BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 23 2008, pages 2684–2690
doi:10.1093/bioinformatics/btn526

Sequence analysis

Position-dependent motif characterization using non-negative matrix factorization

Lucie N. Hutchins[1], Sean M. Murphy[1], Priyam Singh[2] and Joel H. Graber[1,2,*]

[1]Center for Genome Dynamics, The Jackson Laboratory, Bar Harbor, ME 04609 and [2]Bioinformatics Program, Boston University, Boston, MA 02215, USA

Received on May 16, 2008; revised on September 12, 2008; accepted on October 6, 2008

Advance Access publication October 13, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: *Cis*-acting regulatory elements are frequently constrained by both sequence content and positioning relative to a functional site, such as a splice or polyadenylation site. We describe an approach to regulatory motif analysis based on non-negative matrix factorization (NMF). Whereas existing pattern recognition algorithms commonly focus primarily on sequence content, our method simultaneously characterizes both positioning and sequence content of putative motifs.

Results: Tests on artificially generated sequences show that NMF can faithfully reproduce both positioning and content of test motifs. We show how the variation of the residual sum of squares can be used to give a robust estimate of the number of motifs or patterns in a sequence set. Our analysis distinguishes multiple motifs with significant overlap in sequence content and/or positioning. Finally, we demonstrate the use of the NMF approach through characterization of biologically interesting datasets. Specifically, an analysis of mRNA 3′-processing (cleavage and polyadenylation) sites from a broad range of higher eukaryotes reveals a conserved core pattern of three elements.

Contact: joel.graber@jax.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Regulatory sequence identification and characterization remains an important and challenging problem. Many classes of functional sequences, for instance those involved in processing of precursor mRNA (pre-mRNA), are constrained both by sequence content and positioning with respect to a functional site. Standard pattern recognition tools for sequence analysis typically either ignore positioning effects altogether, or else force the positioning to fit a predefined model, such as a Gaussian distribution (Ao *et al.*, 2004). Recent reviews of computational approaches to pattern recognition have pointed out that positioning preference remains an underutilized means of specifying motifs (Li and Tompa, 2006; Tompa *et al.*, 2005). Specific positioning of motifs, especially relative positioning, can be evolutionarily conserved (Vardhanabhuti *et al.*, 2007). Characterization of positioning dependencies have most commonly been approached in variants of positional word counting (PWC) (Fairbrother *et al.*, 2004; Hu *et al.*, 2005; Salisbury *et al.*, 2006), tracking both occurrence and relative position of all of the sequence words of a given size k (k-mers). PWC results in a two-dimensional matrix ($\underline{V}$), in which rows are indexed by k-mer and columns are indexed by position relative to a functional site. The rows of such a matrix can be interpreted as unnormalized conditional probability distributions, reflecting the probability of observing the specific k-mer at a given position relative to the functional site (and conditional on its presence).

Previous analysis based on PWC relied on standard clustering approaches, such as hierarchical or k-means, identifying groups of k-mers with similar positioning probabilities near functional sites (Hu *et al.*, 2005; Loke *et al.*, 2005). Several approaches have been developed to merge the clusters of related k-mers into a single representative positional weight matrix (PWM) (Hu *et al.*, 2005; Loke *et al.*, 2005). These approaches produce reasonable results in the instance of independent, non-overlapping motifs, however, functional motifs often overlap significantly in positioning and/or sequence content. A particularly relevant case can be found in the UG- and U-rich elements that occur downstream of vertebrate 3′-processing sites. These elements are quite degenerate in sequence content (Hu *et al.*, 2005; Salisbury *et al.*, 2006; Zhao *et al.*, 1999), and have overlapping but distinct positioning distributions (Salisbury *et al.*, 2006), conditions that result in a number of U-rich k-mers with multi-modal positioning distributions, reflecting the likelihood that the k-mer can appear as a part of either the U- or UG-rich motif (e.g. Supplementary Fig. 1).

This defines the challenge to be addressed: given the set of positioning distributions for all k-mers (as defined in the PWC matrix), derive the underlying motifs that are most likely to have generated it. To solve this problem, we adapted the non-negative matrix factorization (NMF) algorithm, a dimension reduction technique first developed for image processing (Lee and Seung, 1999) that has subsequently gained popularity in processing high-dimensional biological data, such as microarrays (Carmona-Saez *et al.*, 2006; Kim and Tidor, 2003; Pascual-Montano *et al.*, 2006b). NMF, similar to principal components analysis, generates Òbasis vectorsÓ (referred to here as 'elements') to reduce the dimensionality of large datasets. The original authors found that the non-negativity constraint resulted in elements that reflected distinct components of the underlying data, for instance eyes or a nose in facial images. In our analysis, this is analogous to elements representing specific sequence patterns or putative regulatory motifs.

*To whom correspondence should be addressed.

