

DWE: Discriminating Word Enumerator

Pavel Sumazin^{1,2}, Gengxin Chen¹, Naoya Hata¹, Andrew D. Smith¹, Theresa Zhang³ and Michael Q. Zhang¹

¹Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, ²Computer Science Department, Portland State University, P.O. Box 751, Portland, OR 97207, USA and

³Bioinformatics, Merck Research Laboratories, Rahway, NJ 07065, USA

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 over-represented 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 over represented 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/EBP β) are the most over represented in proximal promoters of liver-specific genes. Our results suggest that HNF-4 has strong synergistic relationships with hepatocyte nuclear factors HNF-1, HNF-4 and HNF-3 β and with C/EBP β .

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

Contact: Pavel Sumazin (ps@cs.pdx.edu) and Michael Q. Zhang (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), that can be used to discover motifs in promoters of co-regulated genes. We use DWE to find over-represented gapped degenerate words (motifs) in proximal promoters of liver-specific genes taken from Liver-Specific Promoter Database (LSPD) [Zhang and Zhang (2000)] against vertebrate promoters from the Eukaryotic Promoter Database (EPD), release 78 [Perier et al. (1998)]. We use TSS position data from DBTSS [Suzuki et al. (2002)] and sequence data from Genebank to collect the promoter sequences.

Related Literature

Classical sequence-based motif discovery algorithms include CONSENSUS by Hertz et al. (1990), MEME by Bailey and Elkan (1995), and the Gibbs sampler by Lawrence et al. (1993); Liu et al. (1995). Other motif discovery algorithms that use word-counting methods are reported by 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 REDUCE by Bussemaker et al. (2001), MDscan by Liu et al. (2002), DMOTIFS by Sinha (2003) and YMF by 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 over represented in the foreground set against a Markov background model that can be estimated from a background sequence set; DMOTIFS searches for over-represented motifs in a foreground set against a background set while maintaining a maximum count per sequence; YMF searches for over-represented 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 by Takusagawa and Gifford (2004), who use a p-value statistic to search for over represented ungapped motifs of length 7 in *Saccharomyces Cerevisiae* promoters.

SYSTEM AND METHODS

We searched for over-represented 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 over representation 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 over-represented 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 over representation.

P-value. The fixed marginal contingency table p -value follows the multiple hypergeometric distribution given in Equation 1; see Agresti (1992) for a review. 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 over representation 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).

$$P = \frac{\binom{N_f}{n_f} \binom{N_b}{n_b}}{\binom{N_f+N_b}{n_f+n_b}} \quad (1)$$

Z test. The Z test of Student (1908) is given in Equation 2.

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

Log frequency ratio. The log frequency ratio (LFR) is given in Equation 3.

$$\text{LFR} = \ln \frac{n_f N_b}{n_b N_f} \quad (3)$$

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 to PFMs using the maximum entropy principle of Jaynes (1957a,b); each IUPAC symbol was converted to a maximum-entropy column with total count equal to the number of foreground occurrences n_f . For example, $M = \{A, C\}$ was converted to $[n_f/2, n_f/2, 0, 0]^T$ and $D = \{A, G, T\}$ was converted to $[n_f/3, 0, n_f/3, n_f/3]^T$. We used a chi-squared test to compare discovered-motif PFMs to TRANSFAC PFMs following the methodology of 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 sub-matrix 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.

Data Set 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 given in Supplementary Information.

