Bioinformatics Advance Access published December 16, 2008

Data and text mining

Predicting the Binding Preference of Transcription Factors to Individual DNA k-mers

Trevis M. Alleyne[1], Lourdes Peña-Castillo[2], Gwenael Badis[1], Shaheynoor Talukder[1], Michael F. Berger[3,5], Andrew R. Gehrke[3], Anthony A. Philippakis[3,5,6], Martha L. Bulyk[3-5,6], Quaid D. Morris[1,2], and Timothy R. Hughes[1,2*]

[1]Department of Molecular Genetics, [2]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON M5S 3E1, [3]Division of Genetics, Department of Medicine, [4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 021156, [5]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138 2, and [6]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115

Associate Editor: Prof. David Rocke

ABSTRACT

Motivation: Recognition of specific DNA sequences is a central mechanism by which transcription factors (TFs) control gene expression. Many TF binding preferences, however, are unknown or poorly characterized, in part due to the difficulty associated with determining their specificity experimentally, and an incomplete understanding of the mechanisms governing sequence specificity. New techniques that estimate the affinity of TFs to all possible k-mers provide a new opportunity to study DNA-protein interaction mechanisms, and may facilitate inference of binding preferences for members of a given TF family when such information is available for other family members.

Results: We employed a new data set consisting of the relative preferences of mouse homeodomains for all 8-base DNA sequences in order to ask how well we can predict the binding profiles of homeodomains when given only their protein sequences. We evaluated a panel of standard statistical inference techniques, as well as variations of the protein features considered. Nearest-neighbour among functionally-important residues emerged among the most effective methods. Our results underscore the complexity of TF-DNA recognition, and suggest a rational approach for future analyses of TF families.

1 INTRODUCTION

Most TFs can be grouped into families of shared conserved DNA-binding structures that are usually identified by common ancestry inferred from sequence homology (Papavassiliou, 1995). Despite the sequence conservation within TF families, individual proteins within the same DNA-binding domain (DBD) family can have radically different DNA-binding specificities (Ekker, et al., 1994). Since the preferred binding sequences within a family can often be changed by mutating only a single DNA-contacting amino acid residue (Damante, et al., 1996), it has been proposed that a recognition code might exist in which affinity to each base in a TF binding site is governed by either additive or combinatorial rules that pair the identities of amino acids at DNA-contacting positions with relative preferences for each of the four DNA bases at each position of the binding site. Conflicting with this view, however, are observations that changes in DBD sequence can alter the arrangement of DNA-contacting residues in ways that seem to be inconsistent with a simple recognition code (Miller, et al., 2003; Pabo and Nekludova, 2000). In addition, study of the DNA-binding specificities of TFs typically employs a position weight matrix (PWM) (Stormo, 2000), and the assumptions of PWMs, such as independence of base positions, do not fit all of the biochemical data (Benos, et al., 2002).

Several high-throughput, unbiased, and semi-quantitative methods for the assessment of TF sequence preferences have been developed, including protein-binding microarrays (PBM) (Mukherjee, et al., 2004), DNA immunoprecipitation microarrays (DIP-chip) (Liu, et al., 2005), and cognate site identifier microarrays (CSI) (Warren, et al., 2006). The data sets associated with these methods provide an opportunity to examine protein-DNA interactions at previously unprecedented resolution and scale. Here, we present an evaluation of how well a panel of inference algorithms can predict TF DNA-binding specificity data derived from PBM experiments, in an effort to gain deeper insight into the mechanisms governing the specificity of protein-DNA interactions, and also to identify a means to project binding preferences to proteins without known binding preferences. We focus on the homeodomain family, because it is large and diverse, and the vast majority of homeodomain-containing proteins have only a single homeodomain. Homeodomains are also one of the most well-studied DBDs, both structurally and biochemically, such that the DNA-contacting residues are known (Kissinger, et al., 1990) and several residues that can alter sequence specificity have been identified (Ades and Sauer, 1994; Ekker, et al., 1994; Hanes and Brent, 1989). We find that a nearest-neighbour approach using TF protein sequences is at least as effective as more sophisticated tech-

*To whom correspondence should be addressed.

niques. This finding has implications for the mechanics of DNA-binding, and for future study of TF-DNA interactions.

2 METHODS

2.1 Dataset

The Z-score transformed relative signal intensities for 168 homeodomains across all 32,896 8-mer DNA sequences were obtained using PBMs (Berger, et al., 2008). Given that methods that overfit the data may give good results in a leave-one-out cross-validation scheme (see below) if a high portion of the data has at least one nearly identical example, we reduced the dataset to 75 homeodomains unique at the 15 amino acid positions described as making contact with DNA in the Engrailed crystal structure (Table 1). A multiple sequence alignment of the 75 homeodomains was obtained by downloading the primary homeodomain family alignment from Pfam-A (Bateman, et al., 2004) (Accession number PF00046) and extracting the pertinent sequences. From the resulting sequence alignment, three subset sequence alignments were derived for purposes of feature selection: all 57 residues of the Pfam alignment (positions 2 to 58 of the Engrailed homeodomain), 15 residues described by Kissinger *et al.* as making direct contact with DNA in the Engrailed homeodomain crystal structure (Kissinger, et al., 1990) (positions 3, 5, 6, 25, 31, 44, 46, 47, 48, 50, 51, 53, 54, 55, and 57), and six residues described as determinants of sequence specificity in the literature (Ekker, et al., 1994; Laughon, 1991) (positions 3, 6, 7, 47, 50, and 54).