We first define terminology: in the abstract analysis of NMF, we refer to a set of features that are measured in samples. In tangible examples, we have pixels measured in images, gene probesets measured in microarray experiments, or sequence words (k-mers) counted in positions or windows. The successful application of NMF relies on the validity of the following assumptions: (1) the measurement of feature i in sample j can be approximated as a linear superposition of r patterns [equation (1)]; (2) the m-th term in the sum is the product of two probabilities: the probability of observing feature i in pattern m and the probability of observing pattern m in sample j and (3) the relative probabilities of observing each feature in the m-th pattern are constant across all samples.

$$V_{ij} = \sum_{m=1}^{r} p(\text{feature}_i|\text{pattern}_m)p(\text{pattern}_m|\text{measurement}_j) \qquad (1)$$

In this article, we investigate the parameters of the NMF method, specifically investigating the robustness of the solutions with variation in the analysis parameters. We demonstrate the use of the NMF algorithm for the detection and characterization of sequence motifs that include constrained positioning, producing models of both sequence content and positioning. We show how variation in the residual sum of squares (RSS) of the approximated data matrix provides a robust estimate of the appropriate number of elements, while simultaneously revealing whether or not an NMF analysis is appropriate for the dataset. Analysis of synthetic datasets demonstrates that NMF can accurately describe both positioning and sequence characteristics of implanted patterns, and that NMF finds patterns that are missed by other approaches to pattern recognition. Finally, we apply the NMF approach to two interesting biological datasets: 3′-processing sites from a phylogenetically broad sampling of eukaryotic organisms and transcription start sites from the fruit fly, *Drosophila melanogaster* (available in Supplementary Materials).

2 MATERIALS AND METHODS

2.1 Synthetic test matrices and sequences

To test the theoretical basis for the use of the RSS to establish the number of elements r, we generated two artificial matrices. The first matrix was generated from pseudo-random draws from a Gaussian distribution. The second matrix was generated to precisely match the conditions that NMF models. Starting with a random background, patterns to be added were generated by pseudo-random draws that assigned (1) the probability of each feature occurring in each pattern (corresponding to the $\underline{W}$ matrix, defined below), and (2) the probability of each pattern occurring in each measurement (the $\underline{H}$ matrix defined below).

To test the complete motif characterization procedure, we generated randomized sequences (based on fixed known dinucleotide frequencies) seeded with known motifs (Table 1). In most tests, motifs were inserted into 95% of the test sequences, however, we also explicitly also generated sets of size $T = 300$ in which all motifs were put into 50% and 25% of the sequences, respectively. In order to better mimic true conditions, we used sequences with length 400 nt, defining the center point as position 0, and used distinct dinucleotide frequencies for the negative and positive portions of the sequences. To test the performance with changes in number of sequences, we generated independent training sets with sizes $T = 30$, 100, 300, 1000, 3000 and 10 000, respectively. Motifs were placed within these sequences according to independent pseudo-random draws from a positioning distribution and a sequence content PWM. All sequence files used in this analysis are available at http://harlequin.jax.org/nmf/.

Table 1. Summary of the six motifs used in artificial sequence sets

Motif	Consensus	*bits/base*[a]	[b]X^2
1	CCCCCC	0.64	2.9
2	GGTGGG	1.5	7.1
3	AATAAA	1.5	23.2
4	ACAC	0.82	51.1
5	TGTGTG	0.82	18.2
6	GCGCGCGC	0.82	1.9

[a] *bits/base* is the summed information context across the motif, divided by the length of the motif.
[b] X^2 is a χ^2 measure of the divergence of the motif positioning from a uniform distribution.

2.2 Biological training sequences

We obtained putative 3′-processing sites from our database PACdb, which uses EST-to-genome alignments to assign probable sites as described previously (Brockman *et al.*, 2005). Putative transcription start sites and surrounding sequences for *D.melanogaster* were obtained from the Supplementary Material from Gershenzon *et al.* (2006).

2.3 NMF decomposition

Starting with a set of training sequences all containing, and aligned on, a common functional site, we generate the PWC matrix $\underline{V}$, a two-dimensional matrix of counts that is indexed by k-mer and relative position, respectively. In practice, k-mers are counted in contiguous windows of size w, where each k-mer is assigned to the window in which it begins. As such, k-mers can span the boundary between adjacent windows, but will only be assigned to one. In the case of small datasets (less than a few hundred training sequences), we find it advantageous to smooth each row of $\underline{V}$ independently, maintaining the total counts across the entire row (for real or artificial data). Pseudocounts are also an available option to compensate for small datasets, however they have not explicitly been investigated in this study. Using the same update and objective function as the original NMF publication (Lee and Seung, 1999), we decompose the PWC matrix $\underline{V}$ according to Equation (2), where V_{ij} = count of the i-th sequence word in the j-th position window, W_{im} = the weight of the i-th k-mer in the m-th element, H_{mj} = activity of the m-th element in the j-th position window, M is the number of k-mers considered, N is the number of positioning windows, and r is the number of elements created.

