IDENTIFYING MUSCLE REGULATORY ELEMENTS AND GENES IN THE NEMATODE *CAENORHABDITIS ELEGANS*

D. GUHATHAKURTA, L.A. SCHRIEFER, M.C. HRESKO
R.H. WATERSTON, G.D. STORMO
Department of Genetics, Washington University School of Medicine
4566 Scott Avenue, Campus Box: 8232, St. Louis. MO 63110. USA
Phone: (314)747−5535; Fax: (314)362−7855
Email: {dg,larry,coutu,stormo}@genetics.wustl.edu; bwaterst@watson.wustl.edu

We report the identification of several putative muscle−specific regulatory elements, and genes which are expressed preferentially in the muscle of the nematode *Caenorhabditis elegans.* We used computational pattern finding methods to identify *cis*−regulatory motifs from promoter regions of a set of genes known to express preferentially in muscle; each motif describes the potential binding sites for an unknown regulatory factor. The significance and specificity of the identified motifs were evaluated using several different control sequence sets. Using the motifs, we searched the entire *C. elegans* genome for genes whose promoter regions have a high probability of being bound by the putative regulatory factors. Genes that met this criterion and were not included in our initial set were predicted to be good candidates for muscle expression. Some of these candidates are additional, known muscle expressed genes and several others are shown here to be preferentially expressed in muscle cells by using GFP (green fluorescent protein) constructs. The methods described here can be used to predict the spatial expression pattern of many uncharacterized genes.

1 Introduction

Establishing where and when a gene is expressed and understanding the underlying regulatory network which guides its expression are critical in understanding gene function in a multicellular organism. The transcription regulatory apparatus which directs temporo−spatial expression of genes is encoded in the DNA, in the form of organized arrays of transcription factor (TF) binding sites[1,2]. These *cis*−regulatory sites are recognized sequence−specifically by cognate TFs which control and guide the expression pattern of genes.

We are interested in identifying muscle−specific regulatory elements, and genes expressed in muscle as tools to study muscle development. Out of the thousands of genes which express in a particular tissue, the most interesting genes to study for understanding its development, differentiation, function and structure, are the ones which are *preferentially* expressed in that tissue. Preferential expression can be either *selective* (expression in a *subset* of tissues or cell types in an organism) or *specific* (expression in *only one* tissue or cell type). Experimentally elucidating novel, additional genes which are expressed specifically or selectively in a tissue, or finding new *cis*−regulatory elements which function only in one tissue type is time consuming as well as challenging. Hence, computational methods which can accurately predict tissue−specific genes and regulatory elements are of great value. Here we describe computational approaches for the identification of muscle−specific *cis*−regulatory elements and genes which are expressed preferentially in the *C. elegans* muscle.

We obtained a list of 35 genes known to be selectively or specifically expressed in the muscle of *C. elegans*. We used a subset of these genes as the training set and the remaining genes as the test set. From the promoter regions of the training set genes we identified several conserved motifs using two different computational methods. We evaluated the significance and specificity of these motifs for muscle−expressed genes using the test and several control sets. The identified motifs describe potential target binding sites for novel transcription factors. Using these motifs, we searched the *C. elegans* genome for genes whose promoter regions have a high probability of being bound by the regulatory factors. These genes were considered as potential candidates for muscle expression. Several identified candidates were known muscle genes (present in both training as well as test sets). We have tested the expression pattern of some of the other candidates using GFP technology and found that several of these genes are indeed expressed preferentially in *C. elegans* muscle.

The methods described here can be used in identifying regulatory elements and genes in other tissues and cell types in *C. elegans* and other eukaryotic organisms.

2 Data

Training set: Upstream regions (−2000 to −1, relative to the translation start) of 19 genes known to be expressed selectively or specifically in *C. elegans* muscle[3,4,5].
Control set 1: Upstream regions of a completely *different* set of 16 genes known to be expressed selectively or specifically in the *C. elegans* muscle[3,4,5].
Control set 2: Upstream regions of 500 genes, randomly selected from the *C. elegans* genome.
Control set 3: Upstream regions of 19 genes known to be expressed selectively or specifically in the *C. elegans* intestine[6,7] (J.D. McGhee, personal communication). None of these genes have any known expression in muscle.

