

Mining ChIP-chip Data for Transcription Factor and Cofactor Binding Sites

Andrew D. Smith^a, Pavel Sumazin^{a,b}, Debopriya Das^a, Michael Q. Zhang^a

^aCold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724

^bComputer Science Department, Portland State University, Portland, Oregon 97207

ABSTRACT

Motivation: Identification of single motifs and motif pairs that can be used to predict transcription factor localization in ChIP-chip data, and gene expression in tissue-specific microarray data.

Results: We describe methodology to identify *de novo* individual and interacting pairs of binding site motifs from ChIP-chip data, using an algorithm that integrates localization data directly into the motif discovery process. We combine matrix-enumeration-based motif discovery with multi-variate regression to evaluate candidate motifs and identify motif interactions. When applied to the HNF localization data of Odom *et al.* (2004) in liver and pancreatic islets, our methods produce motifs that are either novel or improved known motifs. All motif pairs identified to predict localization are further evaluated according to how well they predict expression in liver and islets, and according to how conserved are the relative positions of their occurrences. We find that interaction models of HNF1 and CDP motifs provide excellent prediction of both HNF1 localization and gene expression in liver. Our results demonstrate that ChIP-chip data can be used to identify interacting binding site motifs.

Availability: Motif discovery programs and analysis tools are available upon request from the authors

Contact: {asmith, sumazin, mzhang}@cshl.edu

1 INTRODUCTION

The identification of regulatory signals in genomes, and specifically the discovery of transcription factor and cofactor binding sites, is among the greatest immediate challenges in genome science. Computational discovery of transcription factor binding sites usually proceeds by examination of a set of sequences believed to be bound by the same factor to identify common patterns, either in the form of consensus or position weight matrices. Since many transcription factors bind specifically to sequence elements with particular properties, common patterns represent hypothetical transcription factor binding site motifs that can be tested at the bench.

High throughput experimental techniques, including microarray expression and ChIP-chip, can be used to identify sequences that are likely to contain binding sites for the same or similar sets of factors. Analysis of expression data assumes that co-expressed genes are often direct targets of common factors, and that a rough estimate for the location of main factor binding regions can be made (*e.g.* the proximal promoter). ChIP-chip experiments measure *in-vivo* localization of a particular factor on a known sequence, identifying cross-linking ratios for the factor with putative regulatory regions in chromatin DNA [31]. Factor localization is strongly correlated with binding (direct or indirect), and is usually taken as a measure of binding affinity. Because ChIP-chip data is directly correlated with binding, and because identities of localized sequences are known, ChIP-chip data may be better suited for binding site identification than expression data. To make best use of localization data, we incorporate localization data directly into the motif-discovery process, as opposed to using it to select a sequence set or evaluate motifs that have already been discovered.

Regression-based methods maximize the use of available information and have been widely used to correlate predicted motif occurrences with expression data [14]. Wasserman and Fickett [38] used regression to easily incorporate multiple factors, cooperation rules and spacing constraints in muscle promoters (the same method was applied to Liver by Krivan and Wasserman [22]). Bussemaker *et al.* [4] fit motif counts linearly to the log of the expression ratio to identify regulatory elements. Conlon *et al.* [6] extended the method, using motif scores and a greedy heuristic, to identify sets of interacting motifs through stepwise regression. Still, the exact quantitative relationship between sequence elements and expression data is not known, and a single quantitative formulation may not exist, especially when multiple interacting motifs are considered. To overcome this problem, Das *et al.* [7] introduced MARSMotif which uses Multivariate Adaptive Regression Splines (MARS) [13, 18] to correlate non-linear relationships between multiple motif scores and expression. We use

MARSMotif to identify cooperative motifs, by correlating motif scores and localization data.

The importance of transcription factor synergy in both regulating expression and protein-DNA binding is widely recognized. Algorithms that attempt to model such interactions, and discover interacting motifs include Co-Bind [15] and BioProspector [26], which attempt to identify co-occurring motifs, and Gibbs Recursive Sampler [35], which rewards co-occurring motifs. Close proximity is often required for the cooperative interactions of factors [11], and for the function of enhanceosomes, which form on segments of DNA with length approximately 100 bases or less [5]. Hannehalli and Levy [17] use co-localization to identify cooperative factors by examining motifs with occurrences separated by either at most 50 or at most 200 bases. Wasserman and Fickett [38] study co-occurrence of binding motifs for muscle regulatory elements, and observe that sensitivity and specificity are highest when co-occurrences are localized within 100 bases.

We identify motif pairs with co-occurrences within 200-base regions that are significantly correlated with factor localization. In order to discover motif candidates that correlate with factor localization we use an enumerative algorithm called DME-X. DME-X incorporates localization data with sequence data to identify binding site motifs represented as position-weight matrices. DME-X extends the enumerative algorithm DME [32], which identifies motifs that are over-represented in a foreground set relative to a background set. We identify single and co-occurring motifs using DME-X, and evaluate candidate motifs and candidate interacting motifs using regression.

We applied our method to the localization data from ChIP-chip experiments of Odom *et al.* [28]. We evaluated motifs identified by DME-X, as well as previously characterized binding site motifs from TRANSFAC [27]. We show that all but one of the top motifs identified by DME-X are highly similar to top motifs from TRANSFAC (using Kullback-Leibler divergence [24]), and most provide a better prediction of localization. For comparison purposes, we also evaluated candidate motifs identified by MDModule [6], and show that DME-X and TRANSFAC motifs display stronger correlation to HNF localization than MDModule motifs. To identify interacting pairs among top scoring individual motifs, we evaluated pairs of motifs according to conservation of the relative positions of their occurrences, and the correlation of their co-occurrences with HNF localization. To identify motifs whose occurrences co-localize, we searched the sequence neighborhood of occurrences of top motifs.

