Genet. Res., Camb. (2004), **83**, *pp.* 177–188. *With* 4 *figures.*
DOI: 10.1017/S0016672304006780 Printed in the United Kingdom

Use of regression methods to identify motifs that modulate germline transcription in *Drosophila melanogaster*

EMILY HONEYCUTT[1] AND GREG GIBSON[1,2]*
[1] *Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695-7566, USA*
[2] *Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614, USA*

(*Received 18 July 2003 and in revised form 29 December 2003*)

Summary

Identification of *cis*-regulatory motifs has been difficult due to the short and variable length of the sequences that bind transcription factors. Using both sequence and microarray expression data, we present a method for identifying *cis*-regulatory motifs that uses regression trees to refine results from simple linear regression of expression levels on motif counts. Analysis of expression patterns from two separate datasets for genes showing significant differences in expression between the sexes in *Drosophila melanogaster* resulted in a model that identified known binding sites upstream of genes that are differentially expressed in the germline. We obtained a strong result for motif TCGATA, part of the larger, characterized binding site of *dGATAb* protein. We also identified an uncharacterized motif that is positively associated with sex-biased expression and was assembled from smaller motifs grouped by our model. A regression tree model provides a grouping of independent variables into multiple linear models, an advantage over a single multivariate model. In our case, this grouping of motifs suggests binding sites for cooperating factors in sex-specific expression, as well as a way of combining smaller motifs into larger binding sites.

1. Introduction

The search for DNA regulatory motifs has been the focus of much recent research, with various methods being employed in motif discovery. Detection of transcriptional regulatory motifs in the upstream region of genes has presented a real challenge because transcription factor binding regions tend to be short, discontinuous, and quite variable. *Saccharomyces cerevisiae* has frequently been the organism of choice for development of methods that identify transcription factor binding sites since many binding motifs have already been experimentally characterized in this organism. Nevertheless, most methods of motif detection have found limited success, often resulting in a high rate of false positives (Werner, 2002). In higher eukaryotes, the structure of regulatory motifs is more complex and less well defined, making the development of new methods and verification of results even more difficult.

Before access to the sequence of multiple whole genomes, many motif-detection methods involved statistical approaches to the creation of weight matrices. A weight matrix is derived from a number of short sequences known to be bound by a given transcription factor, and then the matrix is used to search a sequence or a set of sequences for a match to that motif. Examples include MatInd and MatInspector (Quandt *et al.*, 1995) and FastM (Klingenhoff *et al.*, 1999). Searches that use weight matrices have a very high rate of false positives, but results have improved when they are used in combination with another method such as phylogenetic comparison (Guha Thakurta *et al.*, 2002).

Alternative methods for motif detection involve direct comparison of regulatory regions, either between genes thought to be co-regulated or between orthologous genes from closely related species. The Gibbs sampling method, which utilizes a modified Expectation Maximization (EM) algorithm (Lawrence *et al.*,

* Corresponding author. BRC, 1500 Partners II, 840 Main Campus Drive, NC State University, Raleigh, NC 27695-7566, USA. Tel: +1 (919) 5132512. Fax: +1 (919) 5153355. e-mail: ggibson@unity.ncsu.edu

1993), has been used in the AlignACE program to return over-represented motifs in co-regulated gene clusters and has found some success (Hughes *et al*., 2000; Manson-McGuire *et al*., 2000). Advanced application of Gibbs sampling methods in this context continues to hold promise, particularly in microorganisms (Liu *et al*., 2001). With the increasing availability of whole genome sequences of closely related species, the phylogenetic comparison of regulatory regions has increased. Comparisons between human and mouse regulatory sequence showed that phylogenetic footprinting can reduce the sequence space to be searched for transcription factor binding sites (Wasserman *et al*., 2000). Rajewsky *et al*. (2002) recovered approximately 75% of the regulatory sites compiled for *E. coli* using interspecies comparisons. Issues still remain as to how best to choose the species for comparison and how many are required to produce meaningful results. A recent study using proteobacteria takes a formal look at these issues (McCue *et al*., 2002).

A somewhat different strategy for detection of transcription factor binding motifs searches for clusters of motifs in upstream sequences (Berman *et al*., 2002; Halfon *et al*., 2002; Markstein *et al*., 2002; Rebeiz *et al*., 2002). These clustering methods require prior knowledge of characterized sites and are targeted more towards finding genes regulated by factors binding to the clusters rather than identifying the clusters themselves. Another combinatorial approach for finding synergistic motifs by Pilpel *et al*. (2001) also requires knowledge of known regulatory motifs. An underlying assumption in several of the above analyses is that binding motifs are redundant in the promoter region, as in the *Drosophila* yolk protein genes (Piano *et al*., 1999) and the *Drosophila* eve stripe 2 gene (Berman *et al*., 2002).

Capitalizing on this redundancy property, a recent study by Bussemaker *et al*. (2001) fitted a linear model of the logarithm of the expression ratio under two different experimental conditions to the counts of oligomers upstream of a set of genes. By first determining statistically significant motifs with a single-motif model of the data, a model describing the additive effects of multiple motifs can then be created. We incorporate this method by identifying the statistically significant motifs through the single-motif model, but instead of building a single additive model for a given experiment, we use the significant motifs to build regression trees. Our regression trees allow for multiple linear models to describe the data based on the prevalence of certain motifs and have the potential to uncover hierarchical or non-additive relationships between motifs.

Regression trees were originally used to generate predictive models of regression estimates. They were developed to deal with continuous-class learning problems (Quinlan, 1992; Wang & Witten, 1997), and combine a classical decision tree with linear regression estimations at the leaves of the tree. The prediction accuracy of regression trees is competitive with linear regression methods (Breiman *et al*., 1984), but the real advantage of the regression tree method lies in the model representation. The decision nodes and their position in the tree indicate which nodes together significantly affect the predicted values. We show that they can be used to identify prospective regulatory motifs bound by transcription factors, as well as combinations of motifs that aggregate to form larger motifs.