Complete gene lists (with additional references) are available at http://ural.wustl.edu/~dg/PSB02.html. All sequences were downloaded from the WormBase anonymous ftp server: ftp://ftp.wormbase.org/pub/wormbase/.

3 Methods

3.1 Identification of regulatory motifs

Two local multiple sequence alignment methods, Consensus[8] and ANN−Spec[9],

were run on the training set sequences to identify conserved motifs. Both Consensus and ANN−Spec use weight matrix models to represent un−gapped sequence motifs. Since the TF binding sites in a set of similarly regulated sequences are expected to be conserved to a certain extent, conserved motifs identified by these programs represent potential regulatory elements.
Consensus: Consensus[8] uses a greedy algorithm and searches for a matrix with a low probability of occurring by chance, or, equivalently, having a high information content (I.C.)[10]. Version 6.c of Consensus was used. The top scoring results were reported from different runs. Different pattern lengths were tested, and both strands of the DNA were searched for motifs since TFs can bind to either strand. Patterns with high I.C. and the lowest expected frequency were considered.
ANN−Spec: ANN−Spec[9] uses a simple artificial neural network and Gibbs sampling[11] method to define DNA binding site patterns. The program searches for the parameters of a simple perceptron network (weight matrix) which maximize the specificity for binding a positive sequence set (or training set) compared to a background sequence set. Binding sites in the positive data set are found with the resulting weight matrix and these sites are then used to define a local multiple sequence alignment. ANN−Spec Version 1.0 was used. A comparison of ANN−Spec and other related programs has shown that ANN−Spec is able to identify patterns of higher specificity when training with background sequences (C.T. Workman and G.D. Stormo, unpublished observation). Hence, for ANN−Spec, a background sequence set of upstream regions from 3000 randomly picked genes was used. Different motif lengths were tried and both strands of the DNA were searched for motifs. Due to the non−deterministic nature of the algorithm, multiple training runs are performed (100), with each run iterating 2000 times. The results were sorted by their best attained objective function values. Weight matrices corresponding to the ten highest scoring runs were observed. If >5 of these top scoring ten runs give a motif with one consistent pattern consensus, that pattern is considered significant.

3.2 Searching for "sites"in sequences

The Patser program (G.Z. Hertz and G.D. Stormo, unpublished) allows one to score the words of a sequence against a weight matrix. Once the weight matrices for regulatory motifs are obtained by Consensus or ANN−Spec, the matrices can be used as input for Patser to identify high scoring sub−sequences (or "sites") in a given set of sequences. Patser calculates the p−value (or probability) of observing a particular score or higher at a particular sequence position[12]. A "cutoff" score for eliminating low scoring sub−sequences is also calculated numerically. From an alignment of sites in a binding site pattern, the program calculates the cutoff score as follows. The "true" information content of an alignment of sites is given by:

$$I_{\text{sites}} = - \sum_{b} \sum_{k} f(b,k) \ln \frac{f(b,k)}{p(b)} \qquad \text{.......} \ (1)$$

where, f(b,k) is the frequency of observing a base, b, at a particular position, k, in a binding site, p(b) is the prior probability for base b in the genome, k sums over all *l* positions of the pattern (*l* being the length of the pattern), and b sums over all four DNA bases. *ln(probability)* of observing binding sites in a random sequence is related to this true information content[10]: *ln(probability)* $\leq I_{sites}$. The "sample size adjusted" information content of an alignment is the true information content *minus* the average information content expected from an arbitrary alignment of random sites. Patser approximates the target *ln(probability)* of the cutoff score (*i.e.* the probability of observing a score greater or equal to the cutoff score) as −(sample size adjusted information content); the cutoff score can then be calculated from this *ln(probability)* value.