We evaluated the correlation between motif occurrences and gene expression using the microarray expression data of Su *et al.* [34]. Our results support and extend the findings of Krivan and Wasserman [38], demonstrating that HNF localization correlates with expression in liver and that co-occurrences of HNF, C/EBP and Sp1 motifs can be used to improve localization-based expression predictions in islets and liver.

We use the microarray expression data of Su *et al.* [34] to identify motif pairs that correlate with HNF localization and have stronger correlation with expression than HNF localization.

2 METHODS

To identify binding site motifs we use a strategy of generating candidates using sequence and localization data, determining how well the candidates can predict the localization data (alone or in pairs), and focusing the search once more on sequence regions near high scoring candidates to identify additional, possibly more subtle motifs that co-localize with a high scoring candidate. We test motif modules that correlate well with factor localization to determine increased correlation with expression.

2.1 The High Level Procedure

Our method examines a set of sequences $F = \{S_1, \dots, S_m\}$, and makes use of a set of localization values $Y = \{y_1, \dots, y_m\}$ where y_i is the localization value associated with sequence S_i . Given a set $B = \{b_1, \dots, b_m\}$ of *experimental* localization values (which may be p -values or localization ratios), where b_i is the experimental localization associated with sequence S_i , we define $y_i = \log(\theta/b_i)$ with significance threshold θ commonly set to 10^{-3} for experimental localization p -values, or $y_i = \log(b_i/\theta)$ with significance threshold θ commonly set 2.0 for experimental localization ratios. The high level procedure for identifying motifs is composed of the following stages.

Obtain a set of candidates. Applying DME-X to the sequence set F , and the localization values Y , we obtain the set C_1 of candidate motifs. In general C_1 can be supplemented with any set of motifs, and we included previously characterized motifs from TRANSFAC [27] and motifs identified by MDModule [6].

Filter candidates based on predictive ability. Each motif from C_1 is evaluated using regression to determine how well it predicts localization. The result is the set C_2 of top individual predictors.

Recursively search sequence neighborhood. For members of C_2 , the sequence neighborhood of the top occurrences in each sequence is given a more focused search to identify co-localizing binding sites of interacting factors. This search permits the detection of weaker motifs, whose interaction with dominant motifs from C_2 makes them more likely to co-localize. For each motif from C_2 , the set of motifs identified by this neighborhood search forms a set C_3 .

Identify interacting pairs of motifs. Candidates from C_2 and their corresponding C_3 set are further evaluated for their ability to make these predictions in pairs using MARSMotif and relative positional preference (see Section 2.6 for definition). Within each of C_2 and the C_3 sets, all pairs of motifs are considered. Finally, motif pairs that predict the localization data well and show a significant relative positional preference are evaluated to determine if their co-occurrence

better predicts expression than knowledge of HNF localization alone.

2.2 The DME-X Algorithm

The DME algorithm [32] uses an enumerative strategy to discover matrix-based motifs that are overrepresented in a set of foreground sequences *relative to* a set of background sequences. DME identifies motifs with *relative* overrepresentation between two sets of sequences, searches a space constrained by information content of the motifs (information content is a measure of the specificity of a motif [33]), and includes a new local search procedure to replace the conventional local search method of optimizing motifs using EM [3, 10, 29] that does not apply when *relative* overrepresentation is the objective.

DME-X generalizes DME by eliminating the strict requirement for foreground–background sequence classification. DME-X incorporates a weight for each sequence: rather than rewarding and penalizing motifs for occurring in the foreground and background, DME-X rewards for occurrences in proportion to the localization-based weight assigned to the sequence containing the occurrence. The greater the weight on a sequence, the more a motif is rewarded for occurring in that sequence. We note that the algorithm allows arbitrary weights to be associated with the sequences, a feature that makes this algorithm of use in other contexts, such as the analysis of sequences with expression data.

Formally, the set Y of localization values is transformed into a set V of weights, where weight v_i is derived from y_i . Throughout we used two weighting schemes, both used each time DME-X is run with results combined. Neither scheme is superior, as each performs better on some data sets. In both schemes we scale the negative weights by α so that $\sum_{i=1}^m v_i = 0$. This is needed because most values from Y are negative, and we want to avoid identifying matrices purely because they have few occurrences in sequences with negative weights. In the first scheme, if $y_i > 0$, then $v_i = y_i$, otherwise $v_i = \alpha y_i$, and in the second scheme, if $y_i > 0$, then $v_i = 1$, otherwise $v_i = -\alpha$. For each $S_i \in F$, let S_{ij} denote the j -th width- w substring of S_i . For any motif M (treated as the set of parameters of a product multinomial model [25]), the score for M with respect to F is

$$\text{score}(M, F, Y) = \sum_{S_i \in F} y_i \sum_{j=1}^{|S_i|} z_{ij}^{-w+1} \log \left(\frac{\Pr(S_{ij}|M)}{\Pr(S_{ij}|f)} \right),$$

where $z_{ij} = 1$ if and only if $\log \Pr(S_{ij}|M) > 0$, f is a multinomial describing the base composition of F , and $|S_i|$ is the length of S_i . The objective of DME-X is to find a motif M maximizing $\text{score}(M, F, Y)$.