Other biological studies have also capitalized on the classificatory property of regression trees. For example, a recent investigation into the nesting habitats of smallmouth bass used regression trees to give a hierarchical view of habitat conditions that affect the smallmouth bass's choice of nesting site (Rejwan *et al*., 1999). Similarly, they have been used to identify the most predictive variables for patients who undergo angiography (Pilote *et al*., 1996). In this study, regression trees identified age as the most important variable. However, in younger patients availability of the angiography procedure was the next most predictive factor, while age was still the second most predictive factor in older patients. This illustrates the ability of regression trees to separate, or group together, cooperating factors under given circumstances.

In our model, we are using counts of binding motifs as the decision points in the tree. The decision nodes in the tree look at the counts of motifs of length k (k-mers) taken from the upstream region of a given gene. The change in estimated regression values between the leaf nodes indicates whether a combination of motifs is associated with the regulation of genes. As with the aforementioned studies, we are not using the regression tree model in its classical sense as a predictor of response, but instead to identify the predictive variables, namely regulatory motifs.

In this study, we searched for transcription factor binding motifs of genes that show sex-biased expression. Our previous study on sex, genotype and age (Jin *et al*., 2001) (subsequently referred to as the aging dataset) showed evidence for between one-third and two-thirds of the *Drosophila* transcriptome having sex-biased expression. Comparisons with *tudor* mutant animals that lack ovaries and testes have since demonstrated that most of the differences in gene expression between reproductively mature adult male and female flies is due to germline expression (Arbeitman *et al*., 2002; Parisi *et al*., 2003). To obtain a larger number of these differentially expressed genes for our analysis, we supplemented the aging dataset (Jin *et al*., 2001) with data from another experiment that tested the effects of nicotine on gene expression in flies of both sexes (G. Passador-Gurgel and G.G., in

preparation: this dataset is subsequently referred to as the nicotine dataset). Although two different clone sets were used to generate the data, a high concordance in the predicted motifs was observed, and this independent replication confirms that regression tree methods may be a valuable new approach to characterization of regulatory motifs.

2. Materials and methods

(i) *Gene selection from microarray experiments*

The genes used for analysis of sex-biased expression are from two datasets: the aging dataset (Jin *et al*., 2001) and the nicotine dataset (G. Passador-Gurgel and G.G., in preparation). The aging array experiment used a split-plot experimental design and tested for sex as a fixed effect using a mixed-models approach (Wolfinger *et al*., 2001). Array set-up and subsequent analysis for the nicotine experiment was done similarly, with 48 two-sample arrays involving three wild-type genotypes, two sexes and treatment (control versus drug) as fixed effects. The set of genes for the nicotine experiment was 4856 genes from the Drosophila Gene Collection (DGC), which were independently identified and amplified from those of the White collection used in the aging experiment. From each experiment, genes with a P value of $<0{\cdot}0001$ resulting from the test for sex effects were chosen for use in this analysis. The lists of genes from both datasets and their associated expression difference are available at http://statgen.ncsu.edu/ggibson/SupplInfo/SexSpecificList.txt

(ii) *DNA sequence motifs*

All possible motifs of length 6 were generated. Initially, we extracted counts of all possible 7-mers of the 250 differentially expressed genes from the aging dataset (Jin *et al*., 2001). Since five of the eight most significant motifs from the linear regression contained the sequence TCGATA, all subsequent analyses were conducted on 6-mer motifs. Motifs were combined with their reverse complement and the motif having the higher lexicographic order was chosen to represent the pair. No allowance for variability in the motif sequence was made. For each gene selected, the 1000 base-pair (bp) sequence upstream of the translation start site (ATG) was extracted from the Version 2 annotation of the *Drosophila* genome sequence at NCBI (March 2002, http://www.ncbi.nlm.nih.gov). This sequence includes variable lengths of 5′ untranscribed and untranslated leader sequences, which are as yet typically uncharacterized in *Drosophila*. Although enhancers in the fly genome can be several kilobases away from the translation start site, the 1000 bp upstream sequence was chosen for two reasons. First, testis-specific promoters in *Drosophila* are usually close to the start site (Arnosti, 2003). Secondly, as more sequence is added to the analysis, the signal-to-noise ratio of regulatory to non-functional motifs probably drops, and with a large number of genes we surmised that we would be most likely to find common motifs in the upstream 1 kb region. This approach is not intended to identify all the enhancer elements that regulate sex-specific gene expression in *Drosophila*, but rather to focus on those located proximal to the transcription start site.

For each gene, all motifs were counted in the upstream 1 kb sequence (allowing overlap, namely 995 motif counts per gene). All work to extract sequence, generate motifs and count motifs was done via Perl scripts.

(iii) *Single-motif linear regression*

The first stage of analysis uses a simple linear regression model to fit single-motif counts and expression data. The model is defined as:

$$Y=\beta_0+\beta_1 X$$

where Y is the base 2 logarithm of the expression difference between females and males. A positive Y indicates greater expression in females; a negative Y indicates greater expression in males. X is the count of a given motif. All genes chosen as significantly differentially expressed between the sexes (in either direction) were fitted to the model. β_1 is the relative increase or decrease in expression difference caused by each additional copy of the motif in the upstream region of the gene, and β_0 is the grand mean expression difference.

Both the nicotine and the aging datasets were run through simple linear regression. To account for the large number of motifs (2080), application of the Bonferroni correction set the experimentwise significance cutoff from regression of expression level on motif count for $\alpha=0{\cdot}05$ at $P=2{\cdot}4\times10^{-5}$. Permutation tests provided independent verification of the appropriateness of this cutoff, but for some analyses we included simply the top 20 motifs as these included a few motifs that were close to the cutoff in both datasets.

(iv) *Regression and decision trees*

Single-motif linear regression was used primarily as a data reduction technique. Motifs with a P value below the Bonferroni-corrected values were considered most likely to affect sex-biased expression and were therefore used in training and validation of the regression and decision tree models.