2.1.1 Numerical encoding All implementations of the compared methods, except Nearest Neighbour (NN), required numerical inputs. We converted the 6-, 15-, and 57-position sequence alignments to numerical encodings representing amino acid sequences of length l as binary vectors of length l x 20 digits, i.e. the 20 different amino acids were encoded as orthogonal 20 digit vectors and an amino acid sequence was represented by concatenating the binary vectors corresponding to residues at each position. Gaps were encoded as a vector of 20 zeros. Insertions were not considered in this analysis.

2.2 Machine learning algorithms

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be the set of m-residue sequence alignments from the dataset described above, where $m = 6$, 15, or 57, and for a given 8-mer out of the t total exemplar 8-mers, let $\boldsymbol{y}_i$ be the Z-scores for the i-th protein with respect to that 8-mer. We defined the problem of predicting homeodomain Z-scores for a particular 8-mer as the estimation of the function $f: \chi \rightarrow \mathbb{R}$ trained using the n data pairs $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n) \in \chi \times \mathbb{R}$, such that $f(\boldsymbol{x}_i)$ is approximately equal to $\boldsymbol{y}_i$ and f correctly generalizes to most unseen examples; therefore the problem of predicting homeodomain 8-mer Z-score profiles across all 8-mers was defined as predicting all t such functions. In this case, χ was the set of sequences {A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V, –}^{m}, where "–" represents a sequence alignment gap. We formalized both definitions as multiple regression problems in which the $\boldsymbol{x}_i$ were considered as n observations on m predictor variables and the $\boldsymbol{y}_i$ were considered as n observations on a response variable, and accordingly, compared a number of regression techniques from machine learning and statistics (outlined below) for the purpose of quantitatively modeling the relationships between these variables.

2.2.1 Nearest Neighbour Assume that x is the length m amino acid sequence alignment of an unseen protein. In order to predict the 8-mer profile of x, our implementation of the NN algorithm calculates a vector $\vec{n}$ of distances, where each element $\vec{n}_i$ represents the distance $d(\boldsymbol{x}_i, x)$ between protein x and $\boldsymbol{x}_i$ ($i \in \{1, \ldots, n\}$). We defined the distance between two proteins as the proportion of non-identities across all m positions. We also tested distances based on the PAM250 matrix, but the results were inferior (Berger, et al., 2008). Using $\vec{n}$, the algorithm then finds the nearest neighbours of x by computing the set $\{\boldsymbol{x}_i \,|d(\boldsymbol{x}_i, x) = min(\vec{n})\}$. Finally, for all t 8-mers, the algorithm calculates the Z-score of each 8-mer as the mean of the Z-scores for that 8-mer across all of the nearest neighbours.

Table 1. List of 75 mouse homeodomains unique at 15 AA positions that contact DNA

Alx3	Dobox4	Hlxb9	Hoxc12	Lhx6	Pax4	Rhox6
Bapx1	Dobox5	Hmbox1	Hoxc8	Meis1	Pax6	Six1
Barhl1	Duxl	Hmx1	Ipf1	Meox1	Pax7	Six3
Barx1	Emx2	Hmx2	Irx2	Msx1	Pbx1	Six4
Bsx	En1	Homez	Irx3	Nkx1-1	Pitx1	Tcf1
Cdx1	Esx1	Hoxa1	Isl2	Nkx2-2	Pknox1	Tcf2
Cphx	Evx1	Hoxa10	Isx	Nkx6-1	Pou1f1	Tgif1
Crx	Gsc	Hoxa13	Lbx2	Obox1	Pou2f1	Tgif2
Cutl1	Gsh2	Hoxa2	Lhx1	Obox6	Pou4f3	Tlx2
Dbx1	Hdx	Hoxa6	Lhx2	Og2x	Pou6f1	
Dlx1	Hlx1	Hoxb13	Lhx3	Otp	Rhox11	

2.2.2 Random Forests Regression We used the R randomForest package, which serves as an interface to the original random forests (RF) Fortran code developed by Breiman and Cutler (available at http://www.stat.berkeley.edu/~breiman/RandomForests. To predict the 8-mer profile of an unseen protein x, we generated t random forests, by using the set of n observations on m predictors $\boldsymbol{x}_i$, the response variable $\boldsymbol{y}$ for a given 8-mer, and default parameters. We then used this collection of random forests to predict the Z-scores across all 8-mers for the sequence x.

