J. Mol. Biol. (1998) **278**, 167–181

JMB

Identification of Regulatory Regions which Confer Muscle-Specific Gene Expression

Wyeth W. Wasserman and James W. Fickett*

Bioinformatics Research Group SmithKline Beecham Pharmaceuticals, Mail Code UW-2230, 709 Swedeland Road, King of Prussia PA 19406, USA

*Corresponding author

For many newly sequenced genes, sequence analysis of the putative protein yields no clue on function. It would be beneficial to be able to identify in the genome the regulatory regions that confer temporal and spatial expression patterns for the uncharacterized genes. Additionally, it would be advantageous to identify regulatory regions within genes of known expression pattern without performing the costly and time consuming laboratory studies now required. To achieve these goals, the wealth of case studies performed over the past 15 years will have to be collected into predictive models of expression. Extensive studies of genes expressed in skeletal muscle have identified specific transcription factors which bind to regulatory elements to control gene expression. However, potential binding sites for these factors occur with sufficient frequency that it is rare for a gene to be found without one. Analysis of experimentally determined muscle regulatory sequences indicates that muscle expression requires multiple elements in close proximity. A model is generated with predictive capability for identifying these muscle-specific regulatory modules. Phylogenetic footprinting, the identification of sequences conserved between distantly related species, complements the statistical predictions. Through the use of logistic regression analysis, the model promises to be easily modified to take advantage of the elucidation of additional factors, cooperation rules, and spacing constraints.

Keywords: muscle-specific expression; logistic regression analysis; position weight matrix; regulatory region prediction; phylogenetic footprinting

Introduction

Eukaryotes, from yeast to man, maintain diverse batteries of genes whose expression levels can be modulated to satisfy the demands of developmental, environmental or physiological conditions. While eukaryotes can manipulate the abundance of proteins through a variety of mechanisms at the DNA, RNA, and protein levels, the most direct and utilized cellular tool is the alteration of gene transcription rates by the differential binding or modification of factors bound to *cis*-acting enhancers. Numerous transcription factors have been identified with roles in context-specific expression, but the activity of any single activating factor is rarely sufficient to explain a biological response. Recent studies suggest that complex, co-operative protein-protein interactions between transcriptions factors are required to determine gene expression patterns (Arnone & Davidson, 1997). The limited knowledge of the interactions between general transcriptional proteins, co-activators, transcription factors, and the DNA-protein scaffold has prevented the formulation of quantitative models to explain complex gene expression patterns, resulting in an abundance of case studies with little generalization.

Muscle tissues have been extensively studied for regulation of gene expression. These tissues have been categorized into three groups: skeletal, cardiac, and smooth muscle (Stockdale, 1992; Hauschka, 1994). Within each group multiple sub-classifications have been defined, such as fast or slow twitch for skeletal muscle, ventricular or atrial for cardiac muscle, and vascular or non-vascular for smooth muscle. Skeletal muscle expression has been most extensively studied, probably as a result of good cell culture models for differentiation (Buckingham, 1992; Olson, 1992; Taylor & Jones, 1979). Expression studies have identified numerous genes which are expressed in differentiated myo-

Abbreviations used: Mef-2, myocyte-specific enhancer factor 2; PWM, position weight matrix; SRF, serum response factor; LRA, logistic regression analysis; EPD, Eukaryotic Promoter Database.

0022–2836/98/160167–15 $25.00/0/mb981700

tubes, but not in myoblasts (Buckingham, 1994). A few dozen of these genes have been analyzed experimentally, sometimes extensively and always partially, to determine the regulatory elements required for expression in differentiated muscle cells.

Regulatory analysis and muscle differentiation studies have revealed several families of transcription factors which contribute to skeletal muscle-specific expression. MyoD, originally discovered for its ability to convert fibroblasts into myoblast-like cells (Davis *et al*., 1987), is the best known member of the skeletal muscle-specific, myogenin-subfamily (Myf) of basic helix-loop-helix (bHLH) proteins. Myocyte-specific enhancer factor 2 (Mef-2) is a family of transcription factors. The Mef-2 isoforms are expressed predominantly in skeletal and cardiac muscle, with expression of some isoforms observed in brain cells (Pollock & Treisman, 1991). The broadly expressed Serum Response Factor (SRF), which is distantly related to the Mef-2 family (both are members of the MADS superfamily), activates expression through CArG sites which are found in many muscle-specific genes (Vandromme *et al*., 1992). Recently several Tef-1-related, muscle-specific factors which functionally bind to muscle-specific CATT regulatory elements (M-CAT sites) have been identified and cloned (Farrance & Ordahl, 1996; Jacquemin *et al*., 1996). The ubiquitous Sp-1 transcription factor, originally considered to be an activator of ubiquitously expressed housekeeping genes, functionally binds to sites required for muscle-specific expression (Sartorelli *et al*., 1990). Uncharacterized binding activities like Mef-3 (Spitz *et al*., 1997), Trex (Fabre-Suver & Hauschka, 1996), and the reverse CArG binding protein (Gopal-Srivastava *et al*., 1995) have been linked to muscle-specific expression, but further studies are required to confirm broad muscle-specific roles for these proteins. The collection of factors with a demonstrated role in activating muscle-specific expression continues to grow in size.

The individual binding of a transcription factor to a regulatory element is rarely sufficient to confer context-specific expression. Cooperation between multiple factors interacting at multiple sites appears to be essential for muscle gene regulation (Weintraub *et al*., 1990), but the biochemical rules governing these interactions remain largely unknown. The presence of multiple regulatory sites is required for muscle-specific expression of the muscle creatine kinase (Amacher *et al*., 1993), troponin-C (Parmacek *et al*., 1994), and cardiac β-myosin heavy chain (Shimizu *et al*., 1992) genes, but multiple regulatory sites alone are not sufficient to confer context-specific expression as shown in a study of Mef-2 binding sites (Gossett *et al*., 1989). Cooperation may be dependent on spacing constraints, as suggested for the interaction of Mef-2 and Myf proteins (Fickett, 1996a). The identification of such spacing constraints requires elucidation of a large number of functional pairs, or focused spacing studies. The identification of additional muscle regulatory regions would help decipher the cooperativity rules which govern context-specific expression, but the time and expense required for detailed regulatory analysis limits the number of genes which can be characterized in the laboratory.

The growth of skeletal muscle regulatory information combined with recent computational advances has opened new avenues for identification of regulatory sequences. When one or two binding sites are known for a transcription factor, identification of new sites is based on comparison to the known sites and is only minimally effective (Claverie & Sauvaget, 1985; Claverie & Audic, 1996). After several transcription factor binding sites are known or *in vitro* selection data are generated, computational studies can utilize specialized multiple alignment methods and position weight matrices (PWM) to more effectively characterize binding specificity and identify possible novel sites (Staden, 1984; Bucher, 1990). Use of PWMs allows quantitative discrimination of sites, with calculated site scores approximating the binding energy of the profiled transcription factor (Stormo, 1990; Berg & von Hippel, 1987, 1988). A recent study found that 95% of the highest scoring sites identified in the GenBank primate DNA sequence database with a PWM for the liver-specific transcription factor HNF-1 can be bound by HNF-1 *in vitro* (Tronche *et al*., 1997). This finding indicates that a carefully constructed PWM is a highly effective tool for identifying sequences which can be bound by a specific transcription factor (cf. also Fickett, 1996b).

The ability to recognize sites to which a factor binds *in vitro*, however, is only a first step towards accurately identifying regulatory regions within an uncharacterized genomic sequence. A significant portion of identified sites seems to be inactive, as demonstrated by the presence of HNF-1 sites in genes specifically expressed in subsets of cells lacking HNF-1 (Tronche *et al*., 1997). Further analysis depends on combining information from a variety of computational tools to indicate which sites are likely to be functional (Duret & Bucher, 1997).

Here, a new approach is developed for the identification of skeletal muscle-specific regulatory modules within a genomic sequence. PWMs for the well characterized muscle factors are developed and analyzed. Potential binding sites for these muscle factors are found to be more prevalent in muscle regulatory regions than other sequence sets. An analysis of known regulatory regions indicates that multiple muscle-specific sites are concentrated into regulatory modules. Logistic regression analysis on multiple sites provides a model through which potential regulatory regions can be found. Phylogenetic footprinting, the identification of conserved sequences, complements the logistic regression predictions, allowing the identification of regions within genes likely to confer muscle-specific expression. The flexibility of the

logistic regression model allows for continual refinement as new factors are characterized, cooperation rules are established, and genomic sequence data expands.

