Chapter 5. Evolutionary conservation of promoter regions

Like the previous chapter, this part of the project relies on the availability of fairly complete sequences for more than one genome, and also on the basic tools which make it possible to compare them. However, while chapter 4 made a global examination of non-coding regions which were conserved between the mouse and human genome, using a machine learning approach to extract characteristic patterns from them, the analyses in this chapter are far more focused, and concentrate specifically on the conservation of sequence in the promoter regions of already-known genes.

With the availability of high quality sequence data for both the human [IHGSC 2001] and mouse [MGSC 2002] genomes, a lot of interest has turned to the comparison of the two. Automatic gene predictions are available across both genomes – for example, those from the Ensembl project [Hubbard *et al.* 2002]. There is also curated annotation for substantial parts of the human genome [http://vega.sanger.ac.uk], but not yet for much of the mouse sequence, so this was not used here. By comparing the predicted protein sequences from these gene sets and identifying reciprocal best hits, likely pairs of orthologous genes – genes which are evolutionarily related and perform the equivalent roles in the two species – can be found. However, it is clear that protein coding sequences are not the only important pieces of information in the genome, so here I consider the evolution and conservation of the promoter sequences which regulate the expression of those genes.

Initial comparative analyses included in the publication of the mouse genome [MGSC 2002] have shown that there is a degree of sequence conservation between human promoters and those of the corresponding mouse gene. The results in chapter 4 offer further evidence that at least some parts of promoter regions must be conserved between species. Here, the aim was specifically to investigate the conservation of these promoter sequences. There are a number of interesting questions about promoter evolution: to what extent do the functional elements of promoters remain conserved between species? Is conservation between "orthologous" promoters more or less significant than the appearance of shared motifs between promoters for two unrelated genes which happen to require similar expression patterns. And finally, can any clear examples be found where a promoter which drove the expression of one gene in the first species is associated with an entirely different gene in the second.

To investigate this, I first identified a set of promoters pairs, based on orthology of their associated protein-coding regions, then performed an all-against-all promoter sequence comparison and determined how effectively the expected pairs are recovered. In execution, this strategy is quite similar to protocols used to assess protein-alignment methods [Brenner *et al.* 1998]. In this case, the results from an all-against-all comparison will indicate whether the conserved signals are specific to particular promoters, or reflect more general themes which appear in many promoter regions.

In the short history of bioinformatics, a wide range of software has been developed for the comparison and alignment of biological sequences. Most of them have, at their core, some kind of dynamic programming methodology [for review, see Birney 1999], but there are many different implementations, designed and tuned for different purposes. The clearest distinction is between global alignment algorithms (the classic example being the method from [Needleman and Wunsch, 1970]) which attempt to find the best match along the full length of two sequences, and local aligners (e.g. [Smith and Waterman, 1981]) which detect the best-matching portions of two sequences, even if the remainder of the sequence shows no similarity whatsoever. Additionally, some methods, for example the **ssearch** program, perform an exhaustive search of alignment space using a fairly pure implementation of the dynamic programming algorithm, while others such as Blast [Altshul *et al.* 1997] use optimizations – specifically, an initial seeding stage which searches for words which match exactly between the query and subject sequences – to detect the majority of matches quickly, at the expense of some sensitivity. In general, there is a trade-off of speed for sensitivity, with full dynamic programming at the extreme of good sensitivity, and methods like SSAHA [Ning *et al.* 2001] concentrating entirely on speed. Blast falls between these extremes. These optimizations are normally achieved by first searching for a small cluster of symbols which all match exactly, then extending this in both directions using more sensitive methods. A variety of sequence search methods have been evaluated by the MaxBench system [Leplae and Hubbard, 2002], but this is based on performance when aligning protein sequences, which is not necessarily equivalent to aligning regulatory DNA.

In this case, I chose to use the well-known **blastn** nucleotide search and alignment tool [Altshul *et al.* 1997]. Since this is a local alignment method, and is seeded by words which match exactly between the two sequences, it specifically detects well-conserved blocks, rather than attempting to align large pieces of sequence with only marginal similarity. Its default parameter set, which was used here, further emphasizes the detection of highly conserved regions.

5.1. Alignment of promoter regions

Based on human release 8.30a and mouse release 8.3c, the compara databases from Ensembl release 8 [Clamp *et al.* 2003] identified 19914 orthologous gene pairs, based on reciprocal best hits in an all-against-all **blastp** (protein sequence) search. Note that while these results are stored in the same compara database as the nucleotide alignments discussed in chapter 4, the protein comparisons are performed quite independently.

Since the aim was to specifically investigate conservation of promoter regions, it was important to work only with those sequences where it was possible to extract the true promoter region with a good level of confidence, and to avoid contamination of the set with other types of conserved sequence – in particular, unannotated coding regions. I therefore applied a rigorous procedure to select a high-confidence subset of the data. From the initial compara set of 19914 orthologous pairs, I selected those cases where both the mouse and the human gene had only a single predicted transcript. This avoided cases with alternate transcripts, which might have additional upstream exons, and also simplified the evaluation schemes applied later in this chapter by removing scope for double-counting of alternate transcripts with starts close to one another. This left 8989 pairs – still a good-sized dataset for large scale evaluation. I then picked the subset where the human sequence had an EponineTSS [Down and Hubbard, 2002 and chapter 3] transcription start site prediction in the interval [-200:+50] relative to the start of the Ensembl-predicted transcript. This gave a set of 2442 pairs. The agreement of Ensembl's (evidence-based) UTR predictions with the computational results from EponineTSS means that these positions should have a very high probability of reflecting true transcription start sites, which in turn means that sequence upstream of this point is likely to primarily have promoter functionality.

For each selected pair pair, mouse sequence was extracted from -5500 to +500 relative to the Ensembl-predicted gene start, and human sequence was extracted for the 5000 bases upstream from the predicted TSS. Note that the mouse sequence was longer at each end than the human sequence. This means that, when aligning human sequences to mouse sequences, cases where the mouse sequence is slightly longer (perhaps due to a repeat insertion) should still give good alignments, and possible edge effects from the alignment algorithm will be reduced. In the case of two closely-spaced pairs of divergent genes, the extracted windows overlapped. These cases were recorded to ensure that matches to the overlapping regions were not counted as false positives. The human sequences were masked for known repeat sequences using the standard RepeatMasker method [Smit and Green, unpublished], and also for possible extra regions of coding sequence as predicted by the Genscan algorithm [Burge and Karlin, 1997]. These additional coding regions could be alternate first exons from genes with alternate transcription start sites (which may not have been recognized in the Ensembl gene build), or they may be pseudogenes. In either case, it is not appropriate to include them in the sequence comparison when searching for promoter regions.

