BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i72–i79
doi:10.1093/bioinformatics/btm224

RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors

Xiaoyu Chen[1,2,4], Timothy R. Hughes[1,2] and Quaid Morris[1,2,3,*]

[1]Banting and Best Department of Medical Research, [2]Department of Medical Genetics and Microbiology, [3]Department of Computer Science, University of Toronto, Toronto, ON, Canada and [4]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

ABSTRACT

Motivation: The sequence specificity of DNA-binding proteins is typically represented as a position weight matrix in which each base position contributes independently to relative affinity. Assessment of the accuracy and broad applicability of this representation has been limited by the lack of extensive DNA-binding data. However, new microarray techniques, in which preferences for all possible K-mers are measured, enable a broad comparison of both motif representation and methods for motif discovery. Here, we consider the problem of accounting for all of the binding data in such experiments, rather than the highest affinity binding data. We introduce the RankMotif++, an algorithm designed for finding motifs whenever sequences are associated with a semi-quantitative measure of protein-DNA-binding affinity. RankMotif++ learns motif models by maximizing the likelihood of a set of binding preferences under a probabilistic model of how sequence binding affinity translates into binding preference observations. Because RankMotif++ makes few assumptions about the relationship between binding affinity and the semi-quantitative readout, it is applicable to a wide variety of experimental assays of DNA-binding preference.

Results: By several criteria, RankMotif++ predicts binding affinity better than two widely used motif finding algorithms (MDScan, MatrixREDUCE) or more recently developed algorithms (PREGO, Seed and Wobble), and its performance is comparable to a motif model that separately assigns affinities to 8-mers. Our results validate the PWM model and provide an approximation of the precision and recall that can be expected in a genomic scan.

Availability: RankMotif++ is available upon request.

Contact: quaid.morris@utoronto.ca

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

1 INTRODUCTION

Understanding protein–DNA interactions is critical to understanding genome function. Recognition of specific DNA sequences by proteins is central in processes such as control of gene expression and chromosome replication, and as a consequence plays an extensive role in shaping genome sequence. On the basis of domain structure, over 6% of all human genes encode potential DNA-binding proteins (Messina *et al.*, 2004), and the majority of sequence conserved among verterbrates is non-coding and may be regulatory (Pennacchio *et al.*, 2006; Siepel *et al.*, 2005).

Given the importance of protein–DNA interactions in computational biology, it is remarkable that the simple position weight matrix (PWM) model remains a virtually unchanged standard more than 20 years after its introduction (Berg and von Hippel, 1987). The PWM is a simple model that predicts the binding energy of a protein–DNA complex. It assumes that each DNA base in the binding site of the protein makes an independent, additive contribution to the binding energy of the protein–DNA complex. Other representations of DNA-binding preferences, such as position frequency matrices (PFMs), that assume that bases contribute independently to the binding affinity can be transformed into PWMs.

For most transcription factors (TFs), the PWM is derived from a few dozen sequences that the TF is known to bind. These sequences are most commonly obtained from *in vitro* selection (Vlieghe, 2006), a procedure that selects for and reports only the most highly specific binding sequences (Roulet *et al.*, 2002). PWMs are increasingly inferred from ChIP-chip data as well (MacIssac *et al.*, 2006), despite binding sites *in vivo* being influenced by factors other than binding specificity [e.g. protein–protein interactions and chromatin structure (Liu *et al.*, 2005)]. In fact, one recent report suggests widespread use of predicted weak DNA-binding sites *in vivo* (Tanay, 2006), suggesting that stringent *in vitro* selections may not capture all of the information relevant to *in vivo* function of TF DNA-binding activities.

To our knowledge, the additive PWM model has not been rigorously challenged in its ability to correctly predict affinity to all possible binding sequences, largely due to the difficulty of making measurements on an exhaustive collection of sequences *in vitro* and the presence of confounding factors *in vivo*. Until recently, only a relatively small number of sequences had ever been tested for any given TF (Man and Stormo, 2001; Mukherjee *et al.*, 2004). However, development in the last two years of microarray-based techniques for measuring protein–DNA associations (Berger *et al.*, 2006; Liu *et al.*, 2005; Mukherjee *et al.*, 2004; Warren *et al.*, 2006) promises a dramatic expansion in protein–DNA interaction data.

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

These data present new opportunities to revisit both the PWM model of protein–DNA affinity as well as the various algorithms designed to learn this model from binding data. The validity of the PWM model can be tested by measuring its ability to reproduce the protein–DNA interaction data. However, fitting a PWM model, or motif, to these microarray-based measurements of binding affinity is difficult. In these data, a microarray intensity level is a semi-quantitative measurement of the affinity of the TF for the DNA sequence represented by the probe. Traditional motif finding algorithms like Bioprospector (Liu *et al.*, 2001) and MEME (Bailey and Elkan, 1994) mostly ignore this intensity and treat all probe sequences above a given threshold equally. As a result the motifs produced by these algorithms are very sensitive to the *ad hoc* choice of the threshold.