Results

Generation of frequency matrices from known muscle-specific regulatory elements and binding sites independent of muscle genes

As a tool to locate binding sites for transcription factors associated with skeletal muscle-specific expression, we developed position weight matrices profiling the binding sites of Mef-2, Myf, Sp-1, SRF, and Tef factors. While similar in concept to a consensus sequence, PWMs account for the frequency of each base at each position in an alignment of known sites (Staden, 1984). To ensure maximum specificity of the PWM, only those sites were desired for which there was clear and direct evidence both for function and for the identity of the factor bound. Thus sites were collected directly from the experimental literature. Sites were included if required for gene expression in skeletal muscle, even if the conferred expression is limited to skeletal muscle subtypes like fast-twitch, slow-twitch, or embryonic skeletal muscle cells. The experimentally determined sites were aligned using a Gibbs sampling algorithm (Lawrence *et al.*, 1993; Fickett, 1996b). The frequency matrices drawn from the alignments of the muscle-specific sites are presented (Figure 1A).

Construction of specific PWMs requires utilization of the maximum number of elements available, so all known muscle regulatory sequences bound by each transcription factor were included in the binding profiles. The influence of any single element on the function of a PWM is small; nevertheless searching for regulatory regions with matrices including elements drawn from the regions can be considered circular. In order to provide an approach without circularity, matrices were also constructed for each transcription factor using data obtained from *in vitro* binding studies and regulatory sequences from genes not specifically expressed in muscle cells (Figure 1B). High quality *in vitro* binding site selection data were available for Mef-2 (Pollock & Treisman, 1991), Myf (Funk & Wright, 1992), and SRF (Pollock & Treisman, 1990) factors. No site selection data were available for the recently discovered muscle Tef factors, so a muscle-gene independent matrix was produced from five functional binding sites known for the widely expressed Tef-1 protein. While site selection data have been published for Sp-1 (Thiesen & Bach, 1990), the data from this early application of the procedure provided insufficient flanking sequence information for analysis. A set of well characterized Sp-1 sites from genes not specifically expressed in muscle was aligned instead. Binding site selection data for the proto-oncogene c-jun/AP-1 (Pollock & Treisman, 1990) were aligned to produce a matrix for use as a negative control of muscle-gene element detection.

The frequency matrices derived from muscle and independent sources exhibit significant differences. As previously observed (Fickett, 1996b) the Mef-2 matrices are similar at most positions, but the muscle sites show tolerance for adenosine at position 12 and stricter nucleotide preferences at positions 2 and 9. The muscle E-box (Myf) matrix has strict nucleotide requirements at positions 7 and 12 relative to the TSDA-derived matrix. The Sp-1 matrices differ outside the central positions. The Tef matrices differ primarily at position 6, with muscle-derived sites showing a strict requirement for thymidine. The SRF matrices differ from the others in that the site-selection matrix shows stricter binding preferences, particularly at positions 8 to 10. The differences observed between the muscle and site-selection matrices (Mef-2, Myf, SRF) most likely arise from the absence of binding cofactors or differences in the transcription factor isoforms present *in vivo* and utilized in the *in vitro* studies. For example, the binding profiles of c-Jun homodimers are distinct from the profiles for c-Jun/c-Fos heterodimers (Pollock & Treisman, 1990). The Sp-1 and Tef differences may reflect differences between protein isoforms as has been observed for Mef-2 proteins in muscle and brain (Andres *et al.*, 1995). While there are obvious distinctions between the muscle-derived and independent data matrices, the profiles are sufficiently similar for comparison in subsequent analyses.

Generation and analysis of the position weight matrices

The frequency matrices were converted into position weight matrices (PWMs). Using PWMs allows a potential site to be quantitatively evaluated by summing the appropriate entries for the nucleotides observed at each position. The final quantitative site scores can be either positive or negative with a unique score range for each matrix.

To assess the sensitivity and specificity of the PWMs, all were tested against the known muscle regulatory elements. All of the muscle-derived matrices were more effective than the corresponding muscle-independent PWMs at identifying the sites. This partially results from the circular nature of the approach, but, also, likely reflects the above discussed biological differences between the data used to produce the matrices. The score range observed with each PWM for the known, experimentally defined, muscle regulatory sites are presented as a percentage of the individual matrix potential score range (Table 1): $100 \times (\text{observed\_score} - \text{min\_score\_possible}) / (\text{max\_score\_possible} - \text{min\_score\_possible})$. The presented score ranges cannot be directly compared between matrices, as each matrix varies in

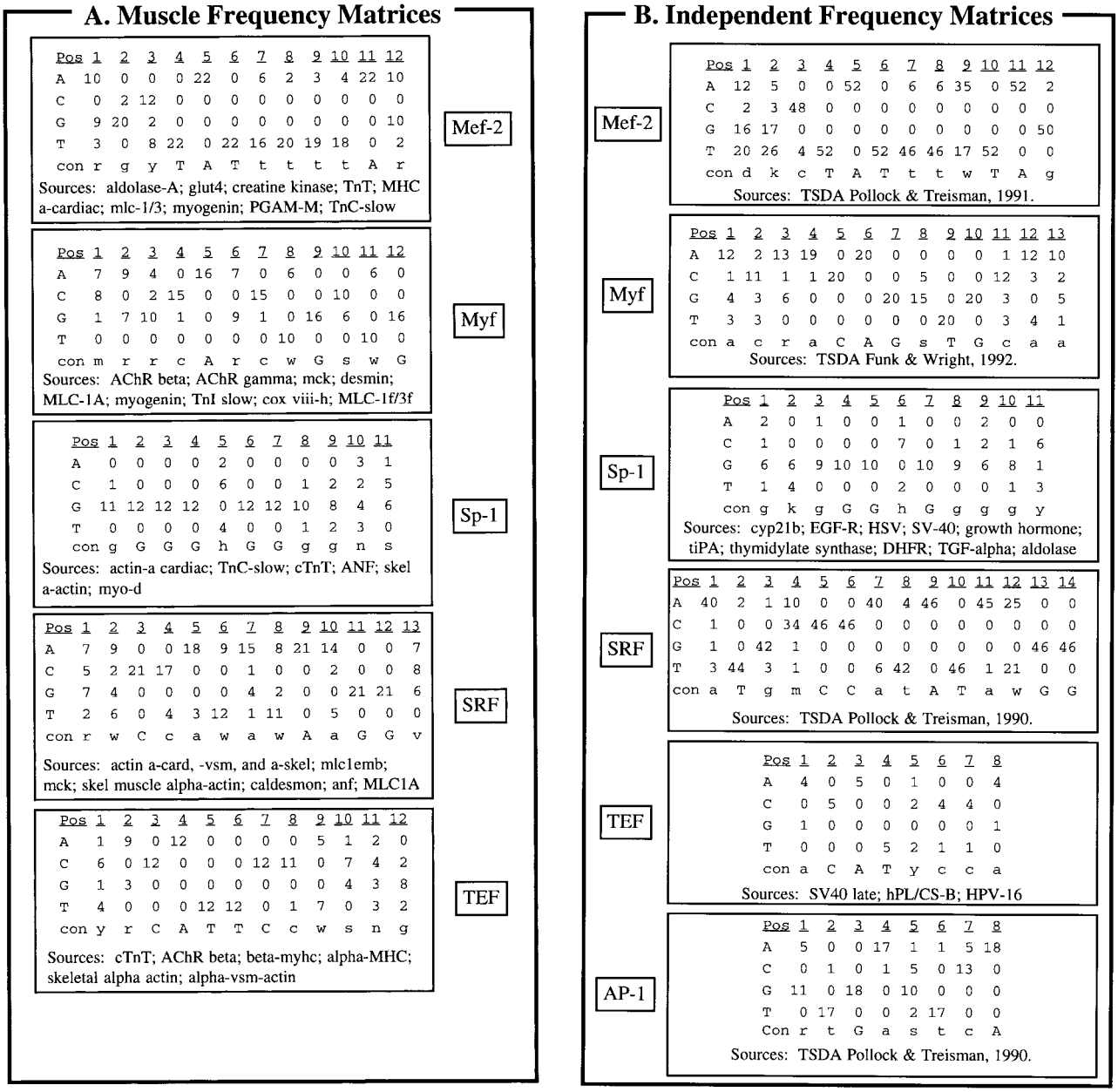

Figure 1. Position frequency matrices. Capital letters in consensus sequences indicate invariant nucleotides. A, Regulatory sites with an experimentally determined role in skeletal muscle gene expression were obtained and grouped by the transcription factors which bind to them. The groups of sequences were aligned and the frequency of each nucleotide at each position is presented in these matrices. B, *In vitro* selected binding sites or functional sites with no role in muscle-specific expression (Sp-1 and Tef-1) were obtained and aligned.

the length of the site, the number of known sites contributing to the matrix, and the base composition. To compare scores, the frequencies of hits in the primate subset of GenBank above a threshold were determined for each matrix (Table 1). The known-site threshold score for each matrix was set to the lowest score determined for any known site. The site frequencies range from 190 bp/hit for the independent Sp-1 matrix to 7100 bp/hit for the muscle-derived Tef matrix. This range of specificity may reflect biological differences in the binding of the individual factors.