2.3 Using Regression to Select Motifs

Each member of the set C_1 of candidate motifs is evaluated for ability to predict localization data. Given a motif $M \in C_1$,

define the set of predictor variables $X = \{x_1, \dots, x_m\}$ such that x_i is the *max score* value for M in S_i , where substring score is the log likelihood ratio of the substring being an occurrence of $M \in C_1$ vs. base composition. Using a linear model [8] with a “don’t care” cut-off ξ , the set of predictor variables X is fit to the set of localization values Y . The form of the model, with cut-off for the low scores, is

$$\hat{y}_i = a \cdot \max(x_i, \xi) + b,$$

where $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_m\}$ is the set of *predicted* binding values. The fit is measured using reduction in variance (RIV) or the corresponding percent reduction in variance (% RIV). RIV is calculated as

$$\text{RIV} = 1 - \left(\frac{\sum_{i=1}^m (\Gamma_i - \bar{\Gamma})^2}{\sum_{i=1}^m (y_i - \bar{y})^2} \right),$$

where $\Gamma_i = y_i - \hat{y}_i$, and \bar{y} and $\bar{\Gamma}$ are the corresponding means. We optimize for ξ , and find \max RIV in $O(m \log m)$ time.

Localization values in the HNF ChIP-chip data are concentrated about the mean. To fit predictor variables to a subset of the data that would amplify the contributions of extreme values, while still considering contributions from values around the mean, we perform regression on randomized sets constructed using a biased promoter-selection scheme. In this scheme, sequence sets are constructed by including (1) r promoters localized with the factor (*i.e.* those with a localization value above 0), (2) r promoters most likely not to be localized with the factor, and (3) $2r$ of the remaining promoters, chosen uniformly at random. The experiment was repeated 20 times, and motif quality was determined using the average rank over the 20 experiments. The top k motifs are produced as the top individual predictors, and also as the set C_2 of candidates to check for interactions.

2.4 Neighborhood Search to Identify Interactions

A more focused search is performed in the *neighborhood* of each motif from C_2 . For each such motif, the top occurrence (with ties broken arbitrarily) is identified in each sequence with a positive localization score. A new set of sequences is constructed consisting of (at most) 100 bases on either side of each top occurrence. We apply DME-X to this new smaller set of shorter sequences. The large reduction in the size of this set, relative to the original set of sequences, enables consideration of motifs with lower information content that would have been rejected due to high false-positive detection in the full sequence set. We conjecture that this computational phenomenon mirrors conditions in the nucleus, where the binding of factors with high specificity helps recruit interacting factors with lower specificity. The motifs identified during this *neighborhood search* form the set C_3 of candidate motifs that co-localize with a motif from C_2 .

2.5 Identifying Interactions

The set C_2 of motifs selected for individual predictive ability and each of the sets C_3 of motifs resulting from neighborhood

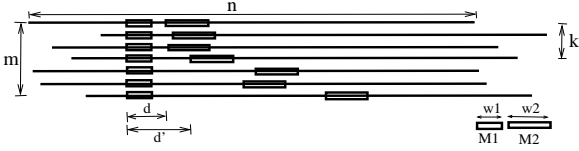


Fig. 1. M_1 and M_2 are within $[d, d']$ distance in k of the m sequences.

searches are examined for interactions using MARSMotif [7]. MARSMotif uses MARS [13, 18] to detect second and third order interactions between motif scores and factor localization values. MARS is a non-parametric and adaptive regression method that builds a set of models using stepwise forward selection and backward elimination in terms of linear splines and their products. From among the set of models, the one with the smallest *generalized cross-validation score* (GCV) is selected. GCV is the residual sum of squares multiplied by a factor to penalize for model complexity, and is a generalization of leave-one-out cross-validation. Let f be a model that predicts binding based on the scores for the set of motifs $M = \{M_1, \dots, M_k\}$ in F . Define $X_i = \{x_{i1}, \dots, x_{ik}\}$ as the set of scores for motifs of M in sequence S_i , and let $X = \{X_1, \dots, X_m\}$. Then the GCV for f with respect to the predictor variables X and the observed localization variables Y is defined as

$$\text{GCV}(f, X, Y) = \sum_{i=1}^m (y_i - f(X_i))^2 / (1 - T(f)/m)^2,$$

where $T(f)$ is the effective number of parameters for the model f , obtained by cross validation [19, 7]. Statistical significance for RIV of models obtained using MARS is determined using an F -test [7].

2.6 Relative Positional Preference (RPP)

To further discriminate true interacting motif pairs, we identify pairs with an unusual relative positional preference (RPP). RPP is defined as a distance range $[d, d']$ between the left-most positions of the best occurrences of two motifs. Given a set of m sequences of length n , the RPP p -value is the probability that the left-most positions of M_1 and M_2 of widths $w_1 \leq w_2$ are within $[d, d']$ distance of each other in at least k of the m sequences (see Figure 1). Assuming that the left-most positions of M_1 and M_2 are taken uniformly at random from the set of permissible positions in the sequence S_i , the probability that these positions are within $[d, d']$ distance of one another is the ratio of the number of position pairs that are within $[d, d']$ distance and the number of permissible position pairs. This probability $p(n, w_1, w_2, d, d')$ is a discretized special case of the r -scan statistics of Karlin and Brendel [21]

and is computed as

$$\begin{aligned} p(n, w_1, w_2, d, d') &= \frac{v + \sum_{i=n-w_2-d'+1}^{n-w_2-d+1} i}{(n-w_2+1)(w_2-w_1) + \sum_{i=1}^{n-w_2+1} i} \\ &= \frac{2v + (d' - d + 1)(2(n-w_2+1) - d' - d)}{(n-w_2+1)(n+w_2-2w_1+2)}, \end{aligned}$$

where $v = \min(w_2 - w_1, d) \cdot (d' - d + 1) + \sum_{i=1}^{w_2-w_1-d} \max(d' - d + 1 - i, 0)$, given that $n > (d' + w_2)$. When M_1 is known to be at the center of each sequence and $n > 2(w_2 + d')$, as in Sections 2.4 and 3.4, the probability calculation is simplified and $p(n, w_1, w_2, d, d') = 2(d' - d + 1)/(n - w_2 + 1)$.