Two newer popular motif finding algorithms, MDScan (Liu *et al.*, 2002) and MatrixREDUCE (Foat *et al.*, 2005, 2006), incorporate intensity data into their search but make different assumptions about the relationship between binding affinity and the microarray intensity. MDScan uses the intensity data to classify bound probes into two groups based on the rank of their intensity. The higher ranked group contributes more to the PWM. MatrixREDUCE uses the microarray intensity as a surrogate for binding affinity and fits a motif model by assuming a linear relationship between the two. Though microarray intensity can be to used to predict the relative binding preferences of a TF for given pairs of probes, the exact functional form of the relationship between the TF-binding affinity for a probe sequence and its intensity level remains unclear and may vary depending on the experimental technique.

To address this problem, algorithms have recently been developed that fit motif models to intensity rankings rather than the intensities themselves (Berger *et al.*, 2006; Chua *et al.*, 2006; Tanay, 2006). One of these, PREGO (Tanay, 2006), was designed originally for ChIP-chip data and fits a PWM model by attempting to maximize the Spearman rank correlation between binding affinities and the microarray intensities using hill-climbing optimization. Another, Seed and Wobble (Berger *et al.*, 2006), was designed for protein binding microarray (PBM) data (Mukherjee *et al.*, 2004) and calculates an enrichment score similar to a Mann-Whitney U statistic for each 8mer and gapped 8mer. A PWM model is constructed by combining the enrichment scores of the best 'seed' 8mer and those of single base 'wobbles' around the seed using a Boltzmann distribution. These and other algorithms based on intensity rankings may be sensitive to noise in the microarray intensity levels that can translate into large changes in the relative rankings of probes with similar TF-binding affinities.

Here we introduce a new motif finding algorithm, RankMotif++, which uses the microarray data to infer relative binding preferences of the TF for pairs of probes and learns a motif model that is consistent with these preferences. By predicting binding preferences inferred from microarray intensities rather than the intensities themselves, RankMotif++ is applicable for a wide variety of relationships between the TF-binding affinity and semi-quantitative readout of that affinity. Because it relies on inferred binding preferences rather than absolute rankings, RankMotif++ can be less sensitive to noise in probe intensity measurements. We evaluate our algorithm and test the general validity of the PWM model of binding affinity using data from a recent study by Berger *et al.* (2006).

Berger *et al.* (2006) used PBMs to analyze five different TFs. These were each analyzed on two different array designs containing independent de Bruijn sequences, such that each 10mer was represented once on each array, but embedded in a different 35-base context from the other array. We have used these data to both gauge the reproducibility of the individual K-mer-binding affinities inferred from the PBM data for the TFs, as well as the accuracy of the PWM model derived from several state-of-the-art approaches. The results of this analysis support the general validity of the PWM model, highlight some cases where it may or may not be advantageous to use individual K-mer scores, and also serve as a guide to the accuracy of genomic scans using either a PWM or a library of K-mers.

2 ALGORITHM

The RankMotif++ algorithm is based on a probabilistic model of binding preferences. The binding preferences of the TF can be inferred from the microarray intensity data using standard statistical techniques. In our model, the probability that one probe will be observed to be preferentially bound by the TF over another probe depends on the ratio of the affinities assigned to each probe sequence by the motif model. RankMotif++ fits PWM by searching for those that maximize the likelihood of the observed preference data under this probabilistic model.

2.1 Probabilistic model of binding preferences

Let a potentially noisy observation of the binding preferences of a TF be represented by a set of binary random variables $X = \{X_{ij}\}_{i,j=1}^{N} \setminus \{X_{ii}\}_{i=1}^{N}$, where N is the number of probes and $X_{ij} = 1$ if probe sequence i is observed to be preferred over sequence j. By definition we set $X_{ij} = 1 - X_{ji}$.

Let the outcome of a particular experimental assay, for example a PBM experiment be represented by a configuration of a subset of the binary random variables in X. Though, in general, a given experiment may result in an observation of only a subset of the possible preferences, to simplify the presentation of the model, we will assume that all possible preferences are observed. Later on we will describe how to relax this assumption. Without loss of generality, we will also assume for the time being that the probes are indexed so that if $i > j$ then $X_{ij} = 1$.

Let s^i be the sequence associated with probe i and $S = \{s^1, s^2, \ldots, s^N\}$ be the set of all probe sequences, and let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ represent the parameters of the motif model. Our model also includes a scaling factor $w > 0$ that governs the effect that changes in predicted binding affinity have on the probability of observing a preference in the microarray data. The value w is also fit by maximizing the log likelihood.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

Given Θ, w and S, our model assumes that the preferences are conditionally independent, so the log likelihood of Θ and w, $L(\Theta, w) = \log P(X|\Theta, w, S)$, can be written as

$$L(\Theta, w) = \sum_{i>j} \log P(X_{ij} = 1|\Theta, w, S)$$

We model the probability of observing a given preference $P(X_{ij} = 1|\Theta, w, S)$ as depending only upon the scaled log ratio of the binding affinities of the motif model for sequences s^i and s^j. If we set $f(s^i, \Theta) = \log g(s^i, \Theta)$ where $g(s^i, \Theta)$ is the binding affinity of motif model Θ for s^i, then we write this probability as:

$$P(X_{ij} = 1|\Theta, w, S) = \sigma(w(f(s^i, \Theta) - f(s^j, \Theta))) \quad (1)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the well-known logistic sigmoid function. Note that we can rewrite equation (1) as

$$P(X_{ij} = 1|\Theta, w, S) = \frac{g(s^i, \Theta)^w}{g(s^i, \Theta)^w + g(s^j, \Theta)^w}$$

In the following when the values of Θ, w and S are obvious, we will use the notation z_{ij} to represent $P(X_{ij} = 1|\Theta, w, S)$.

