BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 1 2005, pages 31–38
doi:10.1093/bioinformatics/bth471

DWE: Discriminating Word Enumerator

Pavel Sumazin[1,2,*], Gengxin Chen[1], Naoya Hata[1], Andrew D. Smith[1], Theresa Zhang[3] and Michael Q. Zhang[1,*]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [2]Computer Science Department, Portland State University, P.O. Box 751, Portland, OR 97207, USA and [3]Bioinformatics, Merck Research Laboratories, Rahway, NJ 07065, USA

Received on September 11, 2003; revised on July 26, 2004; accepted on August 7, 2004
Advance Access publication August 27, 2004

ABSTRACT

Motivation: Tissue-specific transcription factor binding sites give insight into tissue-specific transcription regulation.

Results: We describe a word-counting-based tool for de novo tissue-specific transcription factor binding site discovery using expression information in addition to sequence information. We incorporate tissue-specific gene expression through gene classification to positive expression and repressed expression. We present a direct statistical approach to find overrepresented transcription factor binding sites in a foreground promoter sequence set against a background promoter sequence set. Our approach naturally extends to synergistic transcription factor binding site search.

We find putative transcription factor binding sites that are overrepresented in the proximal promoters of liver-specific genes relative to proximal promoters of liver-independent genes. Our results indicate that binding sites for hepatocyte nuclear factors (especially HNF-1 and HNF-4) and CCAAT/enhancer-binding protein (C/EBPβ) are the most overrepresented in proximal promoters of liver-specific genes. Our results suggest that HNF-4 has strong synergistic relationships with HNF-1, HNF-4 and HNF-3β and with C/EBPβ.

Availability: Programs are available for use over the Web at http://rulai.cshl.edu/tools/dwe

Contact: ps@cs.pdx.edu; mzhang@cshl.edu

Supplementary information: Data and omitted results are available at http://rulai.cshl.edu/tools/dwe/supp

INTRODUCTION

One of the main goals of modern genetics is to decipher the mechanisms of gene expression and regulation. Recent years have seen the generation of a significant volume of data that will help to probe expression mechanisms. Microarray techniques and chromatin immunoprecipitation (ChIP) techniques allow for genome-scale investigation of gene expression and DNA-binding protein localization. These techniques can be used to classify expression by cell environment and transcription factor binding.

Completed or nearly completed genome sequences are publicly available for a growing number of vertebrate species including human, mouse, rat and chicken. Increasingly accurate methods for detecting transcription start sites (TSSs), such as Davuluri *et al*. (2001) and Scherf *et al*. (2000), enable localization of promoter regions. Coupled together, sequence information and TSS location can be used to identify proximal promoter sequences. Proximal promoter sequences have already been well identified for a large number of genes in human, mouse and rat.

We are interested in methods that combine gene expression and sequence information for de novo discovery of transcription factor binding sites (TFBSs) in proximal promoters of co-expressed tissue-specific genes. The annotation of proximal promoters for such genes will advance the understanding of tissue-specific transcription regulation.

We describe a discriminant word counting algorithm, Discriminant Word Enumerator (DWE), which can be used to discover motifs in promoters of co-regulated genes. We use DWE to find overrepresented gapped degenerate words (motifs) in proximal promoters of liver-specific genes taken from Liver-Specific Promoter Database (LSPD) (Zhang and Zhang, 2000, http://cgsigma.cshl.org/LSPD) against vertebrate promoters from the Eukaryotic Promoter Database (EPD), release 78 (Perier *et al*., 1998). We use TSS data from DBTSS (Suzuki *et al*., 2002) and sequence data from GenBank to collect promoter sequences.

Related literature

Classical sequence-based motif discovery algorithms include CONSENSUS (Hertz *et al*., 1990), MEME (Bailey and Elkan, 1995) and the Gibbs sampler (Lawrence *et al*., 1993; Liu *et al*., 1995). Other motif discovery algorithms that use word-counting methods are reported previously (Van Helden *et al*., 1998, 2000; Sinha and Tompa, 2002). Recent motif search algorithms that use sequence and microarray data from expression or ChIP analysis include

*To whom correspondence should be addressed.

REDUCE (Bussemaker *et al*., 2001), MDscan (Liu *et al*., 2002), DMOTIFS (Sinha, 2003) and YMF (Sinha and Tompa, 2000, 2002; Blanchette and Sinha, 2001). REDUCE relates motif occurrence counts to gene expression ratio; MDscan iteratively constructs matrix representations of TFBSs that are overrepresented in the foreground set against a Markov background model that can be estimated from a background sequence set; DMOTIFS searches for overrepresented motifs in a foreground set against a background set while maintaining a maximum count per sequence; YMF searches for overrepresented motifs in a foreground set against a third-order Markov model estimated from a background sequence set. Beer and Tavazoie (2004) describe a method for predicting expression from TFBSs abundance; this method could be extended to include motifs found by DWE. We extend recent work which uses a P-value statistic to search for overrepresented ungapped motifs of length 7 in *Saccharomyces cerevisiae* promoters.