Site frequencies in a variety of datasets

The relationship observed between the log of the frequency of sites scoring above a threshold score and the threshold score is approximately linear (representative graphs are shown in Figure 2). In three sequence collections analyzed, including a set of photoreceptor regulatory regions, the EPD database of eukaryotic promoter regions (Bucher & Trifonov, 1986), and the primate subset of GenBank (January 1997, 5.4×10^7 bp), the linear plots have approximately equal slope and y-intercept for

Table 1. Statistics for construction and performance of the position weight matrices

Factor	Name	Sites source	Number sites	Known site scores (%)	bp/hit[a]
Mef-2	A/T-rich	Nat-mus	22	88–100	3700
Mef-2	A/T-rich	TSDA	52	76–90	500
Myf	E-box	Nat-mus	16	82–99	390
Myf	E-box	TSDA	20	80–100	230
SRF	CArG	Nat-mus	17	86–100	3000
SRF	CArG	TSDA	46	70–83	760
TEF	M-CAT	Nat-mus	21	89–97	7100
TEF	M-CAT	Nat-other	5	83–100	490
Sp1	G/C-box	Nat-mus	12	85–99	570
Sp1	G/C-box	Nat-other	10	77–96	190
AP-1	TRE	TSDA	18	86–100	720

[a] Frequency of potential sites is expressed as the average number of base-pairs between ''hits'' (bp/hit) in the primate division of GenBank, where a hit is any site scoring at least as well as some experimentally verified functional site.

each matrix. A higher frequency of high scoring sites is observed in the collection of minimal muscle regulatory regions. This prevalence of high scoring sites is observed with both muscle site-dependent and independent matrices (Figure 2A and B). The plots for the frequency of sites within the photoreceptor dataset is truncated at the threshold score at which the small size of the dataset resulted in the identification of no sites. The SRF graphs are representative of the graphs produced with Tef, Myf, and Mef-2. Sp-1 sites were found to be most prevalent in the muscle collection, but were also more common in the EPD, reflecting the ubiquitous role of Sp-1 factors in gene expression (data not shown). The PWM for the transcription factor AP-1, which is not linked to muscle gene expression, did not identify a greater representation of high scoring sites in any of the datasets (Figure 2C). In contrast to the muscle factor matrices, a matrix for photoreceptor-specific Tα elements identified a higher frequency of sites within the photoreceptor set, and approximately equal site frequencies in the muscle, EPD, and GenBank datasets (data not shown). The PWMs for the muscle transcription factors find high scoring sites more frequently in sequences linked to muscle-specific expression.

The number of putative sites scoring above the known-site threshold scores indicates that most of the sites are not biologically functional. Sites were identified in the primate subset of GenBank with the muscle SRF PWM at a rate of one per 3000 bp (Table 1), or about three putative sites for every human gene. The union of the EPD sequences containing predicted binding sites for muscle factors reveals remarkably little selectivity for muscle regulatory sequences, with 97% of the promoters containing at least one putative site. Even with the exclusion of the Sp-1 sites which occur frequently in many regulatory regions unrelated to muscle expression, sites were still identified in 60% of the sequences in EPD with the muscle derived PWMs (Table 2). A more accurate methodology is needed for identification of potential regulatory regions.

Clustering of functional sites in regulatory regions

Recent studies of muscle regulatory sequences have indicated that cooperativity between transcription factors bound to distinct elements is a key to generating muscle-specific expression (Weintraub *et al.*, 1990; Amacher *et al.*, 1993; Fickett, 1996a; Firulli & Olson, 1997). Similarly, other context-specific regulatory regions appear to contain multiple, functional binding sites (Arnone & Davidson, 1997). To determine if muscle regulatory elements commonly occur in groups of two or more, a review of extensively studied muscle genes was performed. All muscle regulatory regions analyzed by linker scanning or other comprehensive mutagenesis were found to contain multiple functional elements within 200 bp regions (Figure 3), with most containing two sites within a 100 bp region.

Multiple sites as predictors for muscle regulatory regions

Based on the co-occurrence of elements in muscle regulatory regions, the frequency of pairs of sites scoring above the known-site threshold scores in regulatory sequences was determined. Using the more specific muscle matrices, 48% of known muscle regulatory sequences were identified, but 27% of the EPD sequences were found to contain two non-overlapping sites within a 500 bp region (Table 2). The same 48% of muscle regulatory regions were still detected when the pairs were required to occur within 100 bp, while the EPD hits dropped to 15%. An extension of the multiple sites approach to regions containing three sites resulted in further decreases of sensitivity, without substantial gains in specificity (data not shown).

The dual site approach is improved by taking into consideration the quantitative score for each putative element. When pairs of non-overlapping sites within 100 bp were measured by Poisson probability, the specificity was improved. Only 8% of the EPD sites were found to contain pairs of sites with $p < 0.2$, while 42% of the true muscle

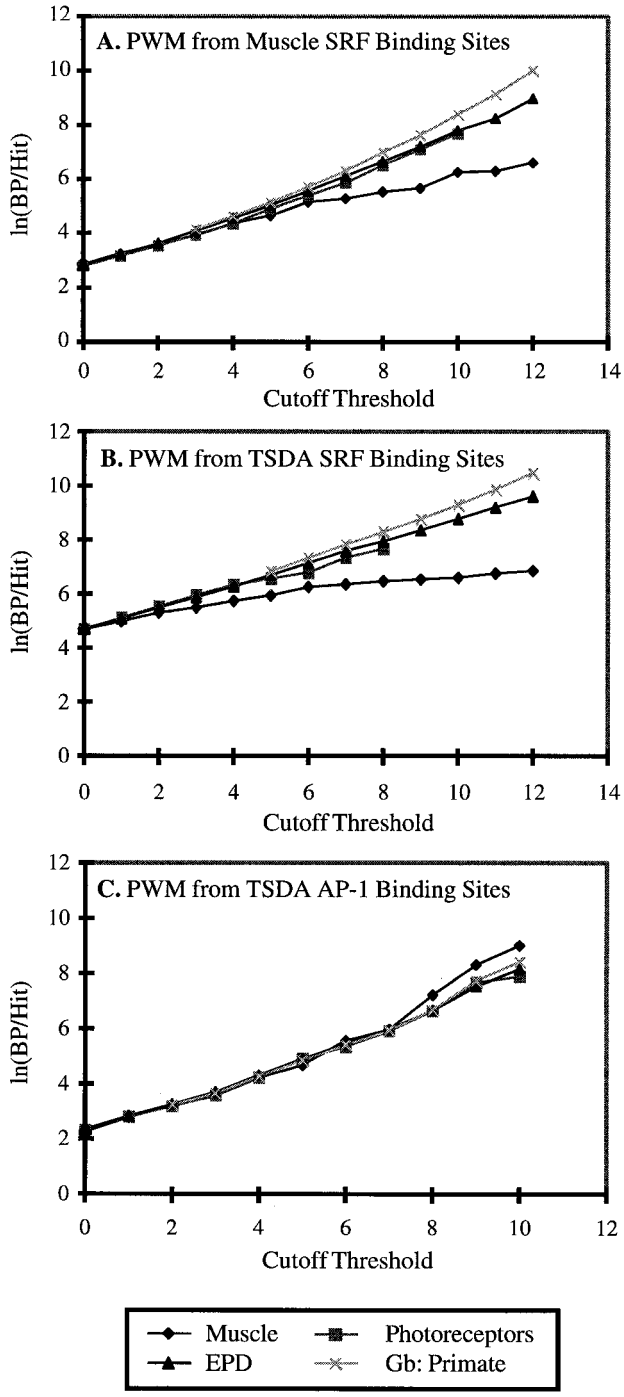

Figure 2. Frequency of sites identified with position weight matrices. Plots show the frequency with which individual matrices identify sites scoring above a threshold score. Data are presented for matrices derived from (A) muscle regulatory sites bound by SRF, (B) sites bound by SRF in an *in vitro* binding assay, and (C) sites bound by AP-1 in an *in vitro* binding assay.

regulatory regions were identified (Table 2). The application of a simple statistical model to the dual site data provides a start towards identification of muscle regulatory sequences.

A model for identification of muscle regulatory regions

A variety of mathematical approaches exist for classifying objects (in our case windows of DNA sequence) with observed attributes (scores for putative transcription factor binding sites) into positive (muscle regulatory regions) or negative classes. Many of these methods are likely to perform with similar effectiveness, and a case could be made for applying any one of them to the problem of predicting regulatory regions. Some methods are prone to overfitting the data, while others provide little insight into the impact of various observations on the predicted outcome. One of the simplest methods to understand, perform, and interpret is logistic regression analysis (LRA; Hosmer & Lemeshow, 1989; Vollmer, 1996). As indicated in Materials and Methods, LRA is based on identifying coefficients for each contributing data field to produce a *logit* value. The coefficients are determined to maximize the discrimination of a positive training set from a negative training set. The maximum likelihood procedure for determining the coefficients identifies them in the order of their contribution to correct classification of the training data. A coefficient for a data field is included in the model only if it significantly contributes to the correct classification of data. The *logit* value is scaled to produce a score between 0 and 1.