Regression and decision tree models were built and trained with publicly available Weka software (Witten & Frank, 1999) available at http://www.cs.waikato.ac.nz/ml/weka/. Data from the nicotine experiment

Table 1. *The most significant sex-specific motifs from single-motif regression for both the nicotine and aging datasets*

	Nicotine dataset			Aging dataset		
Rank	Motif	*P* value	sign[a]	Motif	*P* value	sign[a]
1	**TCGATA**	1.4e-19	+	**TCGATA**	1.2e-12	+
2	**CGATAG**	2.5e-11	+	**ATCGAT**	0·0000013	+
3	**ATCGAT**	3.7e-10	+	**ATATCG**	0·0000024	+
4	GGTCAC	0.000000050	+	ACGACG	0·000065	+
5	**ATATCG**	0·00000019	+	AGTCGC	0·000092	+
6	**ACACTG**	0·00000024	+	CGCAAC	0·00014	+
7	CACGTG	0·00000033	+	**CGATAG**	0·00016	+
8	TAAAAA	0·0000012	+	CCAAAG	0·00021	−
9	GGCGCA	0·0000022	+	GCAACG	0·00021	+
10	CCGTTA	0·0000030	+	**ACACTG**	0·00038	+
11	GTCACA	0·0000032	+	CACGCA	0·00058	+
12	AAGAAG	0·0000032	+	**GCACGC**	0·00063	+
13	CGCACG	0·0000057	+	CCTTTC	0·00066	−
14	AGACTC	0·0000073	−	**AGTGTG**	0·00075	+
15	CGGTAA	0·0000161	+	AGGGCC	0·00099	−
16	TTAAAA	0·000016	+	GTGTGA	0·0013	+
17	AAAATA	0·000019	+	ATCGAC	0·0015	+
18	**AGTGTG**	0·000022	+	ATTCGC	0·0015	+
19	GCGCAC	0·000022	+	AGAAGA	0·0016	+
20	**GCACGC**	0·000028	+	ACTACG	0·0020	+

Motifs in common between the two sets are indicated in bold.
[a] Positive coefficients indicate that the motif is associated with increased transcription in females. Negative coefficients indicate that the motif is associated with increased transcription in males.

were used to train the models and data from the aging experiment were used for model validation. Specifically, the regression trees were built with the M5 software using a $-O\,r$ option. The decision trees were built with the J48 software using the $-R$ option to reduce error pruning and the $-M$ option to vary the minimum number of instances per leaf.

The models were built from motifs as follows:

Model 1: Motifs that were above Bonferroni-corrected significance cutoff from single-motif regression and were seen >4% of the time within 20 bp of TCGATA/TATCGA (8 total).
Model 2: Motifs seen >5% of time within 20 bp of TCGATA/TATCGA (25 total).
Model 3: The most significant motifs from single-motif regression at or below Bonferroni-corrected cutoff (20 total).
Model 4: Combination of 20 most significant motifs from single-motif regression and 20 motifs most often seen within 20 bp of TCGATA/TATCGA.

3. Results

(i) *Identification of female-specific regulatory motifs*

The first stage of the analysis searched for motifs that may contribute to male- or female-specific gene expression in adult flies using linear regression of expression difference against motif count in the promoters of differentially expressed genes (Bussemaker *et al*., 2001). Table 1 shows the top 20 motifs after linear regression with the two different datasets. The significance threshold for regression of motif count on expression difference after Bonferroni correction is approximately $2{\cdot}4 \times 10^{-5}$. Three motifs exceed this threshold in the aging dataset, and 19 in the larger nicotine dataset. Several results stand out. Most noticeably, the two experiments converge on a similar set of motifs, with the three most significant motifs found in the aging dataset also being found within the five most significant motifs resulting from analysis of the nicotine dataset. Three other motifs are also common between each dataset's list of 20 most significant motifs. Additionally, the motif TCGATA/TATCGA is at a much higher significance level than any other motif in both datasets, with a *P* value of 10^{-19}. Lastly, almost all the motifs are associated with female-biased gene expression, and no case of a male-specific motif was replicated in both datasets. Representative linear regression profiles shown in Fig. 1 also highlight the point that none of the motifs is either necessary or sufficient for sex-specific gene expression: some genes with multiple copies of TCGATA are actually male-biased, and many female-specific genes lack the motif within 1 kb of the translation start site.

https://doi.org/10.1017/S0016672304006780 Published online by Cambridge University Press

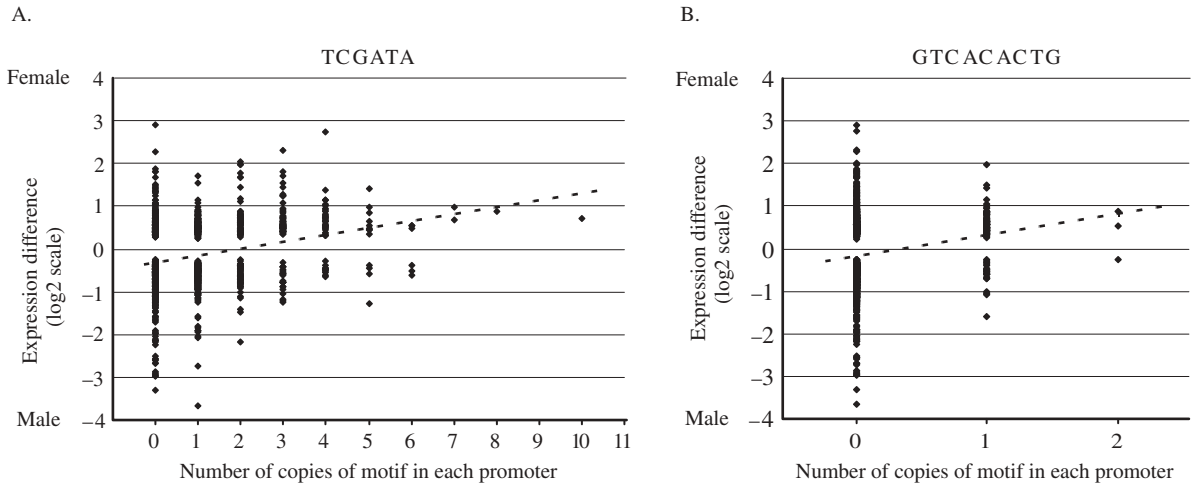

Fig. 1. Linear regression of expression difference on motif counts. Each diamond represents the normalized difference between gene expression in adult females and males on a log 2 scale, given the number of copies of the indicated motif (A: TCGATA; B: GTCACACTG) in the nicotine dataset. Only genes that are significantly different between the sexes are included. On this scale, 1 represents a two-fold difference, 2 a four-fold difference, and so on. Dashed lines shows linear regression fit. Female-biased genes are towards the top.