SYSTEMS AND METHODS

We searched for overrepresented motifs in a set of non-orthologous proximal promoters of genes that are known to have high expression in liver. We also searched for motifs in the consensus sequences of these proximal promoters. We measured the overrepresentation of motifs in these sets against the set of all vertebrate proximal promoters in EPD78, and the set of EPD78 vertebrate proximal promoters whose corresponding genes are not known to be strongly expressed in liver. We report the most overrepresented motifs in these comparisons, and infer the transcription factors most likely to bind to the corresponding TFBSs.

Statistical evaluation

We use three methods to evaluate the significance of motif overrepresentation.

P-value. The fixed marginal contingency table P-value follows the multiple hypergeometric distribution given in Equation (1) for a review see Agresti (1992). The P-value for the table is the sum of the probabilities of all tables that are at least as extreme. In this application we set a P-value for the overrepresentation of a motif in the foreground set against the background set, so that N_f and N_b are the potential occurrences in the foreground and background sets (trials), and n_f, n_b are the number of observed occurrences in the respective sets (successes).

$$P = \frac{\binom{N_f}{n_f}\binom{N_b}{n_b}}{\binom{N_f+N_b}{n_f+n_b}}. \tag{1}$$

Z-test. The Z-test (Student, 1908) is represented by the following Equation (2).

$$Z = \frac{\frac{n_f}{N_f}\frac{n_b}{N_b}}{\frac{n_f+n_b}{N_f+N_b}\left(1-\frac{n_f+n_b}{N_f+N_b}\right)\left(\frac{1}{N_f}+\frac{1}{N_b}\right)} \tag{2}$$

Log frequency ratio. The log frequency ratio (LFR) is as follows:

$$\mathrm{LFR} = \ln \frac{n_f N_b}{n_b N_f}. \tag{3}$$

From TFBS to transcription factor

We searched through TRANSFAC (Knuppel *et al*., 1994) for position frequency matrices (PFMs) that match the motifs found by DWE and PFMs found by MDscan. Transcription factors that are known to bind to the TRANSFAC PFMs are likely to bind to the matching DWE motifs and MDscan PFMs. To facilitate the search, we converted consensus-based motifs into PFMs using the maximum entropy principle of Jaynes (1957a,b); each IUPAC symbol was converted into a maximum-entropy column with total count equal to the number of foreground occurrences n_f. For example, $\mathrm{M} = \{\mathrm{A}, \mathrm{C}\}$ was converted into $[n_f/2, n_f/2, 0, 0]^{\mathrm{T}}$ and $\mathrm{D} = \{\mathrm{A}, \mathrm{G}, \mathrm{T}\}$ was converted into $[n_f/3, 0, n_f/3, n_f/3]^{\mathrm{T}}$. We used a χ^2 test to compare discovered-motif PFMs to TRANSFAC PFMs following the methodology proposed by Schones *et al*. (2004); PFMs are iid observations from a product multinomial distribution and were compared column by column, with the smaller PFM compared at each possible position to a submatrix of the larger PFM and the best match reported. PFMs were said to match when the normalized probability that they are occurrences from the same product-multinomial distribution was better than 0.05.

Dataset and consensus set

We selected LSPD genes that have at least one known ortholog, a known TSS, and sequence information covering the $[-299, 100]$ region relative to the TSS. With the objective of collecting promoters with known sequence information covering the $[-499, 100]$ region relative to the TSS, we selected a longest promoter from each set of orthologs, breaking ties arbitrarily. The resulting Liver-Specific Promoter Subset (LSPS) includes 35 promoters with mean length 549. In contrast, the vertebrate promoter subset of EPD78 includes 2380 promoters with average length 579, and the promoter subset of liver expressed genes in EPD78 includes 103 promoters with average length 558. LSPS includes four promoters that are subsequences or orthologs of Krivan and Wasserman (2001) promoters, including RATAADC01, HUMVITDBP, MMILGF and HUMGLUT201. Promoters of selected LSPD genes, LSPS, mapping from LSPS to EPD78 and mapping from promoters of liver expressed genes in EPD78 to LSPS are provided in the Supplementary information.

We generated a consensus sequence for each ortholog set, and used those consensus sequences to check for the conservation of motifs found in LSPS. To generate a consensus sequence, we first aligned orthologs using CLUSTALW (Thompson *et al*., 1994) with default parameters. We selected a consensus element for each aligned position according to the following procedure. Collect the set of nucleotides

that appear at least twice at this position across the aligned sequences; if any of the sequences contains a gap at this position or if the nucleotide set is empty output a '-', otherwise output an IUPAC symbol that corresponds to the collected nucleotide set. To measure conservation, we report the number of occurrences of each discovered motif and motif pair in the consensus set.

We searched for overrepresented motifs in the consensus set against vertebrate promoters in EPD78 (Table 3). To accommodate for motif discovery programs, which do not accept degenerate nucleotide input, we modified the consensus generation procedure to output the majority nucleotide in a column (and a '-' in case of a tie) instead of a degenerate IUPAC symbol. The modified consensus sequence set has four sequences that are different from the original. Both consensus sequence sets are provided in the Supplementary information.