LRA models were constructed to determine if the procedure can accurately distinguish muscle regulatory sequences. A training set of 200 bp sequences was produced which contained the following non-muscle (negative) sequences: 300 randomly selected promoter sequences from the EPD database; 1500 randomly drawn sequences from the primate subset of GenBank; and four sequences composed of di- and trinucleotide repeats. The positive training set was composed of the 29 regulatory sequences sufficient for skeletal-muscle-specific expression which have been experimentally localized within 200 bp. Muscle gene regulatory sequences less than 200 bp in length were extended to include flanking nucleotides to bring the total length of all training sequences to 200 bp. For all 1833 sequences, the two highest scoring sites were identified for each transcription factor matrix. The site score coefficients obtained for two models are presented in Table 3. The first model, using the muscle-derived PWM scores for the training set, includes coefficient scores for all five muscle matrices. With the exception of the SRF scores, the model utilized only the best scoring site for each transcription factor. The SRF coefficient was applied to the second best site. The coefficients for the first model were selected in the following order: Tef > Sp-1 > Mef-2 > Myf > SRF. The second model, utilizing the independent PWM scores, utilized the best scores for all five factors, but also included the second best Sp-1 score. This second model selected the coefficients in the following

Table 2. Performance of single or multiple site approaches for identifying the presence of muscle regulatory regions

Criteria	Muscle PWMs		Independent PWMs	
	Muscle	EPD	Muscle	EPD
Single sites	45/48 (94%)	768/1285 (60%)	48/48 (100%)	1249/1285 (97%)
Pairs	23 (48%)	342 (27%)	36 (75%)	1119 (87%)
Pairs in 100 bp	23 (48%)	187 (15%)	36 (75%)	960 (75%)
Pairs w/Poisson	20 (42%)	101 (8%)	23 (48%)	371 (29%)

order: Sp-1 (no. 1) > SRF > Mef-2 > Myf > Tef > Sp-1 (no. 2).

A test-set of the experimentally determined muscle regulatory regions not contained in the training set was used to assess the performance of the LRA models. In Table 4, the test-set performances of the LRA models are presented. These performance scores are based on LRA cutoff scores (thresholds) which enabled identification of 66% of the muscle regulatory sequences in the training set. At this threshold, the LRA model based on muscle-site PWMs identified 60% of the test set, while only 4% of EPD sequences contained positive hits. The independent matrices, as with the previous approaches, showed similar sensitivity, but lower specificity (Table 4). The detection of 60% of the positives in the test set is reasonable for a first generation regulatory module detector, since it is likely that additional participating transcription factors are not yet included (e.g. Mef-3, Spitz *et al.*, 1997 and Trex, Fabre-Suver & Hauschka, 1996).

A jack-knife study was performed in which LRA models were built from training sets lacking each one of the 29 positive sequences in turn (data not shown). Within this set of 29 sequences, the rat glucose transporter-4 (Glut4) enhancer was the only one with a change in classification resulting from the training set variation. More comprehensive jack-knife studies, in which sequences were removed from matrix construction as well as model training, were performed for the six weakest positives in the training set. Only the rat Glut4 enhancer was sufficiently impacted to result in misclassification. Overall the model appears robust, with little dependence on individual sequences in the training data. Future efforts will benefit from additional experimentally defined muscle regulatory regions, but the current data are sufficient to generate an effective model.

Detection of muscle regulatory regions in long genomic sequences

The performance of the muscle PWM LRA method was further evaluated by its performance in analyzing long genomic sequences. The 11 human genomic sequences of 2×10^5 bp or longer present in GenBank were collected, masked to remove common repeat elements (Alu, Line, etc.), and putative regulatory regions were identified by LRA analysis. A total of 91 non-overlapping regions were identified as hits (using the cutoff scores for positive classification of 66% of the muscle modules in the training set) for a frequency of one region per 32,000 bp. A review of 33 hits in the fully annotated sequences was conducted to determine if the predicted sites are in the vicinity of genes preferentially expressed in muscle. Since many of the muscle transcription factor families also have a role in brain-specific expression (Supp *et al.*, 1996; Leifer *et al.*, 1994; Jacquemin *et al.*, 1996), genes preferentially expressed in neural cells were also noted. In the set of 33 predicted regulatory regions, 16 of the putative modules were most proximal (within 1000 bp) to known muscle or brain genes (Table 5). The seven sites which were

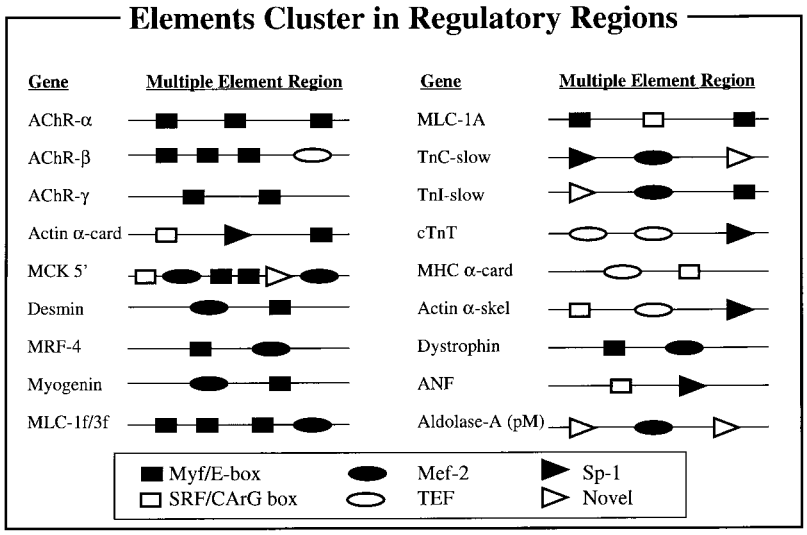

Figure 3. Muscle regulatory sites occur in clusters. The regulatory modules presented contain two or more sites within a 200 bp span.

Table 3. The logistic regression coefficients determined for the optimal classification of the training sequences

	Variable	Coefficient	Wald Chi square score	Pr > Chi-square
Muscle PWMs	Intercept	−20.8	82.4	0.0001
	Tef	0.53	32.0	0.0001
	Sp-1	0.55	30.4	0.0001
	Mef-2	0.39	29.7	0.0001
	Myf	0.38	16.1	0.0001
	SRF[a]	0.40	13.0	0.0003
Independent PWMs	Intercept	−21.1	75.7	0.0001
	Mef-2	0.29	25.1	0.0001
	SRF	0.25	21.2	0.0001
	Myf	0.46	19.9	0.0001
	Tef	0.60	14.2	0.0002
	Sp-1	0.51	9.7	0.002
	Sp-1[a]	0.41	5.1	0.02

[a] The variables indicated were for the second best site for the specified transcription factor found within the 200 bp window.

closest to muscle genes were in the vicinity of a muscle-specific DNase-I-like endonuclease (Pergolizzi *et al.*, 1996; Parrish *et al.*, 1995), the Emery-Dreifuss muscular dystrophy gene emerin (Bione *et al.*, 1994), and the recently identified HXC-26 gene (Toyoda *et al.*, 1996). The generous classification of both brain and muscle hits as positive suggests 17 of 33 (52%) sequences could be incorrectly classified as positive, while a strict muscle-only criterion indicates 26 of 33 (79%) could be mis-classified. Genes are still being discovered and characterized in the genomic sequences (Toyoda *et al.*, 1996), so the numbers should be considered preliminary and conservative. There is no similar algorithm to use in a direct comparison of performance, but these numbers suggest that a predicted muscle regulatory region should receive consideration.

The number of predicted regions appears to be reasonable in quantity. Taking the 79% false positive rate from the genomic sequence analysis and the 40% false negative rate from analysis of the test set into account, the numbers suggest approximately one true muscle regulatory region can be anticipated every 80,000 bp. Assuming an average of two muscle regulatory regions (a promoter and an enhancer) for every muscle gene, the results indicate the presence of approximately 5000 human genes with regulatory regions conferring expression in muscle cells. This is consistent with recent profiles of skeletal muscle gene expression patterns (Pietu *et al.*, 1996; Houlgatte *et al.*, 1995; Lanfranchi *et al.*, 1996). As the LRA performance measures are based on an extremely small sample of the human genome and gene mapping studies indicate muscle genes cluster in the genome (Pallavicini *et al.*, 1997), future studies will be required to reassess the accuracy of the estimates.

