Chapter 4. Learning from comparative genomics

Comparative genomics is the study of similarities and differences between two or more genome sequences. In its simplest form, it is based on the assumption that regions of the genome which perform important biological functions are more likely to be conserved between species than non-functional "junk". This is intuitively quite logical, since at least some proportion of mutations in functional regions would be expected to have deleterious phenotypes, and be selected against, while mutations in non-functional areas should have no phenotype, positive or negative. The availability of the mouse genome, with close similarities to human, has caused much excitement in this field. In fact, there are a variety of genomes which make attractive targets for comparison – for example, the pufferfish *Tetraodon nigroviridis* has been proposed as a useful target for vertebrate comparative genomics, especially because of its small genome size, presumably containing little non-functional DNA [Roest *et al.* 2000].

To date, the main application of comparative genomics, at least in vertebrates, has been in the prediction of coding genes, with approaches such as Exofish [Roest *et al.* 2000] and Twinscan [Korf *et al.* 2001], and DoubleScan [Meyer and Durbin, 2002] utilizing similarity information from a variety of distances to predict genes with a greater confidence that purely *ab initio* methods, while still not requiring direct experimental data. Protein coding regions are among the best conserved regions, and sensitivity can be increased further by using an alignment model such as WABA [Kent and Zahler, 2000] which recognizes that certain nucleotide changes, mainly in the last position of a codon, are synonymous and have no effect on the final protein sequence. This sensitivity allows distant comparisons such as human-fish to yield interesting information. But other types of functional region also show conservation, and other types of comparative genomic application are being developed, for example in the

prediction of functional non-coding RNA genes [Rivas and Eddy 2001, di Bernado *et al.* 2003], where the expectation is that the secondary structure of the RNA will be more conserved than its actual sequence.

The human and mouse genomes show quite substantial similarity. An initial survey of these was published along with the draft mouse genome [MGSC 2002]. Briefly, coding genes are – as expected – among the most strongly conserved regions, but homologous regions can be observed throughout the genome. In total, it is possible to align up to 40% of the mouse genome to human [Schwartz *et al.* 2003], but it seems likely that at least some of this is just random "comparative noise" – regions of sequence which serve no particular purpose but which, purely by chance, have not yet accumulated enough mutations to make their relationship unrecognizable. However, it seems clear that some of the noncoding-but-similar regions, especially those with the highest levels of sequence identity between the two species, must have biological explanations.

Here, I suggest an alternative approach to comparative genomics, and present an example of its application in the analysis of mouse-human homologies. I chose to take a set of so-far-unexplained regions of strong similarity between two species, and try to identify what they have in common. This could be described as an ignorance-driven approach to scientific research, and is significantly different from the traditional approach of proposing hypotheses then testing them. However, it is an attractive way to explore large data sets. To make this strategy feasible, I used a machine learning approach in order to detect significant patterns in the chosen set of sequences. There have been some prior indications that this might prove to be a useful strategy. In the malarial parasite, *Plasmodium falciparum*, unsupervised training of a rather simple hidden markov model can partition the genome into several portions, each modeled by one state of the HMM. Moreover, some of the learned model states effectively identify expressed regions – and the direction of transcription – without having to supply any prior knowledge of what a Plasmodium gene might look like [Pocock 2001].

The HMMs of [Pocock 2001] just detect fairly simple biases in the frequency of either

single nucleotides or pairs of nucleotides in particular regions of the genome. This was effective in one case, but I wished to develop a method which could capture more complex information about regions of the genome, and this motivated the development of the Eponine Windowed Sequence model family.

4.1. The Eponine Windowed Sequence (EWS) model family

Previously, I introduced the Eponine Anchored Sequence (EAS) model, which was a generalized linear model classifier for points within a large sequence. From a biological point of view, this kind of point classifier is a principled way to look at questions like transcriptional activation, since while the signals may be spread over a considerable area, their activity is focused on a single point or small set of points. However, it is often not possible to exactly define the points that are interesting. This is particularly true in this case, where a training set of sequences is being proposed as "interesting" purely on the basis of similarity to another species, with no prior knowledge of function. As well as having no indication of which point (if any) should be used as an anchor point, there will not even be any indication of the orientation in which the sequence should be considered.

The Eponine Windowed Sequence model is a generalized linear model approach to sequence analysis, following many of the same principles as EAS, and sharing many details of practical implementation, but it is directly applicable to regions of sequence. EWS models were designed to be trained with the previously described variational RVM library, using the same sampling-based strategy as EAS. The EWS model as used in this chapter is adirectional by design, but if it is to be applied to data from alternative sources, where the functional orientation of the training sequences *is* known, directionality could be reintroduced with only the most trivial changes in the computation.

In the first version of EWS, each basis function consists of a single position-weight matrix

(see definition, page 19). The basis function score is a normalized sum of the position-weight matrix probability (see page 19) at every position along the window of sequence under consideration:

$$\phi(S) = \sum_{i=1}^{|S|-|W|+1} (\overrightarrow{W}(S,i) + \overleftarrow{W}(S,i)) \frac{4^{|W|}}{|S| - |W| + 1} \tag{4.1.1}$$

Where $|S|$ is the length of sequence S and $|W|$ is the length of PWM W. Dividing the scores by $|S| - |W| + 1$, the number of positions at which a motif of length $|W|$ could begin, makes this scoring function independent of the exact size of the windows under consideration, making training on examples with a range of lengths possible. Like the basis functions of the EAS model, scores are also normalized for motif length: the $4^{|W|}$ term avoids the basis function outputs taking very small values, which could cause numerical precision issues in the training algorithm.

Since the basis function space for this model was somewhat smaller than for EAS, a simpler strategy could be adopted for making initial choices of basis functions. Rather than starting with fragments picked from the training sequence, it was practical to enumerate all possible words of a specified starting length. These were shuffled onto a queue in a random order. When the trainer required a new basis function, a word was taken from the head of the queue and a new PWM proposed which preferentially matches this word. Once the queue is empty, the words are re-shuffled for another cycle through the pack. Once motifs have been proposed, the remaining sampling moves in the training procedure were similar to those used for EWS: weight matrices can be altered by re-sampling the columns from Dirichlet distributions, or can be shortened by dropping a column from either end. In this case, extensions of existing weight matrices are not permitted, so all the PWMs in the final model will have a length less than or equal to the starting length.

The second version extends EWS to capture larger-scale patterns in sequences. In this case, each basis function is a scaffold consisting of one or more PWMs, each with an associated position distribution relative to a scaffold anchor. In principle, distributions such

as the discretized Gaussian extend to infinity, but in practice it is reasonably to apply some cut-off: for instance, only considering the portion of the distribution which includes 99% of the total probability mass. The probabilities of all points outside this region are assumed to be infinitesimal and ignored. Now that the distributions have finite size, for a given scaffold there is a pair of integers, n and m, such that when the scaffold anchor is placed in the interval [n : m], the non-infinitesimal parts of all the position distributions fall entirely within the length of a particular target sequence. A score for scanning this scaffold across the sequence can be given by

$$\phi(S) = Z \sum_{i=n}^{m} \left(\prod_{k=1}^{K} \left(\sum_{j=-\infty}^{\infty} P_k(j) \, (\overrightarrow{W}_k(S, i+j) + \overleftarrow{W}_k(S, i+j)) \right) \right) \tag{4.1.2}$$

where P_k is the k'th position distribution and W_k is the k'th weight matrix in the scaffold. Z is the normalizing constant:

$$Z = \frac{4^{\sum_{k}^{K} |W_k|}}{m - n + 1} \tag{4.1.3}$$

For the case when the scaffold only contains one PWM with a narrow distribution, the results will be the same as those from equation 4.1.1. So this can be considered a direct extension to the basic EWS model that can capture information about sets of motifs with correlated positions. Since the scaffold scores are only evaluated in the regions where the whole scaffold fits onto the sequence window, there is a risk of introducing edge effects. A possible future solution to this would be to use windows that are a little larger than the actual region, of interest, and use "soft boundaries" where scores from the edges of the window are given less weight than those at the centre. To train scaffold-based models, some additional sampling rules are needed:

- Combine the sets of motifs from two scaffolds, with randomly chosen offsets between the two
- Take a scaffold with two or more PWMs and return the scaffold with one of those PWMs

(picked at random) removed

- Alter the position or width of one of the relative position distributions in a scaffold.