ALGORITHM

Given a motif structure, including motif length, gaps and maximum number of degenerate positions, we enumerate all matching motifs using a method similar to that of Waterman *et al*. (1984). Each non-degenerate motif is mapped to an integer by stripping away gaps and converting the resulting word of length ℓ over alphabet of size 4 into an integer ranging from 0 to $4^{\ell+1}-1$. Each motif position and integer representation are recorded, and the operation is repeated for the reverse complement if so specified. Position information is compiled for each permitted degenerate word. The representation of each word and each degenerate word in the foreground is compared with its representation in the background, and the words with foreground overrepresentation above threshold are reported. DWE disregards substrings with characters other than the case insensitive A, C, G, T in the background and foreground sequence sets.

Thresholds are set for *P*-values, LFRs and *Z*-values as described in the Systems and methods section. Comparison conditions such as self-overlap, counting method and motif independence are user specified. When self-overlap is disallowed, the number of potential occurrences (trials) in each sequence set will be set to the maximum number of non-overlapping occurrences. The counting method can be set to word counting or sequence counting. The former refers to counting occurrences independently of their distribution across sequences, and the latter refers to counting sequences that contain at least one motif occurrence. When motif independence is not required, DWE reports all overrepresented motifs above the specified threshold. Such reporting may include similar words that have related sets of occurrences. For example, occurrence sets for degenerate words `CTNTGD` and `CTVTGD` will have a large intersection. When motif independence is required, we use the χ^2-test suggested by Schones *et al*. (2004) to suppress the reporting of lower-quality-dependent words.

Finding synergistic motifs

Given a list of IUPAC motifs and an integer k, DWE will search for motif k-tuples that occur in the same sequences and are overrepresented in the foreground. In the case that overlap is not allowed, the counting procedure is more intricate. When sequence counting is used, the number of trials (potential number of occurrences for a tuple in a promoter set) is the number of sequences in that set, and the number of successes (occurrences of that tuple) is the number of sequences containing at least one set of non-overlapping occurrences of each $x \in X_k$. When word counting is used, the number of trials for a motif k-tuple X_k is given in Equation (4), where $S = \{s\}$ is the set of sequences and $|s|$ is the length of s. We calculate the number of successes for each tuple using a recursion on k. For $k = 2$, the number of successes for $X_2 = \{x_1, x_2\}$ over S is $\sum_{s\in S} x_1^{(s)} x_2^{(s)} - \mathrm{O}(X_2)$, where $\mathrm{O}(X_2)$ is the number of overlapping occurrences of x_1 and x_2, and $x^{(s)}$ is the number of occurrences of x in s. For $k > 2$, the number of overlapping occurrences $\mathrm{O}(X_k) = \sum_{s\in S} \mathrm{O}(X_k, s)$ is given in Equation (5), where $\mathrm{L}(X_k, s)$ is the number of distinct motif k-tuple occurrences that share at least one position in s. The total running time is in the order of $|S| + k \log k\, \mathrm{O}(X_k)$.

$$\mathrm{Trials}(X_k) = \sum_{s\in S} \binom{|s| - \sum_{x\in X_k} (|x| - 1)}{k} \tag{4}$$

$$\mathrm{O}(X_k) = \sum_{s\in S} \Big[-k\mathrm{L}(X_k, s) + \sum_{X_{k-1}\subset X_k} \sum_{x\notin X_{k-1}} x^{(s)} \mathrm{O}(X_{k-1}, s) \Big]. \tag{5}$$

EXPERIMENTS

We used DWE and MDscan to find the most overrepresented motifs in LSPS against EPD. We did not use REDUCE because it is less suitable for discriminating against a background set. Our results on synthetic data suggest that YMF does not perform as well as DWE or MDscan when searching for overrepresented motifs in a foreground set against a background set.

Performance on synthetic sequence data

The sensitivity of motif finding algorithms depends on the total size of the sequence set, motif width and motif degeneracy. We tested the algorithms on synthetic data with dimensions similar to those of LSPS. Foreground and background sets were made of 35 sequences of length 550. We implanted motifs of increasing number and degeneracy in the foreground sets and measured the ability of each algorithm to detect these motifs against background sets. Background sets and non-motif elements in the foreground sets were generated from a background vector with 60% CG. Motifs were generated from position weight matrices (PWMs) that correspond to

