Chapter 1

INTRODUCTION AND LITERATURE REVIEW

1.1 THE FIELD OF RESEARCH CHOSEN

It is well known that an important mechanism by which genes are controlled is the binding of transcription factors (TFs) to transcription factor binding sites (TFBSs). Changes in the control of genes are likely to be an important component of evolution, of which changes to TFBSs will be a part. The evolution of TFBSs was therefore thought to be a worthwhile area within which to work.

As my research work previous to this PhD involved TFBSs within vertebrates, the evolution of such TFBSs was chosen as the subject matter.

If any reader wishes to read a shorter description of the research project before (or instead of) reading the whole thesis, they should read the shorter description included as Appendix D.

1.2 BACKGROUND INFORMATION ON REGULATION OF GENE EXPRESSION IN EUKARYOTES

1.2.1 Regulation by transcription factors

First, a general outline will be given containing basic knowledge of the mechanisms of transcription and its control.

Transcription in eukaryotes is carried out by the molecular complex "RNA polymerase II" (PolII). Also, "at least 50 proteins, in addition to the core polymerase, can be involved in transcribing a gene" ((Ptashne and Gann, 2002), p60-2). Their list of these proteins includes: general transcription factors; factors associated with the TATA-binding protein; proteins of the "mediator" complex; proteins that acetylate histones or de-acetylate them; and nucleosome modifiers (however, some of these are only required for the transcription of certain genes).

Transcription is controlled by "Transcription Factors" (TFs) that can bind DNA and then activate or repress transcription.

Let us note activation mechanisms first. Some argue that "activation-by-recruitment" is the main activating mechanism ((Ptashne and Gann, 2002), p115-118). That is, a TF will bind to DNA and also to one or more of the transcriptional machinery proteins mentioned above; merely by keeping the transcriptional proteins close to the DNA, this increases the probability that the full transcriptional machinery will assemble on the DNA at that point. In higher eukaryotes, the control of a gene will often depend on many TFs; many combinations of different activators can work synergistically - not necessarily by binding together, but by simultaneously helping recruit the transcriptional machinery at various points. It has been speculated that "certain non-acidic activators can recruit only single components of the transcriptional machinery, whereas acidic activators can recruit them all" ((Ptashne and Gann, 2002), p115-118).

Complexes of TFs can form even if the interactions are too weak for this to happen in solution, because the DNA acts as a "crystalisation seed" ((Alberts et al., 2002) p407). These complexes can include "coactivators" or "corepressors" that do not directly bind the DNA.

DNA tends to exist as "nucleosomes", that is DNA wound round histone proteins, and another mechanism of activation is to alter this so as to make transcription easier. The general transcription factors seem unable to assemble on a "conventional nucleosome". The alterations are of two types: (i) chemical - by recruiting "histone acetylase" molecules to acetylate the histones, which can help to bind other transcription factors; (ii) physical - by recruiting "remodellers", which alter how the DNA is wound onto the histones, making the DNA more accessible ((Alberts et al., 2002) p404).

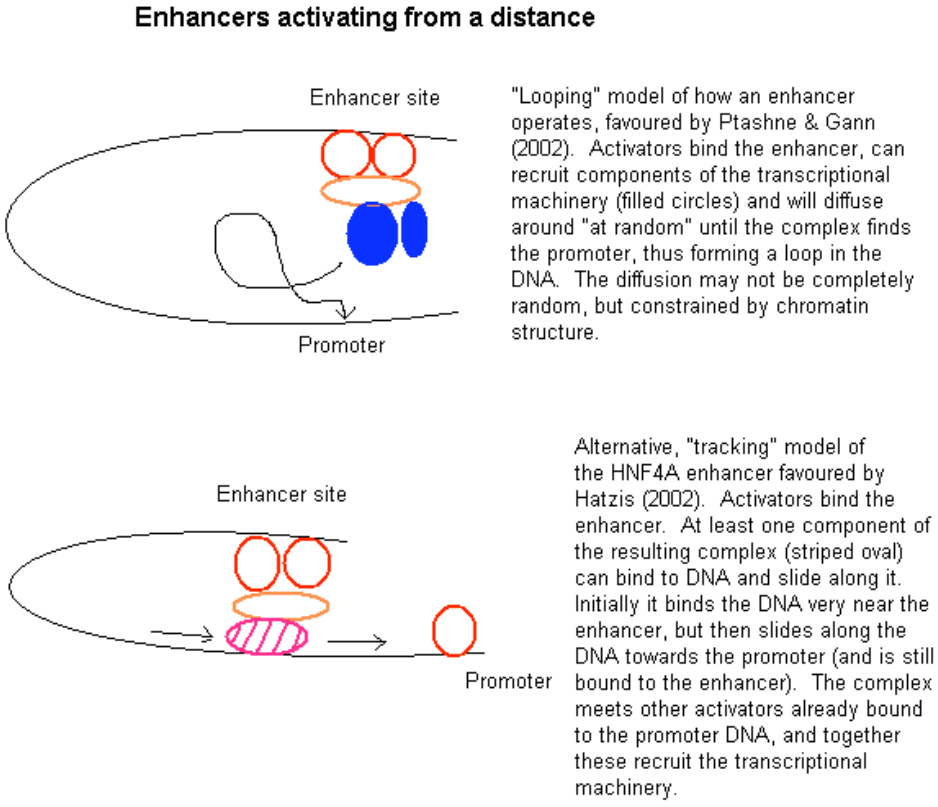
In higher eukaryotes, TFs can bind to "enhancers" and activate a promoter thousands of bases away (Reinitz et al., 2003), but it is a matter of controversy how they do this. Two possible models are presented in figure 1.1. Some argue against the "tracking" model ((Ptashne and Gann, 2002) p129-132), but others present evidence supporting it (Hatzis and Talianidis, 2002).

Turning to repression, there are a number of mechanisms, and "many" eukaryotic repressors use more than one mechanism ((Alberts et al., 2002), p406). The mechanisms include binding to the DNA and thus obstructing any DNA binding sites that overlap; binding to the activation surface of an activator and thus blocking it; interacting with the transcriptional machinery; and altering the nucleosomes, in the opposite ways to those described above for activators.

Reinitz (Reinitz et al., 2003) has proposed classifying repressors into three groups depending on their range of action - (1)repressors that use overlapping binding sites, (2) "quenchers" that repress up to about 150 bps, and (3) repressors that repress at a range of 1kb or more.

"Insulators" in the DNA prevent long distance action - if an insulator is inserted between an enhancer and a promoter, then the enhancer will no longer be able to activate that gene. An insulator is a stretch of DNA that

Figure 1.1: Two models for how enhancers operate
Two alternative models for how an enhancer can activate transcription at a distance of several thousand bases.



binds specialised proteins ((Alberts et al., 2002) p413).

Next, we shall consider how TFs bind to DNA; the places where they do so are called "transcription factor binding sites" (TFBSs). The DNA sequence is important in determining if a TF will bind at that point. Whilst there has been limited success in predicting the DNA sequence to which a given TF will bind (Choo and Klug, 1994), on the whole there is nothing equivalent to the triplet-protein-code. In practise, the DNA sequence bound by a TF is found by experiments on that TF, and not by prediction. Moreover, a particular TF will not have a single exact binding sequence, but rather will bind a number of similar sequences.

TFs very often form "homodimers", consisting of two identical molecules bound together. Both molecules will contact the DNA, and since they have the same binding properties, the preferred DNA sequence will contain the same sequence twice; this sometimes consists of a direct repeat, but often consists of an inverted repeat. For example, the "nuclear receptor" family of TFs contains examples of both types of repeat (Mangelsdorf et al., 1995). The term "half-site" is sometimes used to refer to the DNA sequence bound by a single molecule of the homodimer. The two half-sites may be separated by one, two or more bases of "spacer" DNA whose sequence has relatively little effect on binding - again, the nuclear receptors contain a range of examples - but it would be wrong to say it has no effect at all. For example, the central "spacer" base in the HNF4 binding sequence (which we will meet in more detail later - table 3.1) tends to be the base A, whereas any base can form the central "spacer" base in the HNF1 binding sequence (Tronche et al., 1997).

Bases at the edge of the TFBS may also have a weak effect on binding, which may be difficult to observe with the amount of data available. For example, one group presents evidence that the binding of STAT TFs extends over at least 13 bases, yet they also present a representation of the preferred binding sequence which is only 9 bases long, and evidently omits the weakly binding bases at the edge (Horvath et al., 1995). A representation of this type is sometimes refered to as the "core". Consequently, for a particular TF, the exact size of its binding sequence can vary depend-

ing on where you look it up in. For example, descriptions of the HNF1 binding sequence can be 13 bases long (Crabtree et al., 1992) or 15 bases (Tronche et al., 1997),(Krivan and Wasserman, 2001), and the latter two are displaced by one base, suggesting some uncertainty about the exact location of the edge of a HNF1 TFBS.

TFs can also form heterodimers, where the two molecules that bind are not identical.

As for TF structure, several DNA-binding structures are found in eukaryotic transcription factors ((Alberts et al., 2002) p384-390). The helix-turn-helix (HTH) structure, common in prokaryotes, is also found in eukaryotes, including a particular type (the homeodomain) often used by developmental genes. Other structures include the zinc finger, leucine zipper and helix-loop-helix (HLH) structures. The latter two contain a dimerisation domain and a DNA-binding domain in a single motif. A feature of some zinc fingers is that a "modular design" is possible, with a number of zinc fingers in a single molecule; as an example of a molecule with a large number of zinc fingers, the EVI1 molecule contains ten zinc fingers (Delwel et al., 1993). The "classic zinc finger" is extremely common in humans - in the sense that many genes encode proteins of this type - but appears not to exist in bacteria (Muller et al., 2002).

TFs are themselves regulated. A relatively simple example involves the steroid hormones; a steroid hormone outside the cell (eg, Estrogen) can pass through the cell membrane and bind to a TF (in this example, the Estrogen Receptor TF), activating it (Mangelsdorf et al., 1995). More complex examples exist: for example, an MKKK molecule that has been activated can phosphorylate an MKK molecule; the MKK molecule will then phosphorylate a MAP-kinase molecule, which then phosphorylates a transcription factor, affecting its activity (Whitmarsh, 2002). One reason for having a complex pathway is that the cell may need to respond to signal molecules outside the cell that are unable to enter it. As an example of this (from *Drosophilia*), the "Punt" and TKV molecules lie within the cell membrane; a DPP molecule *outside* the cell can bind to these, and when this happens, it activates the

part of these molecules *inside* the cell. This causes phosphorylation of a MAD, a molecule that sometimes resides in the cytoplasm. However, once phosphorylated, MAD dimerises and then translocates to the nucleus, where it affects the transcription of certain genes (Inoue et al., 1998). Thus, in effect, a "signal" is passed through the cell membrane even though no molecules physically pass through, but several proteins are needed to pass one signal.

1.2.2 Other methods of gene regulation

Genes can be silenced by methylation, which involves methyl groups being attached to C bases. Not all C bases are methylated this way; only a C followed by a G is used.

