BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 4 2006, pages 423–429
doi:10.1093/bioinformatics/bti815

Sequence analysis

A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data

Kenzie D. MacIsaac[1], D. Benjamin Gordon[2], Lena Nekludova[2], Duncan T. Odom[2], Joerg Schreiber[2], David K. Gifford[1], Richard A. Young[2,3] and Ernest Fraenkel[1,2,4,*]

[1]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA, [2]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA, [3]Department of Biology and [4]Division of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received on September 14, 2005; revised on November 14, 2005; accepted on December 1, 2005
Advance Access publication December 6, 2005
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Genome-wide chromatin-immunoprecipitation (ChIP-chip) detects binding of transcriptional regulators to DNA *in vivo* at low resolution. Motif discovery algorithms can be used to discover sequence patterns in the bound regions that may be recognized by the immunoprecipitated protein. However, the discovered motifs often do not agree with the binding specificity of the protein, when it is known.

Results: We present a powerful approach to analyzing ChIP-chip data, called THEME, that tests hypotheses concerning the sequence specificity of a protein. Hypotheses are refined using constrained local optimization. Cross-validation provides a principled standard for selecting the optimal weighting of the hypothesis and the ChIP-chip data and for choosing the best refined hypothesis. We demonstrate how to derive hypotheses for proteins from 36 domain families. Using THEME together with these hypotheses, we analyze ChIP-chip datasets for 14 human and mouse proteins. In all the cases the identified motifs are consistent with the published data with regard to the binding specificity of the proteins.

Availability: THEME is freely available for download.

Contact: fraenkel-admin@mit.edu

Supplementary information: http://fraenkel.mit.edu/THEME

1 INTRODUCTION

Transcriptional regulatory proteins determine the distinct set of genes that are expressed in particular tissues and in response to environmental changes. These proteins are directed to their targets by short, often degenerate, sequence patterns. There is great interest in developing computational analyses of high-throughput data to identify and interpret these regulatory interactions on a genome-wide level (Bar-Joseph *et al*., 2003; Conlon *et al*., 2003; Segal *et al*., 2003; Beer and Tavazoie, 2004; Hall *et al*., 2004; Harbison *et al*., 2004; Hong *et al*., 2005; Kelley and Ideker, 2005; Segal *et al*., 2005; Smith *et al*., 2005). One particularly important data source for these analyses is high-throughput chromatin-immunoprecipitation (ChIP-chip), which identifies regions occupied by regulators *in vivo*. Applying motif discovery algorithms to these data can reveal the sequence patterns present in the bound regions that reflect the binding specificity of each regulator.

Motif discovery algorithms identify common sequence patterns ('motifs') among a set of larger sequences (Lawrence *et al*., 1993; Bailey and Elkan, 1994; Roth *et al*., 1998; Hertz and Stormo, 1999; Stormo, 2000; Liu *et al*., 2001, 2002; Sinha and Tompa, 2002; Bulyk, 2003; Sandelin and Wasserman, 2004; Xing and Karp, 2004). Although these patterns may represent the binding specificity of regulators, the biological meaning of the many identified motifs is often unclear. In the study of Harbison *et al*. (2004), several programs were used to analyze binding data for 203 yeast proteins. Although these programs identified over 68 000 motifs, they failed to recover the specificities of 138 of 203 proteins. The majority of the discovered motifs were probably either non-functional patterns overrepresented by chance or sites for other proteins that function either synergistically, antagonistically or independently of the immunoprecipitated protein.

Identification of functionally relevant motifs in genomes of higher eukaryotes is even more challenging than in yeast. Regulatory regions in higher eukaryotic genes are substantially larger and more complex, and sequence features common in mammalian genomes such as CpG islands further confound motif discovery methods. A recent evaluation of 13 motif discovery tools demonstrated the limitations of these techniques for analyzing mammalian promoter sequences (Tompa *et al*., 2005). There is an immediate need for more powerful and robust approaches, as new ChIP-chip data have begun to emerge for human and mouse tissues (Li *et al*., 2003; Cam *et al*., 2004; Odom *et al*., 2004; Bernstein *et al*., 2005; Boyer *et al*., 2005; Brodsky *et al*., 2005).

We have developed a hypothesis-driven approach that is effective in analyzing ChIP-chip data from human and mouse tissues. Our algorithm, called THEME, is not designed for motif discovery, per se. Rather, it uses principled statistical methods to test hypotheses about the binding specificity of the immunoprecipitated protein. THEME evaluates hypotheses based on their ability to predict accurately which sequences from a held-out test set were bound by the protein and which were not. The most predictive hypothesis

*To whom correspondence should be addressed.

is either accepted or rejected by comparing its predictive value with that of motifs derived by applying the same algorithm to randomly selected input sequences. This approach complements the standard motif discovery techniques.