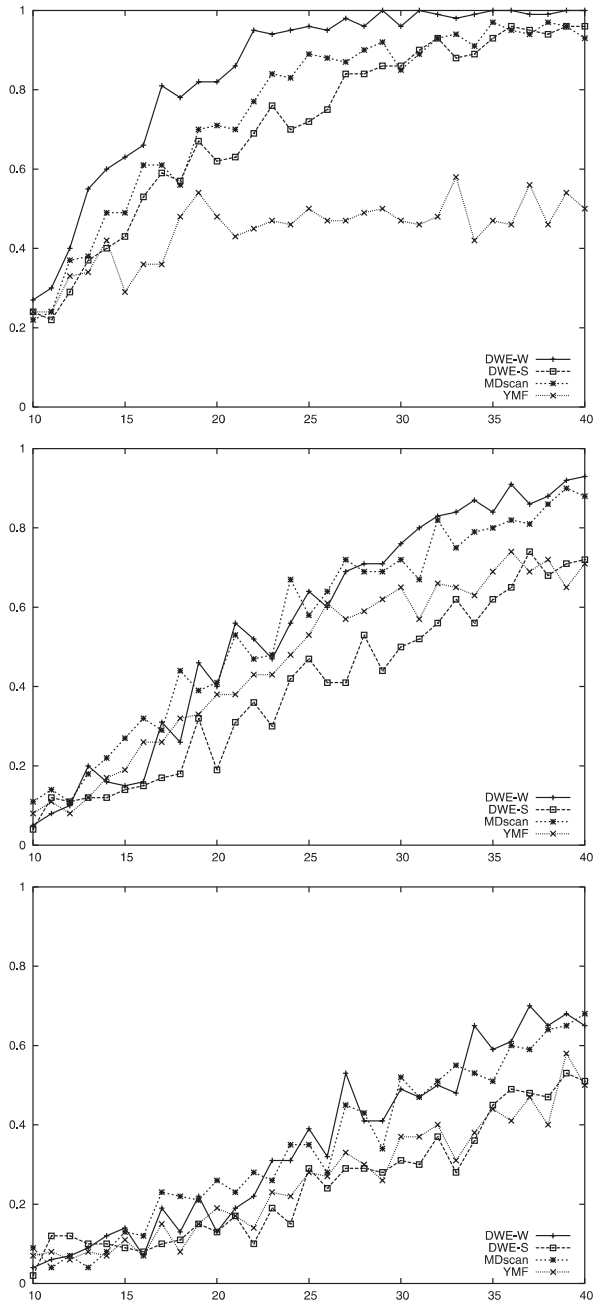

Fig. 1. Detection-quality comparison of DWE, MDscan and YMF when attempting to discover an implanted motif with width six against a vector-generated background sequence set. We plot the frequency (from 0 to 1) of the correct detection in the top five found motifs for each method as a function of the number of implanted motifs (from 10 to 40). Foreground and background sets contained 35 sequences of length 550; motifs are implanted uniformly at random across the set; each data point corresponds to 100 runs of the corresponding algorithm; and DWE-W counts the number of motif occurrences in each set and DWE-S counts the number of sequences containing the motif. We report results for implanted motifs with no degenerate positions (top), one degenerate position (middle) and two degenerate positions (bottom).

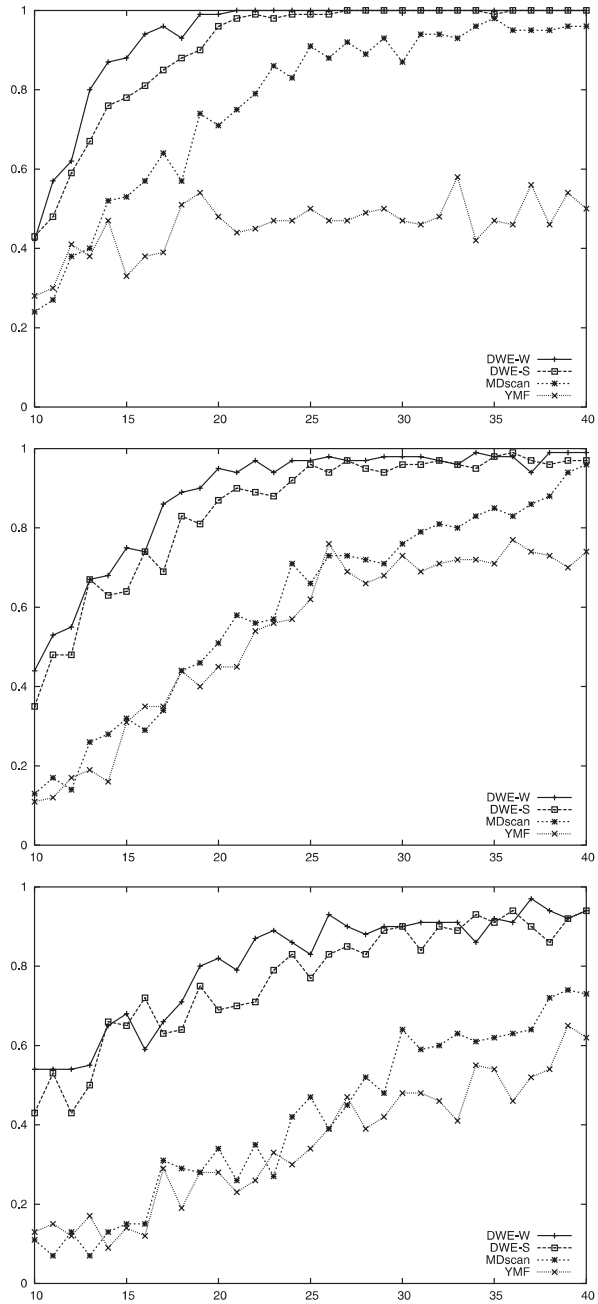

Fig. 2. Detection-quality comparison of DWE, MDscan and YMF when attempting to discover an implanted motif with width six against an augmented background sequence set that is created by adding 35 additional sequences that do not contain the motif to the background set used in the experiments reported in Figure 1. We plot the frequency (from 0 to 1) of the correct detection in the top five found motifs for each method as a function of the number of implanted motifs (from 10 to 40). Each data point corresponds to 100 runs of the corresponding algorithm; DWE-W counts the number of motif occurrences in each set and DWE-S counts the number of sequences containing the motif. We report results for motifs with no degenerate positions (top), one degenerate position (middle) and two degenerate positions (bottom).

uniformly selected IUPAC words with specified number of degenerate positions.