This methylation of Cs has an interesting side effect. A *methylated* C can spontaneously undergo a chemical change into a T, and because this occurs at a high frequency compared to other mutations, a C followed by a G is rarely found in many regions of vertebrates genomes, because most of them have been removed by this mutation. But there are exceptions - notably, the promoters of "housekeeping" genes are not usually silenced by methylation; consequently, in these regions C followed by G will not be removed by this mutation mechanism, and hence occurs more frequently than elsewhere in the genome. Regions with a high frequency of C followed by G are known as "CpG islands" ((Alberts et al., 2002) p434). Since CpG islands can be detected *in-silico*, they can be used to help predict where promoters are in a genome sequence. Examples could be given (Ioshikhes and Zhang, 2000) of promoter-finding software that uses this technique; but an important limitation is that a substantial proportion of genes do not have CpG islands round their promoters, so these will not be detected.

Other methods of regulating genes are associated with the various stages of converting a transcribed gene into a working protein.

The RNA produced by transcription will usually (in higher eukaryotes) be

spliced to remove the introns. For some genes, an exon will sometimes (but not always) be spliced out as well, thus causing two (or more) different versions of a protein to be made by a single gene. This can be regulated by proteins that bind the RNA transcript ((Alberts et al., 2002) p436-439). This form of regulation may be subject to evolution that is rapid enough to produce notable differences between human and mouse; one study considered a sample of 62 genes with alternative splicing, and found that 10 of these genes had a splice variant in mouse that was not found in human (Nurtdinov et al., 2002).

Other stages of RNA processing, which can be controlled so as to regulate the effects of a gene, will be briefly mentioned. They include RNA editing (when extra bases are inserted into a trancript RNA, the insertion locations being determined by another "guide" RNA). Translational control proteins can bind mRNA in such a way as to prevent ribosomes translating it into protein, somewhat analogous to the action of certain TFs (the analogy is with certain repressors, not with activating TFs). Some mRNAs have more than one translation initiation position (either by having more than one start codon, or by having a site that enables ribosomes to commence binding at a position within the mRNA as well as at the start), enabling translation to occur when normal translation has been repressed. The rate at which mRNA is degraded can be altered (the time a mRNA lasts before degradation depends on a mechanism in which its poly-A tail is slowly shortened; therefore, control of this often depends on binding to sites in the mRNA near the poly-A tail). The degradation rate of a mRNA can be altered by proteins that alter the rate at which the tail is shortened, or clip it off altogether, or even by lengthening the poly-A tail of selected mRNAs ((Alberts et al., 2002), p440-450).

RNA interference is a method of gene regulation that has received much attention in recent years, partly because of its use as an experimental technique - however here we are concerned with natural regulatory mechanisms. The process starts with *double-stranded* RNA; this is recognised by the enzyme "dicer", cleaved into single-stranded RNA that is 20-25 bases in

length, and then each of these RNA strands is incorporated in a complex called "RISC" (one RNA strand per complex). RISC will then degrade any RNA molecules it can bind (and this binding will occur when the RNA molecule has a sequence complementary to the RNA incorporated in RISC) (Matzke and Matzke, 2004). Alternatively, especially in animals, the action may involve blocking the movement of ribosomes along the mRNA, rather than degrading it (Zhang et al., 2006).

Thus, the overall effect is that the presence of any double-stranded RNA in a cell will cause the later degradation/repression of single-stranded RNA which includes the same sequence. Whilst double-stranded RNA may be present because of viruses or experimenters, the double-stranded RNA most relevant to this thesis is that transcribed from the cell's own genome ("miRNAs"; transcribed as single-stranded RNA, its sequence will enable it to adopt a hairpin structure that includes double-stranded RNA). Thus transcription of a miRNA can cause degradation of a particular mRNA and, in effect, down-regulate the gene coded by that mRNA. In plants, most of the genes that are regulated by this mechanism are transcription factors (Zhang et al., 2006). It has also been suggested that regulatory RNA molecules can bind directly to DNA (Corey, 2005), and that RNA-directed silencing can cause methylation of histones and thus affect transcription (Matzke and Matzke, 2004) (see page 19 for an outline of histone modification).

Another method of regulation is the post-translational modification of proteins. One type of this is phosphorylation of a protein - for an example, see the MAP kinase example on page 22. Another type is the "small ubiquitin-related modifier" (SUMO), a class of proteins which are small (about 100 residues) and which affect the activity of a protein when SUMO is attached by an enzyme. Many transcription factors have an attachment site for SUMO, and in most of these cases, the transcription factor will be repressed by attaching SUMO (Gill, 2005).

1.2.3 More detailed studies of TFs and TFBSs

Physics of TF binding

An early study of TFBS specificity examined the underlying physics in some detail (von Hippel PH and Berg, 1986). Unambiguous recognition of particular base-pairs depends on hydrogen bonds forming between the bases and the TF protein. They assume each bond has an energy of -0.5 kcal/mol (allowing for the bonds with water molecules that must be removed to allow the protein to bind the DNA). Given the size of a typical TFBS, this gives a protein-to-base binding energy of -6 to -12 kcal/mol of protein bound. If a single hydrogen bond with water is broken and *not* replaced by DNA-protein bond, an unfavourable energy of +5kcal/mol can be added; thus, a single "broken bond" of this type is enougth to severely destabilise the entire DNA-protein interaction. The -6 to -12 kcal/mol is unlikely to provide sufficiently strong binding on its own, but will be assisted by bonding between the phosphates of the DNA and positively charged protein side chains of the TF. Of course, the presence of the phosphates does not depend on the sequence, so this extra energy is not sequence-specific.

Thus, when binding DNA of the "right" sequence, the binding energy is made up of a weak amount of favourable energy from interactions with the bases, which is greatly added to by favourable energy from interactions with the phosphates. When attempting to bind DNA of the "wrong" sequence, the binding energy is made up of a strong amount of unfavourable energy from interactions with the bases, plus the favourable energy from interactions with the phosphates, perhaps resulting in no binding affinity at all. However, in the latter case, the authors suggest the protein may bind with a different conformation in which the only interactions are with phosphates; perhaps this allows the protein to slide along the DNA.

They present equations relating the binding affinities to concentrations of binding sites. There is probably a *maximum* permissible binding affinity, for two reasons. First, the half-life for dissociation needs to be smaller than cell cycle times. Second, for repression by competition, almost 100% occupation

of a TFBS is required; but the total number of molecules of the repressing TF in the cell will fluctuate - so the average total number of these molecules must *exceed* the number of TFBSs, by a sufficient margin to ensure 100% occupation even when the number of molecules of TF fluctuates below average (they give an equation for this). But if the number of molecules of TF is substantially greater than the number of TFBSs, then a finite binding affinity will give almost 100% occupation and there is no point having a higher binding affinity.

They note the problem of "pseudo-sites", by which they mean sequences in the DNA that do not bind as strongly as the genuine TFBSs, yet which bind strongly enough to remove some TF molecules that would otherwise be free. The basic problem is that the number of possible sequences that form a pseudo-site is very large, so that a large number of pseudo-sites will appear in the genome; consequently a number of pseudo-sites will be occupied at any one time. (The low occupancy of an individual pseudo-site is counterbalanced by the large number of pseudo-sites). This could appreciably reduce the number of TF molecules available to occupy genuine TFBSs. To prevent this becoming too serious a problem, the TFBS may have to be longer than it would otherwise be. In their example, if every "incorrect" base-pair in a sequence causes the binding constant to reduce by a factor of 30, then for an *E. coli* sized genome, five molecules will be "lost" by binding to pseudo-sites, if the TFBS is 16 bases long. To be unique in the genome, the TFBS would only have to be 12 bases. However, with a 12-base TFBS, each cell would contain hundreds of molecules of that TF that were "lost" binding to pseudo-sites (in their example), and were thus a waste of protein-production resources. Hence, a 16-base TFBS may give higher fitness than the 12-base TFBS merely because it avoids this waste.

Representing the sequences bound by a TF, and using that representation to score any sequence

One method of representing the preferred binding sequence of a TF is by a "consensus sequence", which in its simplest form, is simply the DNA sequence

that binds with maximum affinity.

However, the real TFBSs for a particular TF are usually similar but not identical, so it is desirable to represent these alternative sequences. One method of doing this is to use a consensus sequence which includes degenerate letters, for example, a W in the sequence means that either an A or a T could occur at that point. But that method is unable to represent more subtle differences in frequency - for instance, a position that was A in 65% of TFBSs and T in 35%. Therefore, consensus sequences are often considered to be an overly-simplified form of representation, and many studies use position weight matrices (PWMs) instead. The work in this thesis was done using PWMs, rather than consensus sequences.

A position weight matrix is the most common way of representing the DNA sequences that a particular transcription factor will bind. For a typical TF, the size of each TFBS is the same, and so the PWM is of fixed length, and does not allow for variation in length. (However, some TFs do have TFBSs of variable length, and methods to deal them have been developed (Roulet et al., 2000), but this is not standard practice). The PWM will be derived from another matrix (which I will call the "frequency matrix"); the frequency matrix shows the frequency with which each DNA base occurs in each position in known TFBSs. In this section, "PWM" will be used only to refer to the derived matrix. Elsewhere in this thesis, however, "PWM" will refer to the frequency matrix.

The method of deriving the PWM from the frequency matrix is not perfectly standardised, so equations from several different authors will be quoted next, to indicate the variations in use.

Usually, each number in the PWM is based on the logarithm of the frequency. An early paper (Staden, 1984) gave the following argument for doing so. When examining a sequence that might be a TFBS, the elements of the frequency matrix "can be thought of as probabilities of each base being part of the recognition sequence and hence their product is the probability that the section of the sequence scanned is a recognition sequence." (Whether that conclusion is really correct will be discussed in a moment). The under-

lying idea is that because the frequencies represent probabilities (he means frequencies expressed as a proportion of the total number), they should be multiplied to give an overall probability P for the sequence:

$$P = \prod(n_{bi}/(n_{Ai} + n_{Ci} + n_{Gi} + n_{Ti})) \tag{1.1}$$

where n_{bi} is a term in the frequency matrix, which gives the number of times base b occurs at the ith position in the binding sequences.

Taking the logarithm gives,

$$\ln(P) = \sum(\ln(n_{bi}/(n_{Ai} + n_{Ci} + n_{Gi} + n_{Ti})) \tag{1.2}$$

and note that the logarithms of the frequencies are being added. Hence the elements in the PWM were based on *logarthims* of frequencies, so that when using the PWM, calculations would be based on addition rather than multiplication. (Where any cell in the frequency table contained an observed value of zero, it would be replaced by a value equal to the reciprocal of the number of sequences used to construct the table).

It should be questioned whether P really represents "the probability that the section of the sequence scanned is a recognition sequence." For example, consider a homodimer TF that binds two half-sites which are separated by a "spacer" base that has little effect on binding. Each of the four bases will appear in the "spacer" position with about 25% frequency, and this alone means the value of P will never exceed 0.25. Thus if P really is the probability, for this TF no sequence could ever have more than a 25% probability of being a binding site, no matter how large the binding affinity of the rest of the sequence. This seems implausible; it suggests that P is not itself the probability that a sequence is a TFBS (though it may be a useful step in calculating an estimate of that probability).

Another author adopted a similar (but not identical) approach (Bucher, 1990), which can be described by this equation for deriving the PWM from the frequency matrix (Bucher, 1990):

$$w_{bi} = \ln(n_{bi}/e_{bi} + s/100) + c_i \tag{1.3}$$

where w_{bi} is an element of the PWM, e_bi is an expected "background" frequency, s is the "smoothing percentage", c_i a column-specific constant, and the rest as defined above. The exact value of s varies but is typically a few percent (though other authors (Tsunoda and Takagi, 1999) use a simple $s = 1\%$ throughout). s is of most importance when n_{bi} is zero, since otherwise (without s) the equation would require the mathematically impossible task of taking the logarithm of zero; and also, it seems unsafe to assume the frequency is exactly zero, merely because we have observed no cases in a sample of limited size.