By deriving initial hypotheses from the binding sites of related proteins in the TRANSFAC database (Matys *et al*., 2003) we can determine whether or not there is a motif that explains the binding data and is consistent with the domain structure of the transcriptional regulator. Most DNA-binding domains show a limited repertoire of sequence specificity, and family members usually recognize variants of the same core sequences. For example, many bZIP proteins bind to variations of the AP-1 site (TGANTCA), the ATF-CREB (TGANNTCA) or the C/EBP site (ATTKC). Similarly, HLH proteins often bind to E-boxes (CANNTG), and differ largely in their specificity for the two middle base pairs and the flanking regions. THEME provides a method for determining if the specificity of the immunoprecipitated protein is similar but not necessarily identical to the prototypes for its family. Using new ChIP-chip results as well as previously published data, we demonstrate the utility of this approach by accurately identifying motifs that correspond to the specificity of 14 mammalian proteins from 10 different domain families.

2 ALGORITHM AND METHODS

THEME is a method for testing hypotheses with regard to the DNA-binding specificity of proteins (Fig. 1). The initial hypothesis consists of a position weight matrix (PWM) (Stormo, 2000) model of the binding specificity, describing the probability distribution for bases at each position of a binding site. Hypotheses can be derived from a variety of sources. Input consists of a set of sequences bound by the protein of interest (the positive data), as well as sequences that are not bound (the negative data). Using cross-validation, hypotheses are refined with training data and evaluated on held-out test data to identify the most predictive motif (Fig. 2). The statistical significance of the best motif is then determined.

2.1 Hypothesis generation—Family Binding Profiles

While hypotheses from any source can be tested, here we derive hypotheses from known binding sites of proteins that belong to the same DNA-binding domain family as the immunoprecipitated protein. Individual members of protein families generally bind related DNA sequences due to structural constraints. These preferences can be represented as PWMs and have been designated Family Binding Profiles (Sandelin and Wasserman, 2004). Family Binding Profiles capture sequence features common to the binding sites of many members of the family, but are consequently poor representations of the specificity of individual family members.

We derive profiles from unaligned binding sites in the TRANSFAC v7.2 database (Matys *et al*., 2003). Pfam hidden-Markov models (Bateman *et al*., 2004) identify 37 families of DNA-binding domains in TRANSFAC that each contain at least 4 proteins and 30 sites. We pool all the binding sites reported in TRANSFAC for members of a family and submit these sequences to two motif discovery programs: AlignACE (Roth *et al*., 1998) and DimerFinder. Our approach successfully identifies one or more motifs for 36 families (Supplementary Table S1 and online Supplementary material). On average, a family is represented by three profiles. In some cases, profiles discovered by AlignACE and DimerFinder are very similar.

AlignACE is run with the default options, except that uniform background base frequencies are used, and for better reproducibility the program is applied 10 times to each family with different random number seeds. All the resulting motifs for each family are grouped together and ranked according to the enrichment score (Harbison *et al*., 2004) of the motif. The top-ranked motif is used as a profile. We iteratively search for additional

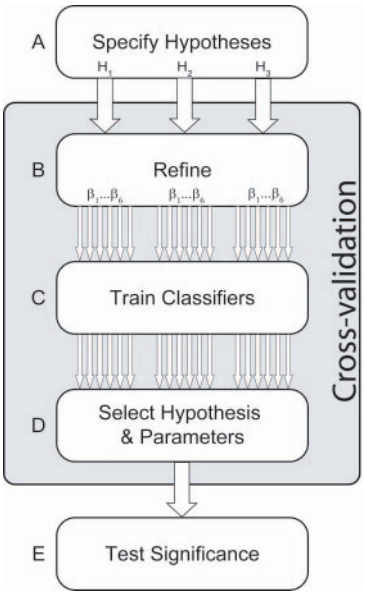

Fig. 1. Overview of THEME. (**A**) THEME requires that one or more binding hypotheses be specified in the form of a PWM. (**B**) The data are partitioned for cross-validation. Using only the training data, the hypotheses are refined using the EM algorithm at varying parameter settings. (**C**) The refined hypotheses are used to train a classifier and the classification error on the held-out test data is evaluated. (**D**) The hypothesis and parameter setting that yields the best mean cross-validation error is identified. (**E**) The statistical significance of the observed cross-validation error is estimated by comparing it with a distribution obtained by applying the hypothesis, with the same parameter settings, to randomly chosen promoter sequences.