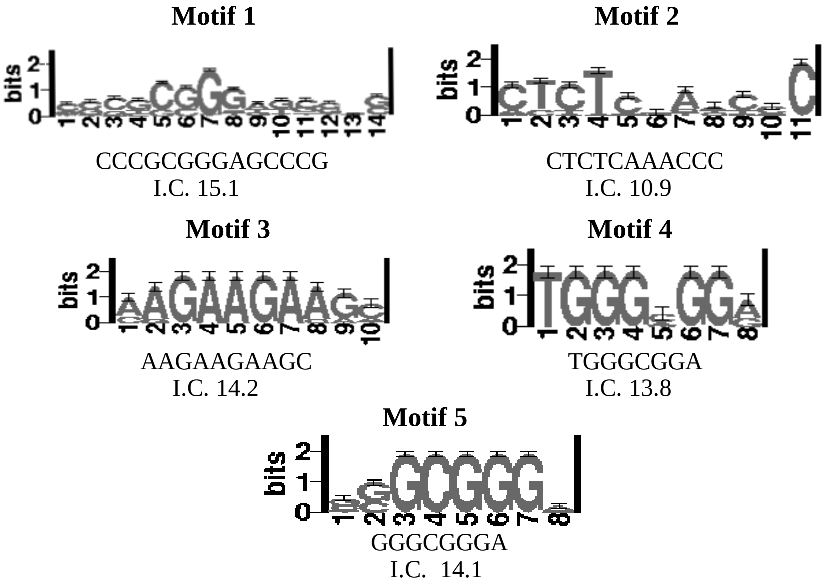
3.3 Establishing spatial expression pattern of genes

A major advance in the attempts to localize gene expression and proteins is the recent advent of green fluorescent protein (GFP) as a reporter molecule in living organisms[13]. GFP is a protein from jellyfish that emits green fluorescence when excited by blue light, even when expressed in heterologous organisms. Here, the promoters (−6000 to −1) of the genes which are predicted to be expressed in muscle are fused with the GFP−coding sequence using genetic recombination, so that the GFP is under the regulatory control of the promoter. Suitable DNA constructs for promoter::GFP are injected into the gonad of the hermaphrodite worms. A portion of the progeny segregating from the injected animals express GFP under the control of the promoter of interest. Green fluorescence from the GFP is observed in the different cells and tissues of these progeny. (Detailed description of the method is available at: http://ural.wustl.edu/~dg/PSB02.html.)

4 Results

4.1 Identification of regulatory motifs

One very strong motif with the consensus CCCGCGGGAGCCCG (Motif 1, Figure 1) was obtained using both Consensus and ANN−Spec. Some shorter motifs were also found which appeared to be parts of the above motif and were ignored. Instances of this motif (sub−sequences scoring above the Patser cutoff value) were identified in the training set. These sites were then deleted from the sequences and Consensus and ANN−Spec programs were re−run which resulted in identification of several other motifs (motifs 2 through 5, Figure 1). We checked to see if the motifs found in our analysis were previously reported. Motifs 4 and 5 are very similar to the G−rich binding sites of the ubiquitous, Sp−1 like, transcription factor which has been shown to regulate the expression of many different classes of genes including housekeeping and muscle genes[14]. Since our objective was to identify muscle−specific regulatory elements, we did not consider motifs 4 and 5 any further. Motifs 1, 2, and 3 were novel since they did not have any matches to any known sites in the TRANSFAC database[15].



Motif 1

CCCGCGGGAGCCCG
I.C. 15.1

Motif 2

CTCTCAAACCC
I.C. 10.9

Motif 3

AAGAAGAAGC
I.C. 14.2

Motif 4

TGGGCGGA
I.C. 13.8

Motif 5

GGGCGGGA
I.C. 14.1

Figure 1: Motifs identified using Consensus and ANN−Spec programs. The motif consensus, information content of the motifs in bits and sequence logos[16] are given.