Finally, each human sequence was searched against the full set of mouse sequences using the **blastn** [Altshul *et al.* 1997] program (release 2.2.2) with its default alignment parameters:

Word size	11
Score threshold	30
Reward for match (N)	+1
Penalty for mismatch (M)	-3

Table 5.1. Default **blastn** parameters used

It should be noted that with this parameterization, the magnitude of the penalty for a single-base mismatch is much greater than the reward for a correct match. This means that alignments will not be extended through regions with a low percentage identity. This implies that the algorithm is running with less than maximal sensitivity, but is desirable in this case: the mouse and human genomes are relatively similar to one another, and a degree of background synteny can be seen across many regions, which does not necessarily indicate any functional reason for conservation but simply the fact that the species diverged recently enough that there have not been enough mutations to make related sequences unrecognizable, even where there is little or no evolutionary pressure to maintain a particular sequences. The parameters used here will highlight the most conserved portions of sequence between the two species.

To investigate the efficiency of gene pairing based on promoter sequences, all pairs of human and mouse sequences for which a **blastn** alignment was produced were scored by the total number of bases in aligned regions in the interval [-2000:0] relative to the human TSS, counting alignments to any part of the 6kb mouse sequences. This somewhat unconventional scoring strategy was motivated by the observation that, in many cases, there were several distinct blocks of aligned sequence. There is no justification in the blast scoring scheme for combining the blast scores for these individual blocks by addition, so in order to count the contributions of all the blocks, it was necessary to use an alternative scheme. This approach also made it easy to calculate scores which only considered alignments of some portion of the sequence, without actually having to re-run the alignment method. The highest-scoring pair in this set had 1308 out of 2001 aligned bases – over 65% of its total length.

Those pairs which matched the previously-defined protein orthology were considered to be "correct" pairings, while all others were considered "incorrect". In two cases, apparently incorrect pairings were seen because the upstream regions of divergent genes overlapped. These cases were ignored. The relationship between number of correct and incorrect pairs as a score threshold was varied is shown in figure 5.1. It can be seen that promoter-based and protein-based pairings agree extremely strongly up to a coverage of just over 40% (which occurs at a threshold of 50 total aligned bases), beyond which correspondence falls rapidly. This indicates the point at which blocks of detectable similarity become comparable in size with the blocks of similarity which occur either entirely by chance, or because of small, common, functional elements – either individual transcription-factor binding sites, or perhaps small clusters of sites. There are, however, a small number of cases where an incorrect pair is detected with a score much higher than noise. Examination of these sequences revealed a number of low complexity regions (in one example, "gaatgaatgcaggatgcagtgag") that were not being masked by the **dust** filter built into the Blast software. A more detailed scan for low complexity sequence was made using the **etandem** program from EMBOSS [Rice *et al.* 2000], searching for tandem repeats between 4 and 24 bases long. Masking these regions and realigning the sequences eliminated a number of probable false positive matches, giving the second trace of figure 5.1.

This process of building an all-against-all score matrix then counting those pairs where the score is greater than a specified threshold is closely analogous to the method of single-linkage clustering. In the case of protein sequences, a number of projects have studied clustering of sequences based on pairwise alignment scores. Clusters are often observed, and have been well studied. In this context they are generally called families, and are presumed to reflect evolutionarily related genes. Ensembl gene predictions are allocated to families using the TRIBE-MCL clustering algorithm [Enright *et al.* 2002]. In mouse release 8.3c, the 22,444