We generated a consensus sequence for each ortholog set, and used those consensus sequences to check for 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 over-represented motifs in the consensus set against vertebrate promoters in EPD78; see Table 9. 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 4 sequences that are different from the original. Both consensus sequence sets are given in 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 to 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 to its representation in the background, and the words with foreground over-representation 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 Systems and Methods. 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 distribu-

tion 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 over-represented 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 chi-squared 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 over represented 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 $\sum_{s \in S} x_1^{(s)} x_2^{(s)} - O(X_2)$, where $O(X_2)$ 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 $O(X_k) = \sum_{s \in S} O(X_k, s)$ is given in Equation 5, where $L(X_k, s)$ is the number of distinct motif k -tuple occurrences that share at least one position in s . The total running time is on the order of $|S| + k \log k O(X_k)$.

$$\text{Trials}(X_k) = \sum_{s \in S} \left(|s| - \sum_{x \in X_k} \binom{|x|}{k} \right) \quad (4)$$

$$O(X_k) = \sum_{s \in S} [-kL(X_k, s) + \sum_{X_{k-1} \subset X_k} \sum_{x \notin X_{k-1}} x^{(s)} O(X_{k-1}, s)] \quad (5)$$

EXPERIMENTS

We used DWE and MDscan to find the most over-represented 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 over-represented motifs in a foreground set against a background set. We chose YMF

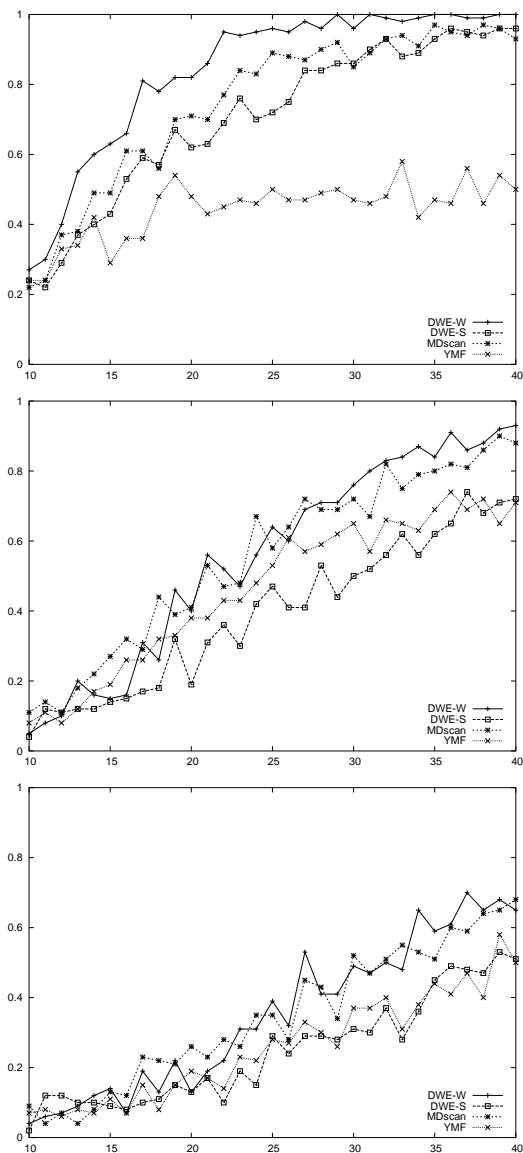


Fig. 1. Detection-quality comparison of DWE, MDscan and YMF when attempting to discover an implanted motif with width six against a vector-generated background sequence set. We plot the frequency (from 0 to 1) of the correct detection in the top 5 found motifs for each method as a function of the number 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; 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).