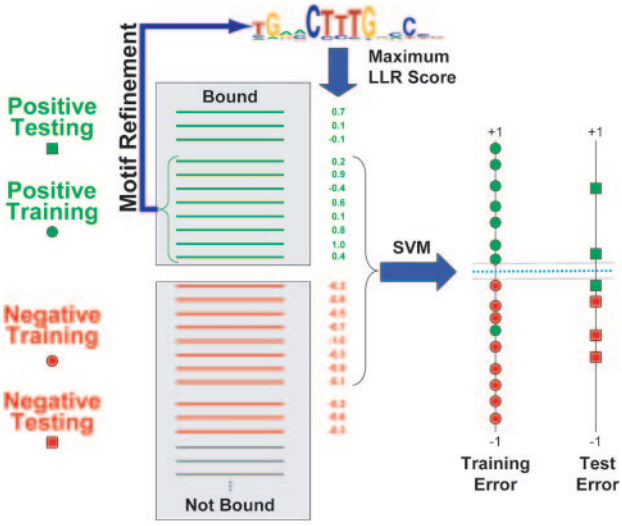

Fig. 2. Hypothesis refinement and cross-validation by THEME. Positive sequences are those that were bound in the ChIP-chip experiments. The remaining sequences on the array are the negative set. Positive and negative data are separated into training and test sets. The positive training examples are used to refine the hypothesis using EM. All training and test examples are then mapped to a one-dimensional feature space by evaluating the LLR score of their best match to the refined hypothesis. A linear-kernel SVM classifier is trained using both the positive and negative training examples. This classifier is then used to evaluate the classification error on the positive and negative test sets.

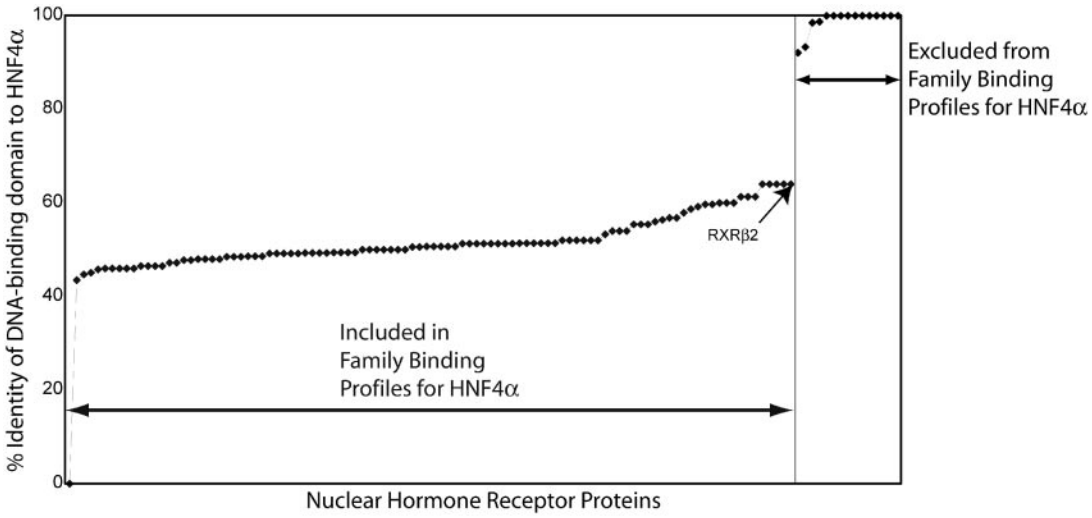

Fig. 3. Similarity of the nuclear hormone receptor DNA-binding domains to HNF4α. The graph shows the percent identity between the DNA-binding domain of HNF4α and each nuclear hormone receptor protein in TRANSFAC. Proteins with >70% sequence identity were excluded when we derived the Family Binding Profiles in Supplementary Table S2. The same threshold was used to generate the restricted profiles for all other proteins.

motifs by removing all binding sites matching this motif and applying AlignACE to the remaining sites. We repeat this process until the enrichment of the top-ranked motif is less than 40.

The profiles discovered by AlignACE do not include all of the known profiles that are characterized in the literature for some families. Many of the missing profiles are for domain families that bind DNA as homodimers. Dimeric proteins typically recognize direct or inverted repeats of short sequences separated by characteristic distances that may differ between family members. We use the motif discovery program DimerFinder to identify motifs with direct or inverted repeats. DimerFinder is a word-counting method that is described in the Supplementary methods. Source code for DimerFinder is available from our website.

2.2 Restricted Family Binding Profiles

Family Binding Profiles can be helpful even in situations where there are no close homologs of a protein of interest. To demonstrate this, we have computed a distinct set of profiles using only data in TRANSFAC for proteins with <70% sequence identity to the DNA-binding domains of the proteins analyzed in this paper (Supplementary Table S2 and online Supplementary material). All reported results use this more restricted set of profiles. For example, the Family Binding Profiles that we use to discover the specificity of HNF4α exclude binding data for all HNF4α, HNF4β and HNF4γ proteins from any species. RXRβ2 is the most similar protein to HNF4α that is included in the profiles (Fig. 3).