4.2 DNA binding probability and significance of identified motifs

A "site" in a sequence is simply a high scoring sub−sequence which is obtained by the Patser program using the motif weight matrix as an input. A term which is proportional to the probability of a TF molecule binding to its sites in a sequence can be obtained from the "site scores" calculated by the Patser program. The free energy of binding ($-\Delta G$) of a TF molecule to a DNA site is proportional to the score (s) of the site[17,18], $\Delta G = -RTs$, where, R is the universal gas constant and T is the absolute temperature. Let us assume that occupancy of DNA sites by a TF follows the Boltzmann distribution under the condition of thermodynamic equilibrium. Then, the probability that a TF molecule occupies a site with score s, is given by:

$$P(s)\ \alpha\ e^{-\Delta G/RT} \qquad \text{..... (2a)}$$

$$\text{or, } P(s)\ \alpha\ e^{s} \qquad \text{..... (2b)}$$

This is called the probability proportionality value (*pp−value*). Since we will use these *pp−values* only for the purpose of comparison between different sequences (see below), the proportionality constant in equation 2 is of no consequence to us, and we assume that the constant is equal to 1. For any motif, *m*, there can be multiple sites in a given sequence scoring above the Patser cutoff; the *pp−value* for binding of a TF molecule to *any* of its several binding sites in a sequence is given by the sum of individual terms:

$$P_{m}^{seq} = \sum_{\mathrm{Sites}} \mathrm{e}^{s} \qquad \text{.... (3)}$$

The average *pp−value* for a TF, corresponding to motif *m*, to be bound to a sequence in a given set of N sequences is:

$$< P_{m}^{seq} > = \frac{1}{\mathrm{N}} \sum_{\mathrm{Seqs}} \sum_{\mathrm{Sites}} \mathrm{e}^{s} \qquad \text{..... (4)}$$

We calculate this *pp−value* for both the training and the random sets (Table 1).

For efficient gene regulation TFs need to bind effectively to the regulatory elements in the promoter region i.e. the binding energy (−ΔG) of the TF to the promoter region of the regulated gene should be higher compared to other (background) sequences. In other words, the probability of binding to the regulated sequences should be higher compared to background sequences, assuming the components of the cells are in thermodynamic equilibrium and the binding events follow the Boltzmann distribution. There are two possible ways in which this may be achieved. First, the binding energy of the TF to one individual site in the upstream region of the regulated gene can be very high (*Mode 1*); alternatively, in the absence of a very strong binding site, there can be multiple weaker sites, with lower binding energies (individually), but the combined effect of these sites may result in high binding probability (*Mode 2*) (see equation 3). A combination of both strategies is also possible. We do not know which mode is more suitable for describing gene regulation by the putative TFs for the identified DNA binding site motifs. Therefore, we determined several relevant parameters for both modes of binding (Table 1) for the training set as well as the control set 2 (random set). First, using Patser, potential binding sites for each motif were determined in both sets. We then calculated: average number of sites per sequence, average score of the binding sites, average of the maximum scoring sites from each sequence, and a measure of the probability of binding of a TF to its sites in a particular sequence. Parameters were determined from (a) only the highest scoring sites in each sequence (*Mode 1*) (b) all sites scoring above the respective Patser cutoff scores (*Mode 2*). For the purpose of comparison, we have also shown the parameters for an unrelated pattern, ACTGATA ("GATA" in Table 1), which is obtained by Consensus and ANN−Spec from the promoters of a list of genes expressed in the *C. elegans* intestine (D. GuhaThakurta, J.D. McGhee and G.D. Stormo, unpublished observation). Sites corresponding to this motif have been shown to be important for intestine−specific expression of the *ges−1* gene in *C. elegans*[6].

Table 1: Potential TF binding site parameters for training and random set. Column 1: average number of sites per sequence above the Patser calculated cutoff, Column 2: average score per site, Column 3: *pp−value* (equation 4) calculated from all sites scoring above the Patser calculated cutoff value. For columns 4 and 5, the highest scoring sites in each sequence were determined using Patser (highest scoring sites in some sequences may have scores below the cut−off values). Average of the highest scores from each sequence is given in column 4. Column 5 shows the *pp−value* based on only the highest scoring sites in each sequence. The values in columns 3 and 5 in both sets are to be multiplied by a factor of 10^4.