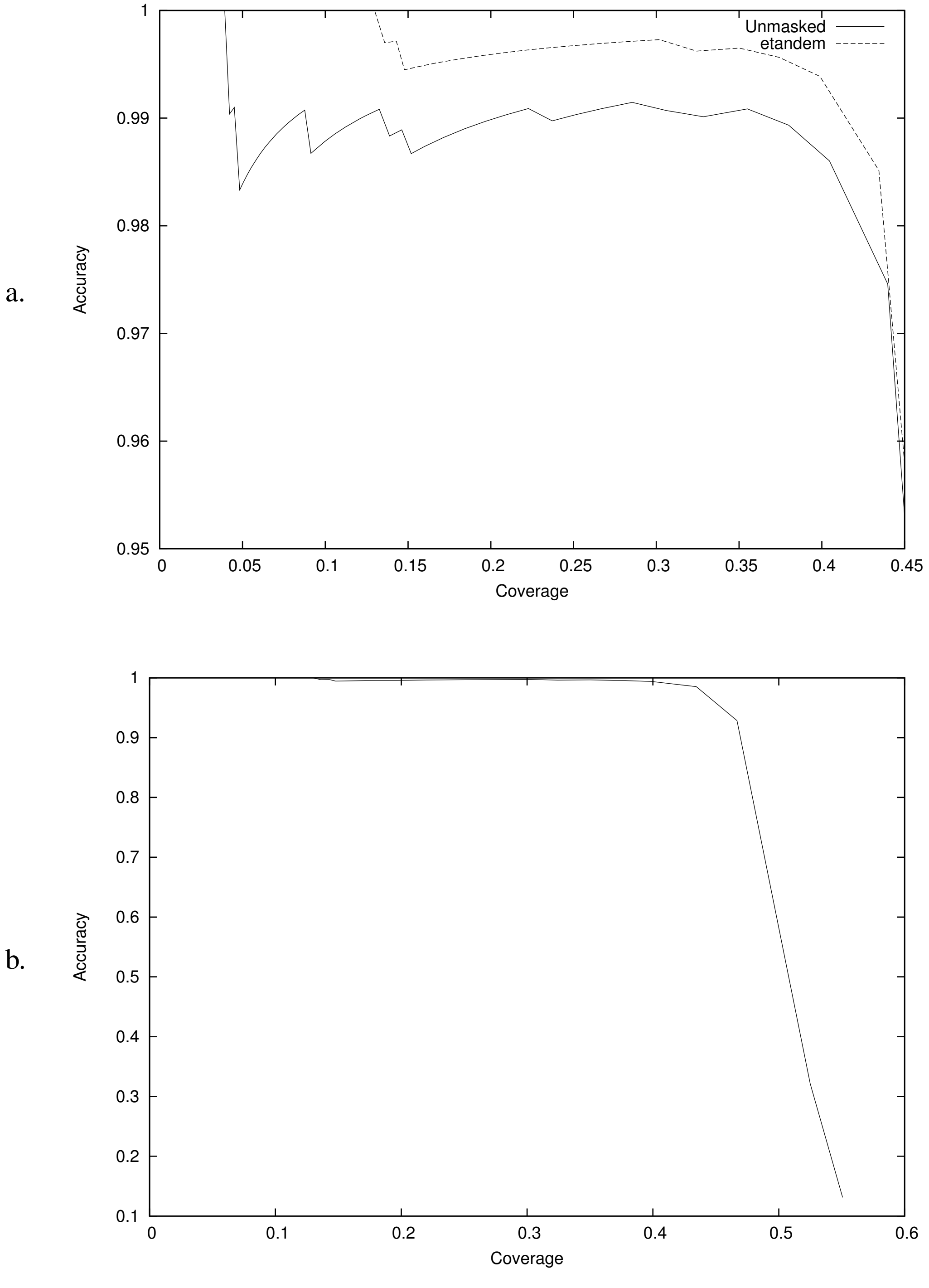

Figure 5.1. Agreement of promoter orthology with protein orthology. Panel a shows coverage (proportion of sequences correctly paired) at high levels of accuracy, with and without additional repeat-masking using **etandem.** Panel b shows a wider range of coverages, and only includes the results with **etandem** masking.

predicted genes were assigned to 11,579 families. If strong clusters of promoter sequences could be observed, these would appear as a gradual fall-off of the accuracy value in figure 5.1 as more distantly related clusters join together. Instead, this plot shows very few strong pairs which do not agree with the protein best-hit relationships, right up to the point where the accuracy collapses. The promoter-pairs which do not agree with the protein-defined orthology were compared with the Ensembl family data for the mouse protein set. At a low threshold of 20 aligned bases, 1283 pairs were detected which were not in the protein orthology set, but just one of these was a match to another mouse promoter for a gene in the same family. Based on this result, I suggest that there is little significant large-scale similarity between non-orthologous promoters, even when they are driving the expression of closely related genes.

Without detailed laboratory experimentation, there is currently no available method to accurately determine the boundaries of promoter regions on bulk genomic sequence, therefore these results are open to the criticism that this strategy might be detecting similarities which remain by chance in regions of sequence which are, after all, believed to be related to one another. For the same reasons, it was not possible to build a corresponding control set of sequences which are close to genes but provably not involved in gene regulation: as well as the extent of promoter regions being uncertain, enhancer regions – which are thought to be similar to promoters in terms of their general architecture, and are certainly likely to be conserved between species – can be found throughout the genome. However, making the assumption that the bulk of regulatory elements, particularly in genes which have simple regulation mechanisms, are situated quite close to the transcription start site, it is reasonable to compare pairing results based on homologies in the previously considered window of [-2000:0] (relative to the TSS) against those further upstream in the range [-4000:-2000]. While the second set is very likely to contain some promoter elements, the number is expected to be rather smaller than in the region immediately upstream. The results of this comparison are shown in figure 5.2. This gives the same basic shape of a plateau followed by a rapid collapse in accuracy, but note that the collapse occurs at around half the coverage seen with a window closer to the TSS. It is, of course, believed that promoter elements still occur (albeit somewhat less frequently) 2kb and more from the TSS, and that this explains the bulk of the 20% of genes which are paired correctly based on the [-4000:2000] window, but this result gives a lower bound on the information contributed by promoter elements. Finally, I note that, although the plateau region shows a generally very good agreement between pairings from protein sequences and [-4000:2000] DNA sequences, the agreement is not quite so strong as that seen for the [-2000:0] region, suggesting that blocks further upstream might be more frequently shared between separate promoters.

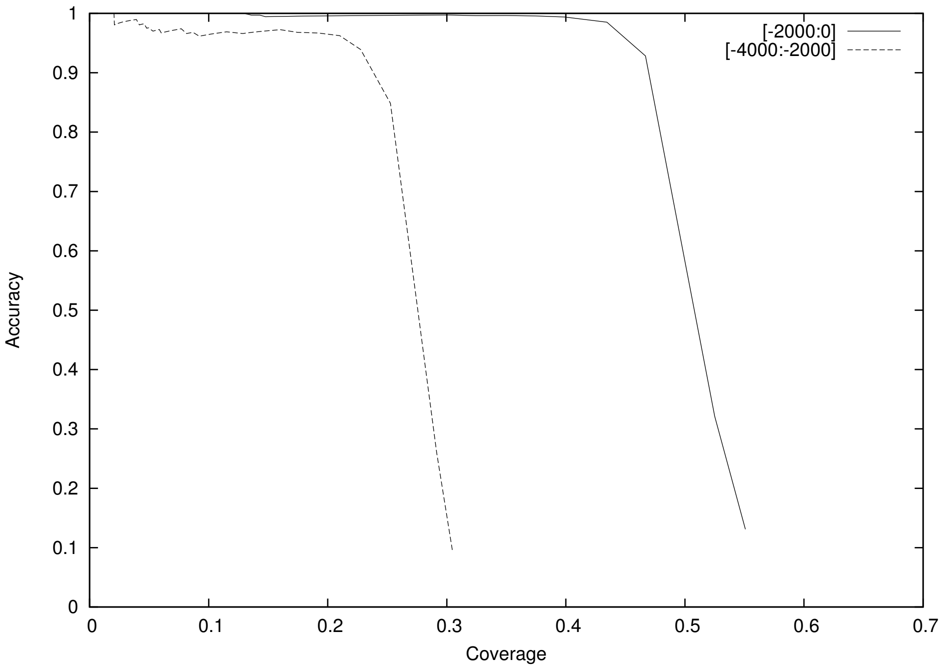

Figure 5.2. Comparison of orthology in the windows [-2000:0] and [-4000:-2000]

5.2. Relationship of promoter alignments to regulatory roles

The extent of alignments between orthologous promoters varies substantially, with many cases having no significant alignment (above **blastn**'s default threshold of 30 bits, which corresponds to a 15-base exact alignment), but others having several large sections of aligning sequence, in some cases covering over 50% of the 5kb region. The set of 2442 alignments include 1312 individual alignment blocks of 100 bases or more. Across this set, there is an average 88% nucleotide identity, with 414 blocks over 90% identity and 81 over 95% identity.