$$\underset{[MxN]}{\mathbf{V}} = \underset{[Mxr]}{\mathbf{W}} \text{ x } \underset{[rxN]}{\mathbf{H}} \qquad (2)$$

We interpret the r basis vectors as representing distinct patterns of sequence content and positioning (Graber *et al.*, 2007), including both specific sequence elements representing functional components, along with specific changes in sequence composition such as found at the junction between UTR and intergenic sequence. In practice, we typically view the positioning data ($\underline{H}$) in its raw form and also in a normalized form, in which each column of $\underline{H}$ is scaled to a common area under its curve. The raw data view enables comparison of the relative activity between elements, whereas the normalized view emphasizes the specificity of positioning for each element. All plots in this article are normalized.

A critical issue in any dimensional reduction analysis is the selection of the reduced number of dimensions (r). Previous attempts to optimize selection of r in NMF analysis have focused primarily on the cophenetic correlation coefficient (CCC) (Brunet *et al.*, 2004; Pascual-Montano *et al.*, 2006a). We find that variation in the RSS between the data matrix ($\underline{V}$) and the NMF-estimated matrix ($\underline{WH}$) provides a natural means for estimation of r, in as much as this plot shows an inflection when r matches the proper number of dimensions. In practice, while an optimal NMF solution requires several hundred random restarts, the determination of r can be made with

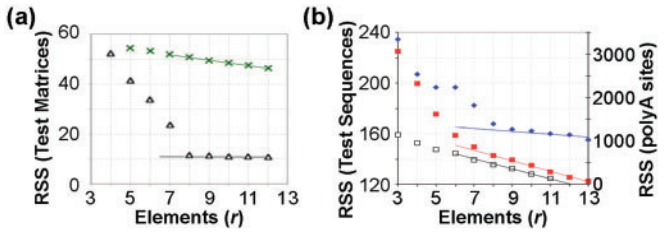

Fig. 1. Variation of the RSS with the number of elements (r) provides a robust estimate of the proper number of vectors. (**a**) Test matrices, with (black triangles) and without (green crosses) inserted patterns. Line plots represent a least squares linear fit of the last five points in each series. (**b**) *RSS* versus r plots for random sequences (open boxes), sequences with six inserted patterns (red squares) and human 3′-processing site flanking sequences (blue diamonds).

a significantly smaller number of solutions, typically 20–30 for tested each value of r (data not shown).

We first investigated variation of RSS with variation of r in two artificial datasets, with and without patterns inserted. The RSS of the purely random matrix shows a roughly linear decrease with increased number of elements (Fig. 1a), whereas the matrix with inserted patterns shows a clear inflection point where r equals the number of patterns. Our interpretation of this result is that while r is less than the number of patterns, there is excess variance in the data that cannot be adequately approximated by the NMF matrix. In contrast, as r equals and surpasses the true number of patterns, the additional reduction of the RSS is minor, since the variation captured by the additional elements is likely only random noise.

Further confirmation of this phenomenon was obtained through NMF analysis of three sequence sets, artificial sequences with and without inserted patterns and 8500 sequences with length 400 nt centered on putative human 3′-processing sites (Salisbury *et al.*, 2006). PWC matrices were generated from each sequence set, then subsequently analyzed via NMF, largely reproducing the results of the artificial matrices (Fig. 1b). The first set, with no patterns inserted, shows a roughly constant and linear decrease in RSS with increasing r. In contrast, the set with known motifs inserted shows an inflection point at approximately the number of inserted patterns. Finally, the analysis of putative 3′-processing sites also displays a disjoint pattern, with an inflection point at approximately $r = 9$. For all analysis in this article, we use the optimal value of r determined in this manner.

The complete NMF analysis requires the selection of a number of free parameters. Of particular note are the selection of window size (w) and k-mers size (k). Larger windows give increased statistical robustness afforded by aggregation but at the cost of less specific positioning. Longer words enable better characterization of long patterns, but at the cost of increased execution time as well as an increased number of sequences needed for training. In the ideal case of a large number of training sequences, k would be at least as long as the smallest expected motif, and $w = 1$, counting each position individually. However, datasets are finite and can number only a few tens to hundreds. Previous studies have provided estimates of the minimum number of sequences required for reasonable estimation of k-mer probabilities (Fairbrother *et al.*, 2002). In practice, we typically select w and k jointly such that the product wT is at least five times greater than 4^k, the number of k-mers under consideration.