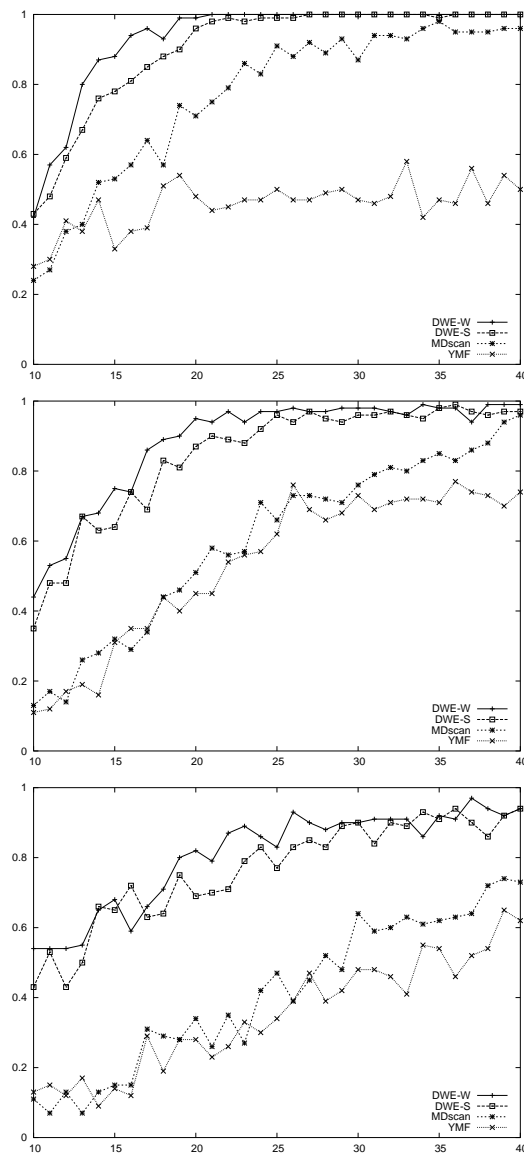


Fig. 2. Detection-quality comparison of DWE, MDscan and YMF when attempting to discover an implanted motif with width six against an augmented background sequence set that is created by adding 35 additional sequences that do not contain the motif to the background set used in the experiments reported in Figure 1. We plot the frequency (from 0 to 1) of the correct detection in the top 5 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).

over DMOTIFS on the recommendation of Sinha (2004) as DMOTIFS is not publicly available.

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 each algorithm's ability 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 that correspond to uniformly selected IUPAC words with specified number of degenerate positions.

We constructed foreground sets with 10 to 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 5 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 dependencies 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; see Figure 1.

We tested the ability of the algorithms to discover implanted motifs that are strongly under represented 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 over represented in LSPS against the vertebrate promoter subset of EPD78 (Table 1), and against that set excluding promoters of liver expressed genes (Table 5). We searched for (3+gap+3)-mers and (4+gap+4)-mers, with rigid gaps ranging from 0 to 7 base pairs and at most 2 degenerate positions.

We also searched for motifs that are over represented in the consensus set against the vertebrate promoter set from EPD78 (Table 9). We repeated these searches using MDscan and report the top 3 motifs of lengths 6,8, and 10 in each experiment; see Tables 10, 11 and 12.

Initially, MDscan reported poly-A and alternating C-T motifs. These motifs are found to be strongly over represented by DWE when motif auto-correlation is not considered. However, the number of occurrences of these motifs decreases substantially when self overlap 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 period 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 for HNF-4 and HNF-1.

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 lower than 0.001 when searching in the consensus against EPD vertebrate promoters. Instead we report motifs by Z -test score; see Table 9.

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 Supplementary Information.

DISCUSSION AND CONCLUSION

DWE is a fast word-counting-based tool for discovering over-represented 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 over-represented motifs against carefully selected background sets. However, the accuracy of DWE decreases with increasing motif degeneracy. In addition to single motifs, DWE can find over-represented 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 over-represented motifs in proximal promoters of liver-specific genes, and found that hepatocyte nuclear factor binding sites and binding sites for CCAAT/enhancer-binding protein (C/EBP β) are the most over represented. This conclusion is largely supported by experiments with MDscan, and agrees with results by Baumhueter et al. (1988); Costa et al. (1989); Xanthopoulos et al. (1991); Thomas et al. (2001); Krivan

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/EBP β	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-3 β	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

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

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/EBP β	2
GTTAATGW	9	44	13.9	2.7e-08	HNF-1	4
GGCWCAAYA	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/EBP β	16
ATAGTYTV	11	93	8.0	9.8e-08	ICSBP	2
MWG●TTA	31	1059	2.0	1.0e-07	HNF-3 β	12
AAMRGT	33	1259	1.8	1.5e-07	PPAR- γ	12

Table 2. Motifs that are strongly over represented (by sequence count) in promoters of liver-expressed genes (LSPS) against promoters of liver-expression independent genes (EPD); see Table 1 for a complete legend.