The inclusion of scaffolds which could, for example, model two transcription factor binding sites with some preferred spacing between them, makes this second variant look more similar to EAS. In the implementation used here, scaffolds were limited to a maximum of three motifs. An example EWS-scaffold model is illustrated in figure 4.6.

4.2. Training from non-coding homologies between human and mouse

Here, I used the EWS model to investigate a set of strong non-coding homologies between the human and mouse genomes. A number of methods have been developed for aligning two genomes [Schwartz *et al.* 2003]. These methods all use optimized sequence-search algorithms which trade some sensitivity for better computational performance. Nevertheless, whole-genome alignment is a computationally very demanding task, so I was keen to use an existing publicly-available set of results rather than running a new set of analyses.

At the time, the main set of publicly downloadable mouse-human alignments was provided by the Ensembl project in their ensembl-compara database [Clamp *et al.* 2003, A. Ureta-Vidal, personal communication]. This data was produced by first using the Exonerate program [G. Slater, unpublished] to perform a very rapid search for strong matches between the two genomes, which were used as "seeds" for the alignment process. When two seed hits occurred close to one another on both genomes, attempts were made to extend the alignments further by running bl2seq [Tutsova and Madden, 1999] (an implementation of the blast algorithm specialized for aligning two sequences rather than searching a database) on the regions of sequence lying between pairs of adjacent seed hits on the respective assemblies.

The compara database from Ensembl release 5 contained results from this protocol on release 5.28 of the draft human genome and release 5.3 of the mouse whole-genome shotgun

assembly. This gave 559,670 regions of similarity, covering a total of almost 93 megabases. While this is a substantial amount of sequence in absolute terms, it represents only 3.4% of the sequenced bases in the mouse genome, and somewhat less than that for the larger, more completely sequenced, human genome. The individual regions ranged from 20 to 8581 bases in length, and from 71% to 100% nucleotide identity. The bulk of the sequences are towards the lower end of this length range, as shown in figure 4.1

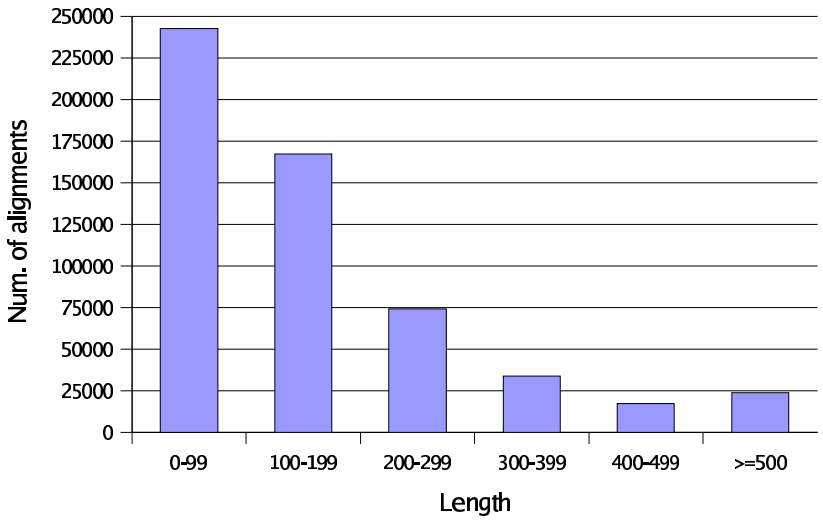

Figure 4.1. Length distribution of sequence regions aligned between human and mouse by ensembl-compara version 5.

One of the strongest contributors to the fraction of sequence in the compara data set was protein-coding genes. These represent a reasonably well-understood fraction of the genome, and were already known to be well conserved, so for the purposes of this particular study they were considered as "uninteresting". Therefore, all similarity regions which overlapped an Ensembl gene prediction in either species were excluded. Note that the UTRs of predicted transcripts were also excluded at this point, but these might in fact make an interesting target for detailed analysis in the future using similar techniques to those presented here. Compara sequences overlapping repetitive elements, as detected by the RepeatMasker program [Smit and Green, unpublished] were also removed at this point, but these represented a very small fraction of the total. This is partially due to the use of repeat masked sequence in the seeding stage of the alignment process, but also since some of the largest families of repeats,

such as the Alu elements of the human genome, proliferated fairly recently in evolutionary time, and therefore only occur in one lineage.

The final restriction which must be applied is a minimum length threshold, since short sequences contain only a small number of sequence words, and are thus unlikely to give interesting results when analysed with the EWS model. For this experiment, I arbitrarily picked a sequence window size of 300 bases. Sequences which were shorter than this were discarded, while longer sequences were trimmed, and only a (randomly selected) 300-base portion from each sequence was used. This clearly removed a large number of sequences from consideration. However, this still left 75332 sequences, so there was little incentive in this case to analyze the shorter regions of similarity, or attempt to extend them with more sensitive alignment methods. In fact, training an EWS model on a dataset of that size would have been prohibitively slow (the limiting factor actually being the vast number of PWM score evaluations, rather than the actual RVM learning algorithm), so for this experiment I picked random subsets of size 2000, 4000, and 6000 to use as training datasets. Since EWS is still a binary classification model, a negative dataset was also required. In this case, randomly chosen non-coding, non-repetitive fragments from human chromosome 20 were used. This gave a roughly equal mixture of intron and intergenic sequence, with a small number of UTR sequences.

Single motif EWS models were trained from each of these sets with proposed words of length 5. The final motif sets from three runs are shown in table 4.1. These sets vary significantly from run to run, raising the serious possibility that the training process was not actually detecting any specific signal. The models were initially tested by scanning a two megabase region of human chromosome 22, considering non-overlapping 300 base windows. The scores for models 1 and 2 from table 4.1 are plotted in figure 4.2. While the two models do not agree precisely, a strong correlation can be seen. This can be quantified by calculating the Pearson correlation coefficient, in this case $r = 0.89$, which confirms the visual impression of a strong, but not perfect, correlation between the models. This correlation value was typical for other pairs of models tested.