	Training Set					**Random Set**						
Motif Index	*Avg. no. of sites per seq.*	*Avg. score per site*	*pp−val* *all sites > cutoff*	*Avg. highest score per seq.*	*pp−val* *with highest scoring sites*	*Avg. no. of sites per seq.*	*Avg. score per site*	*pp−val* *all sites > cutoff*	*Avg. highest score per seq.*	*pp−val* *with highest scoring sites*	*Ratio* $\frac{C^T_3}{C^R_3}$	*Ratio* $\frac{C^T_5}{C^R_5}$
	C^T_1	C^T_2	C^T_3	C^T_4	C^T_5	C^R_1	C^R_2	C^R_3	C^R_4	C^R_5	R_C	R_H
1	3.45	10.5	61.9	11.79	47.8	0.69	9.51	2.55	7.1	1.92	24.3	24.9
2	5.15	7.17	4.19	9.36	3.16	2.68	6.84	0.74	7.04	0.49	5.7	6.4
3	1.6	10.32	10.1	9.3	2.68	0.66	9.15	1.60	7.25	0.51	6.3	5.25
GATA	1.6	6.12	1.1	6.1	0.73	2.15	6.38	2.8	6.54	2.19	0.39	0.33

Ratios, R_C and R_H, are called *discrimination factors* or *R−factors*. Given two sequences, one from the training set and another from the random set, the discrimination factors show how likely it is for the cognate TF to bind a training set sequence as opposed to a random sequence. The R−factors are nearly identical using all sites above the **c**utoff (R_C) or using only the **h**ighest scoring sites (R_H). For the cognate TF corresponding to motif 1, it is about 24−25 times more likely that the TF will bind to a training set sequence. The R−factor for motif 1 is much higher than that of motifs 2 or 3, which might explain why the motif appeared to be the most significant one in our first round of Consensus and ANN−Spec runs. In eukaryotic gene regulation, it is common for multiple TFs to act together and bind DNA in a cooperative fashion[1,2,14,19,20]. If this is the case here, then the combined effect of multiple TFs binding to the sites could be dramatic even though the individual discrimination factors for motifs 2 and 3 are on the order of 5−6. The R−factor for the unrelated GATA motif is less than 0.4.

4.3 Specificity of identified motifs for muscle genes

The *combined pp−value* for multiple motifs is calculated for the upstream sequence of each gene in the *C. elegans* genome. For lack of more specific information regarding the mode of TF binding and interaction at this point, we assume that for selective or specific expression of a gene in the muscle context: *(1) all* relevant TFs (corresponding to the motifs 1, 2 and 3) need to bind to the upstream sequence, and *(2)* if there are multiple sites scoring above the Patser cutoff for a particular motif, *any one* of those binding sites may to be occupied by the corresponding TF. For a particular upstream sequence, the combined *pp−value* for multiple motifs is calculated by taking a product of individual *pp−value*s (from equation 3) for the different motifs:

$$\boldsymbol{P}^{\mathbf{seq}} = \prod_{m=1}^{3} P_m^{seq} \qquad \text{..... (5)}$$

All (19,804) upstream sequences were sorted according to the log of the combined *pp−value,* ln($\boldsymbol{P}^{seq}$) (equation 5). Two sorted lists were generated, viz. list 1, where the combined *pp−values* were calculated for each sequence using only the highest scoring sites corresponding to the three motifs (*Mode 1*); and list 2, where the *pp−values* for each sequence was calculated using all sites for the three motifs scoring above the respective Patser cutoffs (*Mode 2*). Based on the position of the genes in a sorted list, a "*specificity score*" can be calculated for a given sequence set. Suppose the positions of N genes of a given set in the sorted list are: x_1, .. x_n, .. x_N. The probability that a particular gene is at position x_n or higher in the sorted list is given by: (x_n /19,804). The joint probability of observing N genes at positions x_1 ... x_N or higher in the list is given by the product of individual probabilities. We consider the log of this probability, and to have a measure independent of the number of sequences, we define the specificity score as:

$$\text{Specificity Score} = -\frac{1}{N}\left[\sum_{n} \ln\left(\frac{x_n}{19{,}804}\right)\right] \qquad \text{...... (6)}$$