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

Table 3. Top pairs (by occurrence count) of the motifs from Table 1; p -values are 0; Z is the Z -test score; C (conservation) is the sum of the number of non-overlapping co-occurrences in each consensus sequence.

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
GGCWCAAYA ATAGTYTV	7	6	79.3	4.4e-10	1
GWTA●●●●TTDA TTGBAA	13	63	14.0	5.7e-10	4

Table 4. Top pairs (by sequence count) of the motifs from Table 2. C (conservation) is the number of consensus sequences that contain non-overlapping occurrences.

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-3 β	48
AT●●●●CA	162	5521	2.1	1.9e-17	C/EBP β	57
DT●●●AAA	161	6073	1.9	1.1e-13	C/EBP β	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

Table 5. Motifs that are strongly over represented (by occurrence count) in LSPS against EPD vertebrate promoters of genes that are not known to be expressed in liver. See Table 1 for a complete legend.

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/EBP β	2
AT●A●HAAC	17	234	4.7	1.8e-08	HNF-3 β	9
ATA●TWR	28	775	2.4	4.2e-08		10
ATAGTYTV	11	82	8.7	4.6e-08	ICSBP	2
GGCWCAAYA	12	107	7.3	6.1e-08		3
TTGBAA	30	922	2.1	6.4e-08	C/EBP β	16
AGAY●●THTG	13	137	6.2	9.2e-08	HSF1	1
ACATWD	32	1092	1.9	1.1e-07		11

Table 6. Motifs that are strongly over represented (by sequence count) in promoters of liver-expressed genes (LSPS) against promoters of genes that are not known to be expressed in liver.

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

Table 7. Top pairs (by occurrence count) of the motifs from Table 5. p -values are 0; Z is the Z -test score; C (conservation) is the sum of the number of non-overlapping co-occurrences in each consensus sequence.

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

Table 8. Top pairs (by sequence count) of the motifs from Table 6. C (conservation) is the number of consensus sequences that contain non-overlapping occurrences.

Motif	FO	BO	L	Z	TTF
GTAAAT	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
GYAATGA	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

Table 9. Motifs that are over represented (by occurrence count) in the consensus set against EPD vertebrate promoters. Z is the Z-test score; see Table 1 legend for the remaining entries.

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 hepatocyte nuclear factor 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 vs. number of occurrences in the promoter consensus set), and agree with results by 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 Table 2 and Table 6 have low conservation ratios and may be false positives. The majority of motif pairs in Table 4 and Table 8 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. 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 over represented against EPD vertebrate promoters. We found that many of these motifs have high conservation ratios. These motifs are reported in Supplementary Information.

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

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

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

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

Table 11. 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. See Table 10 for complete legend.

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

Table 12. 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. See Table 10 for complete legend.

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 co-linearity 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 (Position Weight Matrix) scores instead of occurrence counts.