These are extremely well conserved pieces of sequence.

To investigate the significance of this variation, I made use of Gene Ontology (GO) terms [The Gene Ontology Consortium 2000], which are applied to the majority of Ensembl gene predictions on the basis of annotations by the GOA project [Camon *et al.* 2003]. Dividing promoters into groups with high (200 bases or more) and low total numbers of aligning bases, I counted GO terms which were overrepresented in the annotation of high-scoring sequences (tables 5.2 and 5.3). This approach is closely related to the comparison of found and unfound promoters in chapter 3. Some clear correlations can be observed in these tables: at the top of the high-alignment list, and overrepresented by factors of around 2, are genes involved in transcription and developmental processes. At the other extreme, genes taking part in the cell's basic metabolic activities are found predominantly in the short-aligning set. Genes annotated with the GO biological process ontology term for 'cell cycle' are also found predominantly in this set.

Focusing on genes annotated with the process term "transcription, DNA dependent" (figure 5.3) it can be seen that there is still a wide variation in number of aligning bases, but the proportions with larger amounts are consistently higher than that for the gene set as a whole. In particular, 8.0% of these promoters have 1000 bases or more of aligning sequence, compared to 2.4% for the set as a whole.

An more detailed way to view the variation is to look at the parts of the sequence which are actually aligning. A compact representation of this for the "transcription, DNA dependent" genes appears in figure 5.4, with each line of the figure representing one of the 250 upstream regions which matched the this term, and the red blocks indicating regions which show strong similarity to the orthologous mouse sequence. This figure also clearly shows that many genes have multiple aligning blocks, sometimes spaced quite widely apart in the 5kb region. For the cases where a moderate amount of sequence is aligning, this is generally, but not always, concentrated close to the transcription start site, fitting in with the view that promoter elements

GO term name	Freq. (short)	Freq. (long)	Diff.
transcription, DNA-dependent	0.088	0.173	0.085
transcription	0.088	0.173	0.085
transcription regulator	0.069	0.142	0.073
developmental processes	0.059	0.127	0.067
embryogenesis and morphogenesis	0.044	0.105	0.060
DNA binding	0.080	0.137	0.057
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.145	0.195	0.049
histogenesis and organogenesis	0.037	0.086	0.048
cell communication	0.218	0.258	0.039
ligand binding or carrier	0.366	0.404	0.038
nucleic acid binding	0.150	0.188	0.038
ectoderm development	0.017	0.054	0.036
cellular_component	0.558	0.590	0.031
biological_process	0.638	0.663	0.025
transcription factor	0.021	0.045	0.024
cell	0.505	0.528	0.023
Gene_Ontology	0.741	0.763	0.022
molecular_function	0.662	0.683	0.020
receptor	0.040	0.061	0.020
transcription regulation	0.022	0.042	0.019
integral membrane protein	0.063	0.079	0.016
protein kinase	0.036	0.052	0.015
defense response	0.013	0.028	0.015
mesoderm development	0.017	0.032	0.015
membrane	0.180	0.195	0.014
cation channel	0.009	0.023	0.014
metal ion transport	0.017	0.032	0.014
chromosome organization and biogenesis (sensu Eukarya)	0.004	0.018	0.013
transcription, from Pol II promoter	0.023	0.037	0.013
voltage-gated ion channel	0.005	0.018	0.013
response to biotic stimulus	0.026	0.039	0.012
DNA packaging	0.004	0.017	0.012
ion channel	0.013	0.025	0.012
protein modification	0.055	0.068	0.012
nuclear organization and biogenesis	0.006	0.018	0.012

Table 5.2. GO terms which are overrepresented in the long-aligning promoter set.

GO term name	Freq. (short)	Freq. (long)	Diff.
enzyme	0.267	0.209	-0.057
cytoplasm	0.153	0.105	-0.047
biosynthesis	0.064	0.028	-0.035
cell cycle	0.055	0.027	-0.028
oxidoreductase	0.033	0.006	-0.026
hydrolase	0.109	0.085	-0.024
catabolism	0.055	0.034	-0.021
nucleus	0.037	0.018	-0.019
transporter	0.089	0.071	-0.017
RNA metabolism	0.025	0.008	-0.017
mitotic cell cycle	0.027	0.011	-0.015
nucleotide binding	0.107	0.091	-0.015
purine nucleotide binding	0.107	0.091	-0.015
protein transport	0.041	0.027	-0.014
DNA replication and chromosome cycle	0.015	0.001	-0.013
macromolecule catabolism	0.038	0.025	-0.013
amino acid and derivative metabolism	0.017	0.005	-0.012
intracellular	0.432	0.420	-0.012
G-protein coupled receptor protein signaling pathway	0.022	0.010	-0.012
RNA binding	0.018	0.006	-0.012
S phase of mitotic cell cycle	0.013	0.001	-0.011
monovalent inorganic cation transporter	0.013	0.001	-0.011
hydrolase, acting on acid anhydrides	0.045	0.034	-0.011
hydrolase, acting on acid anhydrides, in phosphorus-containing anhydrides	0.045	0.034	-0.011
protein degradation	0.036	0.025	-0.011
cell fraction	0.052	0.040	-0.011
RNA processing	0.017	0.006	-0.011
protein binding	0.059	0.049	-0.010
macromolecule biosynthesis	0.025	0.015	-0.010
inner membrane	0.015	0.005	-0.010
ion transporter	0.019	0.010	-0.009
amino acid metabolism	0.011	0.001	-0.009
RNA splicing	0.011	0.001	-0.009
lipid metabolism	0.026	0.017	-0.009
cation transporter	0.017	0.008	-0.009

Table 5.3. GO terms which are overrepresented in the short-aligning promoter set.