The probability of identifying k of m sequences with RPP $[d, d']$ follows a binomial distribution, and the RPP p -value is

$$\Pr(X(m, n, w_1, w_2, d, d') \geq k) = 1 -$$

$$\sum_{i=0}^{k-1} \binom{m}{i} p(n, w_1, w_2, d, d')^i (1 - p(n, w_1, w_2, d, d'))^{(m-i)},$$

Given a significance threshold α , we say that M_1 and M_2 have RPP $[d, d']$ if $\Pr(X(m, n, w_1, w_2, d, d') \geq k) < \alpha$.

3 RESULTS

We verify that HNF localization can be used to predict expression in islets and liver, and demonstrate that occurrences of motif pairs studied by Krivan and Wasserman [22] are better predictors of expression than HNF localization. We identify single motifs and motif pairs that predict HNF localization and expression in islets and liver.

3.1 Correlating Binding and Expression

Guided by established biological knowledge [23, 36], Krivan and Wasserman [22] observed that the presence of motif modules composed of HNF1, HNF3, HNF4, C/EBP and Sp1 can be used to predict expression in liver. They selected 16 genes that are known to be expressed in adult liver, and demonstrated that the corresponding promoters contained occurrences of binding sites for these factors. Odom *et al.* [28] studied the relationship between HNF1, HNF4 and HNF6 localization and RNA Polymerase II (PolII) localization in islets and liver. They showed that the vast majority of promoters localized with HNF4 are also localized with PolII and just under half of the promoters localized with PolII are also localized with at least one of the HNF factors.

We examine the relationship between localization of HNF factors and expression of the corresponding genes in liver and islets using the ChIP-chip data of Odom *et al.* [28] and expression data of Su *et al.* [34]. We refer to the six ChIP-chip experiments of Odom *et al.* [28] as HNF1-Liver, HNF1-Islets, HNF4-Liver, *etc.*

We tested for correlation between HNF localization and expression, and found that in all cases except HNF1-Islets, genes with promoters exhibiting HNF1, HNF4 or HNF6

Factor	Islets							Liver						
	PFG	FG	TP	T	PFG/TFG	TP/T	P	PFG	FG	TP	T	PFG/TFG	TP/T	P
HNF1	30	79	3544	9836	0.38	0.36	0.400	90	174	2670	9836	0.52	0.27	5.9e-12
HNF4	529	1136	3544	9836	0.47	0.36	5.9e-14	496	1250	2670	9836	0.40	0.27	4.0e-13
HNF6	80	161	3544	9836	0.50	0.36	2.6e-04	80	180	2670	9836	0.44	0.27	4.8e-07
PolII	952	1915	3544	9836	0.49	0.36	0	897	2364	2670	9836	0.38	0.27	0

Table 1. Correlation between localization of HNF1, HNF4 and HNF6, and expression of corresponding genes in liver and islets. PFG (Positive Foreground) = # of promoters bound by factor with corresponding gene expressed in tissue. FG (Total Foreground) = # of promoters bound by factor and examined by Su *et al.* [34]. TP (Total Positive) = # of promoters corresponding to genes expressed in tissue. T (Total) = # of examined promoters. P = *p*-value for PFG, FG, TP and T.

localization are significantly more likely to be expressed in the corresponding tissue. To determine statistical significance, we use a binomial distribution (*p*-val is calculated as $\sum_{j>k}^m \binom{m}{j} p^j (1-p)^{m-j}$), where the expression probability *p* is equal to the ratio between the number of promoters with expressed genes and the number of tested promoters, *m* is the number of localized promoters of genes with known expression levels, and *k* is the number of localized promoters of expressed genes. We used a significance threshold of 0.001 (Table 1).

To determine whether motif co-occurrences for factor pairs in HNF1, HNF3, HNF4, HNF6, C/EBP and Sp1 (which were used by Krivan and Wasserman [22]) are better expression predictors than localization of HNF factors alone, we again use a binomial distribution test. We assume that genes with localized promoters are equally likely to be expressed, setting *p* to be the ratio between localized promoters with expressed genes and localized promoters of genes tested by Su *et al.* [34]. Selecting individual motif-score thresholds to minimize *p*-value, *m* is the number of promoters with motif co-occurrences scoring above threshold and *k* is number of expressed genes whose promoters include motif co-occurrences scoring above threshold. We say that a motif pair has *improved prediction of expression* if co-occurrences of the motifs in localized promoters lead to a better prediction of expression than localization alone (binomial distribution as described above; threshold of 0.01). We used TRANSFAC matrices M00132, M00411, M00639, M00770, M00724 and M00931 as binding site models for HNF1, HNF4, HNF6, C/EBP, HNF3 and Sp1, and the results are presented in Table 2.

3.2 Individual Binding Site Motifs

We compared RIV of the top TRANSFAC, DME-X, and MDModule motifs, for each ChIP-chip experiment (Table 3). Top DME-X motifs consistently resemble the top TRANSFAC motifs, while occurrences of motifs produced by MDModule display weaker correlation to the localization of HNF1, HNF6, and HNF4 in islets. Occurrences of TRANSFAC HNF4 and HNF6 motifs, while correlating well with HNF4 and HNF6 localization, have weaker correlation than occurrences of motifs associated with GABP and Clox motifs. This may be due to aspects of our method (*e.g.* method of scoring occurrences) or poor characterizations of binding sites for

those factors, but it may also be an indication that HNF4 and HNF6 localization is greatly influenced by cofactor binding.