Another slightly different equation (Krivan and Wasserman, 2001) that has been used is, in effect:

$$w_{bi} = \log_2(\frac{(n_{bi} + \sqrt{N}/4)/(N + sqrtN)}{0.25}) \tag{1.4}$$

where N is the number of TFBSs the matrix is based on. It will be noted that here, a simple background frequency of 0.25 for each base was used, rather than the more complicated earlier proposals (Bucher, 1990) to base "background" on nucleotide frequencies or even dinucleotide frequencies. Another difference is that the smoothing parameter s has been replaced by the term $\sqrt{N}/4$, which evidently serves the same purpose, but the paper contains no explanation why this term was chosen. Perhaps the reason is as follows: basic statistical theory states that if you take a sample of N items, of which a proportion p is a particular type (in the population), then the number of items of that type in your sample will be subject to a sampling error of $\sqrt{p(1-p)N}$; hence if a DNA base occurs with a lowish frequency of 7% in a particular position, then sampling error will be (by this formula) $\sqrt{0.07 * (1 - 0.07)N} \approx \sqrt{N}/4$. Thus the values the $\sqrt{N}/4$ term gives are likely to be broadly correct, but perhaps the justification for the term depends on some somewhat arbitrary assumptions.

Summarising the PWM formula from the four papers cited above, it can be seen that they are all broadly the same, in that they are slightly elaborated versions of taking the logarithm of n_{bi}. However, there is less agreement about what to do when $n_{bi} = 0$; everyone deals with this by, in effect, adding a small positive number to n_{bi}, but with no agreement on how to choose that small positive number.

A more radical change from this system has been applied to the HNF1 transcription factor (Tronche et al., 1997). In that case, the PWM used was identical to the frequency matrix (except when there was a zero in the frequency matrix). Thus, they omitted the step of taking the logarithm, which sounds like a quite fundamental change mathematically. Each zero in the frequency matrix was dealt with by replacing it by -99, which is also a drastic change from the "add a small positive number" concept met with in all the other papers cited above. Nevertheless, despite these drastic changes, the procedure was successful at predicting DNA sequences which did bind HNF1 (when tested *in vitro*). This suggests that the broad concept of the PWM is sufficiently robust that it will still work (under some circumstances) despite major changes from the standard procedure. Oddly, this particular *in vitro* test has been cited as evidence for the reliability of the more conventional PWMs (Krivan and Wasserman, 2001).

PWMs can be used to search DNA sequences *in-silico*, but a more detailed explanation of how this can be done will be given in a later chapter (table 2.5).

Discovering PWMs using over-represented words

In recent years, one area of research has been PWM discovery using bioinformatic techniques that rely on finding over-represented "words" in genomic sequence. Sometimes the results are not literally PWMs, but rather a collection of short DNA sequences that serve broadly the same purpose (of describing the DNA sequence that a TF binds to). Here is an example of

this type of research, which did use PWMs (other examples of searches for over-represented words will be given later, pages 37 and 38).

The study (Xie et al., 2007) used a combination of two methods: searching genome sequence for over-represented motifs, and cross-species conservation. The search did not cover the entire genome, but only those sequences showing strong conservation in a comparison of 12 mammals. The initial search identified k-mers (k>=12) that occurred more frequently in the conserved sequence than elsewhere. At this stage, each k-mer was a particular DNA sequence, with *no* ambiguity letters - thus the "wobble" often associated with TFBSs was not allowed for. But it was allowed for in the second stage, when over-represented k-mers were grouped together if they had very similar sequences; then, the k-mers in a group would be used to build a PWM (or "motif"). 233 of these motifs were discovered; some matched regulatory elements that were already known.

The statistical testing used a method in which the columns of a PWM were permuted, but in such a way as to preserve any CpG dinucleotides. These were used as "controls".

Sequences matching the motifs were conserved much better than the controls (this was true even for motif-matching sequences that were *outside* the conserved regions used in the initial search). Also, the motifs showed similar patterns of cross-species conservation and within-species conservation. This suggests that, during mammalian evolution, there has been little or no change in the DNA-recognition properties of the molecules that bind the sites.

Although the motifs were 12-22 bases long, they tend to be found in blocks of conserved sequence that are wider than this; the median width of these conserved blocks was 112 bases or 96 bases (depending on the method used). The authors thought this indicates that other regulatory elements are present next to the discovered sites.

Of the individual motifs, the most interesting (and most frequent) was LM2. A wet-lab experiment showed that the LM2 sequence had an affinity to the CTCF protein, out of all the substances in a nuclear extract. CTCF was

known to act as an "insulator", that is, it prevents an enhancer on one side of the insulator from affecting the transcription of a gene on the other side of the insulator.

Matches to the CTCF motif were found in many vertebrates, including pufferfish; the number of matches in each species was similar despite a 5-fold variation in genome size (perhaps because the number of genes does not vary much and CTCF is related to that).

1.2.4 Regulatory DNA: features beyond conventional TFBSs

It has been suggested that the DNA *structure*, rather than the sequence, may be an important feature of regulatory elements (Greenbaum et al., 2007). Although the "backbone" of a DNA molecule forms a helix, the exact shape of the helix can vary. To study this experimentally, DNA molecules were treated with hydroxyl radicals. These cleave DNA, but the ability to cleave the DNA is believed to vary depending on the local structure; hence, by measuring which cleavage fragments are formed, each base can be estimated to be highly susceptible to cleavage/ or moderately susceptible/ etc. This is an indirect measure of the local structure.

Using data of this kind, an algorithm was developed that will predict the cleavage susceptibility of any sequence of DNA. This was then used to predict the cleavage susceptibility of a large sample of vertebrate DNA. These data were then searched with motif-finding software to find frequently occuring motifs. For instance, a motif might be: *the first base is highly susceptible to cleavage - the next base is moderately susceptible to cleavage - the third base is weakly susceptible to cleavage - the fourth base is highly susceptible to cleavage.*

Regions that are hypersensitive to digestion by DNase I often have regulatory or other functions; therefore regions of this type were examined using the motif-finding software; this identified three motifs. "Matches" to these motifs were 5 times more frequent in the hypersensitive regions than elsewhere.

The enrichment was even stronger if the hypersensitive region overlapped a CpG island, or if it was near a TSS. However, there was only slight enrichment for regions of the genome that are strongly conserved in a multi-species comparison. The latter finding suggests the idea that perhaps the DNA structure can be conserved during evolution even if the DNA sequence is altered; consequently, if TF binding depends partly on DNA structure (and not just sequence), TF binding could be conserved despite changes in DNA sequence. Obviously, this type of conservation would be difficult to detect using ordinary alignment procedures.

1.3 INFORMATION FROM THE LITERATURE RELATING TO THE EVOLUTION OF TRANSCRIPTION FACTOR BINDING SITES

1.3.1 Evidence of TFBS evolution in eukaryotes

Studies that identified individual TFBSs

A survey (Dermitzakis and Clark, 2002) found a number of papers that described the regulation of a gene, where experiments had been done for both human and rodent orthologues of the gene. Based on this collection, they estimated that 32%-40% of functional human TFBSs are not functional in rodents. The lower figure was based on a simple count; however, the authors argue that this is an underestimate, as biologists tend to regard conservation as a null hypothesis, and only report divergence if the evidence is strong. The 40% comes from a calculation which takes this into account. On the other hand, I wonder if there is a publication bias which goes in the opposite direction; for several of the papers in their collection, the paper's title emphasizes the human-rodent difference in regulation, so the authors presumably thought that the most valuable finding - perhaps making it more publishable.

Another study (O'Lone et al., 2004) examined TFBSs for the Estrogen Receptor (ER). After reviewing TFBSs for ER known to exist in humans, they compared them *in-silico* with their aligned mouse genome sequence, assessing the TF binding probabilities of the latter using the matrix comparison tool "Possum". They concluded that 60-80% of these TFBSs were not functional in mice. This striking lack of conservation might, they speculate, be due to differences in reproductive physiology. They also considered the 50bp of flanking sequence on either side of each TFBS. The stronger the binding (as measured by matrix score), the more the TFBS sequence was conserved relative to the flanking sequence. Also, for conserved TFBSs, the TFBS was conserved more strongly than the flanking sequence; for non-conserved TFBSs, the TFBS was conserved to about the same extent as the flanking sequence. On a somewhat different topic, they note that ER does not bind DNA for 35% of characterised human target genes. This suggests that often ER activates genes using only protein-protein interactions with other activators.

As is evident from the survey cited earlier (Dermitzakis and Clark, 2002), detailed information is sometimes available on the evolution of particular regulatory regions or TFBSs. Here is one case, which I summarise as an example of a detailed investigation, and also because it was not included in the earlier survey.

The CYP7A1 gene encodes an enzyme (a cytochrome P450) that helps convert cholesterol to bile acids. The feeding of cholesterol may upregulate or downregulate this gene depending on species (Chen et al., 2002) (Xu et al., 2004) (Rudel et al., 1994). Upregulation is thought to be caused by a TFBS for LXR (which is a TF belonging to the nuclear receptor family), which binds the CYP7A1 promoter in rats, but does not do so in humans (Lehmann et al., 1997) (Peet et al., 1998). On the other hand, the human promotor binds HNF-1, whereas the rat promoter does not. HNF-1 is a TF but is not likely to act as a substitute for LXR, as HNF-1 has a quite different

structure and is unlikely to bind cholesterol (Chen et al., 1999).

Thus, these studies of the CYP7A1 promoter lead to several interesting conclusions. First, there has been more than one gain/loss of TFBS (for LXR and HNF-1). Second, these changes are not likely to be compensatory, since the properties of LXR and HNF-1 are so different. Thirdly, the presence/absence of an LXR TFBS appears to have a clear phenotype, namely the ability to respond to cholesterol in the diet. For the current thesis, the most important of these is that more than one gain/loss has occurred, which suggests that either these two events are linked by having some common cause, or (if they are not linked but occured together by chance) that such events are quite common.

Another reported case of TFBS loss concerns a number of yeast TFBSs with the same binding sequence (AATTTT), upstream of a number of genes, where the TFBSs were all lost at about the same time and probably for a single reason (Ihmels et al., 2005). The reason was the evolution of anaerobic growth.

Many yeast (eg, *Candida albicans*) can only grow by using oxygen (aerobic growth); this means their mitochondria will play an important part in their metabolism; therefore, genes encoding mitochrondial ribosomal proteins (MRPs) ought to be upregulated at the same time as other genes involved in growth (they confirm this with expression data).

In contrast, some yeasts (eg *Saccharomyces cerevisiae*) grow anaerobically, which greatly reduces the need for mitochrondrial activity during growth. Consequently, during growth there is little need to upregulate genes that encode *mitochondrial* ribosomes, although, of course, there is still a need to upregulate genes that encode nuclear ribosomes. This confirmed by expression data, which shows little or no correlation between the expression levels of these two types of genes. This was in sharp contrast to the *C. albicans* expression data.