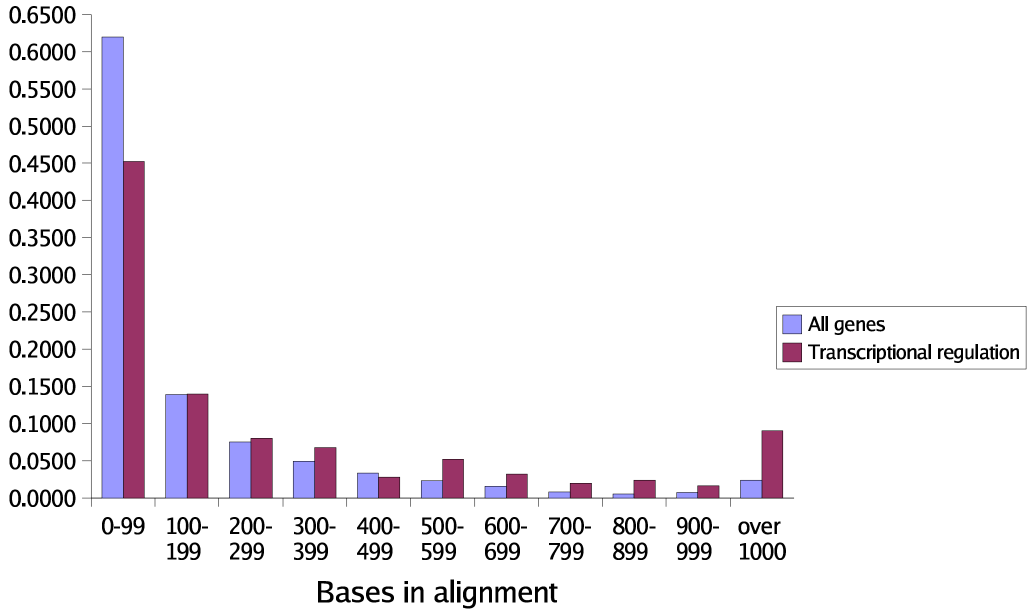

Figure 5.3. Histograms of sequences binned by number of bases of promoter sequence included in blastn alignments to the orthologous promoter.

are close to the transcription start site. It is possible that some exceptions to this rule reflect alternative promoters.

Dot plots provide a detailed view of the similarities between two sequences. A number of mouse-human pairs were examined in this way, using the **dotter** tool [Sonnhammer and Durbin, 1995]. Representative examples from the top, middle, and bottom of figure 5.4 appear in figures 5.5, 5.6, and 5.7 respectively. In all cases, the 5kb region of human sequence is represented on the horizontal axis, while 6kb of orthologous mouse sequence is shown on the vertical axis, and dark dots indicate sequence similarity. The most interesting of these plots is figure 5.7, which shows that similarity does, indeed, continue right across the 5kb upstream region. The line of dark points does not quite follow a perfect diagonal, indicating that there have been some minor insertion or deletion events. However, there is no evidence of either major rearrangement or local inversions.

Of course, "transcription, DNA dependent" is a rather broad category (including over 10% of the gene set considered here). The long-aligning subset is dominated by transcription factors, with four homeobox genes in the top ten. At the other extreme, the group of promoters with no

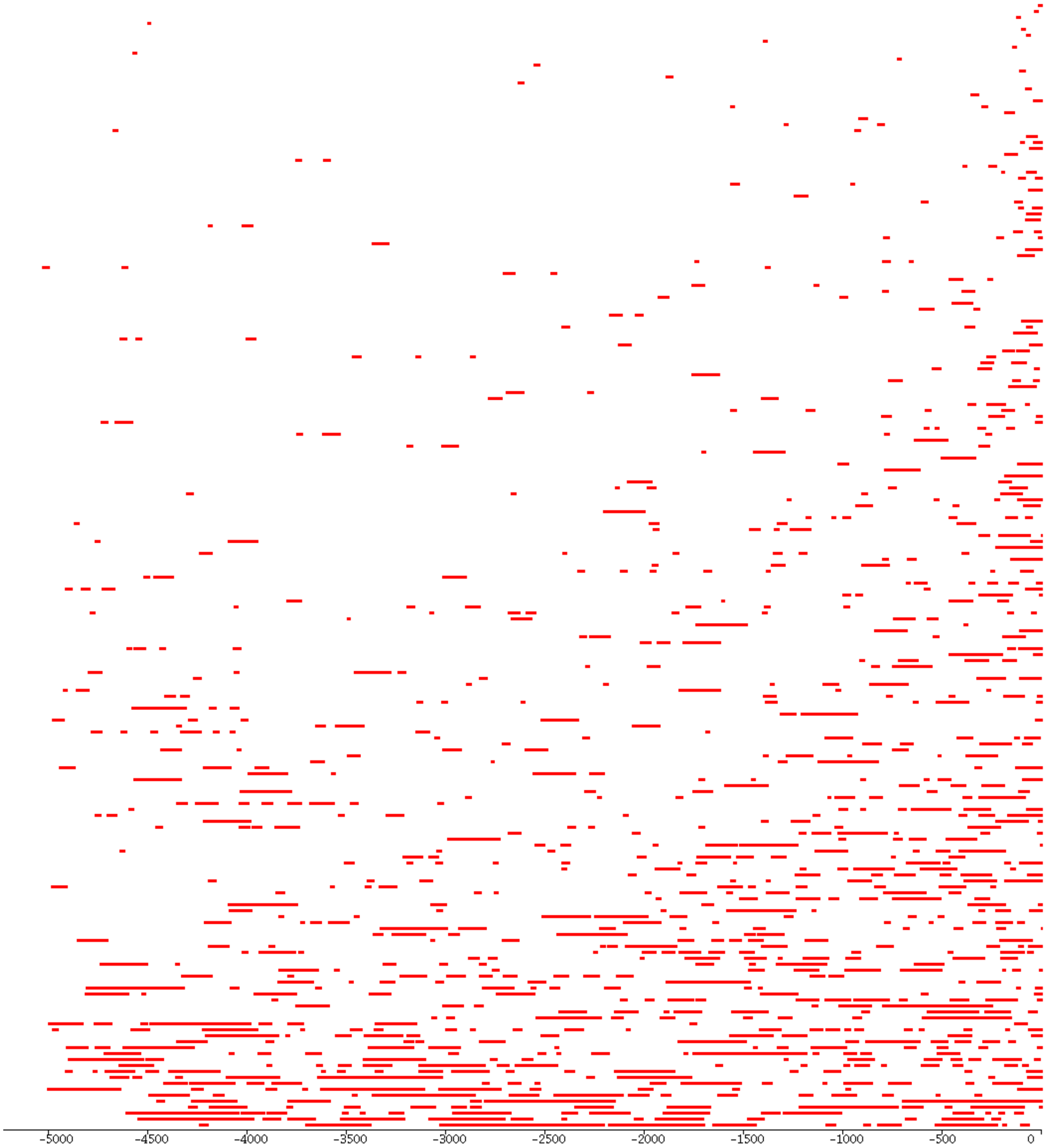

Figure 5.4. Aligned regions from 250 transcription-associated genes, sorted by number of aligning bases.