2.2.3 Support Vector Regression We used the LIBSVM package developed by Chih-Chung Chang and Chih-Jen Lin (available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/) to construct SVMs for every exemplar 8-mer. For each 8-mer, three SVMs were constructed, each using a different kernel: the linear kernel (SVM_L),

$$k(x, x') = \langle x, x' \rangle,$$

the polynomial kernel (SVM_P),

$$k(x, x') = \langle x, x' \rangle^d,$$

or the radial basis function kernel (SVM_R),

$$k(x, x') = exp\,(-\gamma \|x - x'\|^2),$$

where $d \in \mathbb{N}$, $\gamma > 0$, x and x' are alignment encodings, and $\langle x, x' \rangle$ refers to the inner product. All parameters were left at default setting with the following exceptions. For SVM_L, we tried all parameter pairs $[\varepsilon, C]$={ ε, C | $0.1 \leq \varepsilon \leq 4.8$, $2^{-15} \leq C \leq 2^3$}, where ε is the epsilon-SVM precision parameter, which was varied in steps of 0.8, and C is the SVM error penalty parameter. For SVM_P, we tried all parameter pairs $[d, C]$={ d, C | $1 \leq d \leq 6$, $2^{-15} \leq C \leq 2^3$}, where d was varied in steps of 1. For SVM_R, we tried all parameter pairs $[\gamma, C]$={ γ, C | $2^{-15} \leq \gamma \leq 2^3$, $2^{-15} \leq C \leq 2^3$}, where γ was varied by a factor of 2^2. In all cases, C was varied by a factor of 2^2 and the best parameter pair was chosen using five-fold cross-validation.

2.2.4 Principal Components Regression As the encoding strategy that we used produces a much larger number of variables relative to the number of samples (rank deficiency) as well as a large number of correlated variables (multicollinearity), both of which are problematic for linear regression, we used Principal Components Regression (PCR) to simultaneously reduce the dimensionality of the encodings and remove the correlation between variables. PCR was carried out by first applying principal

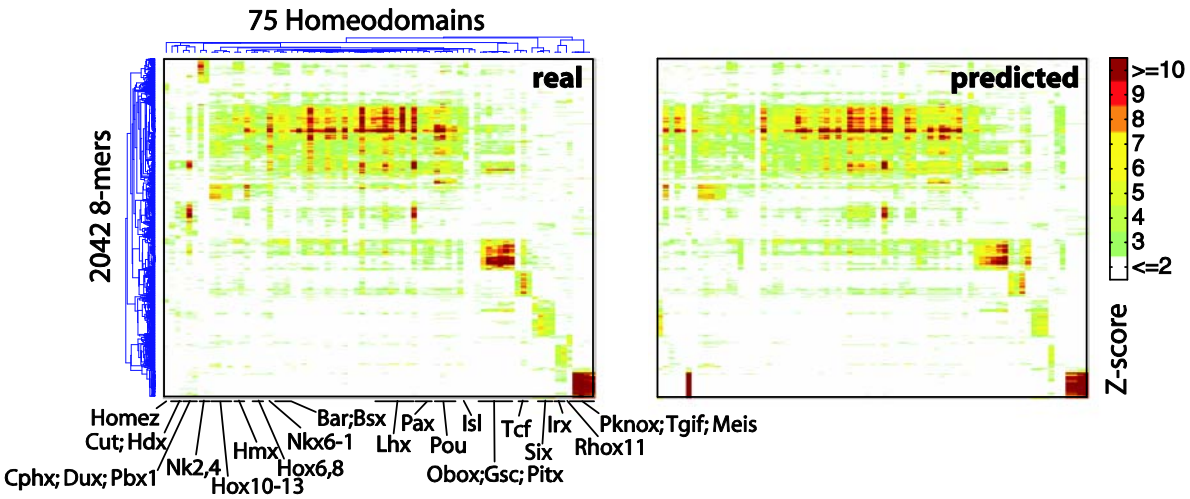

Fig. 1. 2-D clustergram of Z-scores for 2,042 8-mers and 75 mouse homeodomains, as observed in either real PBM data (left) or nearest mneighbour predictions (right), with some of the established classes of homeodomains labeled. Nearest neighbour predictions were made using 6 AA positions and leave-one-out cross-validation. The 2,042 8-mers were selected because they comprise the top 100 8-mers by Z-score over the DNA-binding domains shown.

components analysis to the encodings. The number of principal components retained p was selected using Parallel Analysis (PA) with 1000 shuffles, which is essentially a permutation test that asks whether the N-th principal component explains more of the variance than the N-th principal component would in a permuted version of the same data (reviewed in reference Franklin, et al., 1995). On the basis of PA, we retained 6, 12, and 19 principal components for the 6-, 15-, and 57-position alignments, respectively. For each 8-mer, we then built a regression model using an approach similar to five-fold cross-validation, described as follows:

(1) Randomly partition the sample set into five subsamples.
(2) Retain one subsample as the validation set and aggregate the remaining $k = 4$ subsamples into a matrix of training data, t_{ij} ($i \in \{1, \dots, k\}, j \in \{1, \dots, p\}$).
(3) So that the intercept in the regression model will always be estimated by $\bar{y}$ (Montgomery and Runger, 2007), centre and transform the training data into a new set of variables as:

$$x_{ij} = \frac{t_{ij} - \overline{t_j}}{\sqrt{s_{jj}}},$$

where $S_{jj} = \sum_{i=1}^{k} (t_{ij} - \overline{t_j})^2$.
(4) Compute the ordinary least squares coefficients for the transformed training data and calculate the mean squared error (MSE) of the coefficients using the validation set.
(5) Go back to Step 2 until all subsamples have been used as the validation set and retain the coefficients with the lowest MSE.
(6) Repeat steps 1 to 5 three times.