2.2 Maximum likelihood estimation of the motif model

We maximize the likelihood of the binding model with respect to Θ and w using conjugate gradient descent (Press *et al.*, 2002) on the log likelihood of the binding data. In this section, we describe how the gradient of the log likelihood is calculated.

Since $X_{ij} = 1 - X_{ji}$, we can write

$$z_{ij} = 1 - z_{ji} = P(X_{ji} = 0|\Theta, w, S) \quad (2)$$

and thus

$$L(\Theta, w) = \frac{1}{2} \sum_i \sum_{j \neq i} \log P(X_{ij}|\Theta, w, S). \quad (3)$$

Further noting that

$$P(X_{ij}|\Theta, w, S) = z_{ij}X_{ij} + X_{ji}z_{ji}$$

and that

$$\frac{d}{d\theta_m} z_{ij} = z_{ij}z_{ji}w\left[\frac{d}{d\theta_m} f(s^i, \Theta) - \frac{d}{d\theta_m} f(s^j, \Theta)\right],$$

after taking the derivative of equation (3) with respect to an element θ_m of Θ, we get

$$\frac{d}{d\theta_m} L(\Theta, w) = \frac{w}{2} \sum_i \sum_{j \neq i} (X_{ij} - z_{ij})\left[\frac{d}{d\theta_m} f(s^i, \Theta) - \frac{d}{d\theta_m} f(s^j, \Theta)\right] \quad (4)$$

which can be simplified to

$$\frac{d}{d\theta_m} L(\Theta, w) = w \sum_i (r_i - \hat{r}_i) \frac{d}{d\theta_m} f(s^i, \Theta) \quad (5)$$

where $r_i = 1 + \sum_{j \neq i} X_{ij}$ is simply the rank of probe i in a list of probe intensities sorted in increasing order, and $\hat{r}_i = 1 + \sum_{j \neq i} z_{ij}$ can be interpreted as the expected rank of i under the binding preference model. So the partial derivative of the cost function with respect to θ_m depends upon the difference between the actual and expected rank of probe i, weighted by the derivative of $f(s^i, \Theta)$ with respect to θ_m. Similarly, the partial derivative with respect to w is

$$\frac{d}{dw} L(\Theta, w) = \sum_i f(s^i, \Theta)(r_i - \hat{r}_i). \quad (6)$$

2.3 Definition of the motif model and composition function

None of the above derivation depends on the particular form of our motif model Θ or our method of estimating the binding affinity $f(\cdot, \cdot)$. As such, our binding preference model can easily be generalized to more complex motif models, so long as $\frac{d}{d\theta_m} f(s^i, \Theta)$ exists for all elements θ_m of Θ.

To investigate the accuracy of the PWM model and allow direct comparison against other PWM-based motif finding algorithms, we take Θ as representing a PWM, whose $(b, k)^{th}$ element, Θ_{bk}, is the weight of base $b \in D = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ at position k in the binding site. Each column, Θ_k, of the PWM can be used to define a multinomial distribution over D as follows:

$$P(b|\Theta_k) = \exp(\Theta_{bk}) / \sum_{b' \in D} \exp(\Theta_{b'k}), \quad (7)$$

and the probability of an arbitrary K-mer s under the model is

$$P(s|\Theta) = \prod_k P(s_k|\Theta_k) \quad (8)$$

where s_k is the k-th element of s.

A given sequence can contain multiple K-mers that are good matches to the PWM. A number of methods have been proposed to combine the probabilities assigned to the individual K-mers within a sequence into a composite measurement of the binding affinity of the PWM for the whole sequence. We use a combination mechanism similar to the GOMER (Granek and Clarke, 2005) scoring function. In particular, we define $g(s^i, \Theta) = \exp(f(s^i, \Theta))$ to be the probability under the PWM that the TF binds to at least one of the K-mers that are subsequences of s^i, i.e.

$$g(s^i, \Theta) = 1 - \prod_{t=0}^{L^i - K} 1 - P(s^i_{t+1:t+K}|\Theta) \quad (9)$$

where L^i is the length of sequence s^i, and $s^i_{t+1:t+K}$ is the subsequence of s^i that starts at the $(t+1)$-th element and ends at the $(t+K)$-th element inclusive. Assuming that all the K-mers in s^i can be bound independently, $g(s^i, \Theta)$ is the probability that at least one of them will be bound. Our combination mechanism differs from GOMER in how we calculate the probability of binding to a given K-mer using the PWM. We made this change in order to ensure that our K-mer probability function had continuous partial derivatives.

The partial derivatives of $f(s^i, \Theta)$ with respect to Θ_{bk} are:

$$\frac{d}{d\Theta_{bk}} f(s^i, \Theta) = \frac{1 - g(s^i, \Theta)}{g(s^i, \Theta)}\left[\sum_t \frac{P(s_{t+1:t+K}|\Theta)}{1 - P(s_{t+1:t+K}|\Theta)} \cdot (\delta_{s_{t+k}, b} - P(b|\Theta_k))\right] \quad (10)$$

where $\delta_{s_{t+k}, b} = 1$ if $s_{t+k} = b$, otherwise $\delta_{s_{t+k}, b} = 0$. Inserting equation (10) into equation (5) gives the partial derivatives

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

necessary to complete the computation of the gradient of $L(\Theta, w)$.

