrast, if the "site-density hypothesis" is used, there is no reason to suppose any difference between "gains" and "losses" in respect of how well the flanking DNA sequence is conserved.

Figure 3 - Sequence conservation expected in lost and gained TFBSs

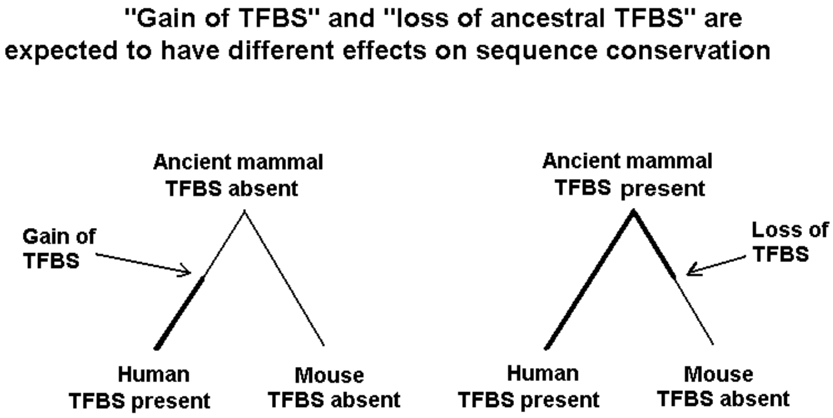

"Gain of TFBS" and "loss of ancestral TFBS" are expected to have different effects on sequence conservation
Ancient mammal TFBS absent
Gain of TFBS
Human TFBS present
Mouse TFBS absent
Ancient mammal TFBS present
Loss of TFBS
Human TFBS present
Mouse TFBS absent

Figure 3 legend: The evolutionary trees show two circumstances that result in apparently the same situation in modern animals, i.e. a TFBS present in humans but not in mice. The evolutionary tree is shown with a thick line when the TFBS is present and, consequently, natural selection will reject mutations that disrupt the TFBS. The tree is shown with a thin line when the TFBS is absent and, conseqently, mutations in the sequence will be tolerated.

It is evident that mutations will have longer to accumulate in the "gain of TFBS" case than in the "loss of TFBS" case. Hence the sequence in the "gain of TFBS" case is likely to be less conserved than with the "loss of TFBS" case. (This assumes that no counter-selection against the TFBS occurs when it is lost).

Distinguishing “gains” from “losses” using phylogenetic trees

To test the prediction made at the end of the previous section, cases where a TFBS was gained during evolution must be distinguished from cases where a TFBS was lost during evolution. (In tables 1-4, there was no distinction as both types were in the “diverged” category). This was done by relating the data to the evolutionary tree of vertebrates (see Methods). For some TFBSs, the data could be more parsimoniously explained as an ancient TFBS that was recently lost in one lineage, rather than as a TFBS that was recently gained. For some other TFBSs, it was the other way round. Unfortunately, many TFBSs had to be left out of this analysis because there was no basis for deciding if they were “gains” or “losses”. The analysis identified almost equal numbers of “loss” events and “gain” events.

Data shows how well conserved was the DNA sequence flanking these TFBSs (table 5). It can be seen that the DNA flanking “losses” tended to be better conserved than the DNA flanking “gains”, by 4.3% on averge. For any particular pair of species, the results will be subject to large error because the sample size was usually extremely small. Even the 4.3% average is 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).

Table 5 - Conservation of flanking sequence, when gain-of-TFBS is distinguished from loss-of-TFBS

For some diverged TFBSs, the results, when compared against the vertebrate phylogenetic tree, enabled the divergence to be attributed to either a gain-of-TFBS or to a loss-of-TFBS. This table shows the conservation of flanking sequence subdivided by gains and losses.

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%	

Thus this is one piece of evidence suggesting that the "correlated-evolution" hypothesis is more plausible than the "site-density" hypothesis.

That conclusion is rather tentative, however, as the significance level indicates the evidence is rather weak. Other possible problems should also be mentioned. A similar analysis of fictional TFBSs did not suggest that problems were likely to be caused by faulty data: for fictional TFBSs, the average difference between "losses" and "gains" was only 0.3% (error: 1.4%). Interestingly, fictional "gain" events were much more frequent than fictional "loss" events, which suggests the sort of result likely to be produced from data that is hopelessly contaminated by incorrect TFBS data (this has also been seen in fly data (Moses et al., 2006)). Another problem is that, in figure 3, "loss" refers to the time when a TFBS ceases to be advantageous, whereas in the data used, a "lost" TFBS is only identified when it has accumulated enough mutations to disrupt binding. Because of the time taken to accumulate mutations, there is likely to be a time interval between these two events, and so the data used is not the ideal data for testing predictions based on figure 3.