2.3 Hypothesis testing by cross-validation

The Family Binding Profiles for each protein are refined and tested using cross-validation to find the hypothesis that best explains the binding data. We define the set of bound probe sequences as our positive set. We produce a negative set by randomly undersampling the set of unbound probes until it is 10 times larger than the positive set. To avoid overfitting motifs to the data, we partition the sequences into test and training sets. We perform THEME hypothesis testing using the following five-step procedure:

(1) Refine the hypothesis on the positive training set

(2) Score each sequence in the training and test data using the refined model

(3) Oversample the positive training and test data

(4) Train an SVM classifier on the training examples

(5) Classify the test examples and report the classification error.

For each hypothesis we perform a grid search over two parameters (the parameter β, which measures the strength of our prior belief in the accuracy of the hypothesis and C, a parameter used as part of the regularization term in the SVM classifier). We repeat the five-step procedure over each parameter setting to determine the setting yielding the lowest 3-fold cross-validation error (Supplementary methods). Each trial takes $\sim$6 min for a typical dataset (grid search for one hypothesis with 380 sequences averaging 898 bases each) on a 2.8–3.2 GHz dual processor Intel Xeon CPU with 4 GB of memory. To evaluate the set of profiles for a family, therefore, takes on average 18 min.

Due to the non-deterministic nature of the sampling procedure, cross-validation results could, in principle, vary among trials with the same input and parameters. We compare hypotheses using three separate THEME trials with different randomly selected negative datasets. The refined motifs did not vary significantly across these trials. We report the average cross-validation errors.

The best refined motif model is the one that has the lowest mean error on the test sets after 3-fold cross-validation. The refinement for this motif is then repeated using the same parameter values and initial hypothesis, but including all the bound sequences to obtain the final reported motif.

2.3.1 Refinement Each hypothesis is refined on the positive data using the expectation-maximization (EM) algorithm with a Bayesian prior. For EM, we used the ZOOPS probability model described by Bailey and Elkan (1994). Since each hypothesis is a probabilistic weight matrix, it can be used directly in the 'E' step of EM. The E and 'M' steps are alternated until the Euclidean distance between the model in subsequent M steps are less than 10^{-3}. In the M step, the PWM model is updated using the expected counts in each position of the matrix. The change in the model during the M step is restrained using pseudocounts added to the matrix in proportions determined by the original hypothesis and the value of the β parameter. β is defined as the fraction of the total counts added to the matrix during the M step that are pseudocounts used to restrain the model. A β of 0.0 indicates that EM refinement proceeds without restraint. When $\beta = 1.0$, no refinement is carried out. Refinements occur in parallel with β values of 0.05, 0.1, 0.33, 0.5, 0.67 and 1.0. The fifth-order Markov background model used in EM was estimated from the set of all sequences represented on the

microarray for a given experiment. Our implementation is written in Python, with core EM routines written in C++ and is based on TAMO (Gordon *et al*., 2005). Source code is available from our website.

2.3.2 Classifying sequences using the refined motif In order to train a classifier and perform cross-validation, the refined hypotheses must be used to define one or more features to score each sequence. In this study, the sequences are evaluated using the log-likelihood ratio (LLR) score of the best match to the refined hypothesis. This score is an intuitive feature that measures our belief that the best match is an instance of the motif, described by the PWM model, after taking into account the single-nucleotide base distribution of the background sequences.

Typically, an arbitrary threshold is used to determine when the LLR is high enough to constitute a match (Harbison *et al*., 2004). We chose a more principled approach, using a linear-kernel support vector machine (SVM) to determine the threshold that best separates the bound and unbound sequences in the training data. The scores of the training data are scaled so that they fall between −1.0 and 1.0, and used to train the SVM at a particular setting of the parameter, C, which is used in the regularization term. The values of C tested are 1.0E−10, 1.0E−4, 1.0E−3, 1.0E−2, 0.05, 0.1, 1.0, 10.0 and 100.0. The test data are then scaled in an identical manner and classified using the SVM. The classification error of the SVM on the test data is evaluated using the optimal value of C determined from the training data.

When building classifiers from datasets with a significant imbalance in the proportion of positive and negative examples, it is important to ensure that the classifier has sufficient sensitivity to the minority class. One solution is to resample the dataset to achieve greater balance between the two classes. We combine undersampling of the negative dataset with SMOTE oversampling of the positive training and test sets so that the number of positive and negative examples is equal. This technique has been shown to improve classification performance on datasets with large disparities in the sizes of the minority and majority classes (Chawla *et al*., 2002).