We constructed foreground sets with 10–40 uniformly-at-random implanted occurrences of motifs with width six, and 0, 1 and 2 degenerate positions. For each motif type and motif number, new foreground and background sets were constructed and the experiment was repeated 100 times. We selected the top five motifs found by DWE when counting motif occurrences (denoted by DWE-W), DWE when counting the number of sequences containing the motif (denoted by DWE-S), MDscan and YMF. We did not remove dependences between the motifs found by the algorithms, potentially allowing for similar motifs in the top-5 set. We report the proportion of trials where the implanted motif matched a top-5 motif. When matching motifs, we matched a degenerate element using all of the nucleotides it represents. Our results suggest that DWE outperforms MDscan on non-degenerate motifs, MDscan outperforms DWE on degenerate motifs, and YMF performs worse than DWE and MDscan (Fig. 1).

We tested the ability of the algorithms to discover implanted motifs that are strongly underrepresented in the background set. We augmented the randomly constructed background sets in our initial experiments with 35 additional sequences of length 550 that do not include any occurrences of the implanted motif. The detection quality of the algorithms when using the augmented background sets is reported in Figure 2. The performance of DWE improved dramatically, while the performance of MDscan and the performance YMF did not improve substantially.

Liver-Specific Promoter Database

We used DWE to discover motifs that are overrepresented in LSPS against the vertebrate promoter subset of EPD78 (Table 1), and against that set excluding promoters of liver-expressed genes (Table 2). We searched for (3+gap+3)mers and (4+gap+4)mers, with rigid gaps ranging from 0 to 7 bp and at the most two degenerate positions. We also searched for the motifs that are overrepresented in the consensus set against the vertebrate promoter set from EPD78 (Table 3). We repeated these searches using MDscan and report the top 3 motifs of lengths 6, 8 and 10 in each experiment; (Tables 4–6).

Initially, MDscan reported poly(A) and alternating C–T motifs. These motif are found to be strongly overrepresented by DWE when motif autocorrelation is not considered. However, the number of occurrences of these motifs decreases substantially when selfoverlap is not permitted, and they are not reported in the top 50. In order to use MDscan more effectively, we masked all substrings that correspond to cycles of periods 1 and 2 and length 8 or greater. The results by MDscan still differ substantially from the results of DWE, but both identify binding sites that are similar to known binding sites for hepatocyte nuclear factors HNF-4 and HNF-1.

Table 1. Motifs that are strongly overrepresented (by occurrence count) in promoters of liver-expressed genes (LSPS) against promoters of liver-expression independent genes (EPD)

Motif	FO	BO	L	P	TTF	C
A•A••T•A	230	7843	2.2	8.4E−27	HNF-4	81
A•T••••A•A	213	8012	2.0	2.9E−20	C/EBPβ	81
T•CA•A	233	9239	1.9	8.8E-19	C/EBP	79
CAA•••T	189	6973	2.0	2.5E−18	HNF-4	74
TAA••HA	149	4950	2.2	2.7E-18	HNF−3β	54
T•T•AA	208	8199	1.9	3.4E−17		73
ACA•ADD	154	5364	2.1	3.5E−17	SRF	44
AT•AA	188	7220	1.9	1.1E-16	HNF−1	90
A•A•AG	254	11078	1.7	1.3E−15	HNF-4	80
CT•TG	284	12933	1.6	3.6E-15	HNF−4	84

FO (foreground occurrences) is the number of occurrences in LSPS; BO (background occurrences) is the number of occurrences in EPD promoters; L stands for LFR; P is the *P*-value; TTF (TRANSFAC transcription factor) is the transcription factor whose binding site PFM in TRANSFAC best matches the motif; and C (conservation) is the number of occurrences of the motif in the consensus set that is generated from an alignment of LSPS promoters with their orthologs.

Table 2. Motifs that are strongly overrepresented (by occurrence count) in LSPS against EPD vertebrate promoters of genes that are not known to be expressed in liver

Motif	FO	BO	L	P	TTF	C
A•A••T•A	230	7271	2.3	1.2E−28	HNF-4	81
TD••TTA	147	4492	2.3	1.2E−19	qa-1F	54
CAA•••T	189	6528	2.1	1.6E−19	HNF-4	74
THT••T•A	168	5781	2.1	1.2E−17	HNF-3β	48
AT••••CA	162	5521	2.1	1.9E−17	C/EBPβ	57
DT•••AAA	161	6073	1.9	1.1E−13	C/EBPβ	62
AAG•••T	191	7808	1.7	7.8E−13	HNF-4	76
HAT••AG	124	4590	1.9	2.9E−11	POU2F1	39
AKTAACCH	16	112	10.2	3.8E−11	HNF-1	6
A•A••G•T	160	6683	1.7	2.5E−10	HSF1	52

See Table 1 for a complete legend.