3 RESULTS

3.1 Comparison of linear and non-linear inference methods

We attempted to learn the Z-score transformed signal intensities for mouse homeodomain DBDs for all 32,896 non-redundant 8-base DNA sequences using PBM experiments (Berger, et al., 2008). We learned the Z-scores rather than PWMs because Z-scores reflect binding affinity (Berger, et al., 2006), whereas PWMs often fail to capture detailed binding activity (Benos, et al., 2002; Chen, et al., 2007) and cannot be aligned with confidence for many homeodomains (Berger, et al., 2008), complicating direct comparisons. To avoid overfitting, we considered a 75-homeodomain subset in which each protein is unique at the 15 amino acid positions described as making contact with DNA in the Engrailed crystal structure (Kissinger, et al., 1990) (Table 1), as we have previously shown that a perfect match at all 15 amino acids yields data comparable to experimental replicates of a single homeodomain (Berger, et al., 2008). All original datasets and supplementary data can be downloaded from http://hugheslab.ccbr.utoronto.ca/supplementary-data/profile_prediction/.

We assessed the performance of a panel of inference algorithms by a leave-one-out cross-validation approach, in which each of the 75 homeodomains was held out from the training set in turn and the remaining proteins were used as training data to predict the Z-score profile of the held-out protein, given its amino acid sequence. We used regression to create linear models via Principal Components Regression (PCR) and linear kernel Support Vector Regression (SVM_L). To create models in which interactions between TF sequence features can be captured, reflecting "combinatorial recognition codes" (Damante, et al., 1996), we also used Support Vector Regression with a polynomial kernel (SVM_P), or radial basis function kernel (SVM_R), Random Forests (RF) (Breiman, 2001) and a nearest-neighbour (NN) approach in which the profile of a held out protein was predicted as the averaged profiles of its nearest (fewest mismatches) sequence neighbour(s) in the training set. With the exception of the NN method, amino acid sequences of length l were numerically represented as binary vectors of length l x 20 digits, i.e. the 20 different amino acids were encoded as orthogonal 20 digit vectors and each protein sequence was represented by concatenating the binary vectors corresponding to residues at each position.

In each of these analyses we also considered three sets of features: (i) the full 57 amino acid homeodomain (omitting insertions), (ii) the subset of 15 amino acids that contact the DNA in the Engrailed structure (positions 3, 5, 6, 25, 31, 44, 46, 47, 48, 50, 51, 53, 54, 55, and 57) (Kissinger, et al., 1990), and (iii) six amino

Table 2. Leave-one-out cross-validation measures for 8-mer Z-score profile prediction algorithms on 32,896 8-mers for 75 homeodomains

Approach	Residues	Top100-overlap (predicted vs real)		Top100-overlap (control)	No. of proteins with top-100 overlap < 50	RMSE (predicted vs real)		RMSE (control)	Spearman (predicted vs real)		Spearman (control)	Median rank (Mean rank)
		median	mean			median	mean		median	mean		
replicates	N/A	86	82.84	80	0	0.63	0.58	-0.49	0.83	0.84	0.16	N/A
NN	15AA	66	58.60	61	18	0.72	0.76	-0.36	0.80	0.77	0.14	3.00 (7.20)
NN	6AA	66	58.65	60	18	0.68	0.77	-0.38	0.82	0.78	0.14	3.50 (6.00)
NN	57AA	69	58.07	62	18	0.75	0.82	-0.36	0.79	0.76	0.13	3.50 (9.00)
NN	top6	66	58.68	58	16	0.72	0.76	-0.35	0.81	0.78	0.13	5.00 (7.00)
NN	top15	69	57.00	63	19	0.75	0.80	-0.34	0.80	0.77	0.13	5.00 (8.70)
SVM_R	6AA	63	55.99	46	23	0.66	0.70	-0.26	0.83	0.81	0.09	5.50 (6.50)
RF	15AA	65	55.85	57	24	0.69	0.71	-0.25	0.83	0.81	0.12	6.00 (6.00)
RF	6AA	63	55.17	54	25	0.71	0.72	-0.25	0.83	0.81	0.12	7.00 (7.70)
SVM_R	57AA	60	51.51	41	28	0.69	0.73	-0.20	0.84	0.81	0.08	7.50 (9.70)
SVM_L	15AA	62	52.40	55	28	0.68	0.73	-0.30	0.82	0.79	0.10	8.00 (9.00)
SVM_R	15AA	63	55.28	50	21	0.66	0.71	-0.28	0.82	0.80	0.09	8.50 (7.40)
SVM_L	57AA	67	55.32	53	23	0.70	0.73	-0.22	0.83	0.79	0.09	8.50 (8.70)
SVM_L	6AA	62	54.51	52	28	0.68	0.73	-0.28	0.82	0.80	0.10	9.50 (8.40)
PCR	6AA	63	54.05	54	25	0.75	0.82	-0.30	0.79	0.75	0.12	10.0 (11.9)
PCR	15AA	63	53.45	55	29	0.72	0.77	-0.28	0.80	0.77	0.11	11.0 (11.0)
SVM_P	15AA	48	41.11	18	39	0.71	0.76	-0.17	0.83	0.81	0.08	11.0 (12.10)
RF	57AA	55	51.53	37	28	0.73	0.75	-0.16	0.84	0.81	0.08	12.0 (10.70)
SVM_P	6AA	49	41.65	16	38	0.70	0.76	-0.17	0.83	0.81	0.07	12.0 (12.6)
SVM_P	57AA	48	38.91	5	39	0.72	0.79	-0.12	0.84	0.80	0.06	15.0 (14.2)
PCR	57AA	60	48.48	51	32	0.77	0.79	-0.19	0.81	0.77	0.09	15.5 (14.3)