2.4 Partial observations of binding preferences

When the difference (or ratio) between the observed intensities of two probes is within the noise level of the assay, we treat the TF-binding preference between those two probes as being unobserved. These unobserved preferences are easily handled in the binding preference model. Because the observed preferences are conditionally independent given the parameters of the model and the probe sequences, unobserved preferences can simply be marginalized out of the likelihood. The only change that needs to be made to the model is in the definition of r_i and $\hat{r}_i$ in equation (6). If T is the set of probe pairs (i,j) for which binding preferences have been observed, then the updated definition of r_i and $\hat{r}_i$ are

$$r_i = 1 + \sum_{\{j|(i,j)\in T\}} X_{ij}$$
$$\hat{r}_i = 1 + \sum_{\{j|(i,j)\in T\}} z_{ij}.$$

Note that if $(i,j) \in T$ then $(j,i) \in T$.

3 METHODS

3.1 PBM intensity data

We downloaded the PBM intensity measures from the Supplementary Material section of Berger *et al.* (2006). These data contain measurements from two different universal microarrays for each of the five transcription factors Cbf1, Ceh22, Oct1, Rap1 and Zif268.

We normalized the microarray intensity data in two steps. We first translated the microarray intensities by adding a constant to each measurement so that the minimum intensity was equal to one. We then log-transformed the translated data. We call these log-intensities the 'normalized intensities' and use y_i to denote the normalized intensity of probe i.

3.2 Deriving protein-binding preferences from normalized intensities

Because microarray intensity data is noisy, the relative probe intensity levels are not always a reliable indicator of transcription factor binding preferences. As such, to define a reliable set of protein-binding preferences, we separately estimated the intensity noise level for each array and only predict a binding preference when one of the probes has a significantly higher normalized intensity than the other.

Typically in microarray-based assays of transcription factor binding affinity, only a small proportion of probes detect sequence-specific binding. As such, we can estimate the SD of the noise in the normalized intensity of non-specific binding using a robust estimate of the SD, σ, of the entire distribution of normalized intensity levels. In our experience, σ is also an empirical upper bound on the SD of the normalized intensity noise affecting probes detecting sequence-specific binding. As such, in the following, we will use σ as an estimate of the SD of the noise affecting all probes.

Our robust estimate is calculated by computing the median absolute deviation (MAD) of the distribution of normalized intensities and setting σ to the MAD divided by 0.6745 (the MAD of the unit normal distribution).

We define the set of positive (i.e. bound) probes as those whose normalized intensity $y_i > m_y + 4\sigma$, where m_y is the median of $\{y_1, y_2, \ldots, y_N\}$. These are probes whose normalized intensity we deem to be significantly different, given the noise level σ, from the distribution of intensities of probes detecting non-specific binding. The proportion of positives varies between 0.9 and 5.4% among the ten arrays. Probes that are not positives are called negatives.

We set $X_{ij} = 1$ if and only if y_i has significantly higher normalized intensity than y_j given σ. We represent that condition as $y_i > y_j + 3\sigma$. If by the former condition neither $X_{ij} = 1$ or $X_{ji} = 1$, i.e. $|y_j - y_i| < 3\sigma$, then we deem the intensity difference between the two probes to be within the noise level of the array. In this case, we represent X_{ij} as being unobserved and thus do not try to predict the binding preference between these two probes.

3.3 Fitting RankMotif++ motif models

We use all the positives and a random subset of the negatives to learn the motif model in RankMotif++. The negative subset is a random subsampling of all negatives with size of 800.

Like many other motif finders, the log likelihood optimized by RankMotif++ is not convex, so different initializations can generate different results. For the experiments reported here, we used three different initialization for each motif length and we found that these three were enough to locate good local optima.

To initialize RankMotif++ for motifs of length seven, we identified the three 7mers most associated high intensity probes. We used the Wilcoxon–Mann–Whitney test on each 7mer to score the distribution of intensities of the probes that did contain the 7mer against the distribution of those that did not. From the three 7mers with the lowest P-values, we generate three initial position frequency matrices to seed the motif search. For each of these 7mers, we initialize a PFM by putting 0.7 for the consensus base in each column and 0.1 for the non-consensus base. We transform the PFM to a PWM by taking the log of each entry. In some but not all cases, the final PWM that our optimization procedure converges on has a similar consensus to the initial PWM.

To initialize the search for motifs of length $K > 7$, we generate two PFMs derived from the highest scoring RankMotif++ PFM of length $K-1$ by adding a column of 0.25 on either side. We also generate one random PWM with weights uniformly distributed between -0.05 and 0.05. Occasionally the search starting from the randomly initialized PWM leads to the highest scoring PWM.

We computed motifs with widths from 7 to 13, and we used the motif model with the highest log likelihood for the test array evaluations.