For HNF1-Liver and HNF1-Islets, the TRANSFAC motif with highest RIV is a known binding site motif for HNF1. The DME-X motifs with highest RIV have RIV similar to that of the TRANSFAC HNF1 binding site motif and strongly resemble this motif. The MDModule motifs for HNF1-Liver and HNF1-Islets have smaller RIV, and while AT-rich, show no resemblance to known HNF1 binding site motifs.

It is not surprising that the motif correlating best with HNF1 localization (for liver and islets) is a known HNF1 motif from TRANSFAC. HNF1 is well studied, it binds with high sequence specificity, and its motif is well characterized. The top DME-X motifs, and the two TRANSFAC HNF1 motifs M00132 and M00790 have a similar pattern. Odom *et al.* [28] used a contingency table test to show that M00790 occurrences have high correlation with HNF1 localization. We found that the 16-position wide M00132 motif has a higher RIV than the 19-position wide M00790 motif, in both liver and islets. We tested the effect of removing the additional 3 positions from M00790, and found the resulting motif to have greater RIV than M00790 in both liver and islets (Islets: 25 vs. 21% RIV; Liver: 16 vs. 15% RIV). We conjecture that M00790 includes unnecessary columns that reduce its predictive ability, and suspect that many TRANSFAC motifs have a similar problem.

For HNF4-Islets, the TRANSFAC and DME-X motifs showed much greater RIV with HNF4 localization than MDModule motifs. The top TRANSFAC motif is associated with Elk-1, and the top DME-X motif strongly resembles a motif associated with GABP. Both GABP and Elk-1 are ETS-class factors, and the shorter GABP motif appears to be contained in the longer Elk-1 motif. Motifs identified by DME-X and MDModule in HNF4-Liver were nearly identical to those identified in HNF4-Islets (8% RIV for both); the top TRANSFAC motif (7% RIV) is associated with E2F1.

Of the three HNF factors, HNF4 occupies the largest number of promoters, binding 1378 and 1521 promoters in islets and liver respectively, compared to 103 to 211 promoters bound by HNF1 and HNF6. Since we associate a larger number of targets with larger functional complexity, we conjecture a greater importance of co-factors for HNF4 binding than for binding of HNF1 and HNF6. Possible co-factors for HNF4

Factor	2nd Factor	CE Islets	CE Liver	Factor	TF2	CE Islets	CE Liver	Factor	TF2	CE Islets	CE Liver
HNF1	HNF4	0.036	0.001	HNF4	HNF1	0.001	0.001	HNF6	HNF1	0.123	0.012
	HNF6	0.037	0.077		HNF6	0.001	0.006		HNF4	0.007	0.038
	C/EBP	0.071	0.017		C/EBP	0.003	0.002		C/EBP	0.123	0.026
	HNF3	0.062	0.009		HNF3	0.001	0.002		HNF3	0.123	0.083
	Sp1	0.008	0.006		Sp1	0.019	0.054		Sp1	0.123	0.008

Table 2. For each ChIP experiment, whether a pair of factors (that includes the immunoprecipitated factor) better predicts expression in liver and islets than the localization of that factor alone. Correlation with expression (CE) is quantified by a p -value as calculated using a binomial distribution (described in Section 3.1).

identified by our analysis include Elk1, GABP, E2F1 and AP2, each having predicted sites that correlate with HNF4 binding.

The top TRANSFAC motifs in HNF6-Islets and HNF6-Liver correspond to the CDP and Clox factors, which are splice variants of the mClox gene [1]. CDP and Clox, like HNF6, are homeo-domain factors and are known to repress transcription in liver by displacing HNF1 binding [2]. The top DME-X motifs also resemble a known Clox motif containing the palindromic ATCGAT pattern, and the top DME-X motif interestingly has much higher RIV in HNF6-Liver. Since the ends of the Clox and CDP motifs appear degenerate, we tested their predictive ability with the ends removed (a similar test is described above for the TRANSFAC M00790 HNF1 motif). Removing the first and last position of the CDP motif M00104 increased % RIV for HNF6-Islets to 20%; removing the first and last two positions of the Clox motif M00103 increased the % RIV for HNF6-Liver to 22%.

3.3 Interactions Among Top Motifs

For each experiment, motifs from the set C_2 of candidate motifs deemed good predictors of binding were examined by MARSMotif, and the results are presented in Table 4. Results are not presented for HNF1-Islets or HNF1-Liver because no significant interactions were identified.

Three pairs of interacting motifs were identified for each of HNF4-Islets and HNF4-liver. For HNF4-Islets, the first interacting pair consists of DME-X motifs, including a motif similar to a TRANSFAC matrix for Elk1, and one with no strong similarity to TRANSFAC motifs that may be novel. The second interacting pair includes TRANSFAC motifs associated with E2F1 and StuAp, which have binding domain homology to HNF3 α . The same StuAp motif was found to interact with a CG-rich motif, identified by MDModule, that resembles a TRANSFAC motif for AP2. For HNF4-Liver, we found interactions between a binding motif for AP2 and both a motif for ZF5 and an MDModule motif that resembles Sp1 (CG-rich). Interactions between AP2 and Sp1 have been observed through an immunoprecipitation experiment [40], and the factors are known to interactively regulate basal promoter activity in liver [37]. We also identified an interaction, that is a significant predictor of expression, between an HNF4 motif and a motif identified by DME-X resembling a TRANSFAC motif associated with Staf.