Algorithms are sorted in descending order of median rank across all columns, where ties are resolved using mean rank. The first row shows the agreement between 19 experimental replicates and their corresponding true Z-score profiles as measured using protein-binding microarrays. Columns labeled 'predicted vs real' show the mean or median performance between each predicted profile and its true, measured Z-score profile. Columns labeled 'control' show the difference between the median predicted vs real performance and the median of the performance between all pairs of predicted and actual profiles. Cells in a given column are coloured according to their position in the range of that column. Rows labeled top6 and top15 represent the result obtained if we use the 6 and 15 most important amino acid positions according to the RF importance score on the 57AA set.

acids that have been demonstrated to influence binding preferences (positions 3, 6, 7, 47, 50, and 54) (Ekker, et al., 1994; Laughon, 1991) (referred to as 6AA, 15AA, and 57AA). We did not consider *de novo* feature selection as part of our training process because feature selection consumes statistical (i.e. training) power, and arbitrary feature selection is NP-hard in the general case (Garey and Johnson, 1979). In Section 3.3 (below), we present evidence that residues scored highest by the RF importance score may be less predictive than literature-derived feature sets.

3.2 Assessing the performance of inference methods

The cross-validation results were assessed using three measures: (a) the number of top-100 8-mers in common, (b) Spearman correlation over all 8-mers, and (c) overall RMSE (Root Mean Squared Error) values between the predicted and the actual Z-score profiles over all 8-mers. As a summary statistic, we also counted the number of proteins with a top-100 overlap <50. As a background control, we calculated the difference between the median of each metric and the median of the performance of all predicted versus all actual profiles, since all homeodomain binding profiles correlate to a degree. Results are tallied in Table 2, which is sorted from best to worst median rank across all of the criteria. Included in Table 2 is the agreement between 19 experimental replicates as a reference for the reproducibility of the assay itself (Berger, et al., 2008) In these replicates, 19 different homeodomains were each analyzed in duplicate, and the numbers reported refer to 19 pairwise comparisons. Since the set of replicates contains some homeodomains not found in the 75 we analyzed, however, the performance values cannot be directly compared to those of the predictions.

Three major conclusions can be drawn from this analysis. First, results of all algorithms are clearly distinct from random (Table 2, columns 5, 9, and 12). Second, the 15AA and 6AA subsets appear to provide a superior training set relative to the 57AA set. Third, presumably due to the importance of non-linear interactions between amino acid positions in defining DNA-binding specificity, methods that can capture interactions and non-linearities have a clear advantage: there is almost always at least one variant of each non-linear method, i.e. NN, RF, and SVM_R, that outperforms every linear method we employed. NN (Figure 1, right panel) in particular has a significantly higher mean top-100 overlap than PCR (95% confidence interval (C.I.) for difference, 3.92-109; Kruskal-Wallis test). NN moreover often shows the greatest difference from random, and has the fewest number of predicted profiles with a top-100 overlap <50. In three instances (Evx1, Irx2, and Lhx1), the 15AA NN-predicted Z-score profiles exhibit Spearman correlation, top-100 overlap, or RMSE values that exceed those of the experimental replicates for these proteins. Therefore, it appears that predicted Z-score profiles can, in specific cases, rival experimental replicates in reproducing the Z-score profile of a given homeodomain. Fig. 2 shows scatter plots of the Z-scores for Evx1, Irx2, and Lhx1, comparing to the predicted and replicate Z-scores.

3.3 De novo feature selection

The feature sets we used were chosen on the basis of biochemical and genetic experiments to ask whether the use of this prior data to select features reduces generalization error. It is also of interest whether automated feature selection identifies the same residues, and whether automatically-selected features perform better than those selected using evidence from laboratory studies. Towards this end, we examined the "node purity" importance scores output by RF run with the full 57AA set. We summarized the importance

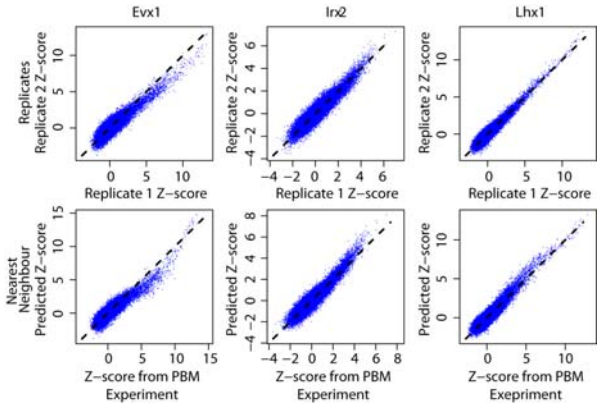

Fig. 2. Comparison of the accuracy of nearest-neighbour predictions versus experimental replicates. Scatterplots show the measured Z-scores for all 32,896 non-redundant 8-base DNA sequences from one PBM versus a second PBM for the same DNA-binding domain (top) or versus the Z-score predicted using NN (6 AA variant; bottom). Median performance metrics are given. Evx1 has a single nearest-neighbour (Hoxa2); Irx2 has a single nearest-neighbour (Irx3); Lhx1 has two nearest neighbors (Alx3 and Lhx3).