We calculate the specificity score for the training and the three control sets. Using list 1, the specificity score for the training set, and control sets 1, 2 and 3 were 4.76, 2.63, 1.04 and 0.63 respectively. Using list 2 the specificity scores for the training set and three control sets (in the same order as above) were 4.22, 2.5, 1.01 and 0.8. The specificity scores for the second muscle gene set (control set 1) are not as high as that for the training set, but still substantially higher than the random (control set 2) or the intestine (control set 3) gene sets. The higher specificity scores for the training set and the control set 1, show that the identified motifs are specific (or, at least selective) for the muscle genes.

4.4 Selection of candidates for testing muscle expression by GFP

We considered the two sorted lists from section 4.3. Several known muscle genes were placed high on the lists (within top 25, Table 2); for reference the highest scoring intestine gene was placed at only 2029 in list 1 and 3943 in list 2. To select a few genes for GFP−expression testing, we took the top−scoring 25 genes from list 1, since the specificity score of muscle genes was higher using this list compared to list 2. To minimize the false positive rate, we checked for the presence these genes in the top scoring 50 genes in the list 2. Genes which score high in *both* lists were considered good candidates for muscle expression. Several of these candidates were known muscle genes (training set or control set 1). The remaining ones were selected for GFP−expression testing (Table 2, Figure 2).

Table 2: the list of top scoring 25 genes from list 1 (see text, section 4.4). genes which were in the training set are in bold. genes which were in control set 1 are in italics. all previously known muscle genes are indicated in column 5 and candidates which have been verified here to have muscle−specific or selective expression are indicated in column 6.

Pos. in List 1	***Pos. in List 2***	***Gene ID***	***Gene name or putative product***	***Previously known muscle expression***	***GFP verified expression in muscle***
1	**1**	**F09F7.2**	**mlc−3**	**Yes**	
2	**4**	**ZK617.1b**	**unc−22**	**Yes**	
3	*13*	*K10B3.8*	*gpd−2*	*Yes*	
4	**9**	**Y105E8B.c**	**tmy−1**	**Yes**	
5	7	F55C7.2	unknown		
6	8	Y44A6D.3	unknown		
7	*12*	*F08B6.4b*	*unc−87*	*Yes*	
8	**25**	**F11C3.3**	**unc−54**	**Yes**	
9	2	C49A1.6	unknown		
10	6	F58F6.3	unknown		
11	11	B0513.1	gei−5		Yes
12	15	F47F6.1	unknown		Yes
13	46	F02E9.2b	lin−28		
14	10	R08B4.2	transcription factor		No
15	19	F14B8.5a	unknown		
16	44	F41E7.6	carnitine octanoyltransferase		Yes
17	14	T22C1.7	unknown		Yes
18	18	W06F12.1a	ser/thr kinase		
19	41	F29C12.1	unknown		
20	*40*	*T22E5.5*	*mup−2*	*Yes*	
21	*31*	*C16C2.2*	*eat−16*	*Yes*	
22	51	C05D11.4	let−756		
23	64	C09D8.2	ptp−3		
24	**36**	**K12F2.1**	**myo−3**	**Yes**	
25	**31**	**D1081.2**	**unc−120**	**Yes**	