2.4 Converting NMF word lists ($W_{.j}$) to motifs

As we reported previously (Graber *et al.*, 2007), we assume that the weighted list of k-mers for each element can be interpreted as the expected distribution of k-mer counts for each element. The optimal motif is inferred as the motif with the maximum multinomial probability of observing the expected distribution. The specifics of the probability calculation were described previously (Graber *et al.*, 2007). Briefly, the probability of observing any specific k-mer within a motif is the sum of the probabilities of observing the motif at all positions within the motif, including those that span the boundaries between the motif and the background. This summation specifically enables characterization of motifs of nearly arbitrary size regardless of the choice of k. To optimize the search of potential motifs, we use a Markov Chain Monte Carlo (MCMC) approach (Gelman *et al.*, 1995). The length of the motif is first sampled from a user-specified range and then nucleotide probabilities at each position are filled randomly. The multinomial distribution for the initial motif is scored against the expected distribution, and then the following procedure is used to update the motif. A specific k-mer (w_i) is sampled according to the distribution returned by the NMF analysis. Then a position j within the current motif is sampled according to the match of the k-mer within the motif. Finally, the motif is either made more or less similar to the k-mer at motif positions j to $j+k-1$, depending upon whether or not w_i is over- or underrepresented in the current model compared with the expected distribution. Following a standard MCMC approach, updated motifs are automatically accepted if the multinomial probability increases, and randomly accepted according to a Boltzmann distribution if the multinomial probability decreases (Gelman *et al.*, 1995). This process iterates until no improvement is observed for a user-specified number of iterations.

In the current implementation of the NMF pattern finder, we have made no attempt to optimize performance. A full analysis, including PWC creation, NMF decomposition and motif construction typically lasts tens to hundreds of minutes, depending on values of T, k, w and r. Preliminary tests incorporating the web-based bioNMF resource (Mejia-Roa *et al.*, 2008) showed significant time savings, while also confirming the analysis results.

2.5 Comparison with other pattern-finding approaches

NMF pattern identification was compared with several other approaches, including the Gibbs Sampler (Lawrence *et al.*, 1993), the Improbizer (Ao *et al.*, 2004), YMF (Sinha and Tompa, 2003), Weeder (Pavesi *et al.*, 2004) and oligo-analysis tools (van Helden, 2003). As with previous studies of pattern identification tools (Li and Tompa, 2006; Tompa *et al.*, 2005), an artificial dataset ($T = 300$) was used as the basis for comparison, specifically to remove uncertainty of the exact patterns present in the test sequence set. The Gibbs sampler and the Improbizer were set to find six nucleic acid motifs, searching on single strand only. The Gibbs sampler was allowed to vary motif size between 5 nt and 8 nt. The Improbizer automatically determines optimal motif size. Background frequency estimates (required for YMF, Weeder and oligo-analysis) were generated through analysis of the randomized sequences with no motifs inserted. Hexamers were counted for YMF and oligo-analysis, and the 'SMALL' setting was used in Weeder. All other settings were left as defaults.

3 RESULTS AND DISCUSSION

3.1 Characterizing the method and exploring parameters with artificial sequence sets

We previously used the NMF method to characterize sequences involved in controlling 3′-processing and *trans*-splicing in the nematode, *Caenorhabditis elegans* (Graber *et al.*, 2007). As described in that work, the complete NMF method is dependent upon several free parameters, including but not limited to, the size of the k-mers in the initial count (k), the position window size (w), the number of elements (r) and the number of training sequences available (T).

To investigate the effects of these parameters and specifically to test the robustness of the results with variation in the parameters, we generated an artificial dataset in which motifs with defined sequence content and positioning were inserted into sequences generated from a defined dinucleotide background. Six different motifs were created

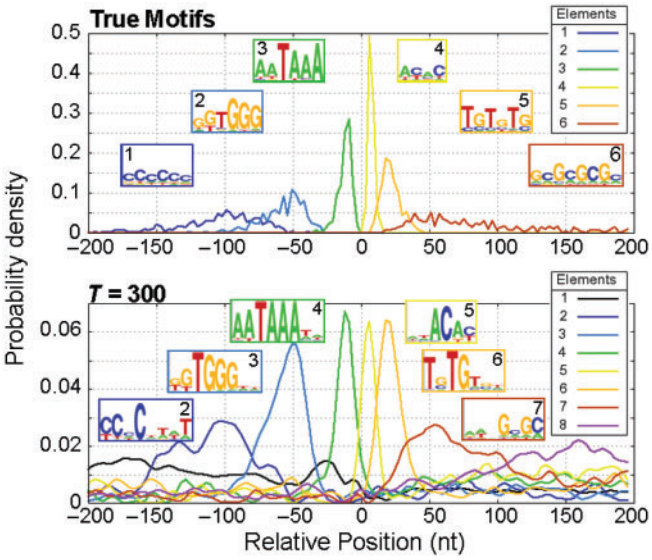

Fig. 2. An example comparing NMF analysis with the actual motifs used to generate the $T = 300$ test set. All sequence logos (Schneider and Stephens, 1990) in this and subsequent figures have a vertical scale of 2 bits and were generated with WebLogo (Crooks *et al.*, 2004). In this and subsequent figures, the colors of line plots and logo frames are matched in corresponding pairs for each element. Patterns that are interpreted as changes in background composition are not shown as sequence logos.

with varying levels of specificity in both positioning and sequence (Table 1). To reduce computational complexity, we varied only one parameter at a time, while holding all others fixed. Figure 2 shows an example analysis, comparing the actual and NMF-derived motifs for the $T = 300$ test set.