For both HNF6-Liver and HNF6-Islets we detected an interaction between motifs for HNF6 and CDP, and in HNF6-Islets we detected an interaction between motifs for CDP and Elk-1. Interaction between Elk-1 and C/EBP β (known to be active in liver) has been demonstrated [16], and Elk-1 has been identified as a regulator in liver and pancreas (we are not aware of previous studies showing interaction between these factors).

3.4 Interactions Identified in Motif Neighborhoods

For each experiment, and each motif from the set C_2 , a neighborhood search was performed, producing sets C_3 of motifs that co-localize with a motif from C_2 . All pairs from a C_3 set with a significant RIV and a significant relative positional preference are presented in Table 5.

For HNF1-Islets, we identified three interacting pairs that include a motif resembling the HNF1 motif (including the HNF1 motif itself). One of these interactions also included a motif associated with C/EBP, and another included a motif resembling the known binding motif for NF- κ B. Both C/EBP and NF- κ B are known to interact with HNF1 [39, 22, 30, 12]. For HNF1-Liver, we identified two interactions, one of which is between motifs associated with HNF1 and CDP. CDP is known to displace HNF1 binding [2], and the interaction between the HNF1 and CDP motifs is one of two that we have identified to improve prediction of expression.

For HNF4-Islets, we found evidence for interactions between motifs produced by DME-X and MDModule. One of the DME-X motifs has a strong resemblance to a TRANSFAC motif associated with GABP (known functional in liver [9]), and the novel palindromic CG-rich MDModule motif weakly resembles the CG-rich AP-2 motif. In both interactions, the motifs are sufficiently distinct with divergence well above our similarity threshold, but their occurrences often overlap. In HNF4-Liver, we identified interactions involving TRANSFAC motifs associated with HNF4 and HNF4 α . Most interesting among these are interactions that involve the HNF4 motif and novel DME-X and MDModule motifs. The MDModule motif is a CG-rich palindrome whose co-occurrence with the HNF4 α motif improves prediction of expression.

For HNF6-Islets we identified interactions between motifs associated with HNF6 and Oct1, and between a motif associated with FOXD3 and a DME-X motif resembling the motif associated with Oct1. For HNF6-Liver we identified interactions between a TRANSFAC HNF6 motif and two other

Experiment	TRANSFAC Motif	%RIV	TF	DME-X Motif	%RIV	TF	MDModule Motif	%RIV	TF
HNF1-Islets		28%	HNF1		28%	HNF1		6%	TBP
HNF1-Liver		16%	HNF1		15%	HNF1		1%	FOXP
HNF4-Islets		16%	Elk-1		20%	GABP		12%	AP2
HNF4-Liver		7%	E2F1		8%	GABP		8%	AP2
HNF6-Islets		18%	CDP		23%	Clox		5%	CDP
HNF6-Liver		19%	Clox		28%	Clox		4%	CDP

Table 3. TRANSFAC, DME-X and MDModule motifs with greatest RIV. For DME-X and MDModule motifs, we give the name of the closest matching TRANSFAC motif, by divergence. Divergences for DME-X motifs range from 0.16 for Clox in HNF6-liver to 0.68 for HNF1 in HNF1-liver. Divergences for MDModule motifs range from 1.22 for TBP in HNF1-Islet to 1.48 for CDP in HNF6-Islet.

TRANSFAC motifs associated with CDP and Oct1. While Oct1 is known to interact with HNF1 [41, 20] we are not aware of any documented interactions between Oct1 and HNF6.

4 CONCLUSION

We presented a comprehensive method for identifying binding site motifs and motif pairs from ChIP-chip data that incorporates several features that are new to ChIP-chip analysis. Our motif discovery algorithm incorporates factor localization data directly into motif search. Regression is used to evaluate how well individual motifs predict factor localization, and multivariate regression is used to evaluate localization prediction of interacting motif pairs. Co-localizing pairs of motifs are identified by searching the sequence neighborhood of top individual motifs, and relative positional preference is evaluated to measure significant conservation of distance between motif occurrences.

We applied our method to data from ChIP-chip experiments of Odom *et al.* [28] on HNF factors in liver and pancreatic islets. Our results demonstrate that, aside from the novel motifs, top individual motifs identified by our method have strong similarity to the best performing known motifs from TRANSFAC, and often provide a better prediction of factor localization. We showed that this method can also be used to identify pair-wise interactions between top motifs and identify weaker co-localized motifs. MARSMotif and the relative positional preference measure can be used identify motif pairs with statistically significant co-localization and prediction of factor localization.

We believe that novel motifs that are similar to previously characterized motifs, but have better correlation to factor localization, provide a better characterization of the binding sites. Known motifs are often derived from a limited number of experimentally verified binding site sequences, and include positions that do not appear to help predict factor localization. Deleting flanking positions from known motifs for HNF1, Clox and CDP improves their ability to predict localization. Our study underscores the importance of using *de novo* motif discovery tools in combination with experimental data, and

indicates that using computational methods in large scale analysis of binding data may provide better characterizations of binding site motifs.

We extended work by Krivan and Wasserman [22], demonstrating that HNF localization is correlated with expression, and showing that occurrences of motif pairs can be used to predict expression in liver and islets with greater accuracy than HNF localization alone. We identified motif pairs whose occurrences are correlated with HNF localization and expression in liver. These pairs include motifs associated with HNF1 and CDP, as well as novel motifs that pair with motifs associated with HNF4 and HNF4 α . Surprisingly, occurrences of HNF4 and HNF6 motifs alone are not the best single motif predictors of HNF4 and HNF6 localization, but occurrences of motif pairs that include these motifs are excellent predictors.