Positive forward	Positive reverse	Negative forward	Negative reverse
aaagc	gcttt	aaaaa	ttttt
aatta	taatt	atttg	caaat
agc	gct	ccacc	ggtgg
catac	gtatg	cgtgc	gcacg
cccac	gtggg	ctagc	gctag
ccgc	gcgg	ctca	tgag
cggta	taccg	cttac	gtaag
gacga	tcgtc	gatgc	gcatc
gcga	tcgc	gcgac	gtcgc
tcaca	tgtga	gggaa	ttccc
tccaa	ttgga	tagga	tccta
		tata	tata

Positive forward	Positive reverse	Negative forward	Negative reverse
aacac	gtgtt	aaaaa	ttttt
atgga	tccat	aaata	tattt
attag	ctaat	attc	gaat
attg	caat	caccc	gggtg
cagc	gctg	ccca	tggg
cagta	tactg	ccga	tcgg
ccgac	gtcgg	ctagc	gctag
cgaaa	tttcg	cttac	gtaag
cgac	gtcg	gacca	tggtc
cgcca	tggcg	gtga	tcac
ctccc	gggag	gttga	tcaac
cttta	taaag	tagca	tgcta
tcaca	tgtga		

Positive forward	Positive reverse	Negative forward	Negative reverse
aacgc	gcgtt	aataa	ttatt
aactt	aagtt	accca	tgggt
aataa	ttatt	ag	ct
agccg	cggct	caac	gttg
agcg	cgct	cctga	tcagg
atatc	gatat	ctagc	gctag
atgac	gtcat	ctttc	gaaag
attag	ctaat	gcgc	gcgc
catca	tgatg	ggaca	tgtcc
cga	tcg		
cgaca	tgtcg		
ctgtc	gacag		
cttta	taaag		
gctga	tcagc		

Table 4.1. Words learned from three EWS models trained from mouse-human homologies.

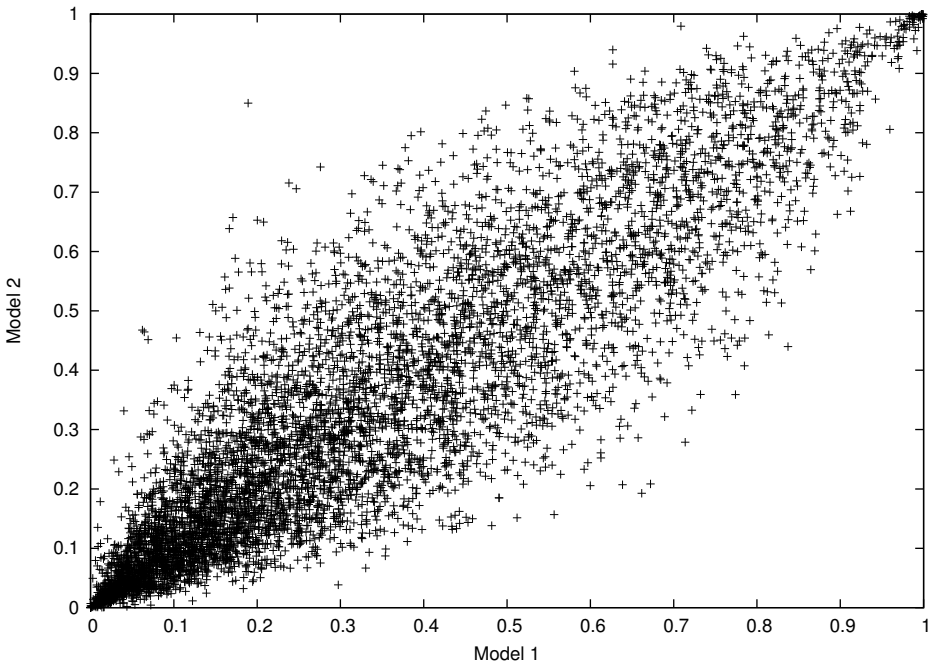

Figure 4.2. Scatter of scores from two different EWS models.

Given these strong correlations, it seems reasonable to assume that the training process locked on to some specific signal in the homology dataset. In order to get some impression of what this might be, I used one of the models to scan 300 base windows, with 200 base overlaps between adjacent windows, across the whole of chromosome 22, and set a threshold of 0.95, giving 985 high-scoring windows – a reasonable number for visual inspection. These were published using a DAS protocol server [Dowell *et al.* 2001] and viewed, together with other types of annotation, using the Ensembl genome browser. Some typical displays are shown in figure 4.3. Note that there is only one track of EWS predictions in each image, since the EWS model is adirectional, and its predictions cannot be placed on one or other DNA strand. Since these results were just taken from human chromosome 22, which is well covered with manually-curated gene annotations, the figures show the curated gene structures (available in Ensembl as "Vega transcripts" [http://vega.sanger.ac.uk/]) rather than Ensembl predictions.

This examination gave some indication of what the EWS models might be learning. Firstly, it is clear that the model has not learned a signal characteristic of coding genes – always a concern, since while homologies overlapping existing Ensembl genes were removed from the