Complementing regulatory region analysis with phylogenetic footprinting

As with any method in biology, confirmatory evidence from independent tests greatly strengthens a hypothesis. The logistic regression model for predicting muscle regulatory regions is effective and promises to be improved as more data and information accumulate. Nevertheless, putative muscle regulatory regions will be computationally identified which turn out to be spurious. In order to reduce the time spent pursuing false leads, the LRA prediction can be supported with phylogenetic footprinting where the sequence data permit. In phylogenetic footprinting non-coding sequences are compared between distantly related species to identify regions of genomic sequence conserved over the course of evolution (Duret & Bucher, 1997; Aparicio *et al.*, 1995). Regulatory sequences are more conserved than non-coding sequences with no sequence-specific function. Phylogenetic footprinting does not provide information on the role of conserved regions, but correlation of the phylogenetic footprints with LRA muscle scores may help elucidate functional muscle regulatory regions.

Analysis of genes with tissue-specific expression patterns demonstrates the strength of combining the muscle regulatory module prediction model with phylogenetic footprinting. The cardiac β-myosin heavy chain gene is preferentially expressed in both cardiac and slow-type skeletal muscle. A regulatory module responsible for skeletal muscle-specific expression has been defined 250 bp from the transcriptional start site (Shimizu *et al.*,

Table 4. Performance of the logistic regression models for the identification of muscle regulatory regions

	Muscle PWMs			Independent PWMs		
	Muscle	Photoreceptor	EPD	Muscle	Photoreceptor	EPD
Pairs w/Poisson (%)	42	15	8	48	7	29
Logistic regression (%)	60	0	4	60	7	13

Table 5. Brain and muscle genes identified in analysis of long genomic sequences as containing one or more muscle regulatory regions by logistic regression model

Gene name	Tissue classification
Emerin	Muscle
DNase-I like	Muscle
HXC-26	Muscle
Iduronate 2-sulfatase	Brain
A-1	Brain
Enolase-2	Brain
DRPLA-1	Brain

1993). Peaks in both the LRA and phylogenetic footprint graphs are apparent in this region in Figure 4A. Additional peaks 5′ of the known site may be true elements, as laboratory studies indicate muscle-specific regulatory enhancers are present far upstream of the gene, but the specific locations have not been determined (Knotts *et al.*, 1994). A gene specifically expressed in liver, apolipoprotein C-III, does not have a predicted muscle regulatory region which corresponds to a conserved region (Figure 4B). Peaks in the phylogenetic footprint graph are observed in the location of a module conferring small intestine expression on the neighboring apolipoprotein A-I gene (Bisaha *et al.*, 1995) and at the apo C-III liver expression module adjacent to the first exon (Mietus-Snyder *et al.*, 1992). A putative muscle region observed in the LRA graph within the first intron does not correspond to a peak or highly conserved region in the phylogenetic footprint, suggesting the region is not functional. These examples demonstrate that the combination of the two analyses can focus attention to regions more likely to confer muscle-specific expression.

Application of the combined approach to a newly sequenced gene, Nspl-1, identifies two putative muscle regulatory modules (Figure 4C). Nspl-1 is expressed in brain and muscle from distinct promoters (Geisler *et al.*, 1998). The brain form is expressed from a promoter adjacent to the first exon, while the muscle form is expressed from an internal promoter adjacent to exon 5. The LRA peaks correspond to regions immediately adjacent to the muscle promoter and in the first muscle intron. These locations are consistent with the positions of regulatory modules in many genes (Tronche *et al.*, 1997). The combination of phylogenetic footprinting with the muscle regulatory module predictor is a powerful tool for the identification of likely muscle regulatory regions when homologous sequences are available.

Discussion

The accurate identification of regulatory regions within a genomic sequence is a difficult challenge, both experimentally and computationally. Vast time and enormous expense are required for laboratory identification of regulatory regions, making bioinformatics approaches attractive alternatives. As the genome projects progress, the increase in uncharacterized genomic sequence will preclude laboratory analysis of each gene's regulatory structure, making computational identification of potential *cis*-acting elements valuable. However, the current methods for analyzing regulatory regions *in silico* are not sufficient. While progress has been made in identifying sequences to which individual transcription factors bind, a significant portion of these putative sites are not active *in vivo*. To increase the specificity of computational predictions, it is necessary to identify regulatory modules composed of multiple, cooperatively acting elements.

By combining profiles of well-characterized muscle regulatory elements into a predictive model, we have generated the first computational means of identifying modules regulating context-specific expression. PWMs which quantitatively score potential sites are effective at identifying sequences bound by a transcription factor (Tronche *et al.*, 1997). Extensive studies of the transcriptional regulatory mechanisms conferring context-specific expression have identified many of the transcription factors required and indicated that multiple binding sites in proximity are required (Arnone & Davidson, 1997; Firulli & Olson, 1997). Based on the regulatory module hypothesis, we have combined transcription factor binding profiles to develop a simple means of computationally identifying skeletal muscle regulatory regions using logistic regression analysis.

Performance

An initial approach for identifying muscle regulatory signals based on the presence of individual binding sites for muscle transcription factors was not specific; in fact putative muscle regulatory sites were found in 60% of the diverse promoters present in the Eukaryotic Promoter Database. By identifying pairs of sites in proximity and applying a simple Poisson statistical test, the specificity was substantially improved, but only 40% of known sites were detected. Logistic regression analysis provided the best performance of the methods tested, identifying 60% of the known sites in a test set, and only 4% of the EPD sequences.

It seems that the logistic regression predictions are correct 20 to 25% of the time. Within the set of ''positive'' EPD sequences, a quarter (13/53) were from genes known to be preferentially expressed in muscle. Within the longer genomic sequences examined, 21% of the putative regulatory sequences were most proximal to muscle genes. Some of the genes have only been partially characterized, so this value should be considered conservative. An additional 29% of the putative regulatory regions were most proximal to brain-specific genes, a number likely related to roles for Tef, Mef-2, and Sp-1 family members in both muscle and brain expression (Supp *et al.*, 1996; Leifer *et al.*, 1994; Jacquemin *et al.*, 1996). In the

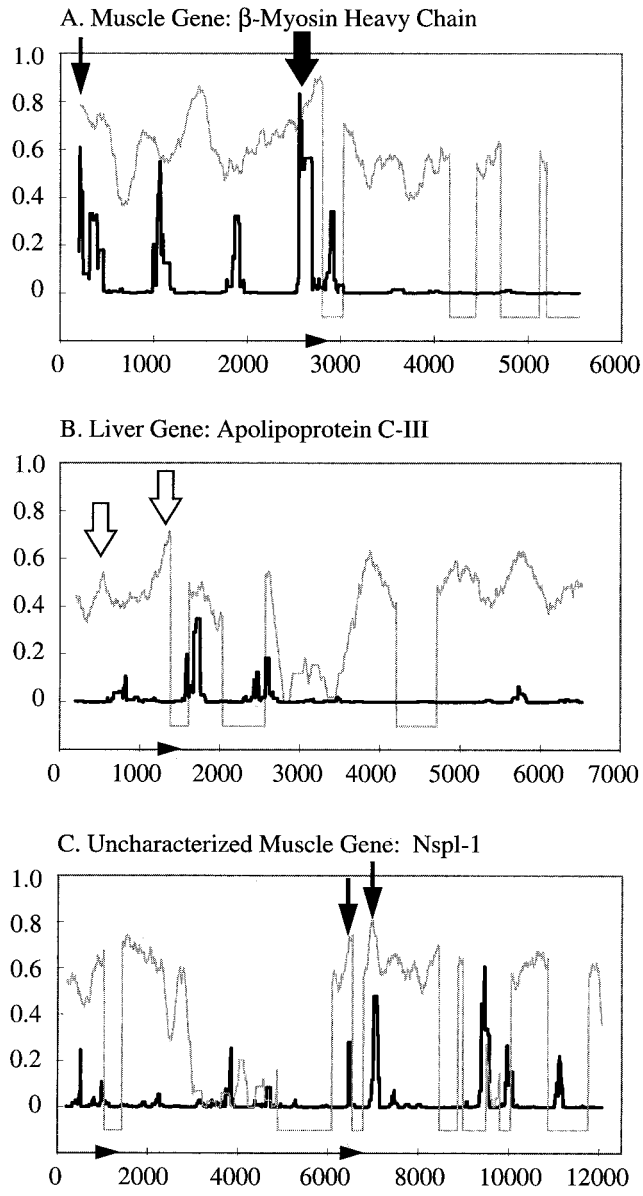

Figure 4. Identification of muscle regulatory modules using logistic regression analysis (dark line) and phylogenetic footprinting (gray line). The phylogenetic footprint values indicate the percentage identity found in 200 bp windows sliding along an alignment of human and mouse genomic sequences. To remove peaks of conservation resulting from the protein coding sequences, the footprint scores for regions containing exons were artificially set to −0.1. The numbers along the *x*-axis indicate the position within the human sequence. Regions with high scores from the logistic regression analysis and evidence of sequence conservation from the phylogenetic footprinting are indicated with dark arrows. The thicker arrows indicate regions with evidence from laboratory experiments, while the thin arrows indicate regions with no published experimental analysis. Thick open arrows indicate experimentally defined regulatory regions with no role in muscle-specific expression. The positions of the transcription start sites are represented on the *x*-axis with triangles. Plots are for (A) the cardiac β-myosin heavy chain gene (human M57965, mouse U86076), (B) the liver-specific apolipoprotein C-III gene (human J00098, mouse L04149), and (C) the Nspl-1 gene (human M89651, mouse submission pending).

long genomic sequences, the LRA model identified putative muscle regulatory regions at a frequency of one per 32,000 bp. The LRA predictions are sufficiently specific to warrant further scrutiny of predicted regulatory modules.