3.1.1 Variation in training set size (T) The number of training sequences (T) is of critical importance, since it is often out of the researcher's control, yet explicitly determines the statistical power to discern subtle signals. Training sets were generated of varying sizes ($T = 30$, 100, 300, 1000, 3000 and 10 000) and NMF analysis was performed with parameters fixed at $w = 3$, $k = 4$ (tetramers) and $r = 8$. Note that while six motifs were inserted, the best results are obtained with $r = 8$. As noted previously (Graber *et al.*, 2007), the NMF analysis identifies both specific motifs and changes in background composition along the length of the sequences. The additional two elements in the NMF decomposition compensate for the change in background at the center point of the test sequences.

The use of an artificial dataset allows explicit comparison of NMF-derived patterns with the motif models. The rows of the $\underline{W}$ matrix, which represent the frequency of occurrence for each k-mer in each motif, were compared with an exact count of the tetramers that were inserted as part of each motif (including those that span the motif-background boundary) using a Pearson's correlation. The best match of each NMF motif to an inserted motif is reported in Table 2(Panel A). Similarly, the columns of the $\underline{H}$ matrix were compared with the exact distribution of starting positions for each motif by Pearson's correlation, with the best match reported in Table 2(Panel B). The NMF approach successfully identifies the positioning for all motifs in the sets with $T \geq 300$ ($r \geq 0.80$), with nearly perfect reproduction for $T \geq 3000$ ($r \geq 0.88$).

Table 2. Improving performance of NMF pattern detection with increase in number of test sequences (T)

	Number of training sequences (T)					
Motif	30	100	300	1000	3000	10 000
Panel A Pattern sequence content						
1	0.14	0.23	0.62	0.90	0.90	0.96
2	0.55	0.67	0.88	0.96	0.98	0.98
3	0.91	0.97	0.99	0.99	0.99	0.99
4	0.37	0.82	0.93	0.95	0.98	0.97
5	0.77	0.85	0.94	0.97	0.98	0.96
6	0.35	0.50	0.70	0.73	0.89	0.91
Panel B Pattern positioning probability						
1	0.75	0.67	0.93	0.97	0.99	0.97
2	0.60	0.96	0.98	0.98	0.97	0.98
3	0.76	0.89	0.91	0.92	0.91	0.89
4	0.31	0.71	0.80	0.85	0.88	0.86
5	0.64	0.90	0.94	0.97	0.97	0.96
6	0.60	0.77	0.95	0.95	0.93	0.94

Table entries list the best-match Pearson's correlations between (Panel A) the exact k-mer counts generated by insertion of the patterns and rows of the NMF $\underline{W}$ matrix; and (Panel B) the positioning distribution used for pattern insertion and columns of the $\underline{H}$ matrix.

Even in the smallest datasets, the worst match (motif 4, $r = 0.31$) is still statistically significant. The sequence content determination varies in a similar fashion. The strongest motif, with consensus AATAAA, is unambiguously matched in all datasets ($r \geq 0.9$). As with positioning, the tetramer content of all motifs are well-matched for all sets with $T \geq 300$ ($r \geq 0.62$). It is interesting to note that motif 2, which has identical sequence information content to motif 3, but a less specific positioning distribution, is less well identified, indicating that both the sequence and positioning specificity can affect our ability to successfully characterize motifs. Full output from these analyses are available in Supplementary Figure 2.

3.1.2 Variation in other parameters and conditions As described in Section 2, the choice of word size (k) and window size (w) are compromises between conflicting needs. Tests on the $T = 300$ dataset with variations of w and k show that the identified motifs are robust with parameter variation. Separate analyses were also performed on datasets with patterns inserted into either half or a quarter of the sequences in the training set. Motifs 3–5 (Table 1) were clearly identified in both sets, motif 2 was identified the 50% set, while the weakest patterns (1 and 6) were poorly matched at best in both sets. Details of these tests are available in Supplementary Figures 3–5.