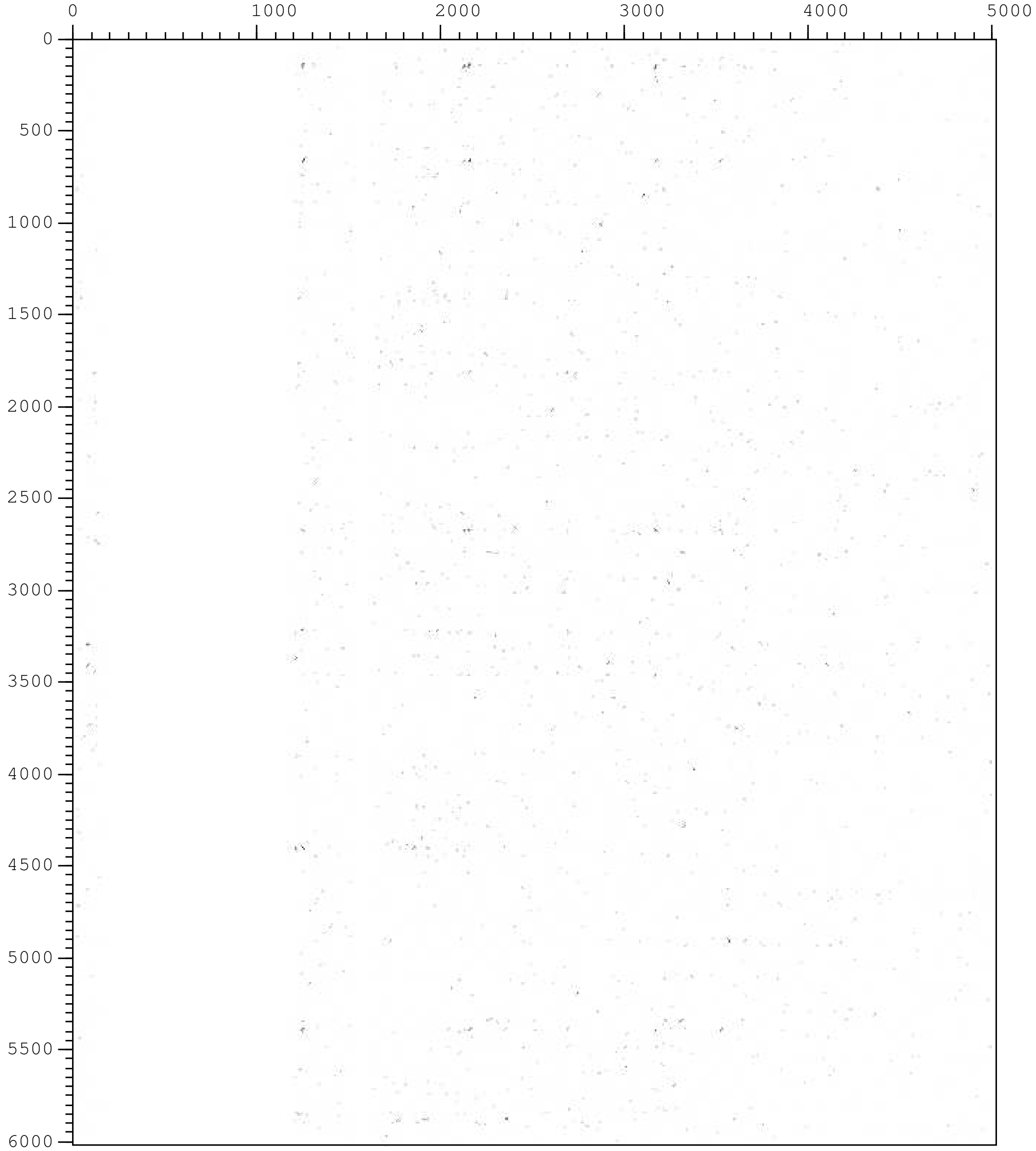

Figure 5.5. Dot plot between human and mouse upstream regions with low similarity.