Related work

While we have developed the first practical algorithm aimed at recognizing all regulatory modules active in a particular context, others have utilized transcription factor binding sites in predictive models. Multiple groups have used binding site distribution to predict the location of gene promoters (Prestridge, 1995; Kondrakhin *et al.*, 1995; Chen *et al.*, 1997; Solovyev & Salamov, 1997). These groups coupled TATA site identification tools with analysis of local binding site density to determine regions with a statistically meaningful concentration of putative regulatory sequences. The methods are designed to recognize promoters in general, without regard to a particular expression pattern. Tools for the identification of sequences containing sites in a specific order have been described (Claverie & Sauvaget, 1985; Frech *et al.*, 1997). These tools, which identify a series of regulatory sites with loose spacing rules, have been applied to well-defined strings of sites like the LTR sequences from retroviruses. A similar approach was used to find putative yeast regulatory sequences within a specified distance of an ATG triplet (Fondrat & Kalogeropoulos, 1996). Since regulatory modules do not have defined positions within genes or orders of binding sites within modules, these approaches would be difficult to apply to the problem treated here. Wagner (1997) calculates the significance of finding multiple imperfect occurrences of a consensus binding sequence (assuming a Poisson distribution of such sites) and, for four particular yeast transcription factors, applies this analysis to suggest genes that may be strongly regulated by each individual factor. A recent approach for identifying potential regulatory regions is based on the presence of ''core motifs'' taken from diverse regulatory elements (Crowley *et al.*, 1997). A primary difference in our approach is that we seek to identify sites which are bound by a set of transcription factors involved in a specific pattern of expression. An emphasis on regulatory modules composed of binding sites for multiple transcription factors is an important evolutionary step in computational gene regulation analysis.

Issues and improvements for future examination

As with any first generation computational method, there are numerous avenues to explore for

improvements. Ideally a regulatory module detection system will accurately identify gene regulatory regions and propose the contexts in which these regions will activate expression. Several areas need to be explored in order to achieve such a system.

While other methods for making predictions have been successfully applied to sequence analysis, for instance hidden Markov models (Eddy, 1996) and neural networks (Hirst & Sternberg, 1992), we elected to use logistic regression for specific reasons. Logistic regression is not necessarily the only or the best means of identifying regulatory regions, but it meets five criteria we sought: (1) predictions are based on biologically meaningful data, (2) it is a quantitative approach, (3) results are easily interpreted, (4) the model is adaptable, and (5) the predictions are sufficiently accurate to be useful. Future efforts should explore other statistical approaches, but we suspect that significant performance improvements are dependent upon increased understanding of the biochemical processes directing gene expression.

The LRA model is advantageous because changes and additional features can be quickly analyzed and incorporated if they improve the overall fit of the model to the data. Several features need to be assessed to determine if their inclusion will improve the model (Figure 5). The current LRA model utilizes 200 bp windows for analysis. This number is arbitrary, and it may be possible to determine a biologically meaningful window size which will improve performance (cf. Crowley *et al.*, 1997). Specific offset distances have been found to be important for cooperating Mef-2 and Myf sites (Fickett, 1996a). These distances are based on the helical turn of DNA, such that the center-to-center distance between two sites is found to equal n-turns $\times$ 10.5 + the fixed offset. A preliminary analysis of neighboring Myf sites suggests offset distances may be important for these pairs as well (W.W.W. and J.W.F., unpublished observation). The inclusion of offset rules may improve the model, but a broad measure of the biological significance of offsets cannot be determined until more regulatory regions are analyzed. Regulatory modules occur more frequently in promoters, first introns, and in the 3′ regions of genes, suggesting that consideration of the gene context of a region could improve the model. The LRA-phylogenetic footprinting studies indicate that site conservation is also a useful measure, and as more homologous gene sequences are produced this could be built into the system. Muscle-linked transcriptional repressors may bind to regulatory modules, so binding profiles of YY-1 (Lee *et al.*, 1992) or Twist (Spicer *et al.*, 1996) could improve detection. Considering the limited data for model training, generation of a complex LRA model could result in overfitting the data, but some of these additional features may be valuable.

An alternative to building a complex LRA system, is to subdivide the targets sought. In our current approach, muscle gene enhancers and promoters are both classified as regulatory modules. In reality these are distinct structures which may require separate tools for identification. The LRA model accurately predicted 88% of regulatory enhancers, but only 50% of the muscle promoters. It may be possible to combine a promoter finding tool (reviewed by Fickett & Hatzigeorgiou, 1997) with muscle transcription factor binding site identification to identify muscle promoters more accurately.

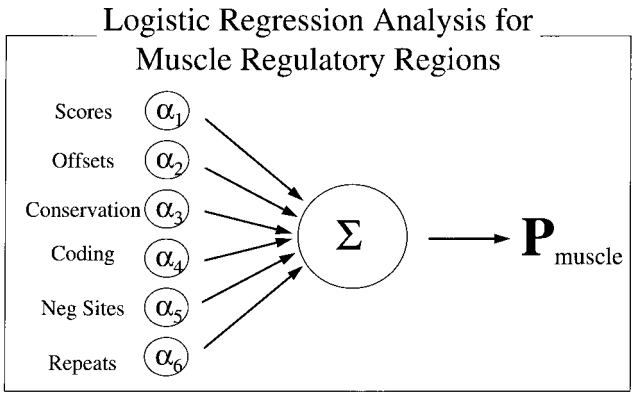

Figure 5. Characteristics to study in future logistic regression models. Logistic regression analysis can be applied to a variety of data to generate predictions.

Extension of the skeletal muscle regulatory module detection system to other tissues will be challenging. In order to generate the muscle gene-based PWMs, a large pool of data from laboratory muscle gene regulatory analysis was needed and extensive time was required to construct the PWMs. For some other extensively studied tissues it may be possible to produce similar systems, particularly for cardiac muscle, liver, brain, and kidney modules. However, the small number of genes comprehensively analyzed is too limiting to extend the current approach to all tissues and contexts. *In vitro* binding assays, as an alternative to compiling regulatory studies, are a powerful tool for the rapid identification of target sites for transcription factors. It is possible that focused TSDA analysis, like a comparison of binding specificity of brain and muscle Mef-2 isoforms (Andres *et al.*, 1995), could provide the needed data to extend regulatory module prediction into other contexts.

Regulatory module identification

The computational identification of regulatory modules is most informative in two instances: (i) when a gene is known to be expressed in a specific context, and (ii) when trying to predict expression patterns for uncharacterized genes identified in genomic sequences. When a gene is known to be expressed in skeletal muscle, approximately 60% of true regulatory regions can be predicted. By combining the LRA model with phylogenetic footprinting, two potential regulatory modules were

identified in the Nspl-1 gene. Phylogenetic footprinting allows proposed regulatory modules to be assessed by their level of conservation across species, but no measures yet exist to assess how much weight should be placed on sequence conservation within putative modules. Identification of regulatory modules within uncharacterized genomic sequences provides insight into the roles of the encoded proteins. Application of the LRA model to genomic sequences identified putative regulatory modules in several skeletal muscle genes, including sites in the DNase-I-like and emerin genes. Since most genes preferentialy expressed in skeletal muscle contain two or more regulatory modules, the probability is high that at least one module will be detected. The results from genomic analysis suggest that given a prediction, the chances that the most proximal gene is highly expressed in muscle is at least one in four. This performance is sufficiently specific to warrant further analysis of all predicted modules.

By harvesting the wealth of data generated in gene regulatory studies, it is possible to develop computational tools for the identification of modules regulating context-specific expression. The specificity of the modular approach results in meaningful predictions.