3.1.3 Comparison with existing tools To further assess the utility of the NMF approach, we compared it with a number of popular pattern recognition tools, focusing on the $T = 300$ test set. The disparate nature of the results returned from these tools complicates comparison of the specific results, therefore we present a summary (Table 3) that focuses primarily on whether or not a reasonable match to each pattern was found. For tools that produce a PWM representation, the match is reported as the sum of the Euclidean distances between columns in the known and inferred motifs (seq d in Table 3) (Gupta *et al.*, 2007), aligned as to produce the

Table 3. Comparison of NMF with other pattern finders on the six motif, $T = 300$ sequence set

	NMF best match			Gibbs sampler			The improbizer			Weeder		Oligo analysis		YMF	
Motif	*L*	seq *d*	pos *r*	*L*	seq *d*	pos *r*	*L*	seq *d*	pos *r*	*L*	seq *d*	*L*	consensus	*L*	consensus
1	8	1.37	0.94	6	0.69	0.91	28	0.94	0.62			6	CCCCCC	6	CCCCCC
2	8	1.1	0.98	6	0.4	0.96	49	0.17	0.8	6	0.24	10	TGGTGGGTAA		
3	8	0.3	0.95									8	AATAAACA		
4	8	0.37	0.79									6	CATACC	6	CACASA
5	8	0.86	0.94	5	1.09	0.83	8	0.48	0.81			8	ATGTGCTA	6	CGYGYG
6	8	2.86	0.94	6	0.73	0.95	18	0.57	0.89	8	1	7	CGCGCGC	6	GCGCGC

minimum distance. For tools that produce positioning data or models, the match is reported as the Pearson's correlation between the actual and inferred positioning distributions. Best match consensus patterns are reported for pattern-based tools. The NMF approach identifies significant matches in both positioning and sequence content for all six inserted motifs in our test set. The best alternative method was the oligo analysis tools, which return patterns that reasonably match the consensus patterns of all six inserted motifs, however, only consensus sequences are returned, with no distinct positioning information. No other tool produces matches to more than four of the inserted motifs. Further comparisons are available in Supplementary Materials.

The principal benefits of the NMF approach come from the focus on specific positioning of motifs relative to functional site. Nearly all pattern recognition tools focus on the identification of patterns that occur more frequently than expected according to a background model that is generated either from the input sequences themselves, or from a known background (typically putative promoter regions reflecting the bias of these studies towards transcription factor binding site identification). The test sequences used here were generated from a relatively AT-rich background. Motifs 3 and 4 (Table 1) do not differ significantly from this background. This is reflected in the poor rate at which these motifs were identified by the alternative tools tested here (Table 3). From these results, we conclude that, for the specific problem of identifying signals with positioning constraint, NMF outperforms alternative approaches.

3.2 Analysis of biologically interesting sequence sets

3.2.1 3′-processing (cleavage and polyadenylation) sites Thirty years of study have demonstrated that the sequences that control 3′-processing in eukaryotic organisms show very distinct positioning about the 3′-processing site. One recent study postulated up to 15 functional elements in the human 3′-processing signal (Hu *et al.*, 2005). Based on our analysis of mouse (Fig. 3), and several other mammals (Supplementary Fig. 7), we find five or possibly six distinct elements based on positioning specificity. We observe clear evidence for two upstream elements: the UGUA-like element and the canonical AAUAAA-like hexamer, shown as mouse elements 2 and 3, respectively in Figure 3. Downstream, we find three elements that can be described as UG-rich (element 5), U-rich (element 6) and G-rich (element 7). In addition, as with previous studies, we find some evidence for a putative upstream U-rich element, evidenced by elements 4 and 6. Elements 1 and 8 are representative of changes