General discussion

Thus, of the four hypotheses considered, "correlated-evolution" is the one most consistent with the evidence. However, it should be realised that there are other possible explanations besides those considered so far.

It has been suggested that some genes need to maintain a precise expression pattern, but others do not (Bilu and Barkai, 2005). Genes in the latter category will be much more tolerant of TFBSs being gained or lost in their regulatory regions. This could result in diverged TFBSs being surrounded by weakly conserved DNA, as found in table 2.

It is also possible that more than one of the hypotheses mentioned is correct. When the "correlated-evolution" hypothesis was introduced earlier, it was remarked that it was easier to imagine it applying to "losses" than applying to "gains". For the "site-density" hypothesis, it was the other way round. This makes it plausible to imagine that the "site-density" hypothesis operates for "gains" whilst at the same time the "correlated-evolution" hypothesis operates for "losses".

The "correlated-evolution" hypothesis, as outlined above, supposed a number of "losses" would occur together *or* a number of "gains" (not a mixture of both). But one might imagine a regulatory region undergoing a number of "gains" and "losses", in roughly equal numbers, at about the same time. However, that would mean that a "lost" TFBS would be flanked by DNA in which the half the changes occuring were "losses" and the other half were "gains". A "gained" TFBS would also be flanked by DNA for which that was true. Hence there would be no difference between "losses" and "gains" in respect to the flanking DNA, leading to a prediction that no differences should be observed in table 5. The difference actually observed argues against this.

This is of interest because the concept of "turnover" is often mentioned in the literature (Ludwig et al., 2000) (Gasch et al., 2004) (Odom et al., 2007) (Doniger and Fay, 2007), meaning that mutations create a surplus TFBS and later destroy another TFBS in the same regulatory region - thereby altering individual TFBSs without much change in the total number of TFBSs in the regulatory region. Figure 4 describes how turnover of individual TFBSs cannot explain the difference in table 5. Thus, that difference is evidence of TFBS evolution being affected by some phenomenon other than turnover of individual TFBSs.

There is evidence that (in yeast) gain-of-TFBSs is more likely to occur in promoters with large numbers of TFBSs (Bilu and Barkai, 2005). That suggests a "site-density" effect occurs, but in the *opposite direction* to the "site-density" hypothesis considered above. So it offers no support at all to the "site-density" explanation offered earlier.

Figure 4 - What models of TFBS evolution predict that the flanking sequence around "losses" differs from the flanking region around "gains"?

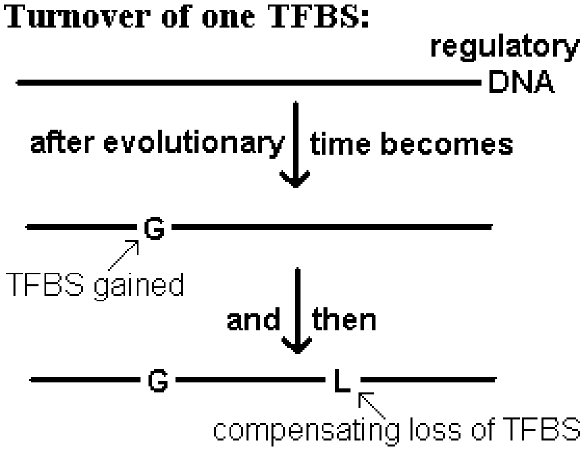


Fig 4A

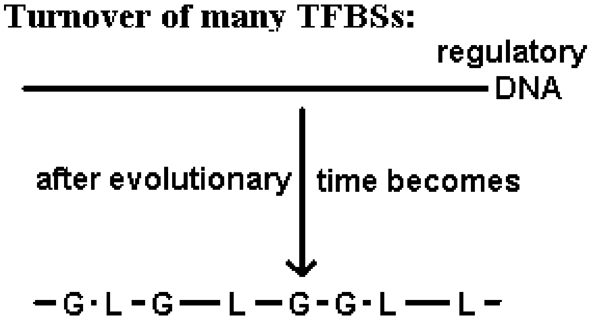


Fig 4B

Figure 4 legend: Here, "L" marks a TFBS that has been lost (compared with the ancestral DNA sequence shown near the top of each diagram), and "G" represents a gain of TFBS. Fig 4A shows a "turnover" model in which a gain-of-TFBS makes an existing TFBS redundant, which is lost afterwards. Fig 4B shows several such turnover events happening in a regulatory region, thus