Materials and Methods

Gibbs alignment of binding sites and PWM generation

The PWM generation and search software have been previously described (Fickett, 1996b). Briefly, transcription factor binding sites were collected from muscle gene regulation literature, some of which has been summarized on the World Wide Web in Muscle-Specific Regulation of Transcription: A Catalog of Regulatory Elements (http://agave.humgen.upenn.edu/MTIR/HomePage.html). Sites for each muscle transcription factor family were aligned using a Gibbs sampling algorithm (Lawrence *et al.*, 1993) as modified for double-strand DNA analysis (Fickett, 1996b). Frequency matrices were produced which contained the number of occurences of each type of nucleotide at each position in the alignment. To create the weight matrices, the PWM entry $m(b,i)$ for base b at position i is calculated from the corresponding frequency matrix entry $f(b,i)$, the number of sites N contributing to the frequency matrix, and the background probability $p(b)$, according to the formula:

$$m(b,i) = \log[(f(b,i) + \mathrm{sqrt}(N)/4)/p(b)]$$

(for a full description see Fickett, 1996b). When PWMs are applied to the analysis of potential sites, the score generated for any particular sequence with a PWM can be interpreted either as an estimate of the free energy of the protein binding to the site, or as the log-likelihood ratio for (i) the hypothesis that the site will be found under the frequency model derived from the alignment of known sites *versus* (ii) the hypothesis that the site will be found under a model derived from the background frequencies of the four bases (Stormo, 1990; Berg & von Hippel, 1987, 1988).

Multiple site analysis

The approximately linear relationship of site scores to the log of the occurrence frequency was subjected to regression analysis to generate an equation to convert scores to expected frequency. The equations are adequate representations of the frequencies for a first approximation. Sequences were analyzed to determine the highest two scoring sites found with each PWM. All possible pairs of these high scoring sites were analyzed to determine the least likely combination. A Poisson probability for at least a pair of sites occurring within the observed distance with the observed scores was determined utilizing the following equation:

$$P = (1 - e^{-\lambda_1})(1 - e^{-\lambda_2})$$

The values for λ_1 and λ_2 were calculated by multiplying the frequency (occurrences per bp) of a site scoring at or above the observed score by the number of base-pairs from the beginning of the first site to the beginning of the second site (sites are modeled as point phenomena, with the abstract point occurrence being at the first base of the actual site). The lowest probability score was reported for each sequence analyzed.

Logistic regression analysis

Logistic regression is similar to the linear regression techniques used to model the dependence of one continuous variable on another, but logistic regression models the dependence of a dichotomous (yes/no) outcome variable on a set of observed (often continuous) variables. In our case the context is a window of DNA sequence, the outcome variable is whether or not that sequence is able to direct muscle-specific expression of a gene, and the observed variables are the best two scores obtained with each PWM, when each PWM is applied at each possible position in the sequence. The outcome variable is modeled as:

$$\pi(x) = e^{logit}/(1 + e^{logit})$$

where the *logit* function is:

$$logit = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_n x_n$$

Here the x_i are the matrix generated scores and the α_i are coefficients determined by a maximum likelihood procedure (in our case using SAS, version 6.11).

The scoring of both genomic and individual gene sequences was performed with a program that identified the highest scoring sites for each PWM within a 200 bp window, and calculated the LRA score using the SAS generated coefficients.

Phylogenetic footprinting

In phylogenetic footprinting sequences are aligned and a regional sequence identity is determined for a window of arbitrary length. To identify the positions of the exons, a mulitple alignment was produced with three sequences: the human and mouse genomic sequences and a sequence composed of the known exons from the human sequence. The progressive multiple sequence alignment software program CLUSTAL W version 1.60 (Thompson *et al.*, 1994) was used to generate the sequence alignments. The gap extension penalty was adjusted to 0 to allow for large gaps in the intronic regions and the gap creation penalty was set at 0.8 to prevent proliferation of gaps. The default match and

mismatch scores were retained. The output from this analysis was submitted to a scoring program which calculated the percentage of nucleotides which were identical between the human and rodent sequences in a sliding 200 bp window. The presence of any nucleotide from an exon sequence within the window resulted in the assignment of a score of −0.1. Based on this scoring system, in the instances where exons are separated by less than 200 bp, the phylogenetic footprint shows a single large region with a score of −0.1.

Sequences

The sequences analyzed in this study for the presence of regulatory modules are available in public databases. The genomic sequences of length greater than 200,000 bp are available with the following GenBank accession numbers (underline indicates sequences which were examined for muscle and brain-specific genes): Y10196, U66061, U66060, U66059, U85195, U91321, AF001549, U91328, AC001228, U91322, U47924, U66082, and L44140.

Most of the sequences used to generate the position weight matrices are available in the public databases, although a small number were obtained from literature sources. All of the muscle regulatory sites used in the matrix construction and the sequences in the muscle regulatory region collection are available on the muscle regulation World Wide Web page at http://agave.humgen.upenn.edu/MTIR/HomePage.html.

Software

All non-commercial software used in these studies was programmed in the language C and implemented on a Sun Unix workstation. The software is available by request from the authors.

Acknowledgments

This work was supported by the Public Health Service grant no. HG00981-01A1 from the National Human Genome Research Institute. We are grateful to John Geisler for providing the Nspl-1 sequence, Laura L. Lopez for aid in gathering muscle regulatory region data, and to Charles E. Lawrence and Pankaj Agarwal for helpful comments and suggestions on the work and manuscript.

References

Amacher, S. L., Buskin, J. N. & Hauschka, S. D. (1993). Multiple regulatory elements contribute differentially to muscle creatine kinase enhancer activity in skeletal and cardiac muscle. *Mol. Cell. Biol.* **13**, 2753–2764.

Andres, V., Cervera, M. & Mahdavi, V. (1995). Determination of the consensus binding site for MEF-2 expressed in muscle and brain reveals tissue-specific sequence constraints. *J. Biol. Chem.* **270**, 23246–23249.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R. & Brenner, S. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl Acad. Sci. USA,* **92**, 1684–1688.

Arnone, M. I. & Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development,* **124**, 1851–1864.

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750.

Berg, O. G. & von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200**, 709–723.

Bione, S., Maestrini, E., Rivella, S., Mancini, M., Regis, S., Romeo, G. & Toniolo, D. (1994). Identification of a novel X-linked gene responsible for Emery-Dreifuss muscular dystrophy. *Nature Genet.* **8**, 323–327.

Bisaha, J. G., Simon, T. C., Gordon, J. I. & Breslow, J. L. (1995). Characterization of an enhancer element in the human apolipoprotein C-III gene that regulates human apolipoprotein A-I gene expression in the intestinal epithelium. *J. Biol. Chem.* **270**, 19979–19988.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578.

Bucher, P. & Trifonov, E. N. (1986). Compilation and analysis of eukaryotic pol II promoter sequences. *Nucl. Acids Res.* **14**, 10009–10026.

Buckingham, M. (1992). Making muscle in mammals. *Trends Genet.* **8**, 144–148.

Buckingham, M. (1994). Muscle: the regulation of myogenesis. *Curr. Opin. Genet. Dev.* **4**, 745–751.

Chen, Q. K., Hertz, G. Z. & Stormo, G. D. (1997). PromFD 1.0: A computer program that predicts eukaryotic polII promoters using strings and IMD matrices. *Comput. Appl. Biosci.* **13**, 29–35.

Claverie, J. M. & Audic, S. (1996). The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.* **12**, 431–439.

Claverie, J. M. & Sauvaget, I. (1985). Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Comput. Appl. Biosci.* **1**, 95–104.

Crowley, E. M., Roeder, K. & Bina, M. (1997). A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.* **269**, 8–14.

Davis, R. L., Weintraub, H. & Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell,* **51**, 987–1000.

Duret, L. & Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399–406.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.

Fabre-Suver, C. & Hauschka, S. D. (1996). A novel site in the muscle creatine kinase enhancer is required for expression in skeletal but not cardiac muscle. *J. Biol. Chem.* **271**, 4646–4652.

Farrance, I. K. & Ordahl, C. P. (1996). The role of transcription enhancer factor-1 (TEF-1) related proteins in the formation of M-CAT binding complexes in muscle and non-muscle tissues. *J. Biol. Chem.* **271**, 8266–8274.

Fickett, J. W. (1996a). Coordinate positioning of MEF-2 and myogenin binding sites. *Gene,* **172**, GC19–32.

Fickett, J. W. (1996b). Quantitative discrimination of MEF-2 sites. *Mol. Cell. Biol.* **16**, 437–441.

Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878.

Firulli, A. B. & Olson, E. N. (1997). Modular regulation of muscle gene transcription: a mechanism for muscle cell diversity. *Trends Genet.* **13**, 364–369.

Fondrat, C. & Kalogeropoulos, A. (1996). Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III. *Comput. Appl. Biosci.* **12**, 363–374.

Frech, K., Danescu-Mayer, J. & Werner, T. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* **270**, 674–687.

Funk, W. D. & Wright, W. E. (1992). Cyclic amplification and selection of targets for multicomponent complexes: myogenin interacts with factors recognizing binding sites for basic helix-loop-helix, nuclear factor 1, myocyte-specific enhancer-binding factor 2, and COMP1 factor. *Proc. Natl Acad. Sci. USA,* **89**, 9484–9488.