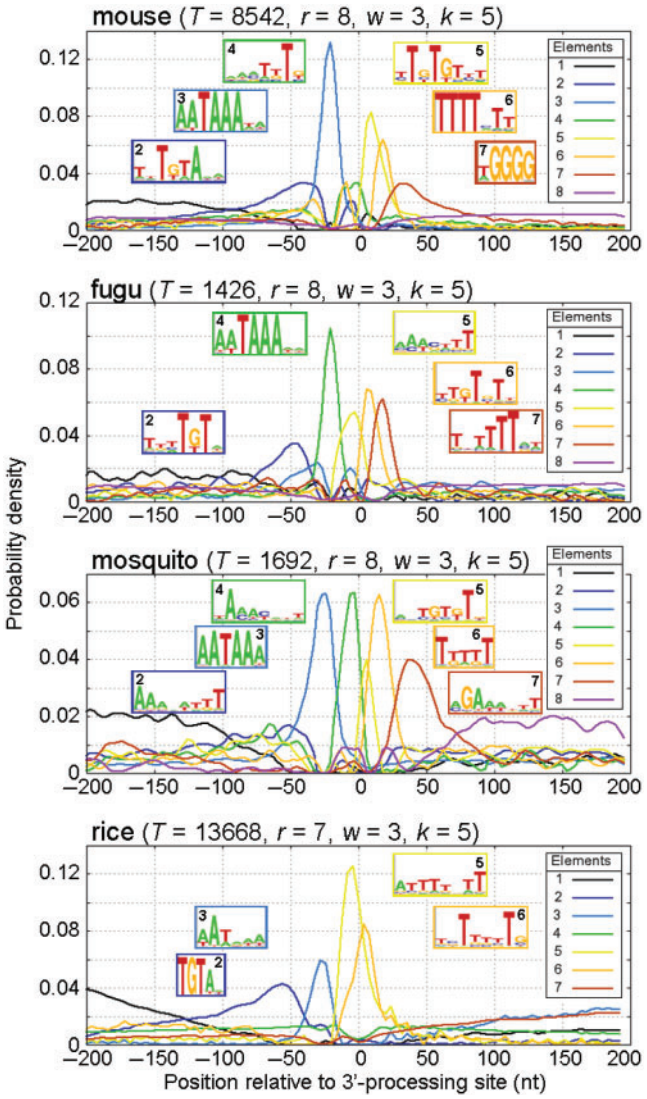

Fig. 3. NMF analysis of the sequence elements that specify 3′-processing sites for a variety of eukaryotic organisms reveals a common arrangement of signals. Plots are shown for mouse (*M.musculus*), fugu (*T.rubripes*), mosquito (*A.gambiae*) and rice (*O.sativa*). NMF patterns that are interpreted as changes in background composition are not shown as sequence logos.

in the background composition in the transition from CDS to UTR to intergenic region.

Previous comparative studies of 3′-processing sites have shown both common and distinct features (Lee *et al.*, 2007; Salisbury *et al.*, 2006) between organisms. We closely examined a number of organisms, and show the results for a phylogenetically broad sampling here (Fig. 3), including mouse, fugu, mosquito and rice. (Analyses of several additional organisms are available in Supplementary Fig. 7). Somewhat strikingly, our inclusion of detailed analysis of positioning constraints reveals a conserved core pattern in essentially all higher eukaryotes, consisting of the upstream UGUA (Venkataraman *et al.*, 2005), polyadenylation signal, with canonical consensus AAUAAA (Proudfoot and Brownlee, 1976), and downstream U-rich elements (Gil and Proudfoot, 1987). While it seems apparent that the balance in importance (and conservation) between these elements has varied, the basic pattern of these elements remains largely the same in both sequence content and positioning (Fig. 3).

We can also use our analysis to highlight where the 3′-processing control sequences diverge. As we previously reported (Salisbury *et al.*, 2006), the downstream UG-rich element is specific to metazoans other than nematode. The organisms that have incorporated a UG-rich element also appear to have a corresponding shift in the positioning of the downstream U-rich element. The downstream G-rich element (mouse element 8 in Fig. 3) appears to be specific to amniotes (Salisbury *et al.*, 2006). Our analysis of invertebrate genomic sequences (mosquito, fruit fly and nematode) indicates the presence of an additional downstream A-rich element (Fig. 3, mosquito element 7) with positioning similar to the mammalian G-rich sequences.

Finally, it is informative to contrast our results with the previous large-scale analysis of human 3′-processing sites (Hu *et al.*, 2005). As stated above, the authors of that study identified 15 distinct motifs determining 3′-processing site placement, whereas we find five or six. The previous study used a less specific positioning classification than ours, initially identifying putative functional sequence words on the basis of their overrepresentation in one of four positioning regions, spanning positions −100 to −40, −40 to 0, 0 to +40 and +40 to +100 relative to the 3′-processing site. These words were then subsequently clustered on the basis of sequence similarity, and positional weight matrix representations were generated from the clusters. A number of their final elements, however, show similar positioning distributions, and sequence patterns that are arguably close enough to represent variants of a single, more general, pattern [e.g. Fig. 4 in Hu *et al.* (2005)]. In contrast, our approach first groups sequence words based on the similarity of positioning at a significantly finer resolution (typically windows of 3–5 nt). In addition, we benefit from the 'fuzzy clustering' nature of NMF, specifically in that each sequence word can contribute to the modeling of multiple elements in a weighted manner. Based on these differences, we believe that our analysis generates a more general, inclusive picture of the constraints of the elements of the 3′-processing site control sequence. We nonetheless leave open the possibility that the subdivision of these elements, such as generated previously (Hu *et al.*, 2005), may represent valid subclassifications of elements.

3.2.2 Fruitfly transcription start sites The elements that comprise the core promoter in the fruit fly (*D.melanogaster*) have been the subject of several recent studies (Gershenzon *et al.*, 2006; Ohler *et al.*, 2002). Many of the identified elements displayed positioning specificity, indicating an appropriate problem for NMF. We analyzed 2561 core promoter sequences (Gershenzon *et al.*, 2006) extending from 250-nt upstream to 100-nt downstream of the transcription start site. Our analysis (Supplementary Fig. 8) identifies a number of patterns that are consistent with the previous studies, while also highlighting possible antagonistic relationships.