The most significant motif, TCGATA, is part of a known binding site for the *dGATAb* (SERPENT) protein, which enhances transcription of yolk proteins in *Drosophila* females (Lossky & Wensink, 1995). The entire binding site has been characterized as GCTATCGATAGC, which highlights the fact that TCGATA and its reverse complement TATCGA have a 4 bp overlap. The combined 8-mer is palindromic, a characteristic that is known to increase the affinity of binding sites for transcription factors but usually associated with head-to-tail dimerization of individual binding sites (Drouin *et al*., 1992). Observations of all TCGATA/TATCGA pairs in the upstream regions of the genes being analysed show that this 4 bp overlap occurs in 29% of these incidences. A chi-square contrast of the incidence of the palindrome in female-biased versus male-biased and non-sex-biased genes provides compelling evidence ($P<0{\cdot}001$) that this palindrome is strongly associated with sex-biased expression, and, specifically, that it is female-specific. A concern is that the prevalence of this overlap artificially inflates the motif counts for TCGATA and enhances its significance in the single motif regression results. However, the overlap of the motif with itself into an 8 bp palindrome creates a more likely binding site, so counting the 6-mer twice simply aids in this discovery.

The high significance of TCGATA could also be a result of its pairing with itself as a composite binding site for a transcription factor pair or for multiple fingers of a zinc-finger binding protein such as SERPENT. Since over half of the DNA-binding proteins in *Drosophila* are zinc-finger proteins (Adams *et al*., 2000), we assumed that close proximity of binding motifs would often allow for the possible binding of multiple-fingers, which prompted us to count all the non-overlapping motifs within 20 bp on either side of TCGATA/TATCGA. TCGATA was found within 20 bp of itself at a greater frequency than any other motif (Table 2), supporting the idea that it often forms a composite binding site.

(ii) *Use of regression trees to identify interacting motifs*

The most significant motifs from the single-motif regression can be used to create an additive model that accounts for the combinatorial nature of cooperative and competitive binding of transcription factors. However, in a single additive model, each included motif is assumed to affect every gene's predicted expression level. This is not always the case. Different combinations of motifs may have dramatically different effects on transcription. Consider a combination of three binding motifs that cause increased binding affinity, and thus an increase in expression levels. If one of those binding motifs is replaced by a different motif, transcriptional repression could result. Regression trees have the potential to account for these types of occurrences. Nodes at the top of the tree indicate motifs that most correlate with expression. As a path is traversed through the tree, a combination of motifs affecting expression is discerned. The values at the leaves of the tree show how the path increases or decreases the expression difference. In our case, an increase in expression difference between paths indicates that transcription tends to be enhanced in females. We are using the regression tree as a model for finding important motifs identified by nodes in the tree. A more conventional use of

https://doi.org/10.1017/S0016672304006780 Published online by Cambridge University Press

Table 2. *Motifs within 20 base-pairs of TATCGA/TCGATA*[a]

Rank	Motif	Number	Percentage
1	TATCGA	202	18·05
2	ATCGAT	155	13·85
3	CGATAG	126	11·26
4	CGATAA	120	10·72
5	AATCGA	110	9·83
6	AAAAAT	94	8·40
7	TAAAAA	88	7·86
8	ATATCG	85	7·60
9	ATCGAA	83	7·42
10	AAAATA	82	7·33
11	AAAATT	80	7·15
12	ATTTTA	74	6·61
13	AAATAT	70	6·26
14	ATAAAA	69	6·17
15	AAAAAA	68	6·08
16	ATAAAT	68	6·08
17	AAAACA	68	6·08
18	CCGATA	68	6·08
19	GATAAC	66	5·90
20	CATCGA	65	5·81

[a] Motifs in this range may form composite binding sites with TCGATA/TATCGA, which was seen a total of 1119 times.

regression trees is as a predictive tool for estimating the values at the leaves of the tree. We instead use the predicted values simply as a test for the direction and amount of change in expression.

As inputs into the regression tree software we used the single motifs identified by simple regression, supplemented by those that occur at elevated frequency within 20 bp of TCGATA. Various combinations of these motifs and corresponding data from the nicotine dataset were used in the creation of four multiple regression model trees using Weka software (Witten & Frank, 1999; see Section 2 for details). The resulting trees were compared via their correlation coefficients, which measure the statistical correlation between the actual and predicted expression level values. These values are shown in Table 3. Models 3 and 4 show the highest correlation coefficients and were rerun with the aging dataset used as a test dataset. The test dataset correlation coefficients were 0·49 for Model 3 and 0·48 for Model 4. These values are higher than those obtained for the training dataset, and thus show strong support for the model.

Models 3 and 4 resulted in very similar regression trees and are shown in Fig. 2. Model 4 had one additional node (GATAAC), a motif found within 20 bp of TCGATA but not found to be significant by simple linear regression. We decided not to use Model 4 as our final regression tree model for two reasons: (i) the motif GATAAC was added because of its proximity to TCGATA in upstream sequences but the node containing the motif was not closely connected to TCGATA in the tree and (ii) GATAAC fell out of the model when we removed AGTGTG from the input dataset because of its 5 bp overlap with ACACTG. Since AGTGTG fits in the overlap with other genes in its path in the tree, we decided to keep that motif in the model and use the resulting tree from the set of significant motifs from single-motif regression.