The DME-X motif discovery algorithm rewards motifs for occurring in sequences according to weights derived from the localization values for the sequences. We used two weighting schemes, both performing well in our experiments, and neither consistently outperforming the other. Further research using a more diverse set of ChIP-chip experiments will be required to determine the appropriate functions for incorporating ChIP-chip localization values into the search process. Finally, we feel that the ability of DME-X to use arbitrary weights assigned to sequences will be effective in other contexts, such as motif discovery from expression data, where experimentally obtained values are associated with the sequences. Use of this algorithm in each different context will require additional research to identify appropriate functions to map the experimental values to sequence weights in DME-X.

ACKNOWLEDGMENT

A.D. Smith and P. Sumazin contributed equally to this work. We thank J. Hogenesch and J. Walker for the tissue-specific expression data, D. Odom and R. Young for the ChIP-chip data and probes used for their custom array, and BIOBASE for providing access to TRANSFAC. This work is supported by NIH grants GM060513 and HG001696, and NSF grants DBI-0306152 and EIA-0324292.

Experiment	Name	Logo	Match	Name	Logo	Match	RPP	CE
HNF4-Islets	DME-X		ELK1	DME-X		-	3 - 61	0.022
HNF4-Islets	M00940		E2F1	M00263		StuAp	1 - 96	0.174
HNF4-Islets	MDModule		AP2	M00263		StuAp	13 - 80	0.191
HNF4-Liver	M00189		AP2	M00716		ZF5	13 - 65	0.062
HNF4-Liver	M00189		AP2	MDModule		Sp1	39 - 203	0.144
HNF4-Liver	M00411		HNF4	DME-X		STAF	88 - 131	0.009
HNF6-Islets	M00104		CDP	M00025		Elk-1	171 - 174	0.030
HNF6-Islets	M00104		CDP	M00639		HNF6	1 - 132	0.122
HNF6-Liver	M00104		CDP	M00639		HNF6	1 - 60	0.017

Table 4. For each ChIP-chip experiment, pairs of motifs that were identified by MARSMotif as statistically significant ($p < 10^{-3}$), and have a statistically significant ($p < 10^{-4}$) relative positional preference (RPP). RPP is defined in Section 2.6 and correlation with expression (CE) is defined in Section 3.1. Motifs accessions are specified for TRANSFAC motifs, but no accessions are available for novel motifs identified by MDModule and DME-X.

Experiment	Name	Logo	Match	Name	Logo	Match	RPP	CE
HNF1-Islets	DME-X		HNF1	M00999		AIRE	35 - 84	0.073
HNF1-Islets	DME-X		HNF1	M00621		C/EBPδ	11 - 15	0.032
HNF1-Islets	M00132		HNF1	DME-X		NF-κB	33 - 37	0.017
HNF1-Islets	M00327		Pax3	DME-X		-	29 - 31	0.276
HNF1-Liver	M00132		HNF1	M00106		CDP	1 - 6	3.2e-4
HNF1-Liver	M00132		HNF1	DME-X		Ik3/Staf	9 - 15	0.162
HNF4-Islets	DME-X		GABP	DME-X		-	1 - 11	0.101
HNF4-Islets	MDModule		AP2	DME-X		-	8 - 10	0.282
HNF4-Liver	DME-X		GABP	M00135		Oct1	6 - 24	0.025
HNF4-Liver	DME-X		GABP	M00770		C/EBP	0 - 0	0.147
HNF4-Liver	M00158		HNF4	MDModule		Sp1	12 - 18	0.091
HNF4-Liver	M00764		HNF4	DME-X		GABP	11 - 11	0.062
HNF4-Liver	MDModule		ETF	M00189		AP2	2 - 16	0.157
HNF4-Liver	MDModule		ETF	M00716		ZF5	1 - 22	0.039
HNF4-Liver	M00411		HNF4α	MDModule		-	7 - 7	0.007
HNF4-Liver	M00411		HNF4α	DME-X		-	0 - 13	0.012
HNF6-Islets	M00639		HNF6	M00138		Oct1	4 - 4	0.122
HNF6-Islets	DME-X		CCAAT	DME-X		Oct1	10 - 13	0.140
HNF6-Islets	M00130		FOXO3	DME-X		Oct1	0 - 11	0.010
HNF6-Islets	DME-X		GATA4	M00096		Pbx1	13 - 13	0.338
HNF6-Islets	DME-X		STAT3	MDModule		-	8 - 13	0.172
HNF6-Islets	DME-X		GABP	DME-X		-	1 - 3	0.015
HNF6-Liver	M00639		HNF6	M00104		CDP	1 - 28	0.017
HNF6-Liver	M00639		HNF6	M00138		Oct1	4 - 11	0.060

Table 5. Pairs with statistically significant RIV ($p < 10^{-3}$) and RPP ($p < 10^{-4}$) that were identified by neighborhood search (*i.e.* motifs from C_3).