per residue for each of the 75 rounds of cross-validation by considering the median importance score for the 2,585 8-mers reported by Berger *et al.* to be bound in at least one experiment using the E > 0.45 criterion (see (Berger, et al., 2008)), reasoning that RF may be learning primarily noise for the remaining 8-mers. A very similar set of importance scores emerged from each of the 75-rounds of cross-validation (Fig 3). Considering the median importance score over all homeodomains over all 2,585 residues as a feature prioritization measure, we obtained the ranking of residues shown in Figure 3. The top 15 AA emerging from this analysis are (in descending order) 50, 6, 46, 54, 7, 56, 14, 28, 4, 19, 43, 22, 29, 36, 37. These residues include only four among our 6AA set (6,7,50,54) and four among our 15AA set (6,46,50,54). Thus, de novo feature selection identifies some, but not all, of the same residues as laboratory studies. We found that the top-6 and top-15 residues selected by the RF importance score did not perform as well in NN (our best performing method) as did the original 6AA and 15AA sets (Table 2). A possible explanation is that *de novo* feature selection is identifying residues that correlate with binding specificity, but without being causative; for example, residues that participate in functions of the homeodomains besides DNA-binding, those that are shared due to common evolutionary descent, and/or those that co-vary due to structural constraints (Clarke, 1995). From these results, and the fact that the 6AA and 15AA sets generally provide better features (Table 2), we propose that use of experimental evidence in the feature selection step can augment training power, by incorporating external information.

3.4 Association between prediction difficulty and number of sequence mismatches

In general, the 8-mer profiles that are difficult for one algorithm to predict are those that are difficult for other algorithms as well. Figure 4 compares the top-100 overlap for all 75 homeodomains for all prediction methods, using the 15AA feature set. The colours of the points reflect the NN distance. There is a clear relationship between the 15AA distance and the top-100 overlap, with the ten proteins with the greatest distance consistently having overlaps <50, indicating that for all methods the difficulty of learning the 8-mer profile for a specific experiment is related to whether there is a similar example in the training set. This trend also holds for other feature sets, and likely explains the success of NN, which does not incorporate any information from more distant profiles.

4 DISCUSSION

Our results show that the full DNA-binding specificity of uncharacterized TFs to individual k-mers can be predicted on the basis of similarity in protein sequence alone, given the sequence specificity of closely-related members of the same TF family, and (preferably) knowledge of the DNA-contacting residues. Our results are likely to underestimate real-world accuracy because we only evaluated homeodomains that are unique at 15 DNA-contacting amino acids. The efficacy of NN makes predicting binding preferences simple to implement and consistent with intuition: it is typically assumed that similarity among functional residues reflects similar protein activity. At least one previous study applied a nearest-neighbour strategy to the inference of PWMs (Qian, et al., 2007), but our NN implementation is more straightforward and provides relative affinity estimates for individual sequences: in contrast, the approached described by Qian *et al.* predicts the consensus motifs of TF binding sites from the TRANSFAC database using the InterPro annotations of the TF of interest and its target genes as training data (Qian, et al., 2007).

Our results are consistent with the "combinatorial code" model of TF binding (Damante, et al., 1996; Suzuki, et al., 1995; Suzuki and Yagi, 1994). In this model, the relative preference of a TF to individual bases in a given DNA sequence is determined by the aggregate identities of a subset of key amino acid residues. In our regime, this model would translate into interaction terms among amino acid residues. Indeed, in our analysis, methods capable of modeling interactions between amino acid positions, such as NN and Random Forests, appear to be best suited to predicting sequence preferences for TFs, or at least for homeodomains. The fact that linear regression is one of the least effective methods among those tested further supports the importance of interaction terms; preferences to individual DNA sequences apparently cannot be taken as a linear combination of the contributions of each amino acid residue.

In addition, the observation that incorporation of the full set of homeodomain residues adversely affects all success measures that were employed here, even using NN (which would not be subject to overfitting), is consistent with a model in which the remainder of the domain structure primarily plays a role as a scaffold, at least with regard to DNA-binding. This is because such a role would provide flexibility in residue identities without impacting DNA sequence specificity.