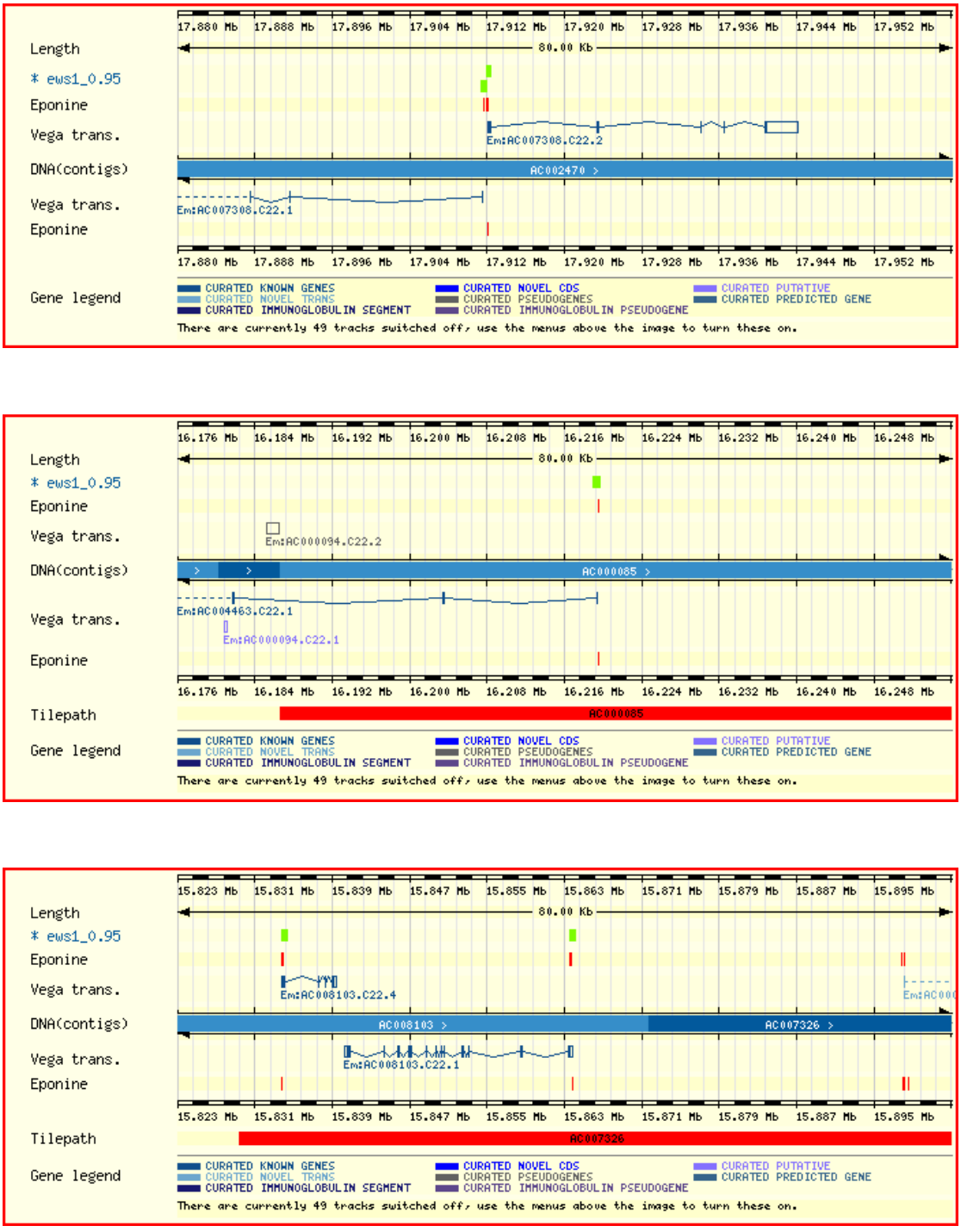

Figure 4.3. Ensembl contigview displays for selected portions of human chromosome 22, showing windows with high scores for one of the homology models (labeled "ews1_0.95"), and predictions from the EponineTSS_2 model (labeled "Eponine").

training set, it seems likely that at least a small number of coding sequences will have remained, reflecting false negatives from the Ensembl gene prediction pipeline. The bulk of predictions do not significantly overlap annotated exons, and for those which do it is almost always the first exon of a gene, which often consists largely or entirely of 5' untranslated sequence. The fact that this experiment did not pick up a protein-coding signal can be seen as evidence that there is probably not a large amount of undiscovered coding sequence in the genome. It is, however, clear that the predictions are concentrated near the start of genes. Moreover, they are strongly correlated with predictions from the EponineTSS_2 model.

Based on this visual inspection, I provisionally concluded that the learned models were acting as promoter predictors, a result which was later backed up by the evaluation in the next section. It is not initially obvious exactly what aspect of the promoter signal is being detected by these models. Considering table 4.1, there are a number of words in this set which include the 5'-cytosine-guanine-3' dinucleotide: 11 such words with positive weights, against just 3 with negative weights. CpG dinucleotides are known to be significantly in promoters, both by inspection of the EponineTSS models, and from previous knowledge of CpG islands. However, given the number of motifs which were learned, it seems unlikely that this method is relying solely on this signal. It seems more plausible to think that the learned set of motifs consists of some CpG island signals, plus additional motifs reflecting common transcription factor binding sites.

One approach to check for transcription factor binding sites in the model is to compare the learned motifs with the TRANSFAC database [Matys *et al.* 2003]. Unfortunately, this proved not to be practical. The set of binding sites in the TFSITE database actually includes all 1024 possible 5-base nucleotide sequences. Obviously, most of these are embedded in longer sequences, but it means that there is no way to rigorously compare TRANSFAC with the set of 5-base motifs listed here. Moreover, moving to slightly longer motifs would not improve the situation significantly: of the 4096 possible hexamers, only 15 cannot be found in TFSITE.

4.3. Evaluating the function of mouse-similarity models as promoter predictors

To test the EWS models as a practical mechanism for predicting promoters in genomic DNA, I took the same approach previously used to evaluate the EponineTSS predictor. The models were used to scan the pseudochromosome described on page 67, again using overlapping 300 base windows, and the set of high-scoring windows was compared with the curated annotation using the same criterion of accepting predictions within 2kb of the annotated transcript start. Once again, receiver operating characteristic curves (accuracy *vs.* coverage) were plotted. As already shown, the learned models varied somewhat between training runs. I therefore trained three models from each dataset and calculated the mean and standard deviation statistics of the accuracy score at a range of coverage figures. Results of this analysis for models trained on two of the datasets are shown in figure 4.4. For higher levels of coverage the models from the 4000 sequence set gave significantly higher accuracy. However, training from the 6000 sequence set did not give any further improvement (data not shown for clarity reasons, since the results closely overlapped those for 4000 sequence models).

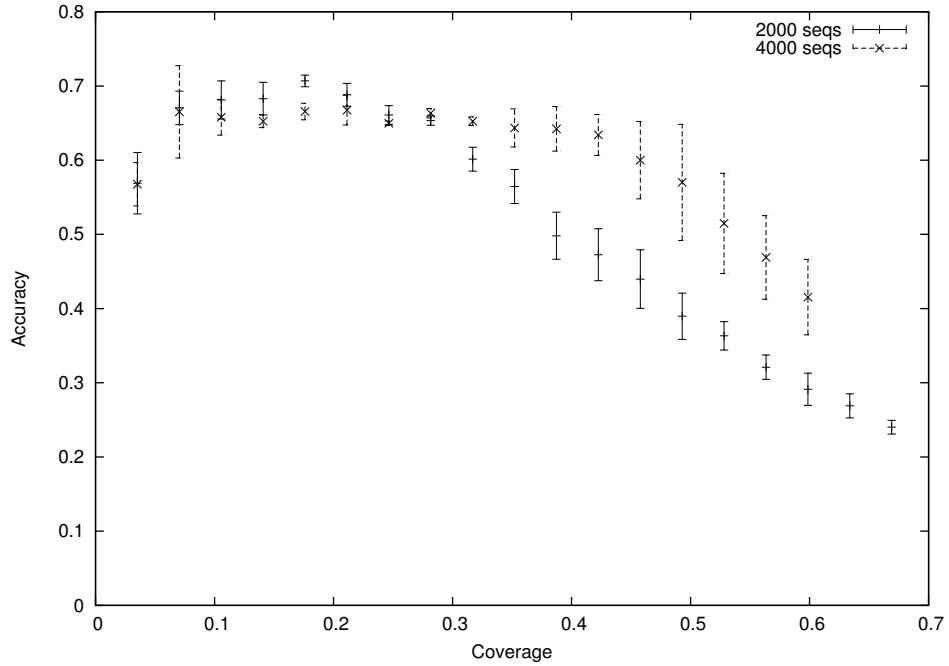

Figure 4.4. ROCs for EWS models trained from sets of 2000 or 4000 human-mouse homologies.

To optimize the initial word-length, a second set of models were trained using proposed words 6 bases in length. These all offered lower accuracy than the 5-base models, but the difference decreased with larger sets of training data. For models trained with 6000 sequences, the differences were minimal. Thus, it seems that models including longer motifs require larger amounts of data to train effectively. But it was not possible to improve the prediction performance significantly significantly above what could be achieved using 5 base words.

One of the 4000-sequence, 5 base word models was selected for further study: from now on, this is labeled as EponineHomol_1. The full set of motifs in this model are shown in table 4.2. At a threshold of 0.97, this gave a coverage on the pseudochromosome of 41% and an accuracy of 68%. These figures are both a little lower than those for the EponineTSS_2 model, but still indicate significant predictive power. In fact, the accuracy is slightly higher than that for the EponineTSS_1 model at a comparable coverage level. Of course, the two are not directly comparable since the EponineTSS models also give information on the actual position of transcription initiation, while the models discussed here simply indicate regions of 300 bases or more which are likely to lie in the vicinity of transcription start sites.