ACKNOWLEDGMENTS

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. *Statistical Science* 7, 131–177.
- Antes, T. J. and B. Levy-Wilson (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(2), 67–74.
- Bailey, T. L. and C. Elkan (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21(1-2), 51–80.
- Baumhueter, S., G. Courtois, and G. R. Crabtree (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 Journal* 7(8), 2485–2493.
- Beer, M. A. and S. Tavazoie (2004). Predicting gene expression from sequence. *Cell* 117(2), 185–198.
- Blanchette, M. and S. Sinha (2001). Separating real motifs from their artifacts. In *Proceedings of the Annual International Symposium on Intelligent Systems for Molecular Biology*, pp. 30–38.
- Bussemaker, H. J., H. Li, and E. D. Siggia (2001). Regulatory element detection using correlation with expression. *Nature Genetics* 27(2), 167–171.
- Costa, R. H., D. R. Grayson, and J. E. Darnell Jr (1989). Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes. *Journal of Computational Biology* 9(4), 1415–1425.
- Davuluri, R., I. Grosse, and M. Q. Zhang (2001). Computational identification of promoters and first exons in the human genome. *Nature Genetics* 29, 412–417.
- Hatzis, P. and I. Talianidis (2002). Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell* 10(6), 1467–1477.
- Hertz, G., G. Hartzell III, and G. Stormo (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* 6(2), 81–92.
- Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review* 106, 620–630.
- Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review* 108, 171–190.
- Knuppel, R., P. Dietze, W. Lehnberg, K. Frech, and E. Wingender (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *Journal of Computational Biology* 1(3), 191–198.
- Krivan, W. and W. W. Wasserman (2001). A predictive model for regulatory sequences directing liver-specific transcription. *Genome Research* 11(9), 1559–1566.
- Lawrence, C., S. Altschul, M. Boguski, J. Liu, A. Neuwald, and J. Wootton (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.
- Liu, J. S., C. E. Lawrence, and A. Neuwald (1995). Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *Journal of the American Statistical Association* 90, 1156–70.
- Liu, X. S., D. L. Brutlag, and J. S. Liu (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* 20(8), 835–839.
- Miura, N. and K. Tanaka (1993). Analysis of the rat hepatocyte nuclear factor (HNF) 1 gene promoter: synergistic activation by HNF4 and HNF1 proteins. *Nucleic Acids Research* 21(16), 3731–3736.
- Perier, R. C., T. Junier, and P. Bucher (1998). The eukaryotic promoter database EPD. *Nucleic Acids Research* 26(1), 353–357.
- Scherf, M., A. Klingenhoff, and T. Werner (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *Journal of Molecular Biology* 297(3), 599–606.
- Schones, D., P. Sumazin, and M. Q. Zhang (2004). Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, In press.
- Sinha, S. (2003). Discriminative motifs. *Journal of Computational Biology* 10(3-4), 599–615.
- Sinha, S. (2004). Personal communication.
- Sinha, S. and M. Tompa (2000). A statistical method for finding transcription factor binding sites. In *Proceedings of the Annual International Symposium on Intelligent Systems for Molecular Biology*, Volume 8, pp. 344–344.
- Sinha, S. and M. Tompa (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research* 30(24), 5549–5560.
- Student (1908). The probable error of a mean. *Biometrika* 6, 1–25.
- Suzuki, Y., R. Yamashita, K. Nakai, and S. Sugano (2002). Dbtss: Database of human transcriptional start sites and full-length cdnas. *Nucleic Acids Research* 30(1), 328–331.
- Takusagawa, K. T. and D. K. Gifford (2004). Negative information for motif discovery. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 360–371.
- Thomas, H., K. Jaschowitz, M. Bulman, T. M. Frayling, S. M. Mitchell, S. Roosen, A. Lingott-Frieg, C. J. Tack, S. Ellard, G. U. Ryffel, and A. T. Hattersley (2001). A distant upstream promoter of the HNF-4alpha gene connects the transcription

-
- factors involved in maturity-onset diabetes of the young. *Human Molecular Genetics* 10(19), 2089–2097.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22), 4673–4680.
- Van Helden, J., B. Andre, and J. Collado-Vides (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281, 827–842.
- Van Helden, J., B. Andre, and J. Collado-Vides (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research* 28(8), 1808–1818.
- Waterman, M. S., R. Arratia, and D. J. Galas (1984). Pattern recognition in several sequences: consensus and alignment. *Bulletin of Mathematical Biology* 46, 515–527.
- Xanthopoulos, K. G., V. R. Prezioso, W. S. Chen, F. M. Sladek, R. Cortese, and J. E. J. Darnell (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. U S A* 88(9), 3807–3811.
- Zhang, T. and M. Q. Zhang (2000). Liver specific promoter database. <http://cgsigma.cshl.org/LSPD>.