3.4 Fitting the MatrixREDUCE motif models

We used MatrixREDUCE to generate one motif of length range between 7 and 13 by setting parameters `motif` between 7 and 13 and setting `max_motif` to 1. In computing this one motif, MatrixREDUCE considers a number of initial seeds and uses the highest scoring one among them, and it automatically chooses the best motif width among the range input. The transcription factors in the PBM study are unlikely to bind as homodimers so we set `max_gap` and `flank` to 0. We use the default settings for the other parameters.

We applied MatrixREDUCE to both the PBM intensity data directly and to the normalized intensities of the positive and negative probe sets used by RankMotif++. The overall performance of MatrixREDUCE on the benchmarks was similar for the two different inputs. The PWMs in Figure 1 are generated from raw PBM intensities but for consistency with other methods, we benchmark MatrixREDUCE on the normalized intensities inputs.

3.5 Fitting the MDScan motif models

We also vary the MDScan motif widths w between 7 and 13. MDScan uses a 'top' set and a refinement set to fit its motif model. We set the size of the top set to 20 which is the largest size used in Liu *et al.* (2002). We set the refinement set size to be equal to the number of positive probes

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

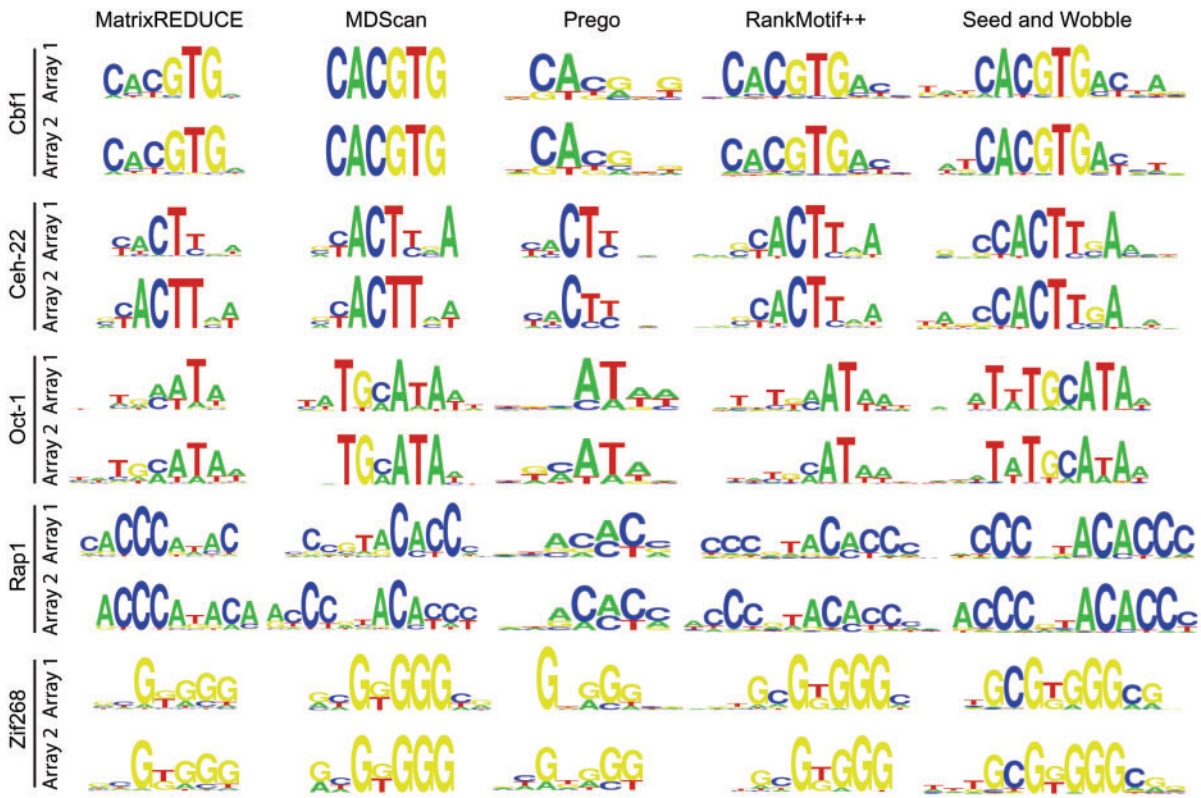

Fig. 1. Sequence logo representations of the PWMs generated by the five motif finding algorithms for the five TFs on each of the two arrays.

on each array, and we use the default settings for all the other parameters. For each length, we only evaluate the motif model assigned the top score by MDScan.

The motif scores provided by MDScan always increased for the shorter motifs. As such, we used an alternative method for determining the optimal motif width. Using the top scoring motif model for each width, we scored each positive probe sequence of the training array using equation (9), and then we selected the motif model whose sequence scores had the highest Spearman correlation with the normalized intensities for the positive probes.

3.6 Fitting the PREGO motif models

The input probe sets for PREGO were the same as those used by RankMotif++.

We set the PREGO parameters `MinK` to 7 and `MaxK` to 14. With these settings, PREGO reports the highest scoring motif of length range between 7 and 13. In order to force PREGO to run on our normalized intensities, we set `SuppressLogWarning` to 1. We use the default settings for the other parameters.