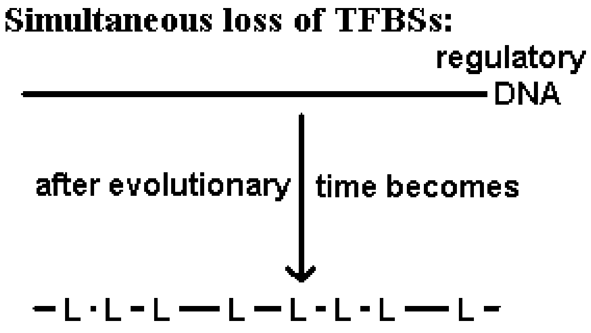


Fig 4C

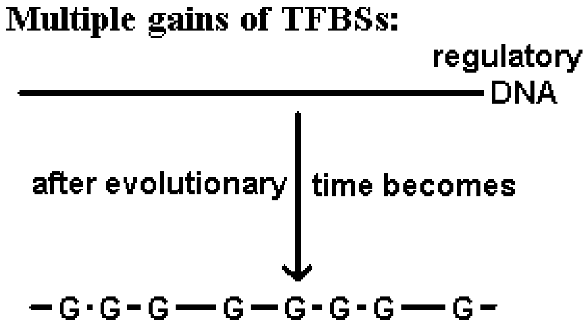


Fig 4D

resulting in equal numbers of gains and losses. Therefore, if we pick any particular lost-TFBS, its flanking DNA will contain equal numbers of gains and losses (approximately). Similarly, if we pick any particular gained-TFBS, its flanking DNA will also contain approximately equal numbers of gains and losses. Hence, in respect of flanking DNA, this "turnover" model does not predict any difference between lost-TFBSs and gained-TFBSs. Whilst turnover of individual TFBSs cannot explain a difference of this kind, it might be explained by other models of TFBS evolution, such as Fig 4C and/or Fig 4D.

Conclusions

By examining TFBS data for a number of vertebrates, it was found that the DNA flanking conserved TFBSs is usually more highly conserved than the DNA flanking diverged TFBSs.

Four different explanations for this were considered. Two of them considered as possible causes incorrect TFBS data, or alternatively TFBS that were under neutral selection. Both these were rejected because, when a database of fictional TFBSs was created then analysed, it failed to reproduce the results seen earlier.

Two other hypotheses involved "correlated-evolution" (when several TFBSs are lost at once), and a "site-density" effect (when the probability of TFBS loss/gain was related to the density of TFBSs in the regulatory region). A means of distinguishing these hypotheses was devised that required "loss-of-TFBS" to be distinguished from "gain-of-TFBS". This evidence suggested that the "correlated-evolution" hypothesis was more plausible explanation than the "site-density" hypothesis, although this is not a conclusive finding, given the limited amount of data available at this stage.

The general procedure adopted has been that evolutionary changes in TFBSs can be related to their flanking DNA. To do this, hypotheses should be made, which predict how evolutionary changes in a TFBS are affected by other nearby TFBSs; these can be tested using methods similar those used in this paper. The hypotheses can be based on ideas about the structure or workings of regulatory regions (for example, the "enhanceosome" structure was used to predict the "correlated-evolution" hypothesis), and so be a means of testing such ideas. Thus a TFBS is not treated as an isolated unit - rather the procedure emphasises how the evolution of a TFBS can be related to the regulatory region of which it forms a part. The actual data analysis does not require detailed knowledge about the structure of regulatory regions, since (apart from the TFBS being studied) the only data required is percent identity of sequence. However, knowledge about regulatory regions can be used to construct explanatory hypotheses, perhaps in a more detailed form than has been described here; for instance, one could include (and thus test) numerical values for the density of TFBSs, etc.

Methods

Finding TFBSs that have evolved within the vertebrate lineage

The bioinformatics method was developed to find a sample of TFBSs that appeared to be conserved in the vertebrates studied, and also a sample of TFBSs that appeared to be diverged. Here "diverged" is defined as a change in sequence that is predicted to disrupt binding. The outline of the method is as follows, and is illustrated by a “worked example” in figure 5.