REFERENCES

- [1] V. Andres, M. D. Chiara, and V. Mahdavi. A new bipartite DNA-binding domain: cooperative interaction between the cut repeat and homeo domain of the cut homeo proteins. *Genes Dev*, 8(2):245–257, 1994.
- [2] T. J. Antes, J. Chen, A. D. Cooper, and B. Levy-Wilson. The nuclear matrix protein cdp represses hepatic transcription of the human cholesterol-7 α hydroxylase gene. *J. Biol. Chem.*, 275(34):26649–26660, 2000.
- [3] J. Buhler and M. Tompa. Finding motifs using random projections. *J Comput Biol.*, 9(2):225–242, 2002.
- [4] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet.*, 27(2):167–171, 2001.
- [5] M. Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, 1998.
- [6] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS*, 100(6):3339–44, 2003.
- [7] D. Das, N. Banerjee, and M. Zhang. Interacting models of cooperative gene regulation. *PNAS*, 101(46):16234–9, 2004.
- [8] D. Das and M. Q. Zhang. Adaptively Inferring cis-Regulatory Architecture in Human Genome, 2005. Submitted.
- [9] K. Du, J. I. Leu, Y. Peng, and R. Taub. Transcriptional Up-regulation of the Delayed Early Gene HRS/SRp40 during Liver Regeneration. INTERACTIONS AMONG YY1, GA-BINDING PROTEINS, AND MITOGENIC SIGNALS. *J. Biol. Chem.*, 273(52):35208–35215, 1998.
- [10] E. Eskin. From profiles to patterns and back again: a branch and bound algorithm for finding near optimal motif profiles. In *Proceedings of the eighth annual international conference on Computational molecular biology*, pages 115–124. ACM Press, 2004.
- [11] J. W. Fickett. Coordinate positioning of mef2 and myogenin binding sites. *Gene*, 172:GC19–GC32, 1996.
- [12] M. S. Figueiredo and G. G. Brownlee. cis-acting elements and transcription factors involved in the promoter activity of the human factor VIII gene. *J. Biol. Chem.*, 270(20):11828–11838, 1995.
- [13] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–142, 1991.
- [14] F. Greil, I. van der Kraan, J. Delrow, J. F. Smothers, E. de Wit, H. J. Bussemaker, R. van Driel, S. Henikoff, and B. van Steensel. Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. *Genes Dev*, 17(22):2825–2838, 2003.
- [15] D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.
- [16] M. Hanlon, L. M. Bundy, and L. Sealy. C/EBP beta and Elk-1 synergistically transactivate the c-fos serum response element. *BMC Cell Biol.*, 1:2:1186–, 2000.
- [17] S. Hannehalli and S. Levy. Predicting transcription factor synergism. *Nucleic Acids Res.*, 30(19):4278–4284, 2002.
- [18] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, 2001.
- [19] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.
- [20] Y. Ishii, A. J. Hansen, and P. I. Mackenzie. Octamer transcription factor-1 enhances hepatic nuclear factor-1 α -mediated activation of the human UDP glucuronosyltransferase 2B7 promoter. *Mol. Pharmacol*, 57(5):940–947, 2000.
- [21] S. Karlin and V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257(5066):39–49, 1992.
- [22] W. Krivan and W. W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, 11:1559–1966, 2001.
- [23] E. Kistaki and I. Talianidis. Modulation of hepatic gene expression by hepatocyte nuclear factor 1. *Science*, 277(5322):109–112, 1997.
- [24] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.
- [25] J. S. Liu, C. E. Lawrence, and A. Neuwald. Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *J Am Stat Assoc.*, 90:1156–1170, 1995.
- [26] J. S. Liu, X. Liu, and D. L. Brutlag. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Proceedings of PSB*, volume 6, pages 127–138, 2001.
- [27] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, 2003.
- [28] D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. *Science*, 303(5662):1378–1381, 2004.
- [29] P. Pevzner and S. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In B. et al., editor, *Proceedings of the Annual International Symposium on Intelligent Systems for Molecular Biology*, pages 269–278. AAAI Press, 2000.
- [30] M. Raymondjean, A. L. Pichard, C. Gregori, F. Ginot, and A. Kahn. Interplay of an original combination of factors: C/EBP, NFY, HNF3, and HNF1 in the rat aldolase B gene promoter. *Nucleic Acids Res.*, 19(22):6145–6153, 1991.
- [31] B. Ren and B. D. Dynlacht. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol*, 376:304–315, 2004.
- [32] A. D. Smith, P. Sumazin, and M. Q. Zhang. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *PNAS*, 102(5):1560–1565, 2005.
- [33] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.
- [34] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–7, 2004.
- [35] W. Thompson, E. C. Rouchka, and C. E. Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, 31(13):3580–3585, 2003.
- [36] F. Tronche, F. Ringeisen, M. Blumenfeld, M. Yaniv, and M. Pontoglio. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.*, 266(2):231–245, 1997.
- [37] C. Uchida, T. Oda, T. Sugiyama, S. Otani, M. Kitagawa, and A. Ichiyama. The role of Sp1 and AP-2 in basal and protein kinase A-induced expression of mitochondrial serine:pyruvate aminotransferase in hepatocytes. *J. Biol. Chem.*, 277(42):39082–39092, 2002.
- [38] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278:167–181, 1998.
- [39] K. Wu, D. Wilson, C. Shih, and G. Darlington. The transcription factor HNF1 acts with C/EBP alpha to synergistically activate the human albumin promoter through a novel domain. *J. Biol. Chem.*, 269(2):1177–1182, 1994.
- [40] Y. Xu, S. Porntadavity, and D. K. St Clair. Transcriptional regulation of the human manganese superoxide dismutase gene: the role of specificity protein 1 (Sp1) and activating protein-2 (AP-2). *Biochem. J.*, 362(Pt 2):401–412, 2002.
- [41] D. X. Zhou and T. S. Yen. The ubiquitous transcription factor Oct-1 and the liver-specific factor HNF-1 are both required to activate transcription of a hepatitis b virus promoter. *Mol. Cell Biol.*, 11(3):1353–1359, 1991.