Next, I investigated which promoters were detected by the EponineHomol_1 model. Of the 284 gene starts on the pseudochromosome, 116 were correctly predicted by the EponineHomol_1 model at the chosen threshold. 107 of these were also detected by EponineTSS_2 (see figure 4.5. If the two methods made predictions entirely independently, the expectation would be that only 62 out of 116 would coincide by chance, so this would seem to be evidence for a significant correlation. As discussed previously, the set of promoters found by EponineTSS_2 could quite plausibly be biased either by the set of promoters which had been submitted to EPD, or by the set of mRNAs which were successfully cloned in the FANTOM project. Neither of these dataset-related biases could have had any effect on the training of the EponineHomol_1 model. Thus, this coincidence suggests that the distinction between the “found” and “unfound” sets of genes discussed in chapter 3 must have some deeper biological significance.

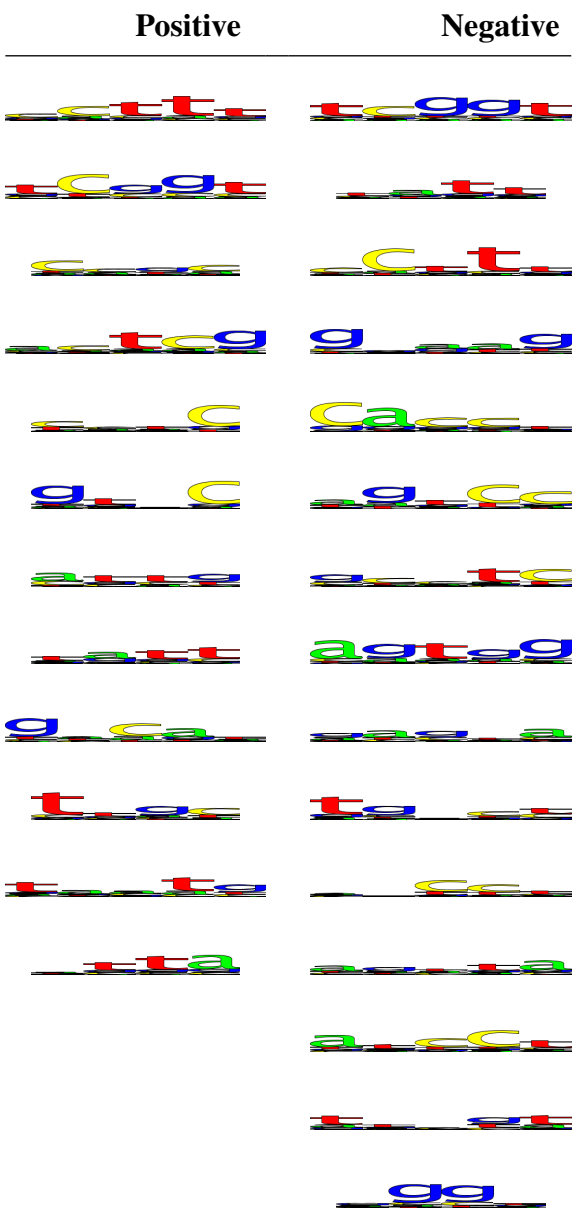

Table 4.2. Logo view of the motifs used in the EponineHomol_1 model.

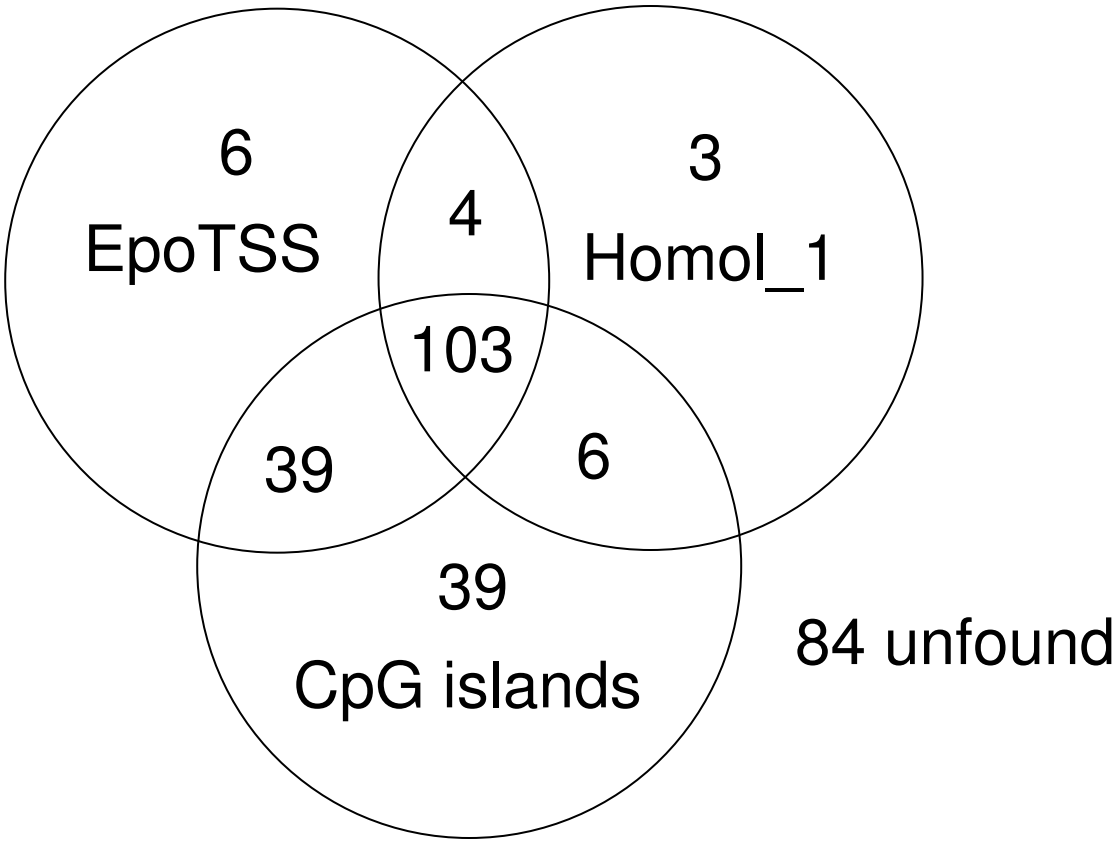

Figure 4.5. Intersection of transcription start sites correctly predicted by the EponineTSS, EponineHomol, and CpG methods.

Given a functional predictive system, it is always interesting to investigate methods to increase the system's predictive accuracy. Two different approaches were taken with the hope of generating better EWS promoter models. The first was to use the extended version of EWS, with extra sampling rules to construct "scaffolds" of several PWMs. Potentially, these scaffolds should be able to capture much more information than simple motifs could: two or more PWMs with very narrow position distributions, close to one another, could effectively model a large transcription factor binding site, while slightly broader and more distant distributions might reflect sites for factors which bind cooperatively. Training on the same datasets as before, scaffold-EWS learns significantly more complex models than those seen previously: the total number of basis functions used in the model was typically slightly higher than for motif-EWS, and around half of these basis functions were scaffolds of two or more motifs. One example is shown in figure 4.6. Unfortunately, the additional information did not appear to improve the

predictive power of the models in the chromosome 22 test system.