An important question is whether the outcome of our comparisons would be different with different feature sets, and whether our results could be improved with more sophisticated approaches. With regard to feature sets, even in a circular regime (selecting amino acids using the same data used to test them) we found no feature sets that offered a substantial improvement over the 6 and 15 AA sets (Figure 3, Table 2, and data not shown), suggesting

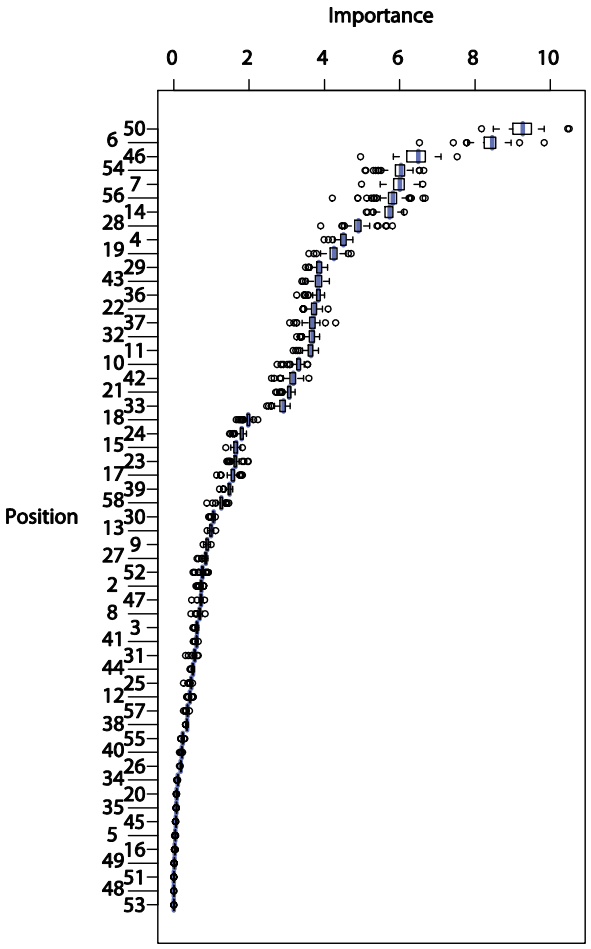

Fig. 3. Node purity importance scores for 57 homeodomain amino acid positions for 75 rounds of leave-one-out cross-validation, sorted by median (purple).

that a single DNA-protein co-crystal structure constitutes a that a single DNA-protein co-crystal structure constitutes a powerful feature selection step, perhaps because it provides information that is not available to the algorithms used here. Nonetheless, it is possible that addition of an automated feature selection step might be advantageous, particularly if it is incorporated into the cross-validation regime, i.e. if the feature selection is done separately at each LOO iteration, and/or if it is done in conjunction with feature selection based on experiments. Due to the large number of permutations, we did not explore such variations in this study, nor did we test every possible variation of the techniques represented. For example, it has been reported that pruned decision trees usually perform better than unpruned trees; this was not an option in the RF implementation that we used but would be worth examining. It may also be beneficial in the future to take advantage of similarity among k-mers. In all of the analyses presented here, each k-mer is treated as a separate learning problem; however, there are relationships among the k-mers in both sequence and affinity for individual proteins. Exploration of these variations could shed light on the biology of DNA-binding in addition to improving prediction results. We note, however, that there are also benefits associated with use of simple inference methods such as NN. While performing as well as other methods, NN is computationally much less intensive than any other method we tested – in our algorithm, NN is determined based on protein sequence alone, so the time complexity of this part of the algorithm does not depend upon the number of k-mers. Also, the success of NN on the full set of k-mer affinities suggests that our NN approach would also work well when the binding preferences of each TF were represented differently, e.g. as a PWM.

Another question is whether better results could be obtained using a training set that more completely samples possible homeodomain amino acid combinations. With regard to sampling depth in the training set, the argument may be academic: the large number of possible combinations would be impractical to survey in the laboratory, and also appears to be sparsely-populated in nature (data not shown) (Berger, et al., 2008). A more extensive PBM–based survey of the binding preferences of naturally-occurring unique combinations among DNA-contacting residues might be the next step towards both theoretical and practical aims. Such a survey would also help clarify the functional evolution of the distinct homeodomain subclasses. One interpretation of the success of NN—coupled with the fact that all algorithms suffer considerably when there is no similar homeodomain in the training set to serve as an example—is that homeodomain groups (described in Banerjee-Basu and Baxevanis, 2001, although the groups we obtained are not always identical) each have distinct DNA-binding modes that cannot be inferred from examples in other groups. Consistent with this notion, there is a strong correspondence between 8-mer binding profiles and sequence groups obtained by ClustalW (data not shown). In fact, we cannot rule out that RF and/or SVM_R are acting in essence as a more sophisticated version of NN, by learning group memberships. We have attempted to improve upon our current results using unsupervised sequence clustering approaches, but have not yet been able to improve upon the NN results (data

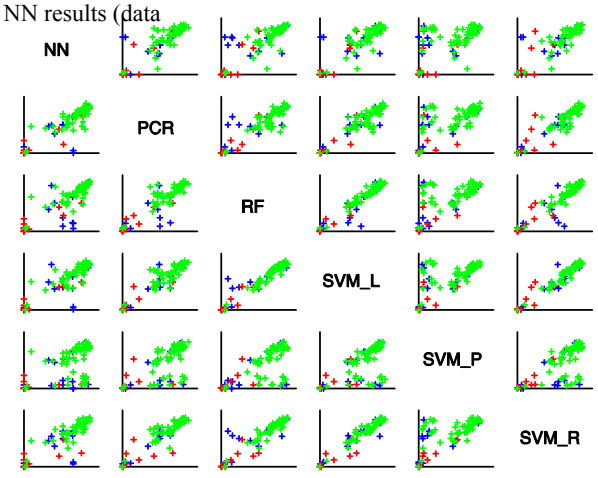

Fig. 4. Association between top-100 overlap scores for pairs of 8-mer profile inference methods. Scatterplots show the top-100 overlap values for 75 homeodomains when Z-score profiles are predicted using one inference method versus another method for the same proteins. All axes range from 0 to 100. The names on the diagonal label the axes. Predictions are made using the 15 homeodomain DNA-contacting residues. Homeodo-

mains are coloured according to whether they have ≥5 (red), 3-4 (blue), or 1-2 (green) mismatches to their nearest sequence neighbor.

not shown). One explanation for this outcome may be that there are no ideal natural subdivisions within these groups; instead, there is variation on a theme within each group, and the variation in amino acid sequence bears a relationship to that seen in the 8-mer binding profiles. If this is the case, then even better inference results might be obtained from a two-stage process in which group assignment is separated from k-mer profile prediction within a group.