2.4 Significance testing

We determine the empirical probability distribution of obtaining a mean cross-validation error for a particular Family Binding Profile under the null hypothesis that the input sequences are unrelated to the profile. We compute the distribution of cross-validation errors by running THEME multiple times using sets of sequences equal in size to the original dataset, selected at random from all those present on the microarray. These calculations are conducted with the same hypothesis and β settings as before. We assume the observed cross-validation errors are normally distributed and perform randomization runs until the standard error on our estimate of the standard deviation is ∼10%. We then compare the observed cross-validation errors with the computed distribution and perform a Z-test to assess the statistical significance of the cross-validation error achieved by a refined hypothesis.

2.5 Comparison to known specificities

The cross-validation error of the best refined hypothesis for each protein is compared with the error obtained using the protein's TRANSFAC motif (where available), which is determined by running THEME with the TRANSFAC PWM, omitting the EM refinement step. The best refined motifs are compared with TRANSFAC motifs for the same protein by calculating an inter-motif distance [as described previously (Harbison *et al*., 2004)]. We also calculate a mean and standard error on the mean of the motif's distance to all the motifs described in the TRANSFAC database for comparison.

2.6 Input sequences

The data sources for each experiment are listed in Supplementary Table S3. We extend the DNA sequences represented on the cDNA-based microarray by 250 bp upstream and downstream to account for hybridization of long shearing products. For tiled oligonucleotide arrays, we use a 500 bp window centered on each probe.

2.7 ChIP experiments

Chromatin-IP experiments were performed using self-printed DNA microarrays containing portions of promoter regions of human or mouse genes as previously described (Odom *et al*., 2004). The arrays contained 13 000 (mouse) and 19 000 (human) promoter regions. Additional details are available in the Supplementary methods. The data have been submitted to ArrayExpress under accession number E-WMIT-8.

3 RESULTS

3.1 Example

HNF4α is an important regulator of transcription in liver, and mutations in this gene cause one form of maturity onset diabetes of the young (Bell and Polonsky, 2001). Odom *et al*. (2004) reported chromatin-immunoprecipitation data for HNF4α obtained from human tissues. To analyze these data using THEME, we derived eight profiles for HNF4α using TRANSFAC data for nuclear hormone receptor proteins that are not closely related to HNF4α (Fig. 3).

Each profile was used as an initial hypothesis and refined using the positive training data for the most strongly bound genes (binding P-value < 0.001). The mean test errors for these hypotheses after 3-fold cross-validation on the HNF4α data are shown in Supplementary Table S2. The refined motif with the lowest mean cross-validation error corresponds to hypothesis H. This motif matches the HNF4α motif reported in TRANSFAC and is statistically significant, with the cross-validation error of 0.30 being over 9 SD below the mean estimated from randomized data. The final classification results indicate that ∼77% of sequences bound by HNF4α in human liver contain this motif.

3.2 Analysis of published human ChIP-chip data

We tested THEME by applying it to published ChIP-chip experiments for 10 additional human transcriptional regulators, which are members of 6 different DNA-binding domain families (Supplementary Table S3). These data are quite diverse and thus constitute a good set of experiments with which to evaluate THEME. Initial hypotheses were generated using Family Binding Profiles derived from the TRANSFAC binding sites, excluding data for close homologs, as described for HNF4α. In each case, the refined hypothesis with the best cross-validation error is statistically significant and agrees with previously reported motifs or binding sites for the protein (Table 1).

3.3 New human and mouse ChIP-chip experiments

To further demonstrate that our approach is applicable to a wide variety of transcriptional regulators, we performed genome-wide chromatin-IP experiments for three additional proteins from three different domain families: the forkhead protein HNF3β, an important liver regulator (Kaestner, 2000); the HLH protein NeuroD1, which causes MODY diabetes when haploinsufficient (Malecki *et al*., 1999) and the winged helix protein E2F4, a central regulator of the cell cycle (Trimarchi and Lees, 2002). As before, we derived the hypotheses for each protein by creating Family Binding Profiles from TRANSFAC sites, excluding data for close homologs. The resulting motifs are statistically significant and agree with the published specificity data for these proteins (Table 1).

NeuroD1 illustrates the power of THEME when there is little prior knowledge with regard to the DNA-binding specificity of a protein or