Traversal of the regression tree should identify binding site combinations that may enhance or repress expression significantly in one sex or the other. On the left side of the Model 3 regression tree, we see that with 0 or 1 copy of TCGATA and 0 copies of GGTCAC, we have an estimated expression difference of $-0{\cdot}346$, indicating that genes lacking these motifs in their upstream regions are more likely differentially expressed in males. We then use $-0{\cdot}346$ as a comparison point. If we have 0 or 1 TCGATA, 1 GGTCAC and 0 copies of AGTGTG, the estimated expression difference is $-0{\cdot}320$ which is not much different from $-0{\cdot}346$. This indicates that the addition of a GGTCAC by itself does not change expression. However, if we find the combination of 0 or 1 TCGATA, 1 or more GGTCACs, 1 or more AGTGTGs and 0 ACACTGs, the expression difference changes to $-0{\cdot}176$, which is a considerable change. This motif combination may cause the gene to be less differentially expressed between the sexes. With the same combination of TCGATA, GGTCAC, AGTGTG, but addition of 1 or more copies of ACACTG, the expression difference becomes positive. This can mean either that ACACTG activates female-specific transcription, or that this motif could be a repressor-binding site for male-specific transcription. Since our analysis has not included genes that are not differentially expressed between the sexes, a change of this magnitude in comparison with our other expression differences most likely indicates up-regulation in females.

From the left traversal of the tree, a motif combination of interest is GGTCAC, AGTGTG and ACACTG. This motif trio combines to form the 10-mer GGTCACACTG that contains the palindromic sub-motif GTCACACTG. Of the 238 GGTCAC–ACACTG pairs found in the upstream regions of sex-biased genes, 84 (or 35%) were found in this overlap. Another chi-square test of motif presence associated with female-biased, male-biased or non-sex-biased genes resulted in strong evidence (P value $<0{\cdot}001$) that this larger motif is associated with female-specific expression. Detection of a larger, overlapping binding site such as this is a direct observation from regression trees. A single multiple-regression model does not provide any type of grouping of motifs that may work together. Regression trees separate independent variables that, together, change the dependent variable and create multiple groupings to explain the data.

https://doi.org/10.1017/S0016672304006780 Published online by Cambridge University Press

Table 3. *Regression model tree results*

Model	Motifs in model	No. of leaf nodes in resulting tree[a]	Training set correlation coefficient
1	Most significant from SLR and seen >4% of time within 20 bp	3	0·3107
2	Seen >5% of time within 20 bp	4	0·2904
3	20 most significant from SLR	6	0·3469
4	20 most significant from SLR plus 20 seen most within 20 bp	7	0·3613

[a] The number of leaf nodes in the resulting tree gives an indication of tree complexity.

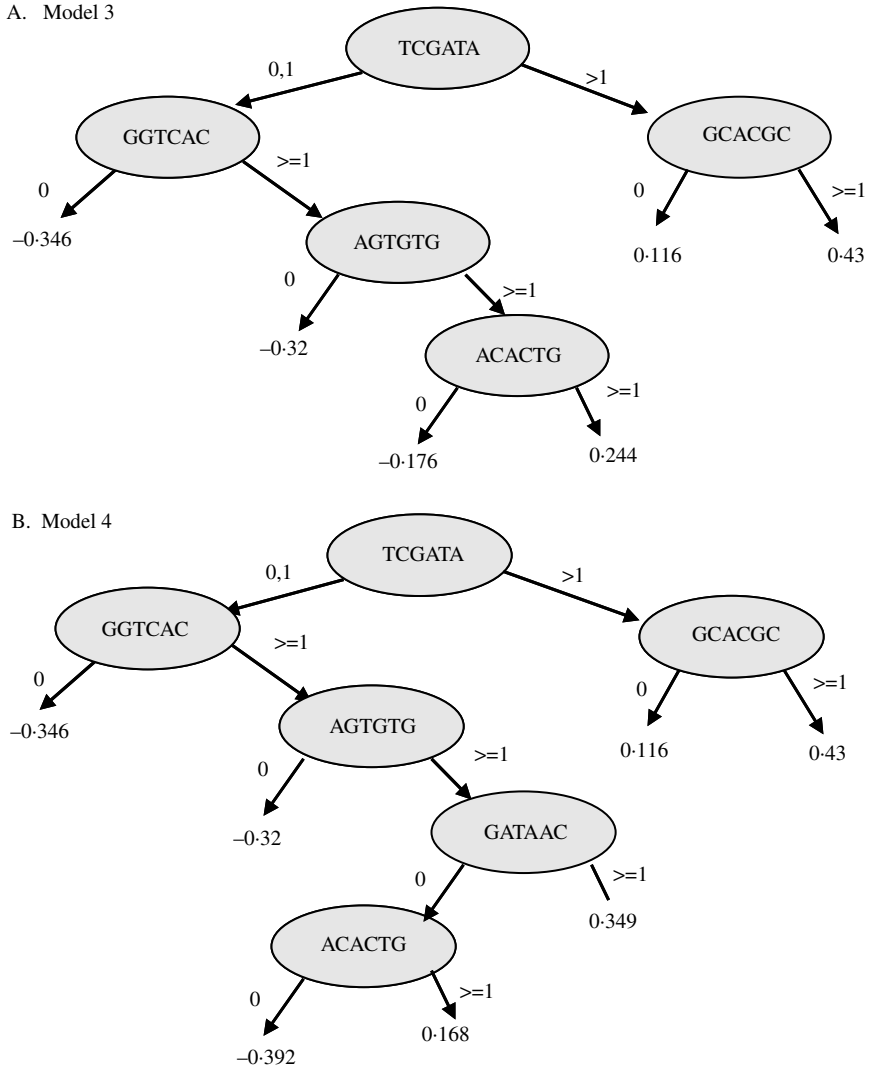

Fig. 2. Regression trees highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster*. See text for details of Models 3 and 4.

This is a distinct advantage over multiple-regression methods.