4 CONCLUSIONS AND FUTURE DIRECTIONS

We have described here a novel approach to the characterization of putative regulatory motifs, using NMF to simultaneously determine both positioning and sequence content. In contrast with other dimension reduction algorithms, the NMF decomposition produces component matrices with direct intuitive interpretations, reflecting the positioning and sequence content of the resulting elements. We also demonstrated that variation of the RSS between the actual and reduced data provides a means of estimating the proper number of elements. Finally, in contrast with other pattern detection algorithms, our analysis is explicitly geared to the detection of motifs with constrained positioning, and consequently outperforms these algorithms for this specific problem.

The motifs generated by the NMF approach are well suited for inclusion in probabilistic predictive models, such as those for transcription start site, splice site, 3′-processing site, or full gene prediction. To this end, we are working to cast the NMF output into a form that can directly be incorporated with the open-source Genie software (Kulp *et al.*, 1997; Reese *et al.*, 1997). We discuss a number of potential improvements to the NMF in Supplementary Materials.

The NMF approach provides a robust and novel approach to characterization of putative regulatory elements with positioning specificity. Given the importance of this additional constraint, our approach will provide benefits to analyses beyond the few examples presented here.

ACKNOWLEDGMENTS

The authors thank Yong Woo, Gary Churchill, Elissa Chesler and members of the Graber Group for helpful comments and critiques. The authors thank Michael Brockman, Carol Bult, Hyuna Yang and Joel Richardson for critical review of the paper.

Funding: NSF 2010 Project (grant DBI-0331497); NIH/NCRR INBRE Maine (grant 2 P20 RR16463); NIH/NIGMS (grant 1R01GM072706).

Conflict of Interest: none declared.

REFERENCES

Ao,W. *et al*. (2004) Environmentally induced foregut remodeling by pha-4/foxa and daf-12/nhr. *Science*, **305**, 1743–1746.

Brockman,J.M. *et al*. (2005) Pacdb: polya cleavage site and 3′-utr database. *Bioinformatics*, **21**, 3691–3693.

Brunet,J.P. *et al*. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Carmona-Saez,P. *et al*. (2006) Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, **7**, 78.

Crooks,G.E. *et al*. (2004) Weblogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Fairbrother,W.G. *et al*. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.

Fairbrother,W.G. *et al*. (2004) Rescue-ese identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.

Gelman,A. *et al*. (1995) *Bayesian Data Analysis*. *Texts in Statistical Science*. Chapman & Hall/CRC.

Gershenzon,N.I. *et al*. (2006) The features of drosophila core promoters revealed by statistical analysis. *BMC Genomics*, **7**, 161.

Gil,A. and Proudfoot,N. (1987) Position-dependent sequence elements downstream of aauaaa are required for efficient rabbit b-globin mnra 3 end formation. *Cell*, **49**, 399–406.

Graber,J.H. *et al*. (2007) C. elegans sequences that control trans-splicing and operon pre-mrna processing. *Rna*, **13**, 1409–1426.

Gupta,S. *et al*. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

Hu,J. *et al*. (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *Rna*, **11**, 1485–1493.

Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.

Kulp,D. *et al*. (1997) Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.*, 232–244.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Lee,J.Y. *et al*. (2007) Polya db 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, **35**, D165–D168.

Li,N. and Tompa,M. (2006) Analysis of computational approaches for motif discovery. *Algorithms Mol. Biol.*, **1**, 8.

Loke,J.C. *et al*. (2005) Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol.*, **138**, 1457–1468.

Mejia-Roa,E. *et al*. (2008) Bionmf: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Res.*, **36**, W523–W528.

Ohler,U. *et al*. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, RESEARCH0087.

Pascual-Montano,A. *et al*. (2006a) Bionmf: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics*, **7**, 366.

Pascual-Montano,A. *et al*. (2006b) Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 403–415.

Pavesi,G. *et al*. (2004) Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.

Proudfoot,N. and Brownlee,G. (1976) 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature*, **263**, 211–314.

Reese,M.G. *et al*. (1997) Improved splice site detection in genie. *J. Comput. Biol.*, **4**, 311–323.

Salisbury,J. *et al*. (2006) A multispecies comparison of the metazoan 3′-processing downstream elements and the cstf-64 RNA recognition motif. *BMC Genomics*, **7**, 55.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Sinha,S. and Tompa,M. (2003) Ymf: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.

Vardhanabhuti,S. *et al*. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.*, **35**, 3203–3213.

Venkataraman,K. *et al*. (2005) Analysis of a noncanonical poly(a) site reveals a tripartite mechanism for vertebrate poly(a) site recognition. *Genes Dev.*, **19**, 1315–1327.

Zhao,J. *et al*. (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mrna synthesis. *Microbiol. Mol. Biol. Rev.*, **63**, 405–445.