Table 1. Comparison of refined THEME hypotheses and previously reported data

Protein	Refined hypothesis				TRANSFAC			
	Motif	Error[a]	Z-score[b]	β	Motif	Error[a]	Distance[c]	Mean distance[d]
c-Rel		0.34 ± 0.00	4.93	0.5		0.38 ± 0.00	0.18	0.44 ± 0.06
HNF4α		0.30 ± 0.02	9.92	0.1		0.34 ± 0.02	0.17	0.40 ± 0.04
HNF6		0.32 ± 0.00	13.50	0.5		0.42 ± 0.00	0.29	0.44 ± 0.04
Nanog		0.42 ± 0.00	9.88	0.5	TAATSGSY[e]	N/A	N/A	N/A
Oct4		0.41 ± 0.00	15.90	0.5		0.42 ± 0.00	0.09	0.43 ± 0.05
P-CREB		0.40 ± 0.00	11.66	0.5		0.42 ± 0.00	0.13	0.43 ± 0.05
p50		0.30 ± 0.00	13.23	0.05		0.30 ± 0.00	0.20	0.43 ± 0.05
p52		0.21 ± 0.01	8.88	0.33		0.33 ± 0.01	0.25	0.46 ± 0.05
p65		0.40 ± 0.00	3.95	0.1		0.40 ± 0.01	0.20	0.44 ± 0.05
RelB		0.30 ± 0.00	9.33	0.67	N/A	N/A	N/A	N/A
Sox2		0.39 ± 0.01	22.37	0.67	AACAA[A/T]G[e]	N/A	N/A	N/A
E2F4		0.34 ± 0.00	16.15	1.0		0.36 ± 0.00	0.20	0.42 ± 0.05
HNF3β		0.39 ± 0.01	5.85	0.05		0.48 ± 0.01	0.20	0.42 ± 0.05
NeuroD1		0.35 ± 0.00	14.71	0.33	CARNTG[e]	N/A	N/A	N/A

[a]Mean 3-fold cross-validation error over three separate trials.
[b]Z-score obtained by comparing the cross-validation errors with those observed for randomization controls.
[c]Distance between the refined motif and the TRANSFAC motif.
[d]Distance between the refined motif and all motifs in TRANSFAC (mean ± SD).
[e]There are no binding sites for Nanog, NeuroD1 or Sox2 in TRANSFAC v7.2. The site listed for Nanog is taken from Mitsui *et al*. (2003), the site for NeuroD1 is from Marsich *et al*. (2003) and the site for Sox2 is from Maruyama *et al*. (2005).

that of its close homologs. The most similar protein that has known binding sites in TRANSFAC (v7.2) is the T-cell acute lymphocytic leukemia-1 protein, SCL/TAL1, which is only 48% identical to NeuroD1 in its DNA-binding domain. Nevertheless, we find a motif, sCAgcTGs, which is statistically significant, present in 97% of the bound probes on the mouse array and consistent with known sites for NeuroD1 in the promoter of Pax6 (Marsich *et al*., 2003).

3.4 Importance of hypothesis testing

Leveraging prior biological knowledge is crucial for successfully identifying the correct motif in complex mammalian datasets. To demonstrate this, we ran the THEME algorithm using an uninformative hypothesis, equal in length to the correct motif, but consisting of background nucleotide frequencies. The uninformative hypotheses produced the correct motif in only one case (HNF6). The cross-validation error for motifs derived from uninformative priors was always higher than when Family Binding Profiles were used.

Of the motif discovery programs that we tested on these data, AlignACE performed the best, discovering motifs consistent with the known specificities in six cases (Supplementary Table S4). The cross-validation errors for AlignACE motifs were always higher than those discovered by THEME. To obtain the AlignACE results, we needed to run the program multiple times using different random number seeds. A typical AlignACE calculation required 21 h to complete, compared with 18 min for THEME.

3.4.1 Deriving hypotheses with limited prior data In the absence of Family Binding Profiles THEME can be used to take advantage of other available data, such as known binding sites. To demonstrate this we derived hypotheses from each of the three known and distinct NeuroD1 binding sites (Marsich *et al*., 2003) by assigning 99% of the probability mass to the nucleotide represented in the sequence and distributing the remaining mass among the other 3 nt at each position. The PWMs were provided as hypotheses to THEME. The refined motifs match the NeuroD1 motif reported in Table 1 and display similar cross-validation errors (Supplementary Table S5).

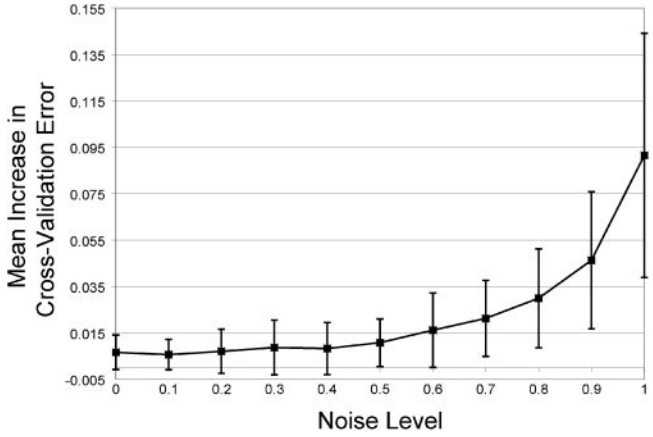

Fig. 4. Effect of noise on cross-validation error. The Family Binding Profiles yielding the lowest cross-validation error for each dataset were corrupted with varying amounts of noise to produce 11 hypotheses of gradually decreasing quality. These were used as hypotheses in THEME. The mean cross-validation error for the refined motif from each hypothesis is compared with the best hypothesis for the same dataset.