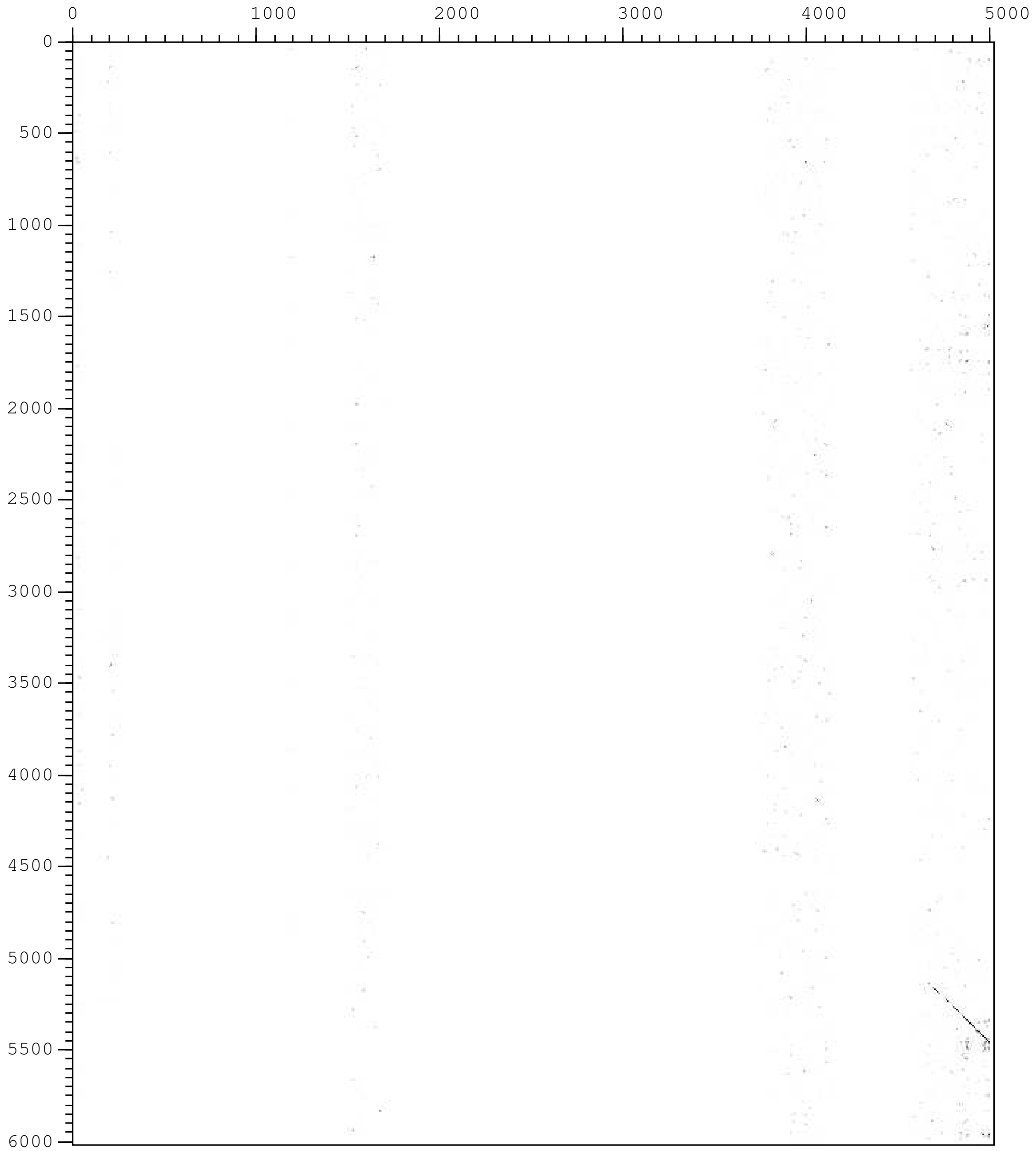

Figure 5.6. Dot plot between human and mouse upstream regions with moderate similarity.

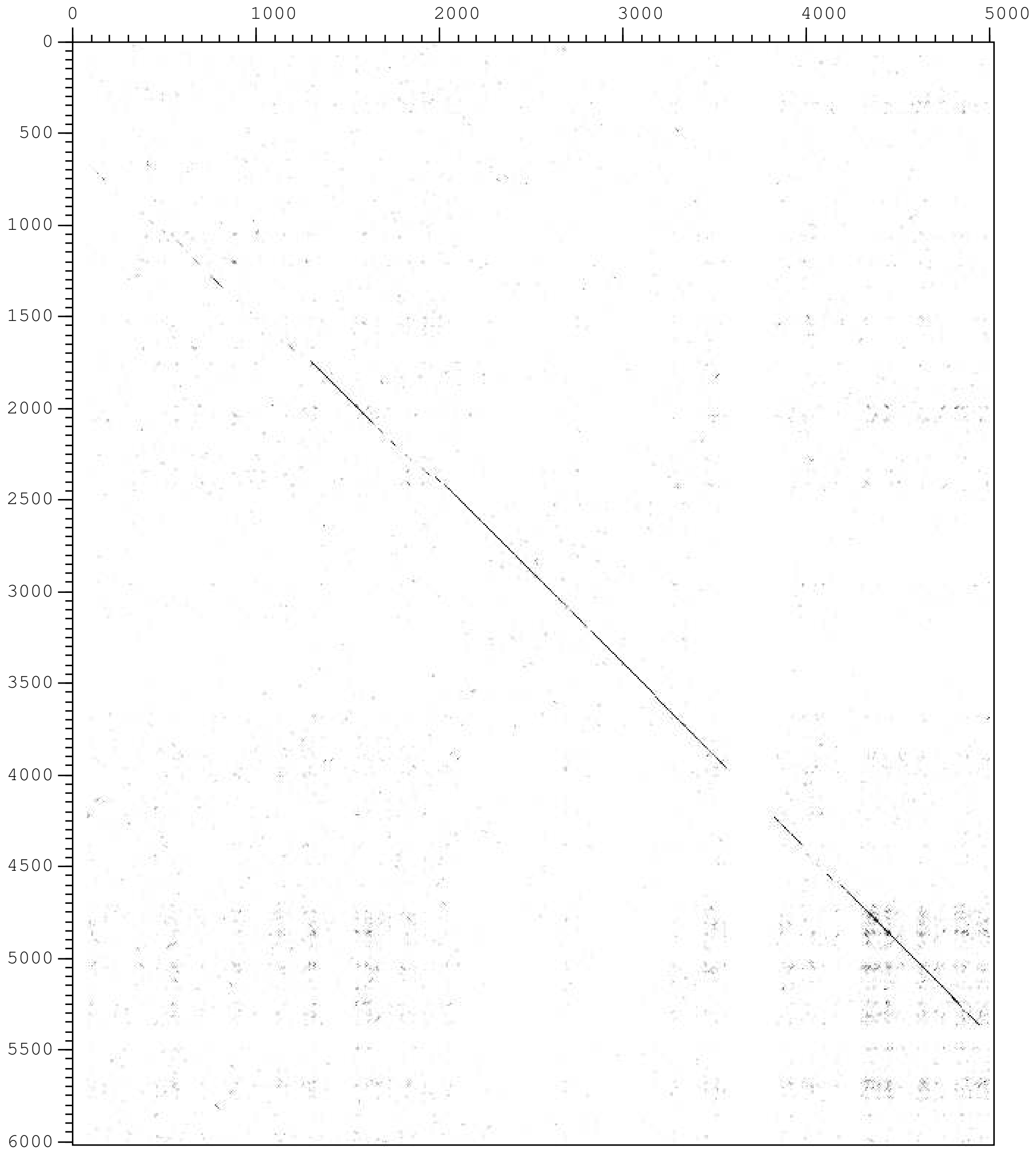

Figure 5.7. Dot plot between human and mouse upstream regions with strong similarity.

detectable alignment at all includes subunits of *polII* RNA polymerase.