A "completely bioinformatics" method would use bioinformatics to both discover TFBSs, and to determine whether they were conserved. This was rejected because bioinformatic TFBSs produce considerable numbers of false matches (Goessling et al., 2001). Instead, the analysis started with a list of TFBSs already known to exist; these came from the TRANSFAC database (Matys et al., 2006). These known TFBSs were then analysed to find examples of TFBS conservation or of TFBS divergence, a principle that has been used in other studies (Sauer et al., 2006),(Dermitzakis et al., 2003).

For a particular TFBS, the TRANSFAC database would provide the name of the gene it controlled. Using this, a orthologous gene in each other vertebrate species was found, using either the HomoloGene or Santa Cruz websites (Wheeler et al., 2006), (Kent et al., 2002), (Kent, 2002). For every species at least 5,000 bases of upstream sequence was requested from the Ensembl (Birney et al., 2006) or Santa Cruz genome websites.

The TRANSFAC database supplied the DNA sequence of the known TFBS, which was used to find its exact location in the genomic upstream sequence. For this purpose, a sequence at least 10 bases long was required from the

Figure 5 - Example illustrating analysis of a particular TFBS

From the version of TransFac that is available free for non-commercial use, it states that 2440 bases upstream of the rat *TAT* gene, there is a TFBS for the transcription factor HNF-3alpha, which has the following sequence:

```
tagaacaaacaagtcctgcgt
```

From the rat genome, 5kb of sequence upstream of the *TAT* gene was obtained. This was searched for an *exact match* to the TFBS sequence specified by TransFac, and the exact match was found:

```
TFBS sequence from TransFac:  tagaacaaacaagtcctgcgt
Matches:                      |||||||||||||||||||||
rat genome seq:  tgtccaacaagactagaacaaacaagtcctgcgtagtcgcctgtcggtttctgggtgtggtgg
```

The sequence from TransFac may be slightly longer than the true TFBS. Certainly it was longer than the HNF-3alpha PWM. By using the PWM to search the sequence from TransFac, plus 7 bases on either side, the following sequence was found to be the TFBS corresponding exactly to the PWM: tagaacaaaca. This was a good match, shown by the high match-score of 0.984, so analysis of this TFBS was continued with.

Sequence from the cow genome, upstream of the *TAT* gene, was obtained and aligned against the rat genome sequence using Blastz. Here is the portion of the alignment that contains the rat TFBS:

```
.                              TFBS in rat, based
.                              on TransFac data:
.
rat genome seq:  tgtccaacaagactagaacaaacaagtcctgcgtagtcgcctgtcggtttctgggtgtggtggt
Matches:         ||||||| |||| | |||     ||                         |||    |  ||
cow genome seq:  tgtccaaaaagagtggaatggctaa-------------------------ctgaacataatg--
.
                            Possible orthologous TFBS?
```

Therefore, tggaatggcta is the cow sequence that is homologous to the TFBS in rat. However, this sequence is a poor match to the PWM for HNF-3alpha, as it only obtained a match score of 0.547. Thus it is unlikely that this sequence is capable of binding HNF-3alpha. Consequently, this rat TFBS does not have an orthologous TFBS in cow.

This rat-cow comparison was therefore classified as a TFBS that had “diverged”.

database, as use of shorter sequences would be too likely to produce a match in the wrong location.

Using Blastz (Schwartz et al., 2003), the upstream sequence from one species was aligned pairwise with the corresponding sequence from the species that contained the known TFBS. If this alignment was successful, then the DNA sequence of a "possible orthologue TFBS" was obtained - that is, the sequence corresponding exactly to the known TFBS. But if this alignment was not obtained, a conclusion of "uncertain" would be given. This was because, although failure to obtain alignment might be due to a regulatory region that was not conserved, it might however be due to a technical problem, such as

- a gene had been wrongly identified as a orthologue;
- the available mRNA sequence was missing exon 1, so the transcription start site was assumed to be in the wrong place;
- the available genome sequence was missing sequence in the upstream region;
- the upstream sequence was conserved but not strongly enough to be detected by the alignment program;

With the procedure actually used, these problems would cause a conclusion of "uncertain", and not an incorrect conclusion of "diverged".