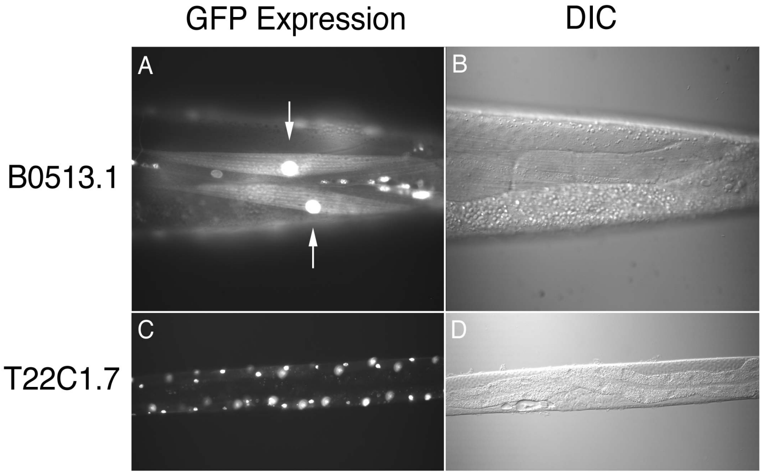


Figure 2: Experimental verification of candidate genes; expression pattern of two genes are shown (additional figures are available at http://ural.wustl.edu/~dg/PSB02.html/). Fluorescence (A, C) and corresponding DIC (differential interference contrast) (B, D) images of transgenic worms expressing GFP under the control of 6 Kb region upstream of B0513.1 and T22C1.7. **A**. GFP−dependent fluorescence is detected in the nuclei (arrows) and in the cytoplasm of bodywall muscle cells. **C**. GFP dependent fluorescence is detected in nuclei of the body wall muscle cells close to the outer edge of the animal. This focal plane shows in−focus nuclei from two quadrants and out−of−focus nuclei are from the other two quadrants.

4.5 Identifying spatial expression of candidate genes

Accurately establishing that a gene is expressed in *only one* cell type using the currently available techniques (e.g. *in−situ* hybridization, GFP) is difficult. This is in part because the observance of expression of in only one cell type does not rule out low or transient expression in other tissues. In addition, depending on the technique used, detection of expression in certain tissues can be problematic. Hence some of the genes which are described in the literature as muscle specific, may actually be expressed in a few other tissues.

We determined the spatial expression of candidate genes in adult and larval worms using GFP−reporter constructs, in which GFP is under the control of the promoter regions of the candidate genes. Complete identification of the spatial GFP−expression pattern is difficult and still ongoing, however, general statements concerning localization can be made. T22C1.7 is expressed predominately in the bodywall muscle (Fig. 2c) and in cells tentatively identified as pharyngeal, vulva and intestinal muscle. Thus, this gene could be muscle specific, but we need to critically identify its expression in other cells before we can make this claim. B0513.1 is clearly expressed in the bodywall (Fig. 2a) and vulval muscle. Its GFP expression also includes a limited set of non−muscle tissue including intestine, neurons and, probably, hypodermis. F41E7.6 has GFP expression in intestinal muscle, sphincter muscle and anal depressor muscle. It is also expressed in non−muscle tissue of the pharyngeal−intestinal valve and a small number of neurons. Muscle expression of the F47F6.1 gene was observed under a GFP−dissecting microscope in the initial progeny (F1) of the injected adult worms. However, no transmitting lines for this gene were established, and therefore further detailed investigation of this gene was not done. R08B4.2 is expressed predominately in neuronal tissue. However, its expression in muscle tissue cannot be ruled out for reasons mentioned above. Transient expression of this gene in muscle during embryogenesis is possible and we are in the process of characterizing the expression pattern during development. There was no observed GFP expression from the promoter regions of genes F58F6.3 and C09D8.2. Possible reasons for this are: experimental error, low level or transient GFP expression, or these genes could be pseudogenes that are not expressed. We are continuing with experiments to determine the expression for the remainder of the genes in Table 2.

5 Discussion

Using computational pattern recognition methods we identified several potential muscle−specific regulatory elements. These putative regulatory elements were then used to predict other genes which might be preferentially expressed in the muscle tissue. Out of the top 25 genes in list 1, 23 score highly (within the top 50) in list 2. Out of these 23, 6 were from the training set, 4 were known muscle genes which were not included in our training set, and 4 more have been experimentally shown to have muscle−selective expression. Thus, checking for consistency in the two lists gives a high true positive rate for identification of muscle−specific or selective genes. We are in the process of checking the expression of some lower scoring genes (e.g. genes at positions 25 through 100). We believe some of these genes will also show muscle−selective or specific expression.