Table 3. Motifs that are overrepresented (by occurrence count) in the consensus set against EPD vertebrate promoters

Motif	FO	BO	L	Z	TTF
GTTAAT	9	323	1.0	8.8	HNF-1
TAAT•ATTR	6	72	3.0	5.5	POU1F1
TMCTGGAA	4	47	3.1	3.7	STAT
GTTA••••TTAA	4	32	4.6	3.6	qa-1F
GTYAATGA	4	35	4.2	3.6	HNF-6
GGHTCATA	3	28	3.9	2.7	LF-A1
CGTGSTGA	3	26	4.2	2.7	SREBP-1
CTAG•CAAK	3	24	4.5	2.7	C/EBP
AMTA••AACC	3	22	5.0	2.6	c-Myb
ACSG••••••GTCA	3	19	5.7	2.6	HNF-4
GAGC••CATC	2	13	5.6	1.7	C/EBPβ

Z is the *Z*-test score; see Table 1 legend for the remaining entries.

Table 4. Top three motifs of lengths 6, 8 and 10 found by MDscan to be over-represented in LSPS against EPD78 vertebrate promoters

Motif	Score	Segments	TTF	C
AGCGCT	4.74	172		0
TTACCT	4.72	154	SREBP-1	6
AGGGCT	4.67	182		4
AGRGCTGG	3.731	98	HEB	2
CTAAGGAA	3.630	77	NERF-1i	1
CCCARCCC	3.607	115	CAC-binding	5
TTAATKATTA	3.044	51	SBF-1	1
RGGGKTGGGG	3.003	67	SREBP-1	0
CTGAGTTCAG	2.978	67	Alx-4	0

Motif is the motif consensus; Score is the total relative entropy score of the motif; Segments is the number of aligned segments used to generate the motif; TTF (TRANSFAC transcription factor) is the transcription factor whose binding site PFM in TRANSFAC best matches the motif; C (conservation) is the number of occurrences of the motif consensus in the consensus set that is generated from an alignment of LSPS promoters with their orthologs.

Table 5. Top three motifs of lengths 6, 8 and 10 found by MDscan to be over-represented in LSPS against EPD78 vertebrate promoters that are not known to be strongly expressed in liver

Motif	Score	Segments	TTF	C
ATGTGT	5.00	131		3
TACATA	4.97	160	VBP	4
TATGTT	4.97	156	HNF-3β	3
AWTAATTA	3.95	67	POU2F1	6
TRATTAAT	3.95	89	HNF-1	3
AATGATTA	3.91	95	Alx-4	1
AATSATTAAY	3.41	49	Vmw65	3
TTAATWATTA	3.36	82	HNF-1	0
GTTAATAATT	3.35	53	HNF-1	1

See Table 4 for complete legend.

Table 6. Top three motifs of lengths 6, 8 and 10 found by MDscan to be over-represented in the consensus set against vertebrate promoters in EPD78

Motif	Score	Segments	TTF
CGTAGG	4.89	112	
CCTATG	4.78	118	HNF-4
CCTACC	4.78	159	
TACCTATG	3.68	88	HNF-4
CGTAGTTA	3.65	80	MYB.PH3
CCGATAAC	3.62	77	GATA-1
SGMTCGRGCG	3.06	51	CUTL1
ATAGGATCGA	3.05	60	CUTL1
GATCGATCGA	3.04	55	CUTL1

See Table 4 for complete legend.

Because the consensus set allows for a very small number of trials for each word structure, and because of the high-false-negative rate when using a consensus, we did not find motifs with P-values <0.001 when searching in the consensus against EPD vertebrate promoters. Instead, we report motifs by Z-test score (Table 3).

For each motif x with n_f occurrences in the foreground set and n_c occurrences in the consensus set, we found all degenerate words having the same structure and the same count in the foreground set, and counted the number of occurrences of these words in the consensus set. Our results suggest that the majority of these words are strongly conserved in the consensus set. These results are reported in the Supplementary information.

DISCUSSION AND CONCLUSION

DWE is a fast word-counting-based tool for discovering overrepresented motifs in one set of promoters relative to another. Our results on synthetic data suggest that DWE outperforms existing methods on a large class of motifs, and is best suited for finding overrepresented motifs against carefully selected background sets. However, the accuracy of DWE decreases with increasing motif degeneracy. In addition to single motifs, DWE can find overrepresented motif tuples. A feature of DWE's P-value motif comparison method is that it allows comparisons of motifs with different structures, and motifs that are found using different foreground or background sets.

We used DWE to search for overrepresented motifs in proximal promoters of liver-specific genes, and found that HNF binding sites and binding sites for CCAAT/enhancer-binding protein (C/EBPβ) are the most overrepresented. This conclusion is largely supported by experiments with MDscan, and agrees with the results of Baumhueter *et al*., (1988), Costa *et al*., (1989), Xanthopoulos *et al*., (1991), Thomas *et al*., (2001) and Krivan and Wasserman, 2001. Our results on synthetic data suggest that DWE has a high degree of accuracy when searching for motifs with structures and frequencies characteristic to the majority of motifs reported.

When searching for co-occurring motif pairs, we found that HNF-4 binding sites have strong synergistic relationships with other HNF-4 binding sites and with binding sites of HNF-1, HNF-3β and C/EBPβ. These relationships are supported by high conservation ratios (number of occurrences in LSPS versus number of occurrences in the promoter consensus set), and agree with the results of Miura and Tanaka (1993), Antes and Levy-Wilson (2001) and Hatzis and Talianidis (2002).