As further verification of our method, we obtained data from a microarray experiment specifically targeting *Drosophila* ovaries and testes, since these reproductive tissues are known to contribute to much of the overall expression difference between adult male and female flies (Parisi *et al*., 2003), and ran it through our analysis. We used genes that showed a four-fold or higher difference in expression between

https://doi.org/10.1017/S0016672304006780 Published online by Cambridge University Press

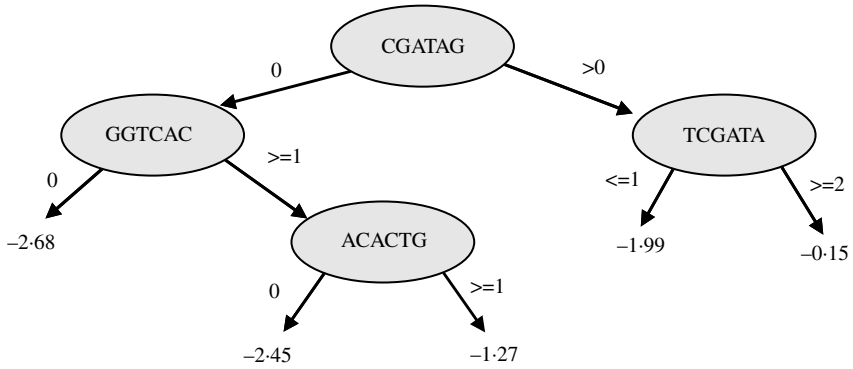

Fig. 3. Regression tree highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster* from the ovaries/testes cDNA microarray dataset. See text for details.

the sexes in order to reduce the dataset to approximately 1600 genes. The most significant motif resulting from the single-motif linear regression was TCGATA/TATCGA, and the six most significant motifs from the ovaries/testes dataset were found in the seven most significant motifs resulting from regression on the nicotine dataset. Again, using the motifs with significance below the Bonferroni-corrected cut-off, we built a regression tree. The resulting tree (Fig. 3) was strikingly similar in structure to the regression tree built from the nicotine dataset. The top node in the ovaries/testes regression tree is the motif CGATAG, which has a 5 bp overlap with TCGATA, and TCGATA is the next node in the tree on the female-biased side. This further supports our theory of overlapping TCGATA motifs enhancing female expression. Additionally, expression becomes more female from left to right among the leaves. This tree further validates our regression tree model obtained from the nicotine dataset. The differences relative to the adult fly trees could either be due to sampling variance, or reflect the additional contribution of somatic tissues to sex-specific gene expression in whole flies.

(iii) *Use of decision trees to predict sex-specific gene expression*

With the identification of motifs affecting sex-specific expression by the regression tree, we wanted to determine whether we could use these same motifs to classify a gene as being differentially expressed in either sex from the motifs found in its upstream region. To do this, we created a decision tree, which, based on motif counts, classified a gene as significantly expressed more in males, females or neither. The structure of a decision tree is very similar to that of the regression tree except that the classification of 'male', 'female' or 'neither' is found at the leaves of the tree instead of a predicted expression difference. Again, various combinations of the significant motifs from the single-motif regression model were used as input. Data from all differentially expressed genes and a subset of genes not differentially expressed in males or females from the nicotine cDNA microarray experiment were used to construct the model tree, again using Weka software (Witten & Frank, 1999). Since the motifs used as input into the decision tree model were determined from analysis of differentially expressed genes between males and females, the expectation for the decision tree correctly classifying the differentially expressed genes from the non-differentially expressed genes was low.

Inputting only the motifs found at the regression tree nodes into the decision tree resulted in a model much more complicated than expected (49 nodes in the tree) but with a correct classification percentage of 47%. After realizing that most motifs occur closer to the promoter, we decided to narrow the upstream region of each gene to 700 bp and construct a tree using motif counts from that smaller region. The resulting tree was similar to our regression tree and highlighted certain motif pairs. It is shown in Fig. 4. This tree also had a correct classification percentage of 47% for our training set. On the entire nicotine array gene set, 67% of the observed male-biased genes and 54% of the observed female-biased genes were correctly predicted. Classification of genes not showing sex bias was low, as expected. To test our decision tree results, we created 1000 decision trees with 20 random motifs selected as input. Our model, with a 47% overall correct classification, ranked within the top 1% of all random trees created.

4. Discussion

(i) *Regression trees and sex-specific motifs in* Drosophila

Regression provides a quantitative method of combining sequence data and expression data. We describe here a two-step method for creating a multifactorial model which links the prevalence of binding motifs to changes in expression. Besides eliminating the need

https://doi.org/10.1017/S0016672304006780 Published online by Cambridge University Press

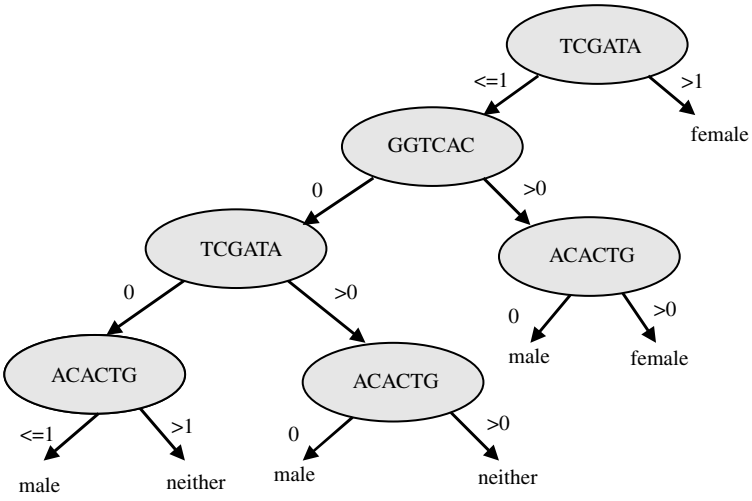

Fig. 4. Decision tree highlighting combinations of motifs that predict sex-biased gene expression in *D. melanogaster*. See text for details.

for clustering of expression data, this technique implies that the presence of multiple motifs in an upstream region is more likely to affect the level of transcription. This concept is starting to be explored in motif-clustering methods (Berman *et al*., 2002; Halfon *et al*., 2002; Markstein *et al*., 2002; Rebeiz *et al*., 2002). However, these motif-clustering methods require prior knowledge of the sequence of the binding sites which are believed to affect expression, and our approach does not. Furthermore, our method provides a straightforward procedure for focusing further analysis on a subset of the numerous significant motifs that may arise using simple linear regression.

