BIOINFORMATICS

Vol. 28 ISMB 2012, pages i84–i89
doi:10.1093/bioinformatics/bts202

Recognition models to predict DNA-binding specificities of homeodomain proteins

Ryan G. Christensen[1], Metewo Selase Enuameh[2], Marcus B. Noyes[2,3], Michael H. Brodsky[2,4], Scot A. Wolfe[2,3] and Gary D. Stormo[1,*]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, [2]Program in Gene Function and Expression, [3]Department of Biochemistry and Molecular Pharmacology, [4]Department of Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA

ABSTRACT

Motivation: Recognition models for protein-DNA interactions, which allow the prediction of specificity for a DNA-binding domain based only on its sequence or the alteration of specificity through rational design, have long been a goal of computational biology. There has been some progress in constructing useful models, especially for C_2H_2 zinc finger proteins, but it remains a challenging problem with ample room for improvement. For most families of transcription factors the best available methods utilize k-nearest neighbor (KNN) algorithms to make specificity predictions based on the average of the specificities of the k most similar proteins with defined specificities. Homeodomain (HD) proteins are the second most abundant family of transcription factors, after zinc fingers, in most metazoan genomes, and as a consequence an effective recognition model for this family would facilitate predictive models of many transcriptional regulatory networks within these genomes.

Results: Using extensive experimental data, we have tested several machine learning approaches and find that both support vector machines and random forests (RFs) can produce recognition models for HD proteins that are significant improvements over KNN-based methods. Cross-validation analyses show that the resulting models are capable of predicting specificities with high accuracy. We have produced a web-based prediction tool, PreMoTF (Predicted Motifs for Transcription Factors) (http://stormo.wustl.edu/PreMoTF), for predicting position frequency matrices from protein sequence using a RF-based model.

Contact: stormo@wustl.edu

1 INTRODUCTION

It is a long-standing goal to predict the DNA-binding specificity of a transcription factor (TF) based only on its protein sequence (Benos *et al*., 2002a). That ability would allow for the inference of regulatory networks from genome sequences alone as well as for the design of TFs with desired recognition sequences. Early hopes for a simple recognition code (Seeman *et al*., 1976) were dashed by the structures of the first few protein-DNA complexes (Matthews, 1988). As the structures of more DNA-protein complexes were determined it became apparent that there were definite preferences for interactions between particular amino acids and base pairs, and that those preferences might vary depending on the class of TF. C_2H_2 Zinc finger proteins, in particular, were heavily studied and a degenerate, qualitative recognition code was shown to be moderately successful at predicting their binding specificities (Choo and Klug, 1994a; Choo and Klug, 1994b; Choo and Klug, 1997; Wolfe *et al*., 1999; Wolfe *et al*., 2000). This general idea of class-specific, degenerate recognition codes was further developed into a quantitative probabilistic code for zinc finger proteins that was more predictive than previous models but sill less accurate than desired (Benos *et al*., 2002b; Kaplan *et al*., 2005; Liu and Stormo, 2008; Persikov and Singh, 2011; Persikov *et al*., 2009).

The homeodomain (HD) family is the second most abundant TF family in mammals and most metazoans (Tupler *et al*., 2001). The HD TF family was first discovered in Drosophila where mutations in some HD proteins caused severe 'homeotic' phenotypes (Lewis, 1978). Homeodomains typically span ~60 residues that fold into a stable bundle of three alpha helices (Gehring *et al*., 1994). The C-terminal helix, or recognition helix, binds in the major groove and an unstructured N-terminal arm binds in the minor groove. This domain provides a favorable family for construction of a predictive recognition model because of similarities in docking geometry for a number of family members (Pabo and Nekludova, 2000; Siggers *et al*., 2005) and the characterization of specificity for many members of this family present in the yeast, drosophila and mouse genomes (Table 1). Using specificity data for 263 HD proteins, mostly determined using new high-throughput methods (Stormo and Zhao, 2010), allowed us to test different machine learning approaches and to assess the ability of recognition models to accurately predict the specificity of HD proteins. Using a cross-validation methodology we demonstrate that both support vector machines (SVMs) and random forests (RFs) based methods produce recognition models that are significantly better than previously published methods.