Having established that a class of genes (MRPs) was differently regulated in the two yeasts, the authors took the genes of this type in *C. albicans* and

searched their upstream regions for any common DNA sequence. AATTTT was found to be common; reporter gene experiments showed that mutating this element would drastically reduce expression, confirming it to be a cis-regulatory element.

S. cerevisiae, in contrast, did not have the AATTTT element upstream of MRP genes. But there were other classes of genes which had this element present in both types of yeast.

Having established that, so far as MRP genes were concerned, the AATTTT element occured in *C. albicans* but not *S. cerevisiae*, then the question arises whether it was gained in the *C. albicans* lineage, or lost in the *S. cerevisiae* lineage. Data from a number of other yeast species indicated that the element had been lost in the *S. cerevisiae* lineage, rather than gained in the *C. albicans* lineage.

In summary, the evolution of anaerobic respiration in *S. cerevisiae* appears to have caused the loss of the AATTTT element from a number of genes.

A genome-wide attempt to study TFBS evolution has obtained TF binding (chip-chip) data for both humans and mice (Odom et al., 2007). Thus, human-mouse comparison of many TFBSs was based on *experimental* data; unlike some other papers reviewed here, there was no need to use bioinformatics to predict the existent or absense of TFBSs in one of the species. Four TFs were used. In 41%-89% of cases, "the orthologous promoters bound by a protein in one species were not bound by the same protein in the second species", suggesting a high proportion of regulatory links are not conserved.

Moreover, even where a regulatory link is conserved, the TFBS is often not conserved: "genomic regions that are bound by the same factors in both species shows that approximately two-thirds of the binding events are not aligned between the mouse and human genomes." These not-aligned cases can be divided into two types: (i) cases where the region bound in one species is aligned to a region not bound in another species; and (ii) cases where the region bound in one species is not aligned to the other species. Type (i) cases

are slightly more frequent than type (ii) cases, and slightly more frequent than cases where TFBSs are clearly conserved. Are the type (ii) cases, which are not aligned, genuine cases of non-conserved TFBSs, or merely caused by an alignment program failing to detect orthologus sequence? The paper does not address this question; this review will mention it later, in the section describing alignment techniques.

The finding that, often, a regulatory link was conserved even though the corresponding TFBS was *not* conserved, is similar to a more specific result in yeast (about the "MCB-box", see page 39).

One finding was that "frequency of conserved motif sequences was lower near binding events that are unique to one species but was still above background" - which suggests that bioinformatic methods may sometimes indicate that a TFBS is conserved even if, in fact, it is not conserved.

A study of TFBS evolution in yeast (Moses et al., 2003) compared the genomes of four species of budding yeast. TFs were considered that already had well-characterised binding properties. TFBSs were found to be more conserved than the sequences surrounding them. Some positions within a TFBS motif tend to be more conserved than other positions. Positions that permit high variability within a genome tend to be the same positions that permit high variability between genomes.

In addition, TFBSs were predicted by taking genes with similar expression patterns and examining their promoters using motif-finding software. These predicted TFBSs showed conservation patterns similar to those mentioned.

Another study of TFBSs in yeast compared six genomes (Cliften et al., 2003), four belonging to the "sensu stricto" group, and two others being more distantly related. The latter were more difficult to align; the method used was to align protein-coding sequences and then use these to align nearby non-coding sequence. The aligned portions of genome sequence were searched for motifs, imposing a requirement that a motif should be precisely conserved in all species in the alignment. The number identified was 8 times greater

than in randomly shuffled versions of the same alignments, suggesting that the motifs found by this method have a real biological purpose.

One might expect genes will have similar expression profiles if they share a regulatory element in their promoters; therefore, a subset of these motifs were identified - for each motif, the conserved aligned examples of it were upstream of genes with a similar expression profile. Curiously, they also noticed at least one example ("MCB box") where genes with an *unaligned* conserved MCB box had a similar expression profile (that is, genes for which an MCB box was present in all four sensu strico species, yet those MCB boxes were not aligned).

They also compared the motif data with ChIP-chip data and found some associations; for example, parts of the genome containing a conserved TGTACGG motif tend to overlap with regions that the ChIP-chip data showed as binding the transcription factor Fhl1.

Comparing the two papers that have just been summarised (Moses et al., 2003) (Cliften et al., 2003), it will be noticed that the first placed much more emphasis on evolutionary changes in TFBSs than did the second (as the latter was more focussed on conserved TFBSs). Even the first, however, focussed on evolutionary changes to TFBS sequence that apparently did not disrupt their function. Thus neither paper focussed on changes to TFBSs that disrupt binding.

Studies which did not identify individual TFBSs

Some studies have aimed to describe some overall properties of evolving regulatory regions, but without producing lists of particular TFBSs that have been identifed as "conserved" or "non-conserved".

A study of TFBS evolution in fly (Moses et al., 2006) concentrated on TFBSs for which there was experimental evidence (from *D. Melanogaster*), in the form of chip-chip data. The data gave 294 regions bound by the TF "Zeste", which (when searched *in-silico*) appeared to contain 1406 Zeste TFBSs. The authors admit that many of these will be false matches to apparent TFBSs

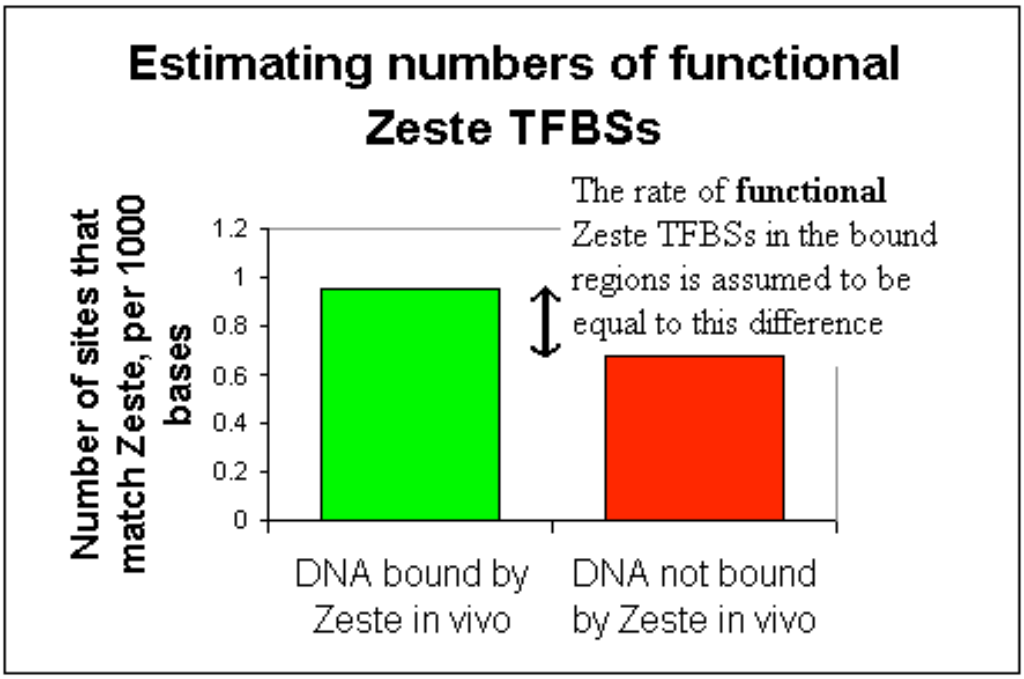
that actually do not function, and they estimate there are 807 functional TFBSs amongst the 1406 apparent TFBSs. When aligned against three other *Drosophila* species, these 1406 TFBSs were conserved more highly than predicted by a model of the "background" rates of substitution, but not so highly as predicted by a model that assumed every TFBS was conserved. This suggests that the functions of the TFBSs were conserved in many, but not all, cases.

Non-conserved TFBSs were identified using a specially-defined statistic (based on the ratio of probabilities predicted by the two models referred to in the previous paragraph), $p<0.01$ being the criteria for being non-conserved. The authors consider alignment problems will not significantly impact their analysis; this was partly because the species are close (<0.1 substitutions per site), and partly because of a modified analysis that allowed a misaligned, conserved TFBS to be identified as conserved (providing it overlapped).

215 of the "Zeste" TFBSs appear to be non-conserved; however, this estimate is reduced to 62 non-conserved functional TFBSs, after allowing for the false matches from the *in-silico* search. This is pictured in figure 1.2. This is quite a drastic reduction, which implies that the false-match rate needs to be estimated fairly accurately; for instance, although the authors do not give this example, it is easy to calculate that if the false-match rate were eventually found to be 20% higher than the authors supposed, then the number of non-conserved TFBSs would be halved. Evidently the false-matches from the *in-silico* search are a serious problem, suggesting the imperfections of chip-chip data are a serious drawback in this kind of work (remember that the *in-silico* search is needed because chip-chip data provides locations that are only accurate to a few hundred base-pairs). Of the 215 apparent non-conserved TFBSs, the method does not identify which are the 62 genuine TFBSs, so the method did not identify particular examples of non-conserved TFBSs. The aim was instead to produce statistical results that would relate to TFBSs "en masse".

Using the phylogenetic tree of the four species of fly used, the authors classify many non-conserved TFBS as being either "losses" or "gains". For "gains",

Figure 1.2: Estimating numbers of Zeste TFBSs
This summarises a published analysis (Moses et al., 2006). The vertical axes shows the rate at which an *in-silico* search finds apparent sites for the TF "Zeste". The green bar shows the rate at which these are found in lengths of the genome that are known (from a chip-chip experiment) to bind Zeste. But it is thought that many of these are not functional, because these sites are also found at a high rate (red bar) in lengths of DNA that are believed to *not* bind Zeste *in vivo*. Thus, in the bound DNA, the rate at which **functional** Zeste TFBSs are found is assumed to be given by the *difference* between these two rates. Using this difference, and the number of bases of DNA that were studied, it was estimated that 62 functional Zeste TFBSs were in the studied region.



the correction for false matches is even more dramatic: 360 apparent gains are corrected down to an estimate of 42 gains of functional TFBSs in the sample, by subtracting an estimated 318 gains that will be produced by false (or non-functional) matches. The number of real gains was greater than the number of real losses, but the size of the difference was not statistically significant, so it is possible that gain events occur at the same frequency as loss events.

Because the experimental binding data came from *D. melanogaster*, it was thought that gains in this species would be more likely to be detected than losses in this species. (Of course, this argument does not apply to all the *Drosophila* species). This prediction appears to be supported by the data, though it is not clear whether this difference is statistically significant. Confusingly, the abstract mentions this asymmetry but does not mention that it is thought to be a consequence of the "asymmetric" experimental data (ie, the binding data was collected from one species of fly only).

They observed 33 cases where a gain and a loss occurred in a single regulatory region, in such a way that the loss might compensate for the gain. Was this evidence of compensatory turnover? The authors did not claim so, as the number of cases was no greater than the number expected to occur by coincidence.

Evolution of cis-regulatory regions has been examined as part of the ENCODE project (King et al., 2007). The data thus comes from high-throughput methods that have been tried using the 1% of the human genome that ENCODE uses as its "test bed". In this case, chip-chip data was used to define a large number of putative transcriptional regulatory regions (pTRR). Most human pTRRs could be aligned with sequence in other placental mammals (that were more distant than primates), but only a minority could be aligned with sequence from marsupials or more distant vertebrates. The authors tried various methods of measuring evolutionary constraint, as the main aim was to find which method was best at predicting the location of regulatory regions. All phylogenetic comparisons were against humans, and the authors do

not seem to have made any general study tracing the fate of each regulatory region on each lineage of the mammalian phylogenetic tree.