In many cases, THEME is able to identify the correct motif, even if the DNA-binding domain or binding sites of the factor are not specified. To demonstrate this, we ran THEME for each factor in Table 1, using every profile across all families as initial hypotheses. We ranked the resulting refined motifs by their cross-validation errors (Supplementary Table S6). In 10 out of 14 cases, we observe

that the correct motif, derived from a hypothesis corresponding to the factor's DNA-binding domain family, has the lowest cross-validation error. Furthermore, in 13 out of 14 cases, the correct motif and the correct family are ranked in the top 5 families (the correct family for Nanog was ranked 8th out of 36 families).

3.4.2 Noise tolerance of hypotheses THEME does not require highly accurate initial hypotheses. To demonstrate this we used THEME to refine noisy versions of the hypotheses that yielded the lowest cross-validation error for each factor. We obtained these hypotheses by combining, in various ratios, 1000 sequences derived from the uncorrupted PWM and from the background base frequencies. Noise levels of up to 40% have little effect on the cross-validation errors (Fig. 4). In 13 of the 14 datasets the motifs obtained with 40% noise are consistent with the known specificities (Supplementary Table S7).

4 DISCUSSION

Genome-wide chromatin-immunoprecipitation experiments provide a snapshot of the physical interactions of a single protein with the genome. Computational analysis of these data has the potential to reveal the sequence motifs that control transcription. However, the statistical criteria typically used in evaluating the motifs produced by motif discovery algorithms cannot directly measure the likelihood that a motif is biologically significant.

We present a hypothesis-driven approach to analyzing ChIP-chip data that differs from motif discovery programs and is complementary to these methods. Unlike motif discovery, we begin by specifying hypotheses and then establish whether these hypotheses are supported by the data. THEME is able to determine whether to accept or reject a hypothesis because it seeks to solve a classification problem. A good motif distinguishes between bound and unbound sequences in the test set. An incorrect hypothesis may produce a motif that appears significant on the training data, but it will be poorly represented in the test data. We note that cross-validation has recently been used as a method for limiting model complexity in an alternative approach to PWMs based on boosting (Hong *et al*., 2005).

4.1 Incorporating biological knowledge

Our hypothesis-testing framework is particularly valuable because it addresses the issue of interpreting motifs. THEME not only assesses whether there is a motif that can distinguish bound and unbound sequences, but also whether that motif is consistent with prior biological knowledge.

When prior biological knowledge is available, either in the form of a known DNA-binding domain or known binding sites (as for NeuroD1), the accuracy of THEME is dramatic. THEME identifies a statistically significant motif consistent with the expected specificity for all 14 datasets we analyzed. By contrast, using the cross-validated approach without an informative prior fails to identify the correct motif in all but one of these mammalian datasets. We note that the THEME motifs are of very high quality. For the proteins with motifs reported in TRANSFAC, in all cases the refined THEME motif had equivalent or better cross-validation error than the corresponding motif from TRANSFAC (Table 1). The quality of the motifs results, in part, from the algorithm's ability to determine the optimal relative weights for the binding data and the prior. No single choice of this relative weight (β) is suitable for all datasets, as can be seen in Table 1.

By testing Family Binding Profiles as hypotheses, we are able to determine whether ChIP-chip data contain a motif that is consistent with the DNA-binding domain in the immunoprecipitated protein. Analyses based on DNA-binding domains have been used previously to aid motif discovery and interpretation (Sandelin and Wasserman, 2004; Xing and Karp, 2004; Mahony *et al*., 2005; Tan *et al*., 2005). Our Family Binding Profiles extend previous efforts by adding new families and more variants within each family. We have derived Family Binding Profiles for 36 of the 37 most common DNA-binding domains. Family Binding Profiles are not appropriate for the C2H2-zinc finger family, whose members have diverse binding preferences. However, it should be possible to analyze individual members of this family by deriving initial hypotheses using the sequence composition of the key base-contacting residues (Benos *et al*., 2002; Kaplan *et al*., 2005).

4.2 Extending THEME

In the absence of information about the DNA-binding domain of the protein, THEME is often able to identify the correct motif by exhaustively testing all available Family Binding Profiles. These results suggest that THEME may be a valuable tool in the analysis of diverse data. As THEME reveals the family of the domain that produced the most predictive motif, it provides insight into regulation that cannot be obtained by motif discovery alone.