A number of additional considerations are likely to increase the efficiency of identification of muscle genes. Here, we started with a partial set of muscle genes for the purpose of cross validation and evaluation against an independent test set; including all known muscle genes in our training set could increase the quality and specificity of the regulatory motif weight matrices and lead to more efficient detection of other muscle genes. A more thorough computational study should also be helpful; e.g. a study of the distance distribution and orientation of the sites can illustrate the possible modes by which the TFs interact with the DNA sites and with each other. This should lead to building better models for the TF−DNA interaction and allow us to identify muscle genes with higher specificity[14,19,20,21]. In addition we need to initiate experiments to test whether the predicted regulatory elements are functional and guide muscle−selective or specific expression. These studies will not only help in more efficient identification of muscle−specific genes but facilitate our understanding of muscle−specific regulatory elements and mechanisms which guide gene expression in this tissue. Clearly, the studies described here can be helpful in understanding the underlying regulatory mechanism, and in identifying new genes which are expressed in other spatial contexts and tissues not only in *C. elegans* but also other eukaryotic organisms. This knowledge may also find applications in gene therapy, where using tissue−specific regulatory elements, one can design promoters for the purpose of gene delivery to specific tissues[22].

Acknowledgments

John Spieth is acknowledged for helpful discussions and advice on our work. Jim McGhee is thanked for providing a list of intestine genes. The work was supported by grant (HG00249) from National Institutes of Health to GDS.

References

1. Arnone, M.I. and Davidson, E.H. *Development*, **124**, 1857, (1997)
2. Yuh, C−H., Bolouri, H., and Davidson, E.H. *Science*, **279**, 1896, (1998)
3. Moerman, D.G., and Fire, A. In Riddle, D.L., Blumenthal, T., Meyer, B.J., Priess, J.R. (eds.) "*C. elegans II*", Cold Spring Harbor Laboratory Press, 417, (1997)
4. Waterston, R.H., In Wood, W.B. (ed.) "The nematode *C. elegans*", Cold Spring Harbor Laboratory Press, 281 (1988)
5. The *C. elegans* consortium, *Science*, **282**, 2012, (1998)
6. Egan, C.R. *et.al., Development*, **170**, 397, (1995).
7. Mochii, M., *et.al. Proc. Natl. Acad. Sci. USA.*, **96**, 15020−15025 (1999)
8. Hertz, G.Z. and Stormo, G.D. *Bioinformatics*, **15**, 563, (1999)
9. Workman, C.T. and Stormo, G.D. *Pacific Symp. Biocomp.*, **5**, 464, (2000)
10. Schneider, T.D., *et.al. J. Mol. Biol.*, **188**, 415, (1986).
11. Lawrence, C.E., *et.al. Science*, **262**, 208, (1993)
12. Staden, R. *Comp. App. Biosci.*, **5**, 89, (1989)
13. Chalfie, M., *et.al. Science*, **263**, 802, (1994)
14. Wasserman, W.W. and Fickett, J.W. *J. Mol. Biol*., 278, 167, (1998)
15. Wingender, E. *et.al. Nucleic Acids Res.*, **29**, 281, (2001)
16. Schneider, T.D. and Stephens, R.M. *Nucleic Acids Res.* **18**, 6097, (1990)
17. Stormo, G.D. *J. Theo. Biol.*, **195**, 135, (1998)
18. Stormo, G.D., and Fields, D.S. *Trends Biochm. Sci.*, **23**, 109, (1998)
19. Fickett, J.W. *Gene*, 172, GC19, (1996)
20. Wagner, A. *Bioinformatics*, **15**, 776, (1999)

21 Klingenhoff, A., *et.al. Bioinformatics*, **15**, 180 (1999)

22. Nettelbeck, D.K., *et.al. Trends Genet.*, **16**, 174, (2000)