Conservation of regulatory regions has been studied (Thomas et al., 2003) by examining a 1.8Mb length of the genome by comparisons of 12 vertebrate species. The length contained 10 genes, including CFTR. Particular attention was paid to regions that were conserved across multiple species ("MCSs"); these were at least 25 base pairs long, and averaged 58 base pairs. Thus, a single conserved TFBS will usually be too small to be detected by this method, but it might detect a group of TFBS close together. 63% of known regulatory elements overlap with MCSs, even though the MCSs only form 3.7% of the region, suggesting that most of the regulatory elements are conserved.

The study examined how many species were needed to detect MCSs. Human-mouse pairwise comparisons were not very effective - in one example it detected less than half the MCSs. Ability to detect MCSs appeared to be strongly dependent on the total-branch-length of the phylogenetic tree of the species used: thus, it made little difference if one omitted one of two closely related species; and, if only one species besides human was used, chicken was most effective. Interestingly, the human-chicken comparison was very good at detecting MCSs in coding sequence, but failed to detect most MCSs in non-coding regions. This suggests that, in a search for conserved regulatory regions, one would choose to study a pair of species more closely related than if one were searching for coding sequence.

Comparing species, there was considerable variation in the quantity of repeats caused by transposons. Large indels (>100bp) caused more sequence difference than any other mechanism. Single-nucleotide changes contributed most of the mutational *events*, but only 33% of the *base* differences, in a human-chimp comparison. In a mouse-rat comparison, their contribution was even smaller, about 17%.

1.3.2 Are unnecessary TFBSs removed by natural selection?

Do genomes contain TFBSs that can bind a TF, yet have no function? One might have thought that natural selection would remove these, yet this does not appear to be the case. One study found a number of apparent binding sites for the TF HNF-1alpha upstream of genes that are never co-expressed with HNF-1alpha (Tronche et al., 1997). These had the DNA sequence characteristic of HNF-1alpha TFBSs, and when some were tested *in vitro* they were physically capable of binding HNF-1alpha. They therefore concluded that "no counter-selection occurred within the rest of the genome". That is perhaps an exaggeration as, intuitively, it is easy to imagine a TFBS in the wrong place causing severe disruption to cellular processes, so some counterselection is likely to occur. Moreover, in a previous project I have suggested some limited regions where counterselection may have occurred (Lockwood and Frayling, 2003). But it seems to be true that counterselection only occurs in special circumstances and does not occur on a general basis.

A consequence is that false-matches easily occur in searches for TFBSs. If we know the DNA sequence to which a TF prefers to bind, it is possible to search any DNA sequence for matches, which can be regarded as possible binding sites. But such searches can generate a large number of false matches. For example, one study (Goessling et al., 2001) searched random DNA using various PWMs, and reported match rates varying from 1 per 430 bases (for the TF that binds the TATA sequence) to 1 per 53000 bases (for the TF called SRF (Serum Response Factor)). Thus searching a human genome containing three billion bases would produce perhaps 100,000 to ten million false matches per PWM, which far exceeds the number of genes.

A popular solution to this problem is to select only matches that are conserved between two or more species, on the grounds that false matches will not usually be conserved, whereas real TFBSs will be conserved. An overview of this field is available (Sauer et al., 2006). However, for the purposes of this

project it is important to find evidence of TFBSs that have changed during evolution, which is rather difficult to combine with a system that only detects conserved TFBSs. It might, however, be possible to do so if data from enough species were available. One would first find TFBSs that are conserved in one pair of species, then examine which of them were conserved/non-conserved in comparison with other species. At the start of the present project, only three vertebrate genomes were then available, so this approach was not pursued.

In principle, only three species are needed to do an evolutionary study that also uses TFBS-detection-by-conservation; and this has actually been done (Donaldson and Gottgens, 2006). In that study, motifs that were conserved in a mouse-chimp comparison were regarded as TFBSs; comparison with the human genome produced a list of candidate TFBSs that were present in mouse *and* chimp *but not* human, and therefore should be TFBSs lost very recently in the human lineage. The number of such TFBSs in a 50kb region, relative to the number of TFBSs conserved in the mouse-chimp comparison, was used as a measure of enrichment. Genes related to the sense of smell were particularly enriched in TFBSs lost in humans. The paper does not contain any information about false-match rates, so it is uncertain whether these may have distorted the results. Whilst the false-match rate will have been greatly reduced by choosing only matches that are present in both mouse and chimp, TFBS false-match rates tend to be so high that even a drastic reduction could still leave more false matches than true ones.

A similar approach was part of another study (Doniger and Fay, 2007), which examined yeast genomes for "semiconserved" TFBSs - which they define as TFBSs that show evidence of conservation along *some* lineages of the phylogenetic tree, but not not along all of them.

Another study (Keighley et al., 2005) highlights TFBSs that appear to not be subject to natural selection. This study did not examine individual TFBSs, but rather regions several kbases long upstream of 1000 genes. The substitution rate observed for these regions was compared with the rate observed for regions (eg introns) believed to be under no selective constraint.

For a rat-mouse comparison, the upstream regions were more conserved than the regions under no selective constraint, suggesting that elements within those upstream regions were being preserved by evolution. But for a human-chimp comparison, the substitution rate of the upstream regions was very close to that of the regions under no selective constraint. In theory, s slightly deleterious mutation can become fixed in the population, but this depends on the population size: the smaller the population, the larger the harm caused by the mutation can be and still have a chance of being fixed. The effective population sizes of humans and mice have been estimated at 20,000 and 450,000-810,000 respectively, so the humans will be much more tolerant of slightly deleterious mutations (whose selection coefficient is $< 1/10000$). So if we suppose nearly all mutations in regulatory regions cause so little harm that they come in this category, then that can explain the lack of constraint in the human-chimp comparison.

Differences in gene expression between humans and chimps were noted, thus supporting the idea that functional changes in the upstream regulatory regions are common. It is not clear how widely these ideas can be applied; since humans and chimps are very close in evolutionary terms (about 1% divergence), the result about lack of constraint on upstream regions has only been shown for a tiny portion of the vertebrate evolutionary tree.

1.3.3 Correlated evolution: the combined effect of the evolution of several TFBS

The evolution of an enhancer in different species of *Drosophila* has been investigated experimentally (Ludwig et al., 2000). During embryo development, the gene "eve" is expressed as a series of stripes, each stripe running across the embryo. The expression of the second of these stripes is driven by a particular enhancer, which is therefore called the stripe 2 enhancer. Comparing species, there are differences in the sequence of this enhancer, yet these do not seem to alter the location where this gene is expressed; when the *D. pseudoobscura* enhancer was used to replace the enhancer in

D. melanogaster, it did not alter the expression pattern. But they also constructed a chimaeric enhancer, whose first half was the same as the first half of the *D. melanogaster* enhancer, and whose second half was the same as the second half of the *D. pseudoobscura* enhancer. This caused a different expression pattern, effectively giving a "stripe 2" that was wider than in the wild type of either species. This suggests that, since the two species diverged (40-60 mya), there have been changes in both halves of the enhancer, which individually would have affected the pattern of expression, but which cancel each other out so the changes have no overall effect.

The exact nature of these changes is not certain. However, in the first half of the *D. pseudoobscura* enhancer, the authors identified a possible binding site for the transcription factor called "Kruppel". In an alignment, this site corresponds to a gap, so it is very likely that there is no corresponding site in the *D. melanogaster* enhancer.

How much of the enhancer is conserved? They record 9 known binding sites in the *D. melanogaster* enhancer. The *D. pseudoobscura* enhancer apparently has not been studied so intensively, but looking for *in-silico* motifs, nearly all of the 9 *D. melanogaster* sites are conserved to some extent; only 1 of the 9 has disappeared completely. Also, although the chimaeric enhancer changes the expression pattern in detail, in broad terms it is still similar (it forms a stripe in roughly, but not exactly, the right location). Therefore, I would also interpret these results as showing that the enhancer is robust enough to retain much of its function after the gain/loss of one or two binding sites.

This work forms an interesting comparison with the CYP7A1 work mentioned earlier (page 35). In both cases, there seems to have been more than one gain/loss of TFBS in a single regulatory region. However, the stripe 2 enhancer changes appear to have compensated for each other, whereas in the CYP7A1 case it was argued (Chen et al., 1999) that the changes will not have compensated for each other.

Two models of enhancers have been proposed (Arnosti and Kulkarni, 2005).

The "enhanceosome" model supposes that a complicated protein structure will assemble on an enhancer, the interactions being such that, if a single TF is absent, then the structure will not form. The "billboard" model, in contrast, supposes that individual TFs (or small groups of TFs) can increase transcriptional activation, irrespective of what is or is not bound to other parts of the same enhancer. They suggest that both types of enhancer exist, but that the proportion of enhancers conforming to each model is not known. Obviously, "billboards" have more scope than "enhanceosomes" for accumulating sequence changes that do not have much affect on the function of the enhancer. Although they do not point it out explicitly, the "enhanceosome" model implies that loss of a single TFBS would make the entire enhancer useless, so that the rest of the enhancer would no longer be conserved.

A study (Gasch et al., 2004) examined 14 species of fungi, many of which were related so distantly that regulatory regions could not be aligned. Therefore they used methods that do not depend on alignment. Using expression data and other data, 264 groups of *S. cerevisiae* genes were identified that were likely to be coregulated. These were searched for matches to more than 80 cis-regulatory elements, nearly all of which had been obtained from the literature. There were 42 cis-elements that were over-represented in at least one gene group. For each gene group, the orthologous genes in another species were identified, and then searched to find if any of those cis-elements were over-represented. After doing this for all gene groups and all species, many examples of conservation were found: for instance, there were 10 species for which GGTGGCAAA was frequently found in "Proteasome" genes. Note that this does *not* necessarily mean that every particular TFBS of this type was conserved - a particular gene might have a TFBS that was not conserved - but the focus was on groups of genes, not individual genes. The more distantly the fungi was related, the fewer examples there were of this type of conservation.

The distribution of locations of these TFBSs were often conserved between species. For instance, GGTGGCAAAA is usually found *less than* 200 bases

from the ORF, and this is true for six species; whereas ACACCCAYACAY is usually found *more than* 200 bases from the ORF, and this is true in four species. Yet the locations of individual TFBSs were often not conserved - "distributions of these elements have been conserved, even though the precise positions of individual elements have not". Presumably there is a limited range of locations that an element can have and still remain functional, and individual TFBSs often change location within that range, but not outside it. There was, however, one example of conserved *relative* location between two TFBSs: upstream of methionine biosynthesis genes, the TFBSs for Cbf1p and Met31/32p tended to be about 100 bases apart, closer than would be expected by chance.

One study (Tanay et al., 2005) examined, amongst other things, a group of ribosomal protein genes which all follow a similar expression pattern. For each of 17 yeast species, the group of ribosomal protein genes was identified and then searched using a cis-element finding algorithm. For *S. cerevisiae*, many genes in this group had a "RAP1" TFBS. The same was true for *A. gossypii*, and it was also true for all the other yeasts studied that were more closely related to *S. cerevisiae* than to *A. gossypii*. On the other hand, "RAP1" sites were not detected for any of the species more distantly related than *A. gossypii*. This suggests that, shortly before the divergence of *A. gossypii*, a gain of "RAP1" sites took place; not merely the gain of a "RAP1" site in a particular gene, but rather a large number of ribosomal-protein genes each gained a "RAP1" site at about the same time. They also show that the Rap1p transcription factor gained a new domain at about this time, thus enabling it to take on a new function in addition to its existing function (of regulating telomere length).