3.7 Fitting the Seed and Wobble models

We were supplied with Seed and Wobble PWMs by Anthony Philippakis. We asked that these PWMs be fit separately to each array. For one of the arrays, Ceh22 Array 1, he supplied us with two slightly different PWM models.

3.8 Fitting the fully specified 8mer motif model

Each 8mer is associated with the median normalized intensity of the probes in which it appears in the training array.

3.9 Scoring test set probe sequences under each motif model

Note that all the motif models were only fit to training arrays, here we describe how these motif models were used to score the probe sequences on the test arrays.

(1) 8mer: The score assigned to a test set probe is simply the maximum of training array median intensities associated with the 8mers present in the probe.

(2) RankMotif++, MDScan, PREGO: We use equation (9) where s is the test probe sequence and Θ is the motif model fit to the training array.

(3) Seed and Wobble: We use equation (9) as above except for the two Ceh22 Array 1 PWMs. For Ceh22 Array 1, we scored each probe separately under the two PWMs [using equation (9)] and then set the composite score for the probe to be $s = 1 - (1 - s_1)(1 - s_2)$ where s_1 and s_2 were its scores under the two PWMs respectively.

(4) MatrixREDUCE: The MatrixREDUCE motif model is neither a position weight matrix nor a position frequency matrix. As such, we use the MatrixREDUCE scoring function $f(s^i, \theta) = \sum_{t=0}^{L^i-K} \prod_{k=1}^{K} \theta_{s^i_{t+k},k}$, where θ is the MatrixREDUCE motif model, s^i is the test probe sequence, L^i is the length of s^i and s^i_{t+k} is the $(t+k)$-th element of sequence s^i.

3.10 Computing performance metrics

To test the generalization ability of each of the motif models, we train each model on intensity data from one of the array designs and tested generalization performance on the other. Each design was used once as a training array and once as a test array.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

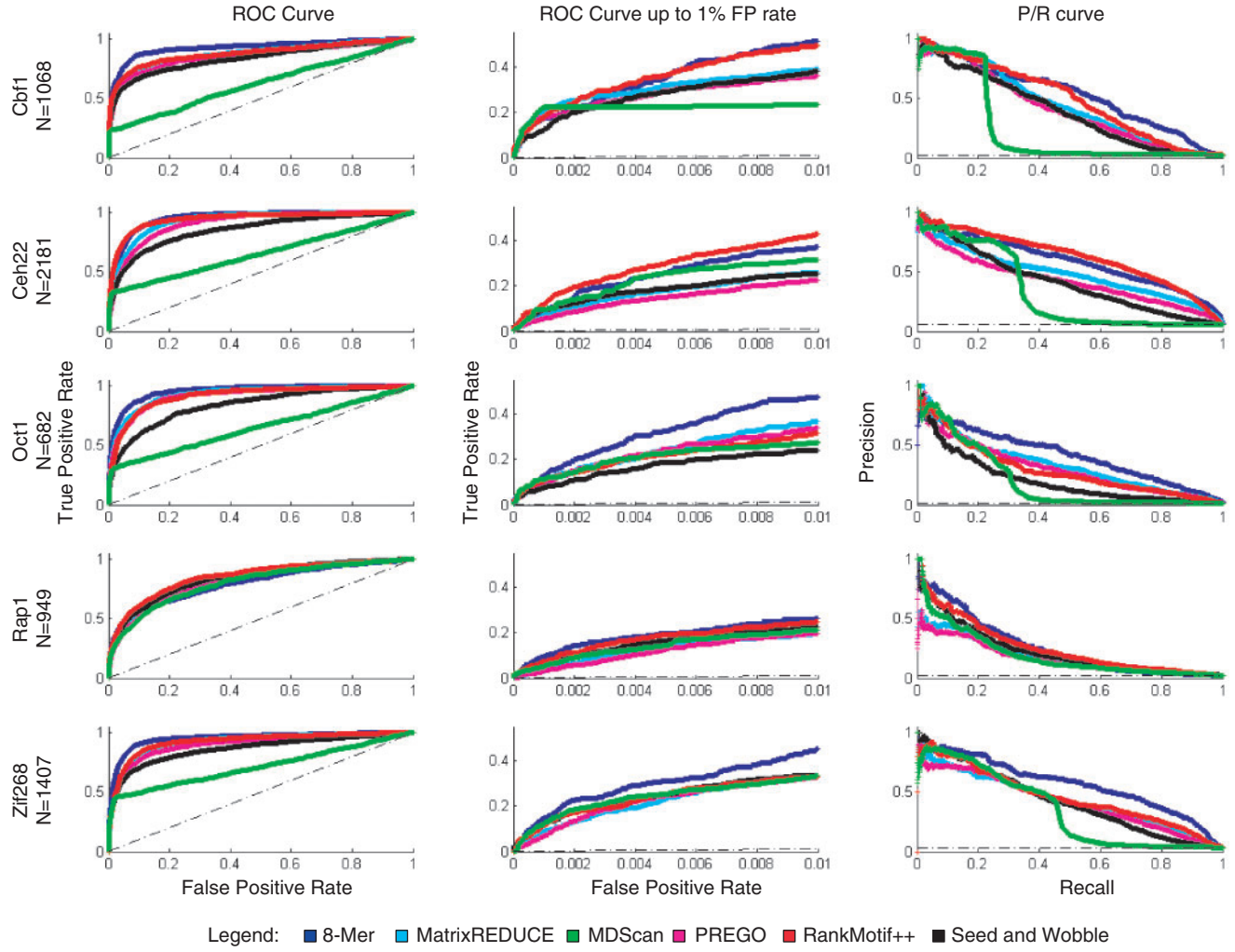

Fig. 2. The ROC curves and the Precision/Recall curves for array #1 for each of the six methods and each of the five TFs.