Finally, our preliminary results (data not shown) suggest that NN will be similarly applicable to other DBD classes. Extension of the work presented here should allow future experimental studies of binding specificity to focus on proteins most likely to possess new DNA-binding activities, and will facilitate more accurate inference of DNA-binding data among proteins with related sequences.

FUNDING

T.M.A. was supported by an Ontario Graduate Scholarship, and L.P.C. was supported by an Ontario Women's Health Postdoctoral Fellowship. Generation of the experimental data analyzed was supported by grants to T.R.H. and M.L.B. from CIHR, Genome Canada through the Ontario Genomics Institute, the Ontario Research Fund, and NIH/NHGRI.

ACKNOWLEDGEMENTS

We are grateful to Harm van Bakel and Jeff Liu for maintenance of computational infrastructure. We thank Gary Bader and Alan Davidson for helpful discussions.

REFERENCES

Ades, S.E. and Sauer, R.T. (1994) Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50, *Biochemistry*, **33**, 9187-9194.

Banerjee-Basu, S. and Baxevanis, A.D. (2001) Molecular evolution of the homeodomain family of transcription factors, *Nucleic Acids Res*, **29**, 3258-3269.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, *Nucleic Acids Res*, **32**, D138-141.

Benos, P.V., Bulyk, M.L. and Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it?, *Nucleic Acids Res*, **30**, 4442-4451.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S.A., Morris, Q.D., Bulyk, M.L. and Hughes, T.R. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences, *Cell*, **133**, 1266-1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd and Bulyk, M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities, *Nat Biotechnol*, **24**, 1429-1435.

Breiman, L. (2001) Random Forests, *Machine Learning*, **45**, 5-32.

Chen, X., Hughes, T.R. and Morris, Q. (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors, *Bioinformatics*, **23**, i72-79.

Clarke, N.D. (1995) Covariation of residues in the homeodomain sequence family, *Protein Sci*, **4**, 2269-2278.

Damante, G., Pellizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S. and Di Lauro, R. (1996) A molecular code dictates sequence-specific DNA recognition by homeodomains, *EMBO J*, **15**, 4992-5000.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E. and Beachy, P.A. (1994) The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins, *EMBO J*, **13**, 3551-3560.

Franklin, S.B., Gibson, D.J., Robertson, P.A., Pohlmann, J.T. and Fralish, J.S. (1995) Parallel Analysis: A Method for Determining Significant Principal Components, *Journal of Vegetation Science*, **6**, 99-106.

Garey, M.R. and Johnson, D.S. (1979) *Computers and intractability : a guide to the theory of NP-completeness*. W. H. Freeman, New York.

Hanes, S.D. and Brent, R. (1989) DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9, *Cell*, **57**, 1275-1283.

Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions, *Cell*, **63**, 579-590.

Laughon, A. (1991) DNA binding specificity of homeodomains, *Biochemistry*, **30**, 11357-11367.

Liu, X., Noll, D.M., Lieb, J.D. and Clarke, N.D. (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity, *Genome Res*, **15**, 421-427.

Miller, M., Shuman, J.D., Sebastian, T., Dauter, Z. and Johnson, P.F. (2003) Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein alpha, *J Biol Chem*, **278**, 15178-15184.

Montgomery, D.C. and Runger, G.C. (2007) *Applied statistics and probability for engineers*. Wiley, Hoboken, NJ.

Mukherjee, S., Berger, M.F., Jona, G., Wang, X.S., Muzzey, D., Snyder, M., Young, R.A. and Bulyk, M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays, *Nat Genet*, **36**, 1331-1339.

Pabo, C.O. and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?, *J Mol Biol*, **301**, 597-624.

Papavassiliou, A.G. (1995) Transcription factors: structure, function, and implication in malignant growth, *Anticancer Res*, **15**, 891-894.

Qian, Z., Lu, L., Liu, X., Cai, Y.D. and Li, Y. (2007) An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization, *Bioinformatics*, **23**, 2449-2454.

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

Suzuki, M., Brenner, S.E., Gerstein, M. and Yagi, N. (1995) DNA recognition code of transcription factors, *Protein Eng*, **8**, 319-328.

Suzuki, M. and Yagi, N. (1994) DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families, *Proc Natl Acad Sci U S A*, **91**, 12357-12361.

Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., Jr. and Ansari, A.Z. (2006) Defining the sequence-recognition profile of DNA-binding molecules, *Proc Natl Acad Sci U S A*, **103**, 867-872.