Our results suggest that the majority of top motifs found by DWE are conserved, but few motifs such as CWGT•••CABA and ATAGTYTV of Tables 7 and 8 have low conservation ratios and may be false positives. The majority of motif pairs in Tables 9 and 10 have weak conservation ratios, but the motif pairs GWTA••••TTDA MWG•TTA, GWTA••••TTDA AAMRGT, GWTA••••TTDA TTGBAA and GDTA••••TTRA TTGBAA have relatively high-conservation ratios, which may indicate a more biologically significant relationship (Tables 11

Table 7. Motifs that are strongly overrepresented (by sequence count) in promoters of liver-expressed genes (LSPS) against promoters of liver-expression independent genes (EPD)

Motif	FO	BO	L	P	TTF	C
GWTA••••TTDA	15	120	8.5	9.4E−11	HNF-4	9
T•ATSA	33	1158	1.9	1.1E−08		21
CWGT•••CABA	17	244	4.7	1.7E−08	C/EBPβ	2
GTTAATGW	9	44	13.9	2.7E−08	HNF-1	4
GGCWCAYA	12	116	7.0	8.7E−08		3
ATA•TWR	28	837	2.3	9.2E−08		10
TTGBAA	30	982	2.1	9.7E−08	C/EBPβ	16
ATAGTYTV	11	93	8.0	9.8E−08	ICSBP	2
MWG•TTA	31	1059	2.0	1.0E−07	HNF-3β	12
AAMRGT	33	1259	1.8	1.5E−07	PPAR-γ	12

See Table 1 for a complete legend.

Table 8. Motifs that are strongly overrepresented (by sequence count) in promoters of liver-expressed genes (LSPS) against promoters of genes that are not known to be expressed in liver

Motif	FO	BO	L	P	TTF	C
GDTA••••TTRA	15	99	9.9	1.4E−11	HNF-4	7
GTTAATSW	11	66	10.8	5.9E−09	HNF-1	5
CWGT•••CABA	17	223	5.0	8.9E−09	C/EBPβ	2
AT•A•HAAC	17	234	4.7	1.8e−08	HNF-3β	9
ATA•TWR	28	775	2.4	4.2E−08		10
ATAGTYTV	11	82	8.7	4.6E−08	ICSBP	2
GGCWCAYA	12	107	7.3	6.1E−08		3
TTGBAA	30	922	2.1	6.4E−08	C/EBPβ	16
AGAY••THTG	13	137	6.2	9.2E−08	HSF1	1
ACATWD	32	1092	1.9	1.1E−07		11

Table 9. Top pairs (by sequence count) of the motifs from Table 7

Motif pair	FO	BO	L	P	C
GWTA••••TTDA MWG•TTA	15	78	13.1	8.6E−11	5
GWTA••••TTDA AAMRGT	15	85	12.0	2.4E−10	5
GWTA••••TTDA ATA•TWR	14	73	13.0	3.3E−10	3
GGCWCAYA ATAGTYTV	7	6	79.3	4.4e−10	1
GWTA••••TTDA TTGBAA	13	63	14.0	5.7e−10	4

C (conservation) is the number of consensus sequences that contain non-overlapping occurrences.

and 12). We note that motifs found by DWE have relatively higher conservation ratios than motifs found by MDscan.

We also examined motifs that had a large number of occurrences in LSPS but were not overrepresented against EPD vertebrate promoters. We found that many of these motifs have high conservation ratios. These motifs are reported in the Supplementary information.

Table 10. Top pairs (by sequence count) of the motifs from Table 8

Motif pair	FO	BO	L	P	C
GDTA••••TTRA ATA•TWR	15	57	17.1	3.5E−12	2
ATAGTYTV GGCWCAYA	7	2	227.8	1.3E−11	1
AT•A•HAAC AGAY••THTG	10	20	32.5	5.4E−11	0
ATA•TWR GGCWCAYA	11	35	20.5	3.6E−10	1
GTTAATSW ATA•TWR	11	39	18.4	9.3E−10	2
GDTA••••TTRA TTGBAA	12	53	14.7	1.5E−09	4

C (conservation) is the number of consensus sequences that contain non-overlapping occurrences.

Table 11. Top pairs (by occurrence count) of the motifs from Table 1

Motif pair	FO	BO	L	Z	C
A•A••T•A A•A•AG	2601	60904	3.4	2.5E+03	1549
A•T••••A•A A•A•AG	2430	63844	3.0	2.3E+03	572
A•A•AG• CT•TG	2414	76222	2.5	2.3E+03	415
A•A••T•A A•T••••A•A	2383	61598	3.1	2.3E+03	577

P-values are 0, Z is the *Z*-test score and C (conservation) is the sum of the number of non-overlapping co-occurrences in each consensus sequence

Table 12. Top pairs (by occurrence count) of the motifs from Table 2

Motif pair	FO	BO	L	Z	C
A•A••T•A DT•••AAA	2107	49913	3.2	2.0E+03	439
A•A••T•A THT••T•A	1871	43255	3.3	1.8E+03	385
A•A••T•A AAG•••T	1845	41300	3.4	1.8E+03	431
A•A••T•A AT••••CA	1813	32682	4.2	1.7E+03	406
A•A••T•A TD••TTA	1752	36261	3.7	1.7E+03	500
A•A••T•A CAA•••T	1693	35538	3.6	1.6E+03	411

P-values are 0, Z is the *Z*-test score and C (conservation) is the sum of the number of non-overlapping co-occurrences in each consensus sequence.