The second approach takes the view that the training data might be the limiting factor in performance of these models. The set of mouse-human homologous sequences is likely to be a relatively complex data set, and while it clearly contains substantial amounts of promoter sequence, it is also likely to include fragments of coding genes which were missed by Ensembl predictions, RNA genes, UTRs, splicing regulators, and probably other types of functional sequence, as well as some regions which are included purely by chance conservation. In an attempt to make a cleaner data set, the EponineHomol_1 model was used to select a training set of 4000 sequences with scores over 0.95. This was used as before to train both EWS-motif and EWS-scaffold models, with initial words of length either 5 or 6, which were then tested on chromosome 22. Once again, the ROC curves were very similar to those seen previously. Despite the enrichment of the training data, it was not possible to increase the classification power of this method. This suggests that performance may be limited by a fundamental problem with classification based on fixed-sized windows of sequence. This is considered further in the discussion section.

4.4. An attempt to discover a second signal in mouse-human homologies

Following the success of discovering one signal in conserved noncoding sequences, I attempted to use the same approach to discover more, distinct, signals. To do this, a new training set was prepared, following the same principles as previously, but including only sequences which received low scores (<0.25) with the EponineHomol_1 model. A set of 2000 sequences was selected, henceforth referred to as the reverse-selected set. Models were trained from this set using the same procedure as before. To confirm that this reverse-selection protocol gives models which do detect novel sequences, the models were tested on the same two megabase region as previously. Figure 4.7 shows the scores for a typical reverse-selected model against those for EponineHomol_1. There is a slight negative correlation between the two axes, so

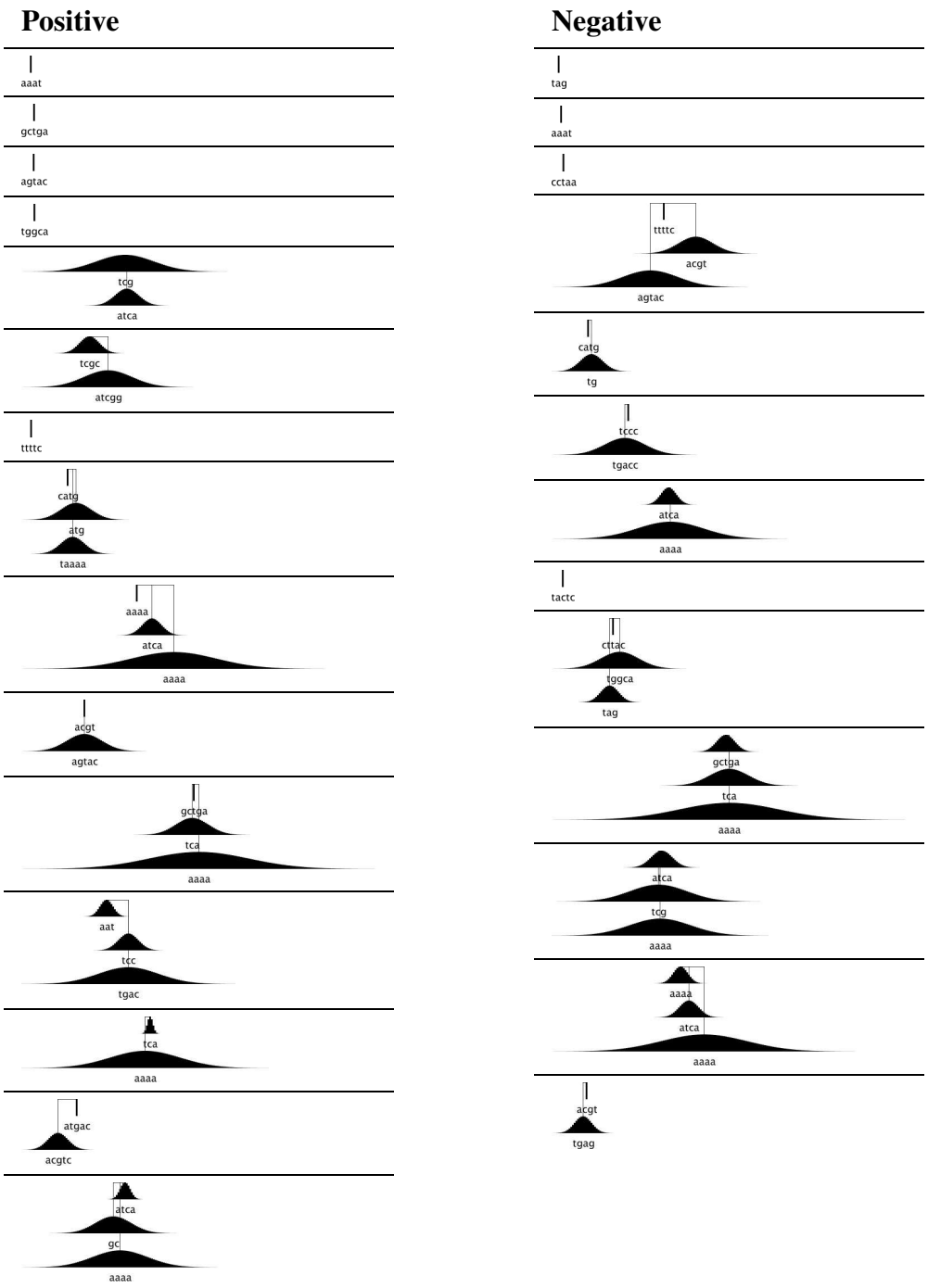

Figure 4.6. A set of basis functions learned by training the EWS-scaffold system on human-mouse homologous sequences. Each cell of this table shows an individual basis function, made up of between one and three sequence motifs.

clearly this model is substantially different from EponineHomol_1, and indeed any of the models which were trained without using the reverse-selection protocol. One of these models, EponineHomol_2, was selected for further characterization. A logo view of the motifs in this model can be seen in table 4.3.

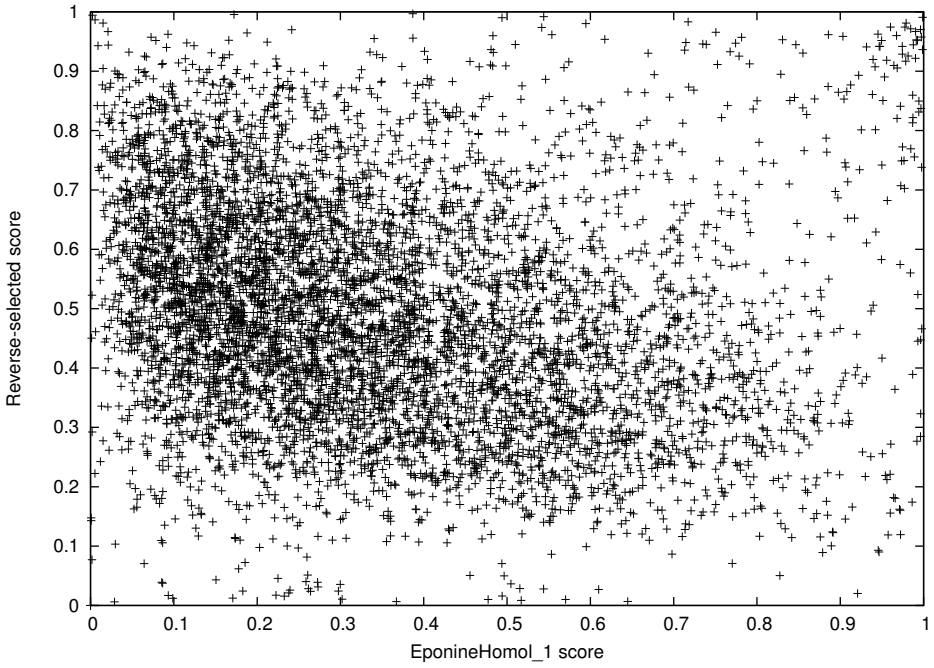

Figure 4.7. Scores for a model trained from the reverse-selected dataset vs. EponineHomol_1.