They also identify another type of cis-regulatory site, the "Homol-D" box, which is present in many ribosomal protein genes of some yeast species, but absent from others. They suggest that the "Homol-D" box was important in ancient yeasts but that, once a "RAP1" site appeared near a gene, then its "Homol-D" box was no longer important, and was often lost eventually.

Thus the gain of "RAP1" sites is thought to have caused the loss of "Homol-D" boxes - although looking at their results, the correlation between these two events is not as convincing as it might be.

1.3.4 "Turnover" of TFBSs

The word "turnover" appears in a number of papers in the literature. The way it is used is not entirely consistent, so a discussion of how it is used will be given.

One paper (Doniger and Fay, 2007) refers to "turnover, where concurrent gain and loss of a binding site maintains gene regulation".

Another paper (Moses et al., 2006) states

"Definition of Turnover - We consider a predicted Zeste binding site to be an example of binding-site turnover if it is bound in *D. melanogaster* but not conserved among the four sequenced species in the melanogaster species group."

This is somewhat different, since a gain-of-TFBS that altered gene expression would be an example of turnover according to this definition - yet it would not be an example of turnover according to the previous definition.

A third paper (Dermitzakis and Clark, 2002) uses the word "turnover" without giving a formal definition, but does say that

"A new site may relax the selective constraint acting on another already present site, allowing for transcription factor binding site turnover."

which appears to be using "turnover" in the same way as in the Doniger paper. However, elsewhere in this paper (Dermitzakis and Clark, 2002), evidence of non-conservation is described as "evidence that there is widespread turnover of transcription factor binding sites", apparently using "turnover" in the same way as in the Moses definition.

A fourth paper (Odom et al., 2007) describes "turnover" as a situation where - when there are two orthologues of a particular regulatory region - a par-

ticular TF binds *both* orthologues *but* the binding positions are not aligned. They do not use "turnover" to describe a situation where one orthologue is bound by the TF but the other orthologue is not. This is similar to the way Doniger uses the word, and different from the Moses definition.

Summing this up, it is evident that the research community has not settled on a single, precise definition of the word "turnover". Some use "turnover of TFBSs" as if it is merely synonymous with "non-conservation of TFBSs" (Moses et al., 2006). But others (Doniger and Fay, 2007) (Odom et al., 2007) use it to mean a particular type of non-conservation in which the expression pattern of a gene is unaltered, because compensatory gains and losses occur at the about the same time (presumably, inspired by the first paper to describe this particular type of TFBS evolution (Ludwig et al., 2000)). One paper appears to use both definitions of the word (Dermitzakis and Clark, 2002).

My own opinion is this: The concept of TFBSs being gained and lost in such a way that gene expression is unaffected, is an interesting concept, and it is useful to have a single word to describe it - "turnover" is appropriate word to do this. In contrast, if "turnover" is used merely as a synonym for "non-conservation", it introduces an extra word of jargon to describe something that is already conveniently described by the word "non-conservation". Therefore I support the use of "turnover" as used by Doniger.

Turning to actual evidence of "turnover", a number of recent papers have examined whether "turnover" occurs on a widespread basis. One *Drosophila* study looked for evidence of frequent turnover but failed to find sufficiently convincing evidence (Moses et al., 2006). However, they did find some cases that could be interpretted as turnover, but as the number of these cases was did not exceed the number they expected to detect by chance, they did not claim these were evidence of widespread turnover.

A high-throughput study of mammals found evidence of widespread turnover (Odom et al., 2007). Amongst regulatory links that were conserved (that is, a gene was regulated by the same TF in both human and mouse), about

one-third of cases conformed to the turnover model (the exact percentage varied slightly depending on which TF was studied, from 31% for HNF1A to 41% for HNF6).

Another group found that, in yeast, 38% or 57% of TFBSs losses can be explained as turnover (depending on the exact method) (Doniger and Fay, 2007). They noted that their method would not detect certain types of turnover (eg, where the gained TFBS bound a different transcription factor than the lost TFBS); this applies to the other studies as well, so all the figures quoted may be underestimates of the true amount of turnover.

Summing up, this suggests that turnover occurs frequently during TFBS evolution. It also suggests that turnover is a major topic of interest in this field.

1.3.5 Must the evolution of TFs be studied as well as the evolution of TFBSs?

A relevant question is whether binding motifs change over evolutionary time. For instance, will a human TF bind a DNA sequence that the orthologous mouse TF cannot?

This can be addressed by considering a TF family for which we know both the phylogenetic tree and the DNA-binding motifs of the individual TFs. The Nuclear Receptor (NR) family is one such, and it has been reported (Mangelsdorf et al., 1995) that most TFs in this family had a RGKTCA binding motif, or a more specific version of it. The main exception was the four steroid receptors (which bind AGAACA), and TLL, and FTZF1. The steroid receptor difference has been explained at the residue level (Zilliacus et al., 1994). To date the steroid receptor change, note this comment on the NR family (Laudet, 1997): "most subfamilies appear to be ancient since they contain receptors in arthropods and vertebrates ... subfamily III (steroid receptors) is more puzzling, since there are no known homologues of steroid receptors in Drosophila. This could suggest that subfamily III is ... a rather young one formed after the arthropod/vertebrate split."

It thus appears that most NRs have bound RGKTCA since before the arthropod/vertebrate split.

In view of this, it is likely to be very rare to find major differences in the core binding sequence of a NR when comparing species as closely related as human and mouse. Other binding properties (dimer orientation, dimer spacing, and DNA contacts outside the "core") seem more susceptible to change but even these are often the same for closely related NRs. Thus, an attempt to distinguish between human and mouse TFs would probably succeed only if it could detect very detailed changes in DNA-binding properties. Nothing of this kind was attempted during this project.

As noted earlier (page 32), a study (Xie et al., 2007) concluded that, during mammalian evolution, there had been little or no change in the DNA-binding properties of the molecules that bound the regulatory elements in their study.

A study that found vast numbers of non-conserved TFBSs (Odom et al., 2007) (see page 37 for a more detailed review) nevertheless determined that these changes did *not* arise from differences in the DNA-binding specifity of the TFs (human-mouse comparison).

The evolution of the bZIP family of transcription factors has been studied (Amoutzias et al., 2007). One of their findings was that: "Major features of the network topology were most probably formed before the genome duplication events that occurred in the vertebrate lineage. Therefore, whole-genome duplication did not significantly change the topology of the network. Rather, it added new paralogues that mostly retained their ancestral dimerization preferences, with new interactions being formed in a few cases. These vertebrate lineage duplications also *did not have any major impact on the evolution of new specificities for recognition by DNA-binding motifs*" (my italics). They classify the bZIPs into "5 pairs of duplicate families" which retained their DNA-binding properties, but diverged in the dimerisation domain, following a hypothesised pre-Cambrian duplication event. For the purposes of this thesis, the most relevant point here is that DNA-binding properties seem to have become fixed very early (pre-Cambrian), so very little change is to be expected in mammals.

A paper mentioned earlier (Gasch et al., 2004) found evidence that the transcription factor RPN4P had evolved a different specificity in *S. cerevisiae* than in *C. albicans*. However, the difference was subtle. In both cases the most preferred binding sequence was GGTGGCAAAA. The difference was that certain deviations from this sequence were tolerated by the *C. albicans* TF, but were not tolerated by the *S. cerevisiae* TF. This was originally inferred from the upstream sequences of a group of genes likely to be regulated by this TF; it was then confirmed by *in-vitro* binding tests that showed GAAGGCAAA being bound by one TF, but not by the other. To put this into context, these two fungi are so distantly related that it is not usually possible to align the upstream regions of orthologous genes. I would therefore interpret these results as emphasising that changes in binding properties are very slow: the fungi have diverged so much that alignment is usually impossible, and yet the changes in the binding properties of this TF are so subtle that they could only be detected by very detailed investigation. This suggests that, where two species are related closely enough for alignment to be possible, for instance human-mouse, changes in TF binding properties are likely to be very small.

A mathematical model has been used (Sengupta et al., 2002) to predict that the binding properties of a transcription factor should change at a rate that is proportional to 1/(number of binding sites). Thus the rate of change should be highest in TFs which have a particularly small number of TFBSs. It seems likely that, in previous years, TFs with very few TFBSs would be less likely to be discovered than those with many TFBSs (though this may have changed now whole-genome projects are common). Thus the most well-known TFs may be exceptionally slow in changing their binding properties, and consequently be the least useful as raw material for the study of changes of such properties.

This mathematical model could, I suggest, make an interesting prediction when applied to whole-genome duplications. Suppose we assume that a whole-genome duplication doubles the number of TFBSs bound by a particular transcription factor. Then, according to the mathematical model,

this will halve the rate at which its DNA-binding properties will change during evolution. Similarly, the rate will be reduced by a factor of 4 (or 8) as a consequence of 2 (or 3) genome duplications in succession. (These predictions assume there is no gene loss following the duplications, and would have to be altered to allow for any such losses). This might be summarised by saying, roughly speaking, that the DNA-binding properties became "frozen" by the genome-duplications. I regard this suggestion as quite speculative, but advance it because it could explain a finding Amoutzias et al made after studying the bZIP family (see above). They suggest the family was affected by two, or even three, whole-genome duplications, and they state that the DNA-binding properties of TFs in the family have been essentially unchanged since that time.

In summary, all these papers support the view that there has been little or no change in the DNA-binding properties of TFs during mammalian evolution.

1.3.6 Alignment and analysis techniques for regulatory regions

If we align the orthologues of a regulatory region from two species, and the alignment produced is incorrect, then it is obvious that TFBSs in the region will often *appear* to have diverged, even if all the TFBS were in fact conserved. The seriousness of this problem has been quantified (Pollard et al., 2006). The method used computer simulation to generate two or more sequences for which the true alignment was known; these sequences were submitted to alignment programs, and the output from the alignment programs was compared with the true alignment. The simulations were designed to mimic part of the *Drosophila* evolutionary tree, the simulated TFBSs were based on real TFBSs in *Drosophila*, and the simulated evolution of sequences caused TFBSs to be better conserved than other sequence.

A surprising conclusion was that, at some divergences, it is very common to get "overlapping" alignments of TFBSs - that is, a TFBS is not correctly aligned, but the size of misalignment is so small that the orthologues of

the TFBS overlap in the alignment. This disproved the authors' intuition - they had supposed that, because TFBSs were more highly conserved than surrounding DNA, the TFBSs would be perfectly aligned except in cases when the alignment was totally incorrect.

Although that paper is focussed on fly evolution, it is relevant to this thesis to ask how the results could be applied to mammalian evolution. The authors themselves briefly do so, noting that the human-mouse divergence is about 0.5 substitutions per site (which agrees with other studies, see page 60). It is not a simple task to apply this figure, because nearly all the results presented show the total divergence distance of a multiple aligment of a 4-species tree, rather than a pairwise alignment. However, let us suppose that a 4-species tree including human and mouse would have a total divergence of 0.75. For a divergence of 0.75, and examining the results presented for the "Blastz/Tba" alignment program, about 6% of TFBSs have an overlapping aligment, and a few percent are not aligned at all, suggesting that over 90% of TFBSs are aligned perfectly. This looks quite comforting compared against claims that 30% or more of human TFBSs are not conserved in mice (see page 34), as it suggests that genuinely diverged TFBSs will greatly outnumber misaligned TFBSs. On the other hand, pairwise alignments might not give so favourable a result as alignments from a 4-species tree.