Our consensus construction method can be used to filter out false-positive detections, but in its current state it is error-prone. Consensus construction through ortholog alignment requires promoter alignment tools and consensus construction tools that are not yet perfected. Our method is very conservative when aligning ortholog promoters from distant species, and has little impact on false-positive filtration when aligning ortholog promoters from close species. Moreover, by using CLUSTALW we impose a colinearity constraint and do not consider inversions or TFBS birth and death events.

We used DWE to discover liver-specific *cis*-regulatory elements. Of course, DWE can be used to discover motifs in promoters of any co-regulated genes. To improve its performance in detecting more degenerate motifs, DWE should be modified to use PWM scores instead of occurrence counts.

ACKNOWLEDGEMENTS

We thank Zhenyu Xuan, Debopriya Das and Saurabh Sinha for useful discussions. This work is supported by NIH grant GM060513 and NSF grants DBI-0306152 and EIA-0324292.

REFERENCES

Agresti,A. (1992) A survey of exact inference for contingency tables. *Stat. Sci.*, **7**, 131–177.

Antes,T.J. and Levy-Wilson,B. (2001) HNF-3 beta, C/EBP beta, and HNF-4 act in synergy to enhance transcription of the human apolipoprotein B gene in intestinal cells. *DNA Cell Biol.*, **20**, 67–74.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, **21**, 51–80.

Baumhueter,S., Courtois,G. and Crabtree,G.R. (1988) A variant nuclear protein in dedifferentiated hepatoma cells binds to the same functional sequences in the beta fibrinogen gene promoter as HNF-1. *EMBO J.*, **7**, 2485–2493.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Blanchette,M. and Sinha,S. (2001) Separating real motifs from their artifacts. In *Proceedings of the Annual International Symposium on Intelligent Systems for Molecular Biology*, Copenhagen, Denmark, pp. 30–38.

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

Costa,R. H., Grayson,D.R. and Darnell,J.E.,Jr (1989) Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes. *J. Comput. Biol.*, **9**, 1415–1425.

Davuluri,R., Grosse,I., and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.

Hatzis,P. and Talianidis,I. (2002) Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell*, **10**, 1467–1477.

Hertz,G., Hartzell,G. III, and Stormo,G. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Jaynes,E.T. (1957a) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.

Jaynes,E.T. (1957b). Information theory and statistical mechanics. II. *Phys. Rev.*, **108**, 171–190.

Knuppel,R., Dietze,P., Lehnberg,W., Frech,K. and Wingender,E. (1994) TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.*, **1**, 191–198.

Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.

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

Liu,J.S., Lawrence,C.E. and Neuwald,A. (1995) Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**, 1156–70.

Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Miura,N. and Tanaka,K. (1993) Analysis of the rat hepatocyte nuclear factor (HNF) 1 gene promoter: synergistic activation by HNF4 and HNF1 proteins. *Nucleic Acids Res.*, **21**, 3731–3736.

Perier,R.C., Junier,T. and Bucher,T. (1998). The eukaryotic promoter database EPD. *Nucleic Acids Res.*,**26**, 353–357.

Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.

Schones,D., Sumazin,P. and Zhang,M.Q. (2004) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, doi: 10.1093/bioinformatics/bth480

Sinha,S. (2003). Discriminative motifs. *J. Computat. Biol.*, **10**, 599–615.

Sinha,S. and Tompa,M. (2000) A statistical method for finding transcription factor binding sites. In *Proceedings of the Annual International Symposium on Intelligent Systems for Molecular Biology*, Copenhagen, Denmark, Vol. 8, pp. 344–344.

Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **30**, 5549–5560.

Student (1908) The probable error of a mean. *Biometrika*, **6**, 1–25.

Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length CDNAS. *Nucleic Acids Res.*, **30**, 328–331.

Thomas,H., Jaschkowitz,K., Bulman,M., Frayling,T.M. Mitchell,S.M. Roosen,S., Lingott-Frieg,A., Tack,C.J., Ellard,S., Ryffel,G.U. and Hattersley,A.T. (2001) A distant upstream promoter of the HNF-4alpha gene connects the transcription factors involved in maturity-onset diabetes of the young. *Hum. Mol. Genet.*, **10**, 2089–2097.

Thompson,J.D., Higgins,D.G. and 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. *Nucleic Acids Res.* **22**, 4673–4680.

Van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Van Helden,J., Andre,B. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.

Waterman,M.S., Arratia,R. and Galas,D.J. (1984) Pattern recognition in several sequences: consensus and alignment. *Bulletin of Mathematical Biol.*, **46**, 515–527.

Xanthopoulos,K.G., Prezioso,V.R. Chen,W.S., Sladek,F.M., Cortese,R. and Darnell,J.E.J. (1991) The different tissue transcription patterns of genes for HNF-1, C/EBP, HNF-3, and HNF-4, protein factors that govern liver-specific transcription. *Proc Natl Acad Sci., USA*, **88**, 3807–3811.

Zhang,T. and Zhang,M.Q. (2000) Liver specific promoter database.