Provided the "possible orthologue TFBS" was obtained, it was then necessary to decide if it would bind the transcription factor. In principle, this could be done either by comparing the sequence with a consensus sequence, or else by scoring it against a "position weight matrix"(PWM). This research used PWMs from the TRANSFAC database - a different PWM for each transcription factor - those used being based on known TFBSs that exist in a genome. Comparing a possible orthologue TFBS to a PWM produced a "score" between 0 and 1, indicating how closely it matched. The TFBS would be classified as "diverged" if it met both the following criteria: (i) the score was below a threshold; (ii) the score was substantially less than the score of the known TFBS. If the possible orthologue TFBS met one but not both of these criteria, a conclusion of "uncertain" would be given. If it met neither criteria, it would be labelled a "conserved" site. The numerical values used in these criteria were set using the scores of the known TFBSs upon which the matrix was based; for criterion (i), the threshold was set as the 20 percentile score minus one standard deviation; for criterion (ii), "substantially less" was set as twice the standard deviation.

Thus, for each known TFBS, the process just described would cause a label of "conserved" or "diverged" or "uncertain" to be given to each of the seven vertebrate species that were used as comparisons.

In a high proportion of cases, no definite conclusion was drawn, partly due to missing data, and partly because many cases were put in the “uncertain” category. Of 1697 TFBSs retrieved from TRANSFAC for the main study, only 431 were compared against other species, as the other 1266 could not be analysed due problems with the data; the most common reasons for this were that (i) the TFBS from TRANSFAC was not found in the genomic sequence used, (ii) the match-score of that TFBS against the PWM was too low, (iii) failure to identify any othologous genes in the other vertebrates, (iv) failure to retrieve genomic sequence from the location expected to contain the TFBS.

431 TFBSs were compared against other species; since eight species were used, each TFBS could in principle be compared against seven other species, thus giving 3017 comparisons (in principle). In fact, data for all eight species was not available in every case, so only 2666 comparisons were actually possible. Of these, 1780 were put into the “uncertain” category. The underlying policy was that a case should be put into the “uncertain” category if there was anything to cause doubt about putting it in the “conserved” or “diverged” categories.

When a “possible orthologue TFBS” contained gaps, it was not automatically classified as “diverged”; a gapfree version of the possible orthologue TFBS was examined using the PWM, and if the match-score was too high to meet the criteria for a diverged TFBS, then it was classified as “uncertain”.

It was realised that a slightly misaligned sequence could cause a TFBS to appear to be diverged when it was, in fact, conserved (Pollard et al., 2006). As a precaution against this, the sequence around the “possible orthologue TFBS” was searched (50 bases either side) and a high-scoring match to the PWM would cause it to be classified as “uncertain” when it would otherwise have been classified as “diverged”.

So that findings from the exploratory phase of the study could later be verified, a "reserve list" was compiled of PWMs that were not to be used except for verification; each PWM represented a transcription factor and a list of TFBSs that were only to be used for verification. Apart from the first few PWMs, PWMs selected for use in this study were put into pairs; purely at random, one PWM in a pair would be selected for use in the main part of the study, and the other placed on the reserve list, producing 25 "main study" PWMs and 14 "reserve list" PWMs.

Fictional TFBSs

To produce a "control" sample, which would mimic the effects of incorrect data, a collection of fictional PWMs and fictional TFBSs was produced. Starting with a real PWM, shuffling the rows at random would produce a "shuffled matrix" (a technique used by others (Tronche et al., 1997)). This is similar to taking a consensus sequence and shuffling the letters around at random. Effectively this describes the preferred binding sequence of a fictional transcription factor, yet some of its characteristics will be the same as those of the real one it was based on; eg, any preference for AT-rich sequences will be the same, and (for those familiar with information theory (Schneider et al., 1986)), the "information content" will be the same. For statistically reliable results, it was usually necessary to produce a large number of shuffled matrices from a single real one.

A fictional TFBS would be produced as described in the following example: Starting with a known real TFBS for the transcription factor CREB, a shuffled-matrix based on CREB would be used to search the DNA sequence 50 bases either side of the real site. If a match was found (and the match score exceeded the threshold described earlier), the matching sequence would be added to the "raw" list of fictional TFBSs. Thus, each fictional TFBS was related to a particular transcription factor, and was near a genuine TFBS for that real transcription factor. These “raw” fictional TFBSs tended to have average match scores that were lower than the corresponding average score for real TFBSs; to correct for this, some fictional TFBSs were removed (the probability of removal depended on the score), to produce a “filtered” list of fictional TFBSs. The list of “filtered” fictional TFBSs was then analysed in the same way as the list of real TFBSs. A result from the "real TFBS" analysis would be regarded as unreliable if an identical result was obtained from the "fictional TFBS" analysis.