Few binding sites for sex-specific expression have been identified in *Drosophila*. Almost all the motifs identified by our single-motif regression were associated with female-biased gene expression. Therefore, the motifs selected by the regression tree model were mostly female-specific. Verification of the function of the three major motifs that are highlighted in the regression trees was achieved by scanning TRANSFAC and the literature, which revealed that each of these motifs has previously been shown to form part of known binding sites for transcription factors during oogenesis. Most interesting is the TCGATA/TATCGA motif that forms the core of the SERPENT binding site, GCTATCGATAGC, in the promoters of the *yp1* and *yp2* genes (Lossky & Wensink, 1995). Similarly, GGTCAC/GTGACC is part of the extended TAGTGTATATAGGTCACGT binding site for chorion factor II in the chorion protein *s15* promoter during oogenesis (Shea *et al*., 1990), and ACACTG/CAGTGT is the core of the CCTACACTGTAAG binding site for DEP3 in the ovarian promoter of *Alcohol dehydrogenase* (Bayer *et al*., 1992).

Very few male-specific motifs were found by any of our single-motif models, and between the datasets, the male-specific motifs that tested with higher significance were different. Although it was surprising that our regression tree did not find any male-specific or antagonistic binding site combinations, it was encouraging that known female-specific motifs were selected and used as decision nodes in the regression tree. Since we only looked at mature adults, our motifs are actually associated with germline (ovary- and testis-specific) expression. Notwithstanding the empirical evidence discussed above that GGTCAC and ACACTG are part of female-specific enhancers, a possibility suggested by the regression trees is that the presence of these motifs is sufficient to contribute to repression of male-specific transcription. It is known, for example, that repressor binding sites in mRNA actively inhibit translation in the male germline (Crowley & Hazelrigg, 1995; Blumer *et al*., 2002). Extra power can be obtained by fitting regressions over a developmental time course, and this had led to the detection of male-specific elements as well (K. P. White & H. J. Bussemaker, personal communication).

A multiple regression model including all the significant motifs was also built on the same sets of motifs as the regression trees and resulted in a model with a correlation coefficient of 0·44. Even though this was similar to the correlation coefficient for our regression tree, the associated model does not uncover all the salient features revealed by our regression tree approach. The TCGATA motif stands out the most from all our analyses as it was always at the root of both the regression and decision trees, indicating that it is the most highly correlated motif in sex-biased expression. Additionally, the TCGATA motif was found overlapping with itself in an 8 bp palindrome 29% of the time, and this overlapping motif tested

positively for association with sex-biased expression. Because TCGATA is found in these situations so often, the motif seems to be involved somehow in regulation and deserves further investigation. Overlap of GGTCAC, AGTGTG and ACACTG into a larger motif is also highly suggested by our results.

(ii) *Advantages and drawbacks of regression trees*

There are at least two situations in which regression trees are expected to outperform direct multiple regression. As documented above, one is where the short motifs overlap and combine to perform a single binding site. Multiple linear regression does not suggest any grouping of motifs, but merely gives partial regression coefficients indicating the contribution of the motif to the change in expression. In fact, overlapping motifs will tend not to add significance to the overall model fit once the most strongly associated motif has been accounted for. The second situation where regression trees should provide an advantage is where multiple different combinations of motifs give rise to similar expression patterns. Though not strongly indicated here, most likely because only a short section of each promoter was examined, in theory combinations of motifs that act together should generate their own arms of the regression tree. It should even be possible for the same motif to appear on different arms at different frequencies, as for example TCGATAT in our decision tree, and for repressor and activator functions to be distinguished.

The utility of regression trees is thus more likely to lie in the perspective they provide concerning the relationship among motifs, rather than superior performance in identifying single motifs. The major factors restricting the application of regression trees relate to the enormous range of possible ways of combining and formulating motifs. While 8-mer and longer motifs may often be functional, perfect matches will often be rare in promoters of co-regulated genes so statistical power is reduced, particularly given that the increased number of possible longer motifs requires more stringent significance thresholds. Similarly, formulation of trees that combine motifs of different lengths, or link motifs in two different regions of a gene (for example, putative promoter and distal enhancer elements), creates so many possible combinations that it will be difficult to assess *a priori* which trees are more or less probable. If the number of co-regulated genes for which a regulatory motif is sought is less than 20 or so, it may never be possible to use regression-based approaches since P values of the order of 10^{-5} would require an unreasonably tight relationship between motif count and transcript abundance. Nevertheless, systematic simulation studies and statistical modelling, including use of other evidence to define candidate regulatory regions within which motifs may lie (Wasserman *et al*., 2000), should improve the performance of regression trees in the context of regulatory motif detection.

(iii) *Do computational approaches identify enhancer elements?*

The standard approach to confirmation that a motif actually regulates gene expression is to demonstrate that it is sufficient to drive expression of a reporter gene in the predicted pattern in a transgenic organism. In our case, the expression data themselves demonstrate, however, that the identified motifs are insufficient to drive female-specific expression, since a large number of genes with each motif combination are expressed more strongly in males than females. Several other recent studies have failed to confirm that sequences identified using bioinformatic approaches are functional. For example, Halfon *et al*. (2002) extracted 34 potential dorsal mesodermal enhancers consisting of multiple binding sites for known transcription factors, but only 8 of the 18 of these for which data are available appear to drive transcription in embryonic *Drosophila* mesoderm. They concluded that there can be a high false-positive identification rate associated with computational strategies.

Given the extremely high significance associated with particular test statistics, it should also be considered that some potential regulatory motifs are not classical enhancers, but rather define a class of 'modulator' elements that act in a more probabilistic manner. Either the effects of individual elements are too subtle to detect in transgenic assays, or the elements act in a context-dependent manner. Promoter-proximal elements such as those characterized in this study are likely to require distal true enhancer sequences, as regulatory regions in flies typically extend over tens of kilobases. The corollary may also be true, that enhancers require the context of modulator elements, such as those identified here, more commonly than generally recognized.