Traditional motif discovery methods could be used together with THEME to develop a more complete understanding of transcriptional regulation. For example, our analysis identified a motif for HNF4α present in 77% of the bound sequences that is likely to direct HNF4α to its highest affinity targets through direct protein–DNA contacts. Previous analysis of these data did not discover the HNF4α motif, but identified enriched motifs for several other proteins (Smith *et al*., 2005). These other proteins may serve to recruit HNF4α as a coactivator to liver- or pancreas-specific genes that do not contain matches to the HNF4α motif (Eeckhoute *et al*., 2004).

THEME could be extended to incorporate additional information or to explore higher order feature combinations. Phylogenetic conservation information, for instance, could be incorporated using one of several previously reported conservation-based motif discovery tools (Wang and Stormo, 2003; Moses *et al*., 2004; Li and Wong, 2005) in the refinement step. In addition, conservation metrics could be used as an additional feature in training the SVM classifier.

The hypothesis-based approach of THEME provides a principled method for interpreting high-throughput data sources. Combined with other techniques, THEME will allow researchers to discover the mechanistic basis for regulation in mammalian systems.

ACKNOWLEDGEMENTS

We acknowledge Oliver King, Robin Dowell, Esti Yeger-Lotem and Alan Qi for helpful comments; Tommi Jaakkola and Adrian Corduneanu for help with an earlier version of the algorithm; and Elizabeth Herbolsheimer and Elizabeth Jacobsen for technical assistance. The work was supported by the following funds: NIDDK grants DK-68655 (D.T.O. and R.Y.) and DK-70813 (D.T.O), NIH grant 1R01 HG002668-01 (K.D.M.) and NIH/NIGMS NRSA award

(D.B.G.). E.F. is a Whitehead Fellow and was funded in part by Pfizer.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bar-Joseph,Z. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32** (Database issue), D138–D141.

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

Bell,G.I. and Polonsky,K.S. (2001) Diabetes mellitus and genetically programmed defects in beta-cell function. *Nature*, **414**, 788–791.

Benos,P.V. *et al*. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.

Bernstein,B.E. *et al*. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.

Boyer,L.A. *et al*. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Brodsky,A.S. *et al*. (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.*, **6**, R64.

Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.

Cam,H. *et al*. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell*, **16**, 399–411.

Chawla,N.V. *et al*. (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.

Conlon,E.M. *et al*. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.

Eeckhoute,J. *et al*. (2004) Hepatocyte nuclear factor 4alpha enhances the hepatocyte nuclear factor 1alpha-mediated activation of transcription. *Nucleic Acids Res.*, **32**, 2586–2593.

Gordon,D.B. *et al*. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, **21**, 3164–3165.

Hall,D.A. *et al*. (2004) Regulation of gene expression by a metabolic enzyme. *Science*, **306**, 482–484.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hong,P. *et al*. (2005) A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636–2643.

Kaestner,K.H. (2000) The hepatocyte nuclear factor 3 (HNF3 or FOXA) family in metabolism. *Trends Endocrinol. Metab.*, **11**, 281–285.

Kaplan,T. *et al*. (2005) *Ab initio* prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.

Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.

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

Li,X. and Wong,W.H. (2005) Sampling motifs on phylogenetic trees. *Proc. Natl Acad. Sci. USA*, **102**, 9481–9486.

Li,Z. *et al*. (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci USA*, **100**, 8164–8169.

Liu,X. *et al*. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138.

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

Mahony,S. *et al*. (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, **21** (Suppl. 1), i283–i291.

Malecki,M.T. *et al*. (1999) Mutations in NEUROD1 are associated with the development of type 2 diabetes mellitus. *Nat. Genet.*, **23**, 323–328.

Marsich,E. *et al*. (2003) The PAX6 gene is activated by the basic helix–loop–helix transcription factor NeuroD/BETA2. *Biochem. J.*, **376**, 707–715.

Maruyama,M. *et al*. (2005) Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells. *J. Biol. Chem.*, **280**, 24371–24379.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Mitsui,K. *et al*. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, **113**, 631–642.

Moses,A.M. *et al*. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.*, 324–335.

Odom,D.T. *et al*. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.

Roth,F.P. *et al*. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Segal,E. *et al*. (2005) From signatures to models: understanding cancer using microarrays. *Nat. Genet.*, **37** (Suppl), S38–S45.

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

Smith,A.D. *et al*. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics*, **21** (Suppl. 1), i403–i412.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tan,K. *et al*. (2005) Making connections between novel transcription factors and their DNA motifs. *Genome Res.*, **15**, 312–320.

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

Trimarchi,J.M. and Lees,J.A. (2002) Sibling rivalry in the E2F family. *Nat. Rev. Mol. Cell Biol.*, **3**, 11–20.

Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.

Xing,E.P. and Karp,R.M. (2004) MotifPrototyper: a Bayesian profile model for motif families. *Proc. Natl Acad. Sci. USA*, **101**, 10523–10528.