ChIP-chip data

As an alternative to the data obtained from TransFac, use was also made of ChIP-chip data. ChIP-chip data for the HNF-4 transcription factor was chosen as it suggested that HNF-4 “is associated with an unusually large number of promoters” (Odom et al., 2004). The PWM for HNF-4 was compiled from a published list of TFBSs (Ellrott et al., 2002). Unfortunately, ChIP-chip data only locates a TFBS to an accuracy of a few hundred bases, so the exact location was found by taking the best match to the PWM in a 500-base wide search region. This introduces the risk that the wrong location will be found. To reduce this risk, the 500-base region was also searched with shuffled PWMs, and the best match to the real PWM was only accepted if its match-score exceeded 90% of the best match-scores obtained by the shuffled PWM. This procedure had also been used with the TransFac data, but with a much smaller search region; with the ChIP-chip data, the much larger 500-base search greatly increased the risk of choosing an incorrect location.

Distinguishing “gains” from “losses” using a phylogenetic tree

This procedure used a phylogenetic tree of vertebrates that was already known. For each TFBS, it found the most parsimonious explanation for the pattern in which that TFBS was conserved or non-conserved in the various species. Here the “most parsimonious” explanation meant the explanation with the fewest gain/loss events during the evolution of that TFBS.

For some TFBSs, there might not be a single parsimonious explanation. In many cases the data could be explained equally simply either as a single gain, or as a TFBS that was present in ancient mammals and lost once in a particular lineage. Only if there was a single parsimonious explanation would an example of a non-conserved TFBS be labelled as caused by a “gain” event or caused by a “loss” event.

The phylogenetic tree used, fig 6, is supported by both molecular sequence evidence (Murphy et al., 2001) and morphology (Shoshani and McKenna, 1998).

Figure 6 - Phylogenetic tree of vertebrates used to distinguish “gain-of-TFBS” from “loss-of-TFBS”

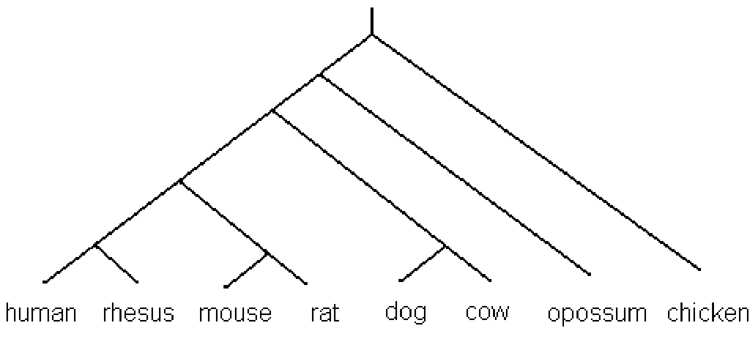


Figure 6 legend: This shows the branching order only; the branch lengths are not to scale.

Statistical estimation by bootstrapping

The standard errors and p-values were estimated by a bootstrapping method. A bootstrap sample would be chosen by selecting an experimentally known TFBS at random, and putting that TFBS in the bootstrap sample; then another TFBS would be chosen at random, and so on until the bootstrap sample contained as many TFBSs as the real sample. Thus a particular TFBS might be copied into the bootstrap sample more than once, or once, or not at all. It is important to note that the "unit of selection" was the experimentally determined TFBS: for example, if the procedure selected a particular TFBS experimentally shown to exist in humans, then all the information for that TFBS (including the results of the human-mouse comparison, the human-dog comparison, etc) would automatically become part of the bootstrap sample. Generally 100 bootstrap samples (or sometimes 1000) were used. For each bootstrap sample, the value of a particular statistic would be calculated, for instance the average %id for flanking DNA. The standard deviation of the 100 values from the 100 bootstrap samples would be used as the standard error of the corresponding value from the real sample.

Authors contributions

CRL designed the research, analysed the data and wrote the paper. SO supervised the project.

Acknowledgements

CRL received financial support from the BBSRC.

The research used publicly available genome sequence data provided by the following projects:

Chicken genome: (Hillier et al., 2004)
Cow genome: The Bovine Genome Project,

http://www.hgsc.bcm.tmc.edu/projects/bovine/
Dog genome: The Broad Institute, and Agencourt Bioscience
Human genome: The Human Genome Sequencing Consortium
Mouse genome: (MGSC, 2002)
Opossum genome: The Broad Institute
Rat genome: The Rat Genome Sequencing Consortium
Rhesus genome: Baylor College of Medicine Human Genome Sequencing Center, and the Rhesus Macaque Genome Sequencing Consortium,
http://www.hgsc.bcm.tmc.edu/projects/rmacaque/

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)