The problem remains as to how to confirm the biological function of statistically significant motifs. One approach is to ask whether the motifs are polymorphic in the promoters of genes that show variable expression within and among species. We sequenced the promoters of 10 wild-type strains of *D. melanogaster* for eight genes that differed between genotypes in the level of sex-specific transcription in our microarray studies. Nine of the 72 SNPs and indel polymorphisms were located within the top 10 motifs described here, but this fraction is not greater than expected given the motif frequencies in the sequenced regions. Nevertheless, polymorphism in modulator elements is an intuitively appealing mechanism for quantitative variation in gene expression that could contribute to gradual evolution of gene expression.

Phylogenetic shadowing (Boffelli *et al*., 2003; Kellis *et al*., 2003), namely extensive genomic comparison of promoter sequences in multiple sibling species among which tissue-specific gene expression diverges, is likely to aid in the functional footprinting of subtle regulatory motifs.

We thank John Doyle for encouraging us in the application of regression and decision trees to motif detection, Brian Oliver for discussions concerning sex-biased gene expression, and Kevin White for communicating unpublished data. Rebecca Riley-Berger and Jennifer King sequenced the promoters of the eight genes. E.H. is the recipient of an IGERT training fellowship in Genome Sciences, and microarray research in G.G.'s laboratory has been supported by the David and Lucille Packard Foundation and NIH award P01 GM45344.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, P. G., *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W. & White, K. P. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275.

Arnosti, D. N. (2003). Analysis and function of transcriptional regulatory elements: insights from *Drosophila*. *Annual Reviews in Entomology* **48**, 579–602.

Bayer, C. A., Curtiss, S. W., Weaver, J. A. & Sullivan, D. T. (1992). Delineation of *cis*-acting sequences required for expression of *Drosophila mojavensis Adh-1*. *Genetics* **131**, 143–153.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the USA* **99**, 757–762.

Blumer, N., Schreiter, K., Hempel, L., Santel, A., Hollmann, M., Schafer, M. A. & Renkawitz-Pohl, R. (2002). A new translational repression element and unusual transcriptional control regulate expression of don juan during *Drosophila* spermatogenesis. *Mechanisms of Development* **110**, 97–112.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L. & Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall/CRC.

Bussemaker, H. J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics* **27**, 167–171.

Crowley, T. E. & Hazelrigg, T. (1995). A male-specific 3′-UTR regulates the steady-state level of the exuperantia mRNA during spermatogenesis in *Drosophila*. *Molecular and General Genetics* **248**, 370–374.

Drouin, J., Sun, Y. L., Tremblay, S., Lavender, P., Schmidt, T. J., de Lean, A. & Nemer, M. (1992). Homodimer formation is rate-limiting for high affinity DNA binding by glucocorticoid receptor. *Molecular Endocrinology* **6**, 1299–1309.

Guha Thakurta, D., Palomar, L., Stormo, G. D., Tedesco, P., Johnson, T. E., Walker, D. W., Lithgow, G., Kim, S. & Link, C. D. (2002). Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Research* **12**, 701–712.

Halfon, M. S., Grad, Y., Church, G. M. & Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated computational model. *Genome Research* **12**, 1019–1028.

Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**, 1205–1214.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G. & Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics* **29**, 389–395.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.

Klingenhoff, A., Kornelie, F., Quandt, K. & Werner, T. (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**, 180–186.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.

Liu, X., Brutlag, D. L. & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing*, 127–138.

Lossky, M. & Wensink, P. C. (1995). Regulation of *Drosophila* yolk protein genes by an ovary-specific GATA factor. *Molecular and Cellular Biology* **15**, 6943–6952.

Manson-McGuire, A., Hughes, J. D. & Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research* **10**, 744–757.

Markstein, M., Markstein, P., Markstein, V. & Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the USA* **99**, 763–768.

McCue, L. A., Thompson, W., Carmack, C. S. & Lawrence, C. E. (2002). Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Research* **12**, 1523–1532.

Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S. & Oliver, B. (2003). Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* **299**, 697–700.

Piano, F., Parisi, M. J., Karess, R. & Kambysellis, M. P. (1999). Evidence for redundancy but not *trans* factor–*cis* element coevolution in the regulation of *Drosophila Yp* genes. *Genetics* **152**, 605–616.

Pilote, L., Miller, D. P., Califf, R. M., Rao, J. S., Weaver, W. D. & Topol, E. J. (1996). Determinants of the use of coronary angiography and revascularization after thrombolysis for acute myocardial infarction. *New England Journal of Medicine* **335**, 1198–1205.

Pilpel, Y., Sudarsanam, P. & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29**, 153–159.

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* **23**, 4878–4884.

Quinlan, J. R. (1992). Learning with continuous classes. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, World Scientific, pp. 343–348.

Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002). The evolution of DNA regulatory regions for Proteo-gamma bacteria by interspecies comparisons. *Genome Research* **12**, 298–308.

Rebeiz, M., Reeves, N. L. & Posakony, J. W. (2002). SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. *Proceedings of the National Academy of Sciences of the USA* **99**, 9888–9893.

Rejwan, C., Collins, N. C., Brunner, L. J., Shuter, B. J. & Ridgway, M. S. (1999). Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology* **80**, 341–348.

Shea, M. J., King, D. L., Conboy, M. J., Mariani, B. D. & Kafatos, F. C. (1990). Proteins that bind to *Drosophila* chorion *cis*-regulatory elements: a new C2H2 zinc finger protein and a C2C2 steroid receptor-like component. *Genes and Development* **4**, 1128–1140.

Wang, Y. & Witten, I. H. (1997) Inducing model trees for continuous classes. *Proceedings of the European Conference on Machine Learning*.

Wasserman, W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000). Human–mouse genome comparisons to locate regulatory sites. *Nature Genetics* **26**, 225–228.

Werner, T. (2002). Finding and decrypting of promoters contributes to the elucidation of gene function. *In Silico Biology* **2**, 249–255.

Witten, I. H. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Diego, CA: Morgan Kaufmann.

Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.