5.3. Discussion of promoter conservation

I this chapter, I have shown that 40% of human promoter sequences have an equivalent with at least one strongly conserved block in the mouse genome which can be detected and distinguished from background noise using the common **blastn** method. Once low-complexity sequences are masked, there are very few cases where a comparably strong match is observed to a sequence other than the promoter of the orthologous gene. I believe that promoter sequences evolve in a manner which is closely linked to the genes they control, and do not see evidence to suggest that a widespread exchange of promoters between genes has occurred, at least in the timescale of the divergence between human and mouse. However, between more distantly related genes, as identified by the TRIBE-MCL clustering, there is little or no identifiable promoter similarity. This suggests that gene duplication (presumably followed by a change in either function or expression pattern of one copy) is likely to be accompanied by fairly radical changes in the promoters. This is interesting but not entirely surprising, since, following a gene duplication, if both copies are to remain active, one must rapidly change to fill some alternative role in the organism, otherwise there is unlikely to be any selection pressure against an accumulation of mutations in one of the copies, eventually leaving a pseudogene. There do not appear to be any large contiguous conserved elements – for example, very complex regulatory modules – which appear in multiple promoters, since these would appear as non-orthologous pairings in this analysis. This is not, of course, inconsistent with the existence of short conserved elements (such as individual transcription factor binding sites, or perhaps even small modules consisting of two or three such sites).

Given the fall-off of similarity when looking further from the transcription start site, I am confident that the bulk of the alignments produced by this method reflect *bona fide* functional sequence, rather than background synteny. The observed correlation between alignment length and gene function adds further support to this view.

As previously mentioned, there is currently no purely computational method to delineate promoter regions (as opposed to transcription start sites) – while PromoterInspector and the EponineHomol models from chapter 4 do predict regions rather than specific TSS points, given the methods used for training these models there is no reason to think that the predicted region accurately matches the region of sequence which has biological regulatory significance. Similarly, experimental delineation of full promoter regions cannot yet be performed in a high-throughput way across the genome. However, the extremely high level of nucleotide identity in many of the blocks detected by this approach, combined with the increased number of conserved blocks closely upstream of the transcription start site, gives a strong indication that these blocks represent functionally important elements – presumably promoter signals. This seems to be further reinforced by the results from chapter 4 which show that a promoter signal can be learned from a random selection of blocks of mouse-human similarity. This means that these high-similarity blocks would seem to be good initial targets for further computational work to discover individual regulatory motifs and modules and learning more about regulatory mechanisms.

The utility of having an approach to highlight promoter regions is not confined to computational methods. One currently promising experimental method of investigating DNA-protein interactions in a relatively high-throughput manner is ChIP-on-chip: chromatin immunoprecipitation followed by hybridization with probes on a DNA microarray [Ren *et al.* 2000]. While it is possible to place a large number of probes on a microarray, reasons of cost and convenience mean it is still important to be selective. From figure 5.4, it is clear that even among the promoters for one family of genes, some only have large conserved blocks close to the transcription start site, while others show conservation right across the 5kb region studies here. In fact, re-aligning longer sequences for a few individual promoters suggested that some of these might extend 8kb or more upstream of the TSS. Clearly, it is worth including probes much further upstream for these genes.

And additional interesting observation from this study is that there is a fairly clear correlation between the length of aligning sequence in a promoter region and the function of the gene regulated by that promoter. I can envisage two possible causes for this. Firstly, in order to fulfill their roles, some genes require many more regulatory "inputs" than others – in a graphical view of a regulatory network [Pilpel *et al.* 2001], some genes will appear as hubs with many connections, while others will appear as leaf nodes. Large numbers of inputs seem likely to correspond with large regions of functional promoter sequence, containing many different transcription factor binding sites. Secondly, even within the functional regions, some promoters may be more labile than others – radical changes can occur without compromising the basic viability of the organism. Both these factors may be significant. I note that cell-cycle genes tend to have short-aligning promoters, yet they represent a well-conserved biological process, where correct regulation is important for survival. However, the regulation of these genes may be quite simple: they need only be expressed in response to very specific signals at one particular point in the cell cycle. In budding yeast, a set of nine cell-cycle regulators have been identified, and most cell-cycle-regulated genes (including these regulators themselves) only respond to a small subset of these regulators [Simon *et al.* 2001]. If the situation in vertebrates is similar, then is seems likely that a relatively compact promoter, with a small number of conserved transcription factor binding sites, could be sufficient to express the required regulatory logic. In conclusion, I suggest that promoter size gives a good indication of a gene's "regulatory complexity", and gives some prediction of the gene's function.

Overall, I believe comparative genomics presents interesting opportunities for research into promoters and regulatory regions, especially in terms of defining the boundaries of functional regions so that they can be studied in more detail with other methods, such as the learning techniques from previous chapters. In the future, it will be interesting to consider more sensitive methods for aligning promoters. However, simply increasing the sensitivity is not a panacea, since it increases the risk of detecting similarities between sequences where there is not actually any functional reason for conservation. There has been some interest in methods to distinguish between alignments covering regulatory regions and alignments between regions of change conservation. In [Elnitski *et al.* 2003], a number of classification methods are used to distinguish between a training set of known regulatory elements and some conserved neutral sites (actually ancient transposons). Some of these methods are effective, but it seems likely that they will only be applicable to relatively long aligned regions – not the short, very strongly conserved blocks discovered by the **blastn** alignments here. The biggest benefits from this approach may be for identifying regulatory possible regulatory regions – such as enhancer elements – far from known genes and transcription start sites, rather than in the analysis is proximal promoters.

Another possible direction of research is to investigate promoter-specific sequence comparison methods, which might give superior performance to generic alignment techniques: for instance, conventional aligners would not give good a good score for a pair of sequences where localized rearrangements or small inversions had taken place, yet these two sequences might actually include the same repertoire of transcription factory binding sites, and have similar effectiveness and specificity as promoters. In defense of "simple" alignment algorithms, however, inspection of the dot plots accompanying this chapter, and several similar plots not shown here, did not show strong evidence for rearrangements or inversions. Finally, it may be that mouse-human comparisons do not represent the optimal evolutionary "distance" for detecting promoter regions. As more genomes are completed, other species may take over as targets for regulatory comparative genomics, or perhaps comparisons of more than two species will be used to improve the confidence of the results. Data from the ENCODE project, which will select various regions from the genome then sequence their equivalents in a large number of different vertebrates [ENCODE, http://www.genome.gov/Pages/Research/ENCODE/], should provide a good testbed to determine the most informative pairs of genomes to compare, and the added value of considering more than two species at once.