Another study compared five species of *Drosophila*, focusing on changes in sequences that could be revealed by alignments (Bergman et al., 2002). The study examined eight regions of the genome, totalling 500k of sequence (in *D.melanogaster*. The divergence of these species (relative to *D.melanogaster*) was up to 2.6 substitutions per site (at silent sites), which is high compared with human-mouse divergence. Microsyntenic order was highly conserved, but not completely. Some rearrangements were attributed to retrotransposition. Predicted genes showed a higher rate of amino-acid substitution than known genes; they view this as evidence that well-known genes are not a random selection of all genes. Amongst conserved sequences, the ratio of protein-coding sequences to non-protein-coding-sequences increased with di-

vergence time.

They also examined the hypothesis that non-coding sequences are conserved because they are mutational cold-spots (in contrast to the usual assumption that they are conserved by functional restraint). One argument they give for rejecting this hypothesis is that the "spacer distance" between two conserved segments is conserved; the evidence for this is that, if the spacer distance in *D.pseudoobscura* is plotted against the corresponding distance in *D.melanogaster*, the resulting scatterplot shows a correlation between these two distances. (I do not find this convincing, as presumably there would be some correlation between the two even if there was no functional constraint).

They conclude that *D.erecta* is too closely related to *D.melanogaster* to produce comprehensive cis-regulatory element prediction by comparative genomics, and recommend *D.pseudoobscura* intead. (It is therefore perhaps worth noting that the divergence distances involved are 0.366 and 1.830 (measured as the observed substitution rate per site at silent sites)).

Another study examined four alignment techniques using mammalian genomic data (Margulies et al., 2007). This was not focused solely on regulatory regions, but the examination of non-coding regions is of interest. One difficulty they point out is that there is no "gold standard" for evaluating genomic alignments (in contrast to alignments of protein-coding regions, where 3-D structures provide an independent test of whether aligned nucleotides are really orthologous). But even with coding exons, which they believe are likely to be shared amongst all the species they analyse, alignment programs can only align 72% of exons at best (human-mouse comparison), depending on the program. This suggests that a substantial proportion of orthologous sequences will fail to be aligned by currently available alignment programs (in a human-mouse comparison). As >28% of exons were not aligned, it seems plausible to speculate that >28% of regulatory regions will not be aligned (even when an orthologue exists) - though of course it is not at all certain that both numbers will be equal.

The conclusion just mentioned is particularly interesting, as it suggests there will be many cases where a TFBS is (in fact) conserved, and yet we are unable to *prove* it is conserved, because the alignment program is unable to locate the orthologous sequence. Using the figures from the previous paragraph, one can speculate that the ratio of clearly conserved TFBSs to TFBSs that are unaligned (but actually conserved) will be 72:28. This makes an interesting comparison with the study mentioned on page 37 (Odom et al., 2007), where clearly conserved TFBSs were about as frequent as unaligned TFBSs. It suggests that a substantial proportion of unaligned TFBSs could be conserved TFBSs where the alignment program failed to find the orthologous sequence. This suggests that one should not assume that unaligned TFBSs are non-conserved TFBSs.

False alignments were assessed using *Alu* insertion elements. *Alu* elements occur only in primates, and therefore *Alu* elements in humans should never be aligned with mouse sequence. In practice, the best aligner by this test was TBA/Blastz, which aligned 3% of *Alu* elements with mouse sequence; the other three alignment programs were less satisfactory, aligning up to 10% of *Alu* elements. TBA/Blastz was also best in a second test, which examined the ability to avoid incorrect alignment of protein-coding sequence.

Substitution rates in neutral sequences were measured using two types of neutral sequence. It was found that rates estimated from Ancestral Repeats were slightly higher (2%-13%) than rates estimated from fourfold synonymous sites.

Three methods of identifying constrained (ie conserved) regions were tried: phastCons, GERP and binCons. All of them identifed 5-6% of the genome studied as being under constraint. However, this did not mean there was total agreement between the methods, as only 3-4% of the genome was identified *by all three methods* as being constrained. Of the constrained sequence, 40% was exonic, 20% overlapped regions that were experimentally identified as having some other function, and 40% had no known function.

Here is an outline of the method used by the Blastz system, which was one of the best in the survey just mentioned. Finding the best possible alignment can be very expensive in computer time, so is not attempted. Instead, like other alignment programs, Blastz first looks for "seed" alignments (that is, short alignments, of a pre-set length, which do not contain variable-length gaps) which can be found relatively quickly. Each seed alignment is then extended into a longer ungapped alignment, and if that is successful enough, then it is further extended into a lengthy gapped alignment (Schwartz et al., 2003).

One novelty of Blastz is the "seed" alignment. Traditional BLAST looks for "seed" alignments consisting of 11 consecutive perfectly matched bases, whereas Blastz follows a suggestion (Ma et al., 2002) of looking for "runs of 19 consecutive nucleotides in each sequence, within which the 12 positions indicated by a 1 in the string 1110100110010101111 are identical". Fig 1.3 illustrates this. Moreover, when searching for a "seed" alignment, transitions (a/t and c/g substitutions) are treated the same as perfect matches. Also, in the scoring system for an extended alignment, transitions are given a much smaller penalty score than other substitutions.

Figure 1.3: Finding a "seed" for a Blastz alignment
Example of a "seed" that could be the basis of a Blastz alignment. Ignore for a moment the letters against a grey background: because the letters showing through the six "windows" are all matching in human and mouse, Blastz can recognise this as a "seed" alignment. Some of the letters between the windows are non-matching, but that doesn't affect the ability of Blastz to recognise this as a "seed". For Blastz, the size and spacing of the windows is always exactly the same as shown here.

A different study also points out the lack of "gold standard" data for assessing tools that align regulatory regions, but addresses the problem using simulation (Pollard et al., 2004). Each simulation would generate two sequences for which the true alignment was known; these sequences would be submitted to alignment programs, and the results of these programs compared with the true alignment. The simulated sequences were designed to represent 10kb lengths of *Drosophilia* non-coding sequence. A number of performance measures are used but, disappointingly, no measure of specificity is mentioned (that is, if a tool is given two unrelated sequences, will it refuse to align them?). Attention is instead given to coverage and sensitivity, the latter being the proportion of truly orthologous bases that are identified as such by the tool. Thus, the study focuses on the problem of alignment tools "losing data" by failing to align bases that ought to be aligned; it does not focus on the problem of alignment tools producing misleading results.

It is easy to imagine that some mutational events will damage two adjacent bases, thus raising the question of whether, in reality, it is common to find double-base mutations, or even triple-base mutations. One study found these to be a substantial minority of mutation events (Whelan and Goldman, 2004); single-base mutation events, double-base events, and triple-base events were estimated to occur with frequencies in the ratio 0.83:0.07:0.10 respectively (based on mammalian data).

Another paper studied, amongst many other things, how the substitution rate varies across the genome, using a human-mouse comparison (MGSC, 2002). The genome was divided into 5-Mb windows and the substitution rate at fourfold synonymous sites was estimated for each window. The average rate was 0.447 substitutions per site but, more interestingly, the standard deviation was 0.067, indicated that the rate varies somewhat across the genome. This variation is linked with GC content - regions of high GC content tend to have the highest rates, whereas regions with a GC content of 40% have the lowest rates. The relationship seems to be non-linear, so it would be wrong to

extrapolate this and imagine that regions with very low GC content will have very low substitution rates. Also, these variations in substitution rate are correlated with variations in the rates of other types of mutation; for example, regions with higher substitution rates tend to also have higher deletion rates, as well as being regions where recombination is more likely to occur.

1.3.7 Evolution of TFBSs in network research

A study (Teichmann and Babu, 2004) examined the growth (during evolution) of regulatory networks, with a particular emphasis on the mechanisms for adding interactions between TFs and genes.

One possible mechanism is duplication: if a gene (including its regulatory regions) is duplicated, then the genome acquires additional TFBSs. Moreover, examples where this happened should be detectable, in the form of data showing a TF regulating both genes in a paralogous pair. 166 such interactions were detected in yeast (out of a total of 851 interactions with potential information about homology). Thus, 20% of interactions show evidence of being descended from an event in which TFBS(s) were duplicated.

Another possible mechanism is duplication of a gene that makes a TF, creating two homologous TFs that bind the same TFBSs; this can create new *interactions* without creating new TFBSs. 188 such interactions were detected in yeast.

Another category includes cases where there was no evidence that an *interaction* had duplicated, even though the target gene had been duplicated (in some cases), or even though the transcription factor gene had been duplicated (in other cases). There were 365 such interactions, apparently cases where there had been duplication of genes without duplications of interactions. In these cases, the authors assumed that many of the observed interactions were interactions that had been gained post-duplication. The reasoning behind this assumption was not discussed in detail in the original paper, though an imaginary example was given in which duplication, followed by gain and loss of TFBS, produced a final outcome like that seen in real data (see fig

1.4). But I can also imagine the same final outcome being produced without any gain of TFBS occurring (fig 1.5). Perhaps the authors rejected the fig 1.5 scenario as unlikely because it involves the number of TFBSs per gene being reduced - whilst this might happen occasionally, if it happened following *most* duplications it would produce a genome-wide reduction in TFBSs per gene, which seems unlikely. Presumably that is the reason for rejecting fig 1.5, and regarding cases which resemble the final outcome in fig 1.4 as evidence of "gain of TFBS".

Another category consisted of interactions that could not be related to any duplication event, which were called "pure innovations", 101 such cases being found.

From the figures quoted, it can be seen that 88% of interactions could be related to some kind of duplication event (of the gene or of the TF). Yet, only 20% of interactions were descended from a duplication event where the TFBS itself was duplicated. Often, the authors claim, interactions will be gained *following* a duplication event - in other words, a new TFBS will be gained, but we do not have to suppose this gain is by a duplication mechansism - it may be that "duplication" is only relevant in the sense that the earlier duplication of a gene (or TF) may perhaps have created a selection pressure favouring a gain-of-TFBS.

One can imagine a TFBS being duplicated, but one of the duplicates being overlooked because of incomplete data, giving the appearance of a non-duplicated TFBS. Given this problem, it is perhaps best to look on the 20% figure in the previous paragraph as a minimum estimate.

In summary, the paper indicates that a substantial minority of TFBSs were gained by duplication. It does not indicate what the most common mechanism is for gaining TFBSs.

Figure 1.4: Gene duplication followed by gain and loss of TFBSs
This shows gene duplication followed by gain and loss of TFBSs, and is based on an illustration in the original paper (Teichmann and Babu, 2004). Two different types of TF, "A" and "B", are shown bound to their TFBSs.