We score each model's ability to correctly identify and rank the positive probes (as previously defined) in the test array using the methods as described in Section 3.9.

For each TF and each method, we calculate the Spearman correlation between the sequence scores and the intensities for the positive probes, which measures how well each motif model can reproduce the rankings of the intensities of the positives.

By thresholding the sequence scores at a given threshold, we can make predictions of which test set probes will be positive. For each threshold, we can calculate the true positive rate (TP/(TP + FN), also called the recall or the sensitivity), the false positive rate (1 − TN/(TN + FP)) and the precision (TP/(TP + FP)). By decreasing the sequence score threshold and computing the true and false positive rates and the precision at these decreasing thresholds, we plot the ROC curves and the precision/recall curves shown in Figure 2.

4 RESULTS

Figure 1 shows the PWMs produced by each method for the five different transcription factors on each of the two array designs using sequence logos generated using enoLogo (Workman *et al.*, 2005). There is a general overlap in the consensus sequences of the PWM but also consistent differences between the PWMs from different methods. In general, MatrixREDUCE and RankMotif++ produce more degenerate PWMs than MDScan. Seed and Wobble motifs are almost always the widest and the PREGO motifs are always between 7 and 8 columns. In general, PWMs from the same method run on data from different arrays are more similar than PWMs from different methods run on the same array data.

To evaluate whether these differences in PWMs would lead to differences in genome-wide predictions of TF binding, we evaluated the accuracy with which each PWM could reproduce the ranking of probe intensities based on the probe sequence.

Each probe sequence was scored by each PWM, and the rankings of the probe scores were compared to the rankings of the measured probe intensities. In order to judge the validity of the position-independent model, we included a fully specified model of 8mer TF-binding preferences in our evaluation.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

Table 1. Spearman rank correlation coefficients

TF	Array	8-Mer	MatrixREDUCE	MDScan	PREGO	RankMotif++	Seed and Wobble
Cbf1	#1	0.647*	0.634	0.512	0.619	***0.636***	0.527
	#2	0.657*	0.604	0.496	0.58	***0.64***	0.49
Ceh-22	#1	0.487*	0.373	0.36	0.366	***0.485***	0.304
	#2	0.408	0.3	0.324	0.278	***0.427****	0.275
Oct-1	#1	0.327*	0.263	0.286	0.281	0.244	***0.315***
	#2	0.446*	***0.308***	0.264	0.272	0.291	0.213
Rap1	#1	0.238	0.273	0.338	0.261	***0.382****	0.372
	#2	0.275	0.239	0.254	0.205	***0.359****	0.357
Zif268	#1	0.421*	0.293	0.265	0.292	***0.336***	0.276
	#2	0.346*	0.279	0.246	0.196	***0.308***	0.25

Correlations were computed between the intensities of the positive probes in each array (row) and the TF-binding affinities predicted for the probe sequence by each of the six motif models (columns). For each experiment, a bold italicized entry indicates the best correlation among the PWM motif models, and a starred entry indicates the best overall correlation. Correlation levels are categorized by color.

Table 2. True positive rates at 1% false positive rate

TF	Array	8-Mer	MatrixREDUCE	MDScan	PREGO	RankMotif++	Seed and Wobble
Cbf1	#1	0.515*	0.39	0.231	0.362	***0.493***	0.383
	#2	0.459*	0.348	0.202	0.336	***0.424***	0.284
Ceh-22	#1	0.37	0.26	0.316	0.225	***0.427****	0.254
	#2	0.257	0.226	0.293	0.2	***0.332****	0.251
Oct-1	#1	0.474*	***0.365***	0.274	0.339	0.315	0.239
	#2	0.382*	***0.31***	0.213	0.274	0.24	0.202
Rap1	#1	0.257*	0.197	0.213	0.197	***0.247***	0.226
	#2	0.277	0.171	0.32	0.179	***0.325****	0.28
Zif268	#1	0.449*	0.332	0.335	0.328	0.33	***0.336***
	#2	0.431*	0.297	0.314	0.301	***0.389***	0.313

At a fixed 1% false positive rate (99% specificity), true positive rates (sensitivities) were computed for the prediction of the positive probes in each array (row) using the binding affinities computed by each method (column). For each experiment, a bold italicized entry indicates the highest sensitivity among the PWM motif models, and a starred entry indicates the best overall sensitivity. True positive rates are categorized by color.

To ensure that the performance measures were not influenced by probes not bound by the TF, we defined a set of 'positive' probes that had intensities at least four SDs above the median intensity. For each PWM, we determined how highly the positive probes ranked in the score rankings (Table 1 and Figure 2) and how well the scoring ranking reproduced the observed intensity rankings (Table 2).