Predicting high scores for different localized sequence windows does not necessarily mean that the model makes very different predictions when considered on a genomic scale. In fact, once again this model appears to localize in promoter regions (see some examples in figure 4.8). The predictions shown in the third panel are somewhat anomalous: the majority of EponineHomol predictions were accompanied by an EponineTSS prediction. When evaluating the EponineHomol_2 model as a promoter predictor on the pseudochromosome, it gave an accuracy of 42% at a coverage of 50%, less effective than EponineHomol_1 but still useful, and comparable to EponineTSS_1.

Perhaps more surprisingly, the set of promoters detected by EponineHomol_2 is correlated with that found by EponineHomol_1. As figure 4.8 shows, predictions from the two models

Positive	**Negative**

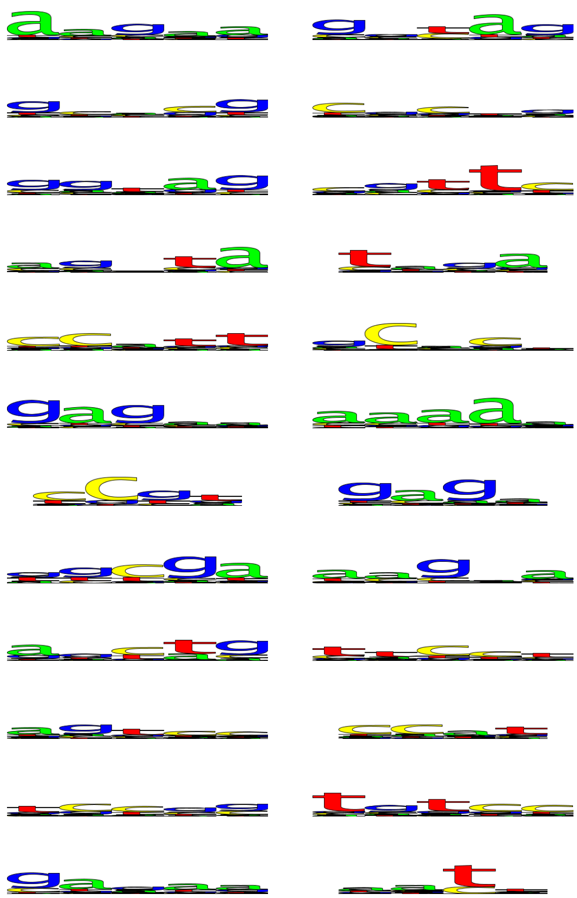

Table 4.3. Logo view of the motifs in the EponineHomol_2 model.

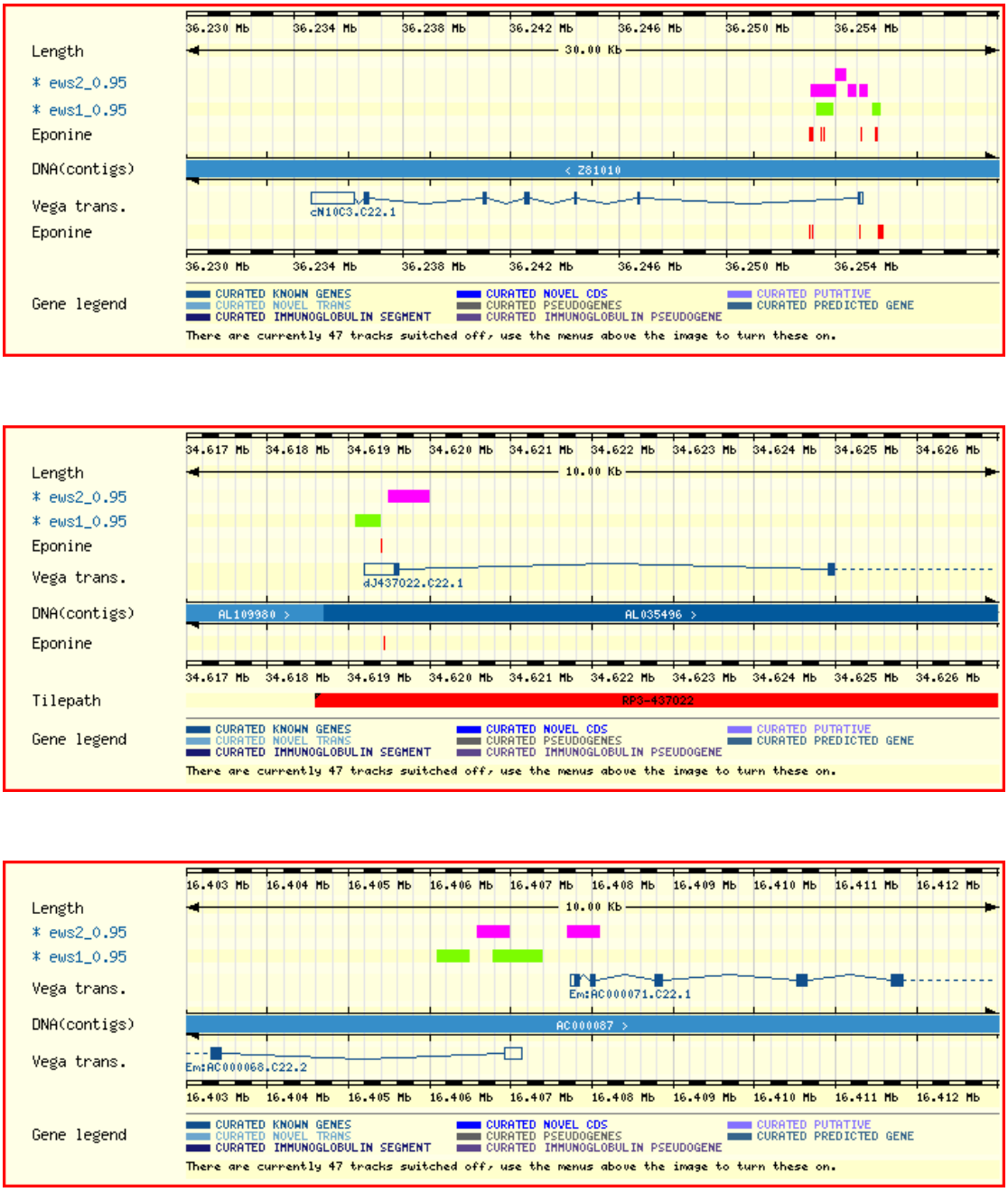

Figure 4.8. Contigview displays showing both EponineHomol_1 and EponineHomol_2 predictions.

are often very close together (although generally not precisely co-localized). The strong intersection between the sets of genes found by the two models (and also EponineTSS_2, and CpG islands) can be seen in figure 4.9. So while it is interesting to find that it is possible to learn a second promoter model, it appears not to be a great help in completing the task of predicting transcription start sites for every gene in the genome.