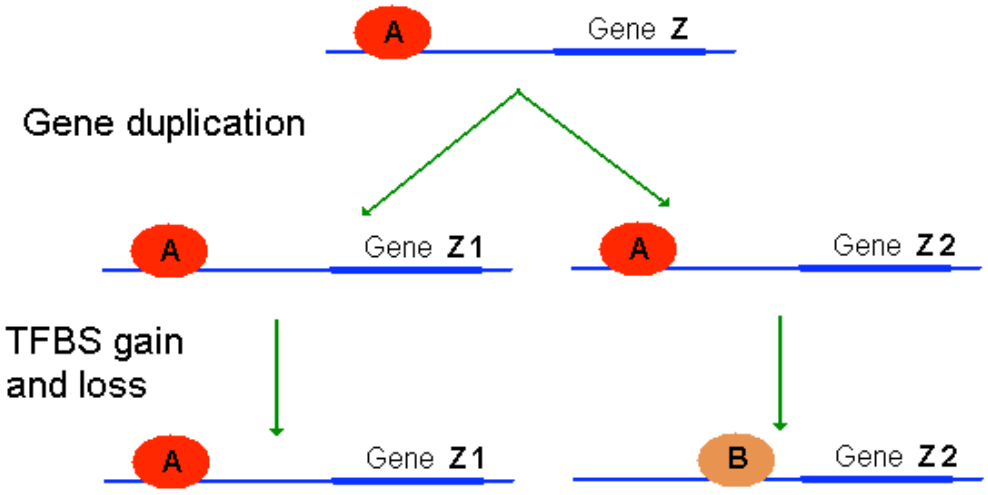
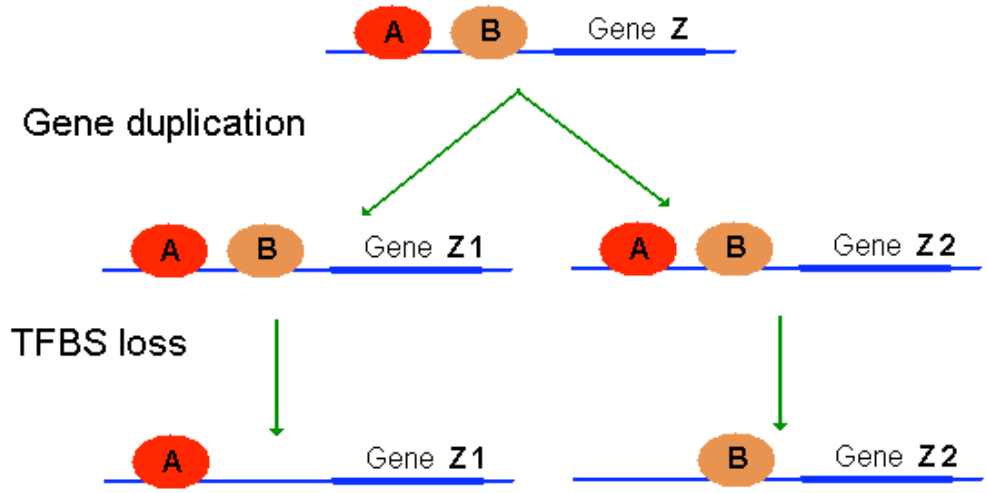


Figure 1.5: Gene duplication followed by loss of TFBSs
This shows gene duplication followed by loss, but *not* gain, of TFBSs. Notice that the final situation is identical to the final situation in fig 1.4.



1.4 SOME POSSIBLE SCENARIOS FOR HOW REGULATORY REGIONS MIGHT EVOLVE

To stimulate ideas on how regulatory regions might evolve, figure 1.6 shows some different possibilities. Each graph shows how a particular regulatory region evolves during the evolution from an early mammal to humans - hence the horizontal axis is marked with a scale of 200 million years.

Figure 1.6 (a) shows "occasional change"- for a long time there is no gain or loss of TFBSs at all, then a single TFBS is gained, then there are no further changes.

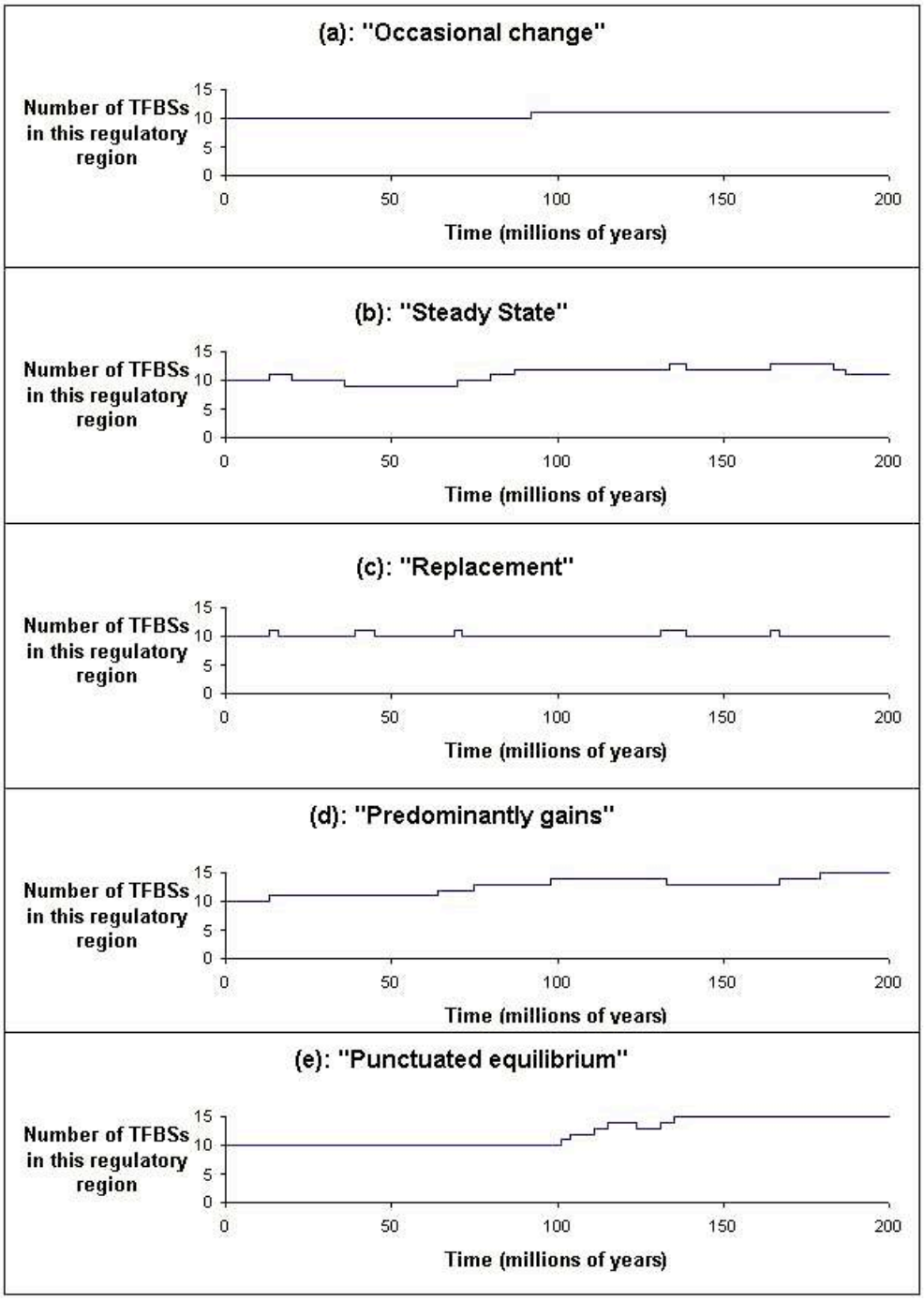
Figure 1.6 (b) shows "steady state" - some gain-of-TFBS events and some loss-of-TFBS events can be seen, apparently scattered at random throughout evolutionary history, but the number of TFBSs in the region stays roughly the same.

Figure 1.6 (c) shows a number of gains, each rapidly followed by a loss. The assumption is that this particular gene gives maximum fitness if there is one TFBS per TF, but that having two TFBSs for the same TF gives only a slight loss of fitness. Thus if a mutation creates a second TFBS for a TF that already binds the regulatory region, it will not be rejected immediately by natural selection. After this second TFBS has been created, a mutation that destroys the original TFBS will be tolerated (and indeed weakly selected for). Thus, each gain tends to be soon followed by a loss.

All the previous graphs have shown the number of TFBSs in the region remaining roughly constant, but figure 1.6 (d) illustrates the possibility that gains occur far more frequently than losses. Obviously this gives an overall increase in the number of TFBSs; intuitively, it seems possible that such an increase would occur during the evolution of very complex animals. Nevertheless one can imagine that regulatory complexity could increase by other means - for instance, duplication of entire regulatory regions - even if the number of gains by point mutations did not exceed the number of losses.

Figure 1.6 (e) is similar to the previous graph, except that all the changes

Figure 1.6: Possible ways a regulatory region might evolve
Some different, speculative scenarios for the evolution of a regulatory region, which are commented on on page 64.



occur within a relatively short period of time, with no changes for long periods before and after. It thus belongs to the punctuated equilibrium theory of evolution - although that theory was originally proposed as applying to obvious changes of phenotype, which will not necessarily follow the same pattern as TFBSs.

There does not seem to be any strong evidence to rule out any of these possibilities, except some surveys (Dermitzakis and Clark, 2002) (Odom et al., 2007) suggest that changes occur more frequently than shown in (a) "Occasional change".

1.5 DEFINITION OF "EVOLVED TFBS"

The definition of "evolved" should be made clear. "This regulatory region has changed during the last 200m years of evolution" is a vague statement that could have several different meanings, such as

1) The DNA sequence of the TFBSs in the regulatory region has changed, but without disrupting the ability of the TFBSs to bind their TFs, so the number and order of the TFBSs has not changed. This type of change has, sometimes, been the main focus of a study of TFBS evolution (Moses et al., 2003).

2) Some of the TFBSs have been removed (or their ability to bind has been disrupted by mutations) and replaced by similar TFBSs elsewhere in the regulatory region, thereby changing the order of TFBSs. There has been no change in the identity of the TFs bound by the region, nor in the number of TFBSs for each TF, nor in the expression pattern produced by the region. Experiments indicate that moving TFBSs around within a regulatory region does not necessarily alter the expression pattern (Arnosti et al., 1996) (although, admittedly, there are *some* cases where spacing is important (Fickett, 1996)).

3) The regulatory region has changed so as to bind a different set of transcription factors, but the expression pattern produced by the region has not changed.

4) The regulatory region has changed in a way that alters the expression level of the regulated gene, but this does not produce any change in the phenotype. One study found a genetic network that is so robust that, in their simulation, many parameters could be altered (within a factor of 10 range, or more) without disrupting the developmental pattern produced by the network (von Dassow G et al., 2000).

5) The regulatory region has changed in a way that alters the phenotype.

All these meanings are reasonable and, I suspect, are in use under various circumstances. Since this project will refer to TFBSs that have "not evolved" during a certain period of time, the definition that will be used here must be made clear.

For this project, definition (2) is the one that will be used: I will say a TFBS has "evolved" if mutations or indels have destroyed (or created) the ability of the site to bind its TF, even if the destroyed site has been replaced by a similar one elsewhere in the promoter. I will say a TFBS has "not evolved" if it has retained the ability to bind its TF, even if the actual DNA sequence has changed.

1.6 AIMS

The initial project aim was to develop an in-silico system that would find examples of TFBSs that had evolved, in the sense that a particular TFBS was known to be present in one species but absent in another. These would then be compared to TFBSs that were conserved between two species, and various characteristics examined, to find ways in which TFBS that had diverged differed from those that remained conserved.

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)