Geisler, J. G., Stubbs, L. J., Wasserman, W. W. & Mucenski, M. L. (1998). Molecular cloning of a novel murine gene with predominant muscle and neural expression. *Mammalian Genome,* **9**, 274–282.

Gopal-Srivastava, R., Haynes, J. I. & Piatigorsky, J. (1995). Regulation of the murine alpha B-crystallin/small heat shock protein gene in cardiac muscle. *Mol. Cell. Biol.* **15**, 7081–7090.

Gossett, L. A., Kelvin, D. J., Sternberg, E. A. & Olson, E. N. (1989). A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes. *Mol. Cell. Biol.* **9**, 5022–5033.

Hauschka, S. D. (1994). The embryonic origin of muscle. In *Myology, Basic and Clinical* (Engel, A. G. & Franzini-Armstrong, C., eds), pp. 3–73, McGraw-Hill, New York.

Hirst, J. D. & Sternberg, M. J. (1992). Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry,* **31**, 7211–7218.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression,* pp. 1–36, John Wiley & Sons, New York.

Houlgatte, R., Mariage, Samson R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B. & Auffray, C. (1995). The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5**, 272–304.

Jacquemin, P., Hwang, J. J., Martial, J. A., Dolle, P. & Davidson, I. (1996). A novel family of developmentally regulated mammalian transcription factors containing the TEA/ATTS DNA binding domain. *J. Biol. Chem.* **271**, 21775–21785.

Knotts, S., Rindt, H., Neumann, J. & Robbins, J. (1994). In vivo regulation of the mouse beta myosin heavy chain gene. *J. Biol. Chem.* **269**, 31275–31282.

Kondrakhin, Y. V., Kel, A. E., Kolchanov, N. A., Romaschenko, A. G. & Milanesi, L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Appl. Biosci.* **11**, 477–488.

Lanfranchi, G., Muraro, T., Caldara, F., Pacchioni, B., Pallavicini, A., Pandolfo, D., Toppo, S., Trevisan, S., Scarso, S. & Valle, G. (1996). Identification of 4370 expressed sequence tags from a 3′-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. *Genome Res.* **6**, 35–42.

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.

Lee, T. C., Shi, Y. & Schwartz, R. J. (1992). Displacement of BrdUrd-induced YY1 by serum response factor activates skeletal alpha-actin transcription in embryonic myoblasts. *Proc. Natl Acad. Sci. USA,* **89**, 9814–9818.

Leifer, D., Golden, J. & Kowall, N. W. (1994). Myocyte-specific enhancer binding factor 2C expression in human brain development. *Neuroscience,* **63**, 1067–1079.

Mietus-Snyder, M., Sladek, F. M., Ginsburg, G. S., Kuo, C. F., Ladias, J. A., Darnell, J. E., Jr. & Karathanasis, S. K. (1992). Antagonism between apolipoprotein AI regulatory protein 1, Ear3/COUP-TF, and hepatocyte nuclear factor 4 modulates apolipoprotein CIII gene expression in liver and intestinal cells. *Mol. Cell. Biol.* **12**, 1708–1718.

Olson, E. N. (1992). Interplay between proliferation and differentiation within the myogenic lineage. *Dev. Biol.* **154**, 261–272.

Pallavicini, A., Zimbello, R., Tiso, N., Muraro, T., Rampoldi, L., Bortoluzzi, S., Valle, G., Lanfranchi, G. & Danieli, G. A. (1997). The preliminary transcript map of a human skeletal muscle. *Hum. Mol. Genet.* **9**, 1445–1450.

Parmacek, M. S., Ip, H. S., Jung, F., Shen, T., Martin, J. F., Vora, A. J., Olson, E. N. & Leiden, J. M. (1994). A novel myogenic regulatory circuit controls slow/cardiac troponin C gene transcription in skeletal muscle. *Mol. Cell. Biol.* **14**, 1870–1885.

Parrish, J. E., Ciccodicola, A., Wehhert, M., Cox, G. F., Chen, E. & Nelson, D. L. (1995). A muscle-specific DNase I-like gene in human Xq28. *Hum. Mol. Genet.* **4**, 1557–1564.

Pergolizzi, R., Appierto, V., Bosetti, A., DeBellis, G. L., Rovida, E. & Biunno, I. (1996). Cloning of a gene encoding a DNase I-like endonuclease in the human Xq28 region. *Gene.* **168**, 267–270.

Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage, Sampson R., Houlgatte, R., Soularue, P. & Auffray, C. (1996). Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* **6**, 492–503.

Pollock, R. & Treisman, R. (1990). A sensitive method for the determination of protein-DNA binding specificities. *Nucl. Acids Res.* **18**, 6197–6204.

Pollock, R. & Treisman, R. (1991). Human SRF-related proteins: DNA-binding properties and potential regulatory targets. *Genes Dev.* **5**, 2327–2341.

Prestridge, D. S. (1995). Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249**, 923–932.

Sartorelli, V., Webster, K. A. & Kedes, L. (1990). Muscle-specific expression of the cardiac alpha-actin gene requires MyoD1, CArG-box binding factor, and Sp-1. *Genes Dev.* **4**, 1811–1822.

Shimizu, N., Prior, G., Umeda, P. K. & Zak, R. (1992). cis-acting elements responsible for muscle-specific expression of the myosin heavy chain beta gene. *Nucl. Acids Res.* **20**, 1793–1799.

Shimizu, N., Smith, G. & Izumo, S. (1993). Both a ubiquitous factor mTEF-1 and a distinct muscle-specific factor bind to the M-CAT motif of the myosin

heavy chain beta gene. *Nucl. Acids Res.* **21**, 4103–4110.

Solovyev, V. & Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organism genome sequences. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A., eds), pp. 294–302, AAAI Press, Menlo Park.

Spicer, D. B., Rhee, J., Cheung, W. L. & Lassar, A. B. (1996). Inhibition of myogenic bHLH and Mef-2 transcription factors by the bHLH protein Twist. *Science,* **272**, 1476–1480.

Spitz, F., Salminen, M., Demignon, J., Kahn, A., Daegelen, D. & Maire, P. (1997). A combination of MEF3 and NFI proteins activates transcription in a subset of fast-twitch muscles. *Mol. Cell. Biol.* **17**, 656–666.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucl. Acids Res.* **12**, 505–519.

Stockdale, F. E. (1992). Myogenic cell lineages. *Dev. Biol.* **156**, 281–298.

Stormo, G. D. (1990). Consensus patterns in DNA. *Methods Enzymol.* **183**, 211–221.

Supp, D. M., Witte, D. P., Branford, W. W., Smith, E. P. & Potter, S. S. (1996). Sp4, a member of the Sp-1-family of zinc finger transcription factors, is required for normal murine growth, viability, and male fertility. *Dev. Biol.* **176**, 284–299.

Taylor, S. M. & Jones, P. A. (1979). Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine. *Cell,* **17**, 771–779.

Thiesen, H. J. & Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP-1 protein. *Nucl. Acids Res.* **18**, 3203–3209.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Toyoda, A., Sakai, T., Sugiyama, Y., Kusuda, J., Hashimoto, K. & Maeda, H. (1996). Isolation and analysis of a novel gene, HXC-26, adjacent to the rab GDP dissociation inhibitor gene located at human chromosome Xq28 region. *DNA Res.* **3**, 337–340.

Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.* **266**, 231–245.

Vandromme, M., Gauthier, Rouviere C., Carnac, G., Lamb, N. & Fernandez, A. (1992). Serum response factor p67SRF is expressed and required during myogenic differentiation of both mouse C2 and rat L6 muscle cell lines. *J. Cell. Biol.* **118**, 1489–1500.

Vollmer, R. T. (1996). Multivariate statistical analysis for pathologists. *Am. J. Clin. Pathol.* **105**, 115–126.

Wagner, A. (1997). A computational genomics approach to the identification of gene networks. *Nucl. Acids Res.* **25**, 3594–3604.

Weintraub, H. L., Davis, R., Lockshon, D. & Lassar, A. (1990). MyoD binds cooperatively to two sites in a target enhancer sequence: occupancy of two sites is required for activation. *Proc. Natl Acad. Sci. USA,* **87**, 5623–5627.

Edited by G. Von Heijne

(Received 8 December 1997; received in revised form 5 February 1998; accepted 6 February 1998)

Note added in proof: The computationally predicted upstream regulatory module in the cardiac β-myosin heavy chain gene (Figure 4A) has recently been identified as functional in an independent laboratory analysis.

Huang, W. Y., Chen, J. J., Shih, N. L. & Liew, C. C. (1997). Multiple muscle-specific regulatory elements are associated with a DNase I hypersensitive site of the cardiac beta-myosin heavy-chain gene. *Biochem J.* **327**, 507–512.