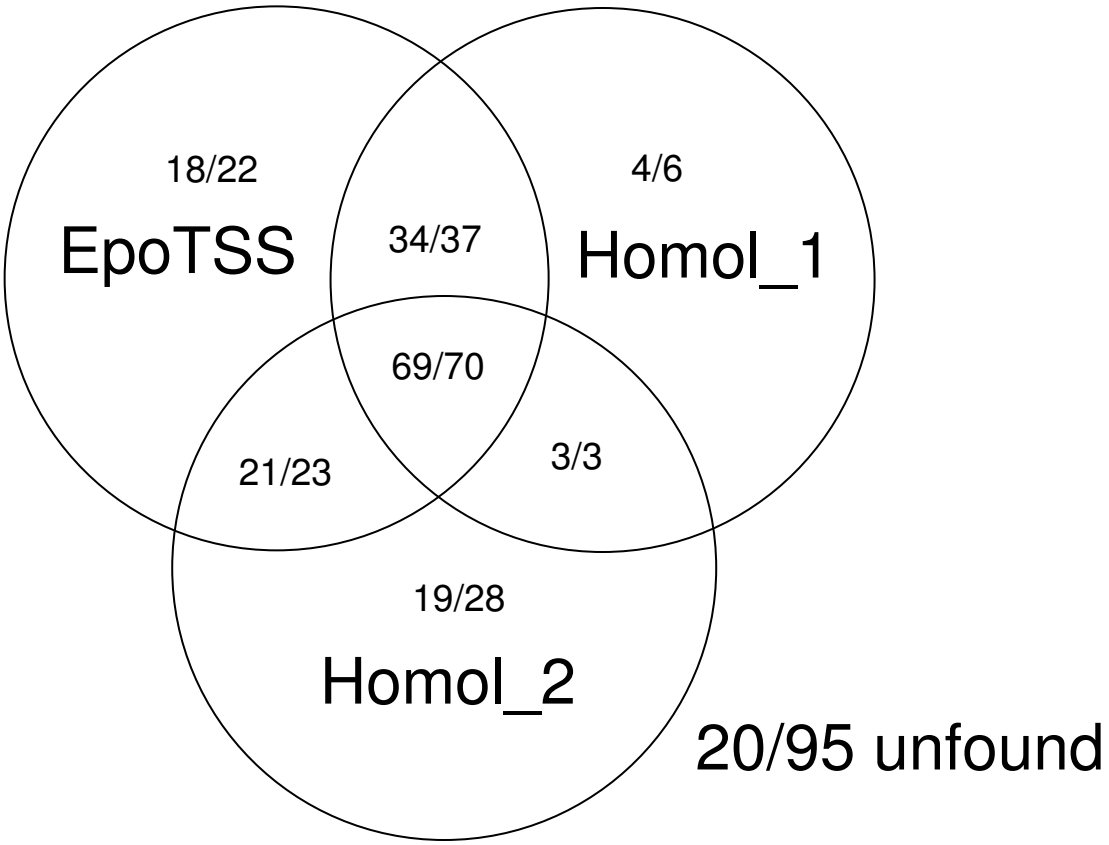

Figure 4.9. Intersection of "correct" promoter predictions from EponineTSS, EponineHomol_1, and EponineHomol_2. In each compartment, the first figure indicates the number of start sites which were also detected by a CpG island predictor, while the second figure gives the total number.

4.5. Comparative Genomics Discussion

This method began as an ignorance-driven approach to find interesting signals in pieces of sequence conserved between the human and mouse genomes. But visual inspection of results, followed by comparison with annotated gene starts on the pseudochromosome, makes it clear that the learned models are effective promoter predictors. Based on the criteria of accuracy

and coverage on the pseudochromosome, the models described in this chapter are slightly less effective than the EponineTSS_2 model from chapter 3. Moreover, they do not give predictions of the actual transcription start site, one of the defining features of EponineTSS. Nevertheless, the results in this chapter are impression, and potentially significant for the future of genomics. The EponineTSS model represents a "shoulders of giants" approach to promoter prediction: the model is, in effect, distilled information from the 339 EPD sequence entries which were used in the training process. Each of these represents a considerable amount of time, effort, and expense in laboratory work.

The models described in this chapter, however, are based on purely computational analysis of genome sequences. While the difficulty of sequencing a genome should, of course, not be underestimated, a number of genomes have now been completed, and techniques such as preparation of shotgun libraries and assembling the resulting data are now quite well developed. In just a few years, whole-genome shotgun sequencing in bacteria has become an almost routine operation, and it seems hard to imagine that this will not eventually happen for higher organisms, too. To date, most eukaryotic genome sequencing (at least in terms of public and academic projects) has focused on the so-called model organisms, species which have a well-established tradition of experimentation – often (but not always) including laboratory molecular biology work. But once genomes are sequenced in organisms which do not have this tradition and background, it becomes more important to be able to "bootstrap" biological knowledge on the basis of genome sequences. The results in this chapter show this kind of bootstrapping in action, learning information about an organism's promoters with rather limited *a priori* information. Such techniques will be particularly applicable if sequences are produced for several different species in a previously uninvestigated clade.

This experiment was also interesting in terms of the surprisingly close agreement of promoter-detection results on chromosome 22 between the EponineTSS_2 and EponineHomol_1 prediction systems. Since the training of the EponineHomol_1 model was entirely independent of the databases used to train EponineTSS, it is now possible to discount

the possibility that the set of promoters found by these models is biased by the contents of the training set. The remaining explanations are either a number of the chromosome 22 annotations being significantly truncated at the 5' end, or that many of them are really psuedogenes with no functional promoter, or that promoters can be split into two classes, only one of which is detected by these prediction methods. If the latter hypothesis is true, it seems likely that promoters in the second class have far less in common with one another than those in the first, hence the difficulty training models to detect them.

Given that a rather simple model gave good specificity as a promoter predictor – perhaps surprisingly good, given the rather *ad hoc* nature of the training, and especially the fact that a substantial fraction of the training data was probably non-promoter sequence – it is interesting that various attempts to train better models using longer motifs, sets of motifs on scaffolds, and enriched training sets did not give a further improvement. This suggests that the training may not, in fact, be the limiting factor defining this method's performance. An alternative view is that there are inherent limitations in the use of fixed-size windows. Some promoters might have all their signals focused into a very small region of sequence – perhaps significantly less than 300 bases – while others are likely to be more diffuse. Therefore, there is no ideal window size to catch all the signals of all promoters while not diluting the signals of the smaller promoters. The ideal solution to this issue is to use windows of variable width. This has the added benefit that, given a suitably comprehensive model, it could offer at least some information about the boundaries of promoter regions. One algorithm for scanning a sequence with variable sized windows was proposed in [Byng 2001] (although this used a rather simpler sequence model, just counting occurrences of specific nucleotide words). It would be be interesting in the future to combine a variable-window approach to scanning sequences with a sparse Bayesian training method.

The final part of this project was an attempt to peel away the first signal detected by this approach, and retrain the same type of model on a new dataset from which that first signal had been removed. This exercise was successful to the extent that it produced a new model whose

scores were not substantially correlated with the first signal. It is interesting that the second signal to be extracted from this data also appeared to come from near the transcription start site. It is also significant that this second promoter-area model includes motifs containing the CpG dinucleotide, and appears to be making predictions for the same class of CpG-associated promoters. Predictions from the second model often lie very close to, but not necessarily overlapping, the predictions from the first model, suggesting that they are modeling different parts of large promoter regions. This failure to detect large numbers of non-CpG promoters adds more weight to the hypothesis that non-CpG promoters do not form a single group with some set of strong motifs in common with one another. It may, in fact, be necessary to treat each of these individually, rather than identifying stereotypical signals which are found in all promoters.