Table 1 shows the Spearman correlation coefficients between the probe scores and the probe intensities for each of the six methods. All of these correlations are statistically significant under a one-sided test, the largest *P*-value of any element in the table is less than 10^{-5} and only four *P*-values were larger than 10^{-10}. RankMotif++ has the highest correlation among the PWMs in eight out of the ten conditions with MatrixREDUCE and Seed and Wobble having the highest correlation for both Oct1 arrays. The low MDScan correlations are likely due to the lack of degeneracy of the MDScan PWMs: between 50 and 75% of positive probes do not have any hits to the MDScan PWMs and are thus assigned the same score as most of the non-positive probes. The degeneracy of the MDScan model is governed by one of the algorithm's parameters. Though we set this parameter to the highest level used in Liu *et al.* (2002), it is possible that the performance of MDScan would improve with a different setting.

Table 2 shows the sensitivity (true positive rate) of each PWM at 99% specificity (1% false positive rate). RankMotif++ has highest sensitivity in seven out of the ten conditions. At this specificity level, the MDScan PWMs still have hits in the positive probes. Figure 2 provides a more detailed view of the predictive performance of each method on the positive probes from the array design #1.

RankMotif++ recovers probe rankings better any of the other four methods though does not perform quite as well as the fully specified 8mer model. However, RankMotif++ has higher correlations and higher sensitivities than the 8mer model in three of the ten conditions, particularly for Rap1 which has binding profiles with more than eight conserved bases. On Cbf1 and Ceh22, 8mer median and RankMotif++ have similar correlations and sensitivities. Though the 8mer median performance is better on Oct1 and Zif268.

5 DISCUSSION

We have made an investigation into the general validity of the PWM model of transcription factor DNA binding, and in the same analysis introduced a new method for learning PWMs. As a gold standard, we used a fully specified model of 8mer-binding preferences.

Surprisingly, the predictive accuracy of the 8-mer model at predicting the 35mer probe intensities was not as high as we had expected. Even for Cbf1, which has a core motif of six bases, a value of 50% recall is associated with only 60% precision by 8mer scores. Spearman correlations among the highest intensity spots (positives) are also far from absolute, suggesting that our stringent definition of positives is not causing a threshold effect. This phenomenon could represent a performance limit imposed by the reproducibility or noise level of the assay or the way the 8mer scores are calculated, or an effect of sequence or structural properties of flanking bases in the 35mer probe. Indeed Berger *et al.* (2006) report an effect of both the position of the consensus on the probe and its sequence context on the intensities of probes containing the Zif268 consensus. However, they also report that 8mer median intensities calculated from different arrays are highly reproducible. In our hands, the Spearman correlations between the highest intensity 8mer scores calculated separately on each array average 0.7593 over the five TFs. However, because the 8mer scores each are calculated from 16 separate 35mer intensity measurements, this correlation may represent an overestimate of their performance on a genome-wide scan.

By several criteria, our method, RankMotif++, performed favorably in comparison to currently used tools. In some circumstances, RankMotif++ even performed better than the 8mer model, for example in the case of Rap1. We note that the binding profile of Rap1 is greater than 11 bp, so the PWM models are inferring the binding affinities of Rap1 to 12 and 13mers not represented in the training array and thus cannot be fully captured by the 8mer model. On the whole, PWMs learned by RankMotif++ on PBM microarray array data are more accurate predictors of transcription factor binding preferences than those learned by two widely used motif finders for use with semi-quantitative data: MatrixREDUCE and MDScan and two new algorithms that use intensity rankings: PREGO and Seed and Wobble. We suggest that these differences in performance are due differences in assumptions that each method makes about the relationship between transcription binding affinity and semi-quantitative readout and differences in each method's sensitivity to intensity noise.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018

ACKNOWLEDGEMENTS

We would like to acknowledge helpful discussions with Anthony Philippakis, Mike Berger and Martha Bulyk and thank Anthony for supplying us with Seed and Wobble PWMs. X.C. was supported by a grant from Genome Canada through the Ontario Genomics Institute. This work was supported in part by an NSERC operating grant to Quaid Morris.

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.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Berger,M.F. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcriptionfactor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Chua,G. *et al.* (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl. Acad. Sci. USA*, **103**, 12045–12050.

Foat,B.C. *et al.* (2005) Profiling conditionspecific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci.* USA **102**, 17675–17680.

Foat, B.C. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

MacIsaac,K.D. *et al.* (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome. Biol.*, **6**, R87.

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

Liu,X, *et al.* (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity, *Genome Res.*, **15**, 421–427.

Liu,X. *et al.* (2006) Whole-genome comparison of Leu3 binding in vitro and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, **16**, 1517–1528.

Man,T.K. and Stormo,G. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.

Messina,D.N. *et al.* (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.

Mukherjee,S. *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

Press,W.H. *et al.* (2002) Numerical Recipes in C++, 2nd edn. Cambridge University Press.

Roulet,E. *et al.* (2002) Highthroughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, **8**, 831–835.

Pennacchio,L.A. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome. Res.*, **8**, 1034–1050.

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome. Res.*, **16**, 962–972.

Vlieghe,D. *et al.* (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.

Warren,C.L. *et al.* (2006) Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. USA.*, **103**, 867–872.

Workman,C.T. *et al.* (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/13/i72/237219
by University of Illinois Music Library user
on 22 March 2018