2 METHODS

2.1 Protein alignment

Table 1 lists the number of HD proteins from each of five species (including 13 variants of fly HD proteins), the experimental method used to determine their specificity and the reference for the datasets. All wild type protein sequences were obtained from UniPROBE (Newburger and Bulyk, 2009) or FlyFactorSurvey (Zhu *et al*., 2011). The hmmsearch program from the HMMER suite (Bateman *et al*., 1999) was used to extract the DBD for every protein using the homeobox Pfam hmm model (Pfam ID: PF00046) (Finn *et al*., 2010).

Protein binding microarray (PBM) data are available for 168 mouse HD proteins, but 14 of these (Dbx1, Hoxb5, Hoxb6, Hoxc6, Irx5, Lhx5, Lhx9, Lmx1b, Obox2, Pax6, Phox2a, Six6, Titf1, Tlx2) were removed from the dataset because their scaled BEEML-PBM PWMs (position weight matrices; see below) had a total information content (IC) <3 std below the mean IC for the entire combined set of motifs.

*To whom correspondence should be addressed.

Table 1. Source of HD motifs

DataSource	Species	Number	Reference
PBM	Mouse	154	(Berger *et al*., 2008)
B1H, SOLEXA	Fly	84	(Zhu *et al*., 2011)
B1H, Sanger	Fly directed mutants	13	(Noyes *et al*., 2008)
B1H, Sanger	Human	8	(Noyes *et al*., 2008)
PBM	Yeast	4	(Zhu *et al*., 2009)

The HD DBDs were aligned using MAFFT (Katoh *et al*., 2005), which gave higher quality alignments, with fewer gaps, than other programs. Perhaps the best studied HD protein is the Drosophila engrailed protein (Fraenkel *et al*., 1998; Kissinger *et al*., 1990; Liu *et al*., 1990; Sato *et al*., 2004). For consistency with previous studies, HD positions are numbered with respect to the engrailed HD and all columns in the multiple sequence alignment that contained insertions relative to engrailed were removed. Only a small minority of the proteins in the dataset had short insertions relative to engrailed. The majority of these insertions (26) correspond to the three residue (TALE) insertion between positions 21 and 22 (Burglin, 1997). Two other types of insertions only occur in the mouse proteins Hdx, Hmbox1, Tcf1 and Tcf2. Figure 1 displays a sequence logo for the set of aligned HD proteins used in this study (Crooks *et al*., 2004).

2.2 HD motif scaling and alignment

Specificities for each HD protein were initially represented by PWMs or position count matrices (PCMs; Stormo *et al*., 1982). For bacterial one-hybrid (B1H) datasets the PCMs were obtained from the FlyFactorSurvey database (Zhu *et al*., 2011) for fly proteins. The PCMs for human and mutant fly proteins were obtained from Noyes *et al*. (2008). For PBM datasets, PWMs were generated using the BEEML-PBM program which provides more accurate PWMs than other analysis methods (Zhao and Stormo, 2011). Although Alleyne *et al*. (2009) were not able to align motifs for HD proteins based on their analysis of PBM data, the PWMs that we obtained could be aligned confidently (see below). For purposes of alignment and recognition modeling the PWMs were converted to position frequency matrices (PFMs), where the elements at each position are the probabilities of each base occurring, using:

$$P_{i,b} = \frac{e^{-W_{i,b}}}{\sum_b e^{-W_{i,b}}}$$

However, we noticed when comparing the PWMs from orthologous HD proteins between flies, obtained using the B1H method, and mouse, using the PBM method, that there was a difference in scaling. A set of close mouse and fly homologs was assembled for comparison. For each mouse HD protein, the fly protein that was identical at key recognition residues 5, 47, 50, 51, 54, 55 (Noyes *et al*., 2008) and that was the smallest Hamming distance away was chosen as the homolog likely to have the most similar motif. Only mouse and fly HD pairs with a Hamming distance <15 were used. For each of the 104 homologous pairs we determined the optimal scaling factor. The mean of all of the optimal scalar values for all pairs was 2.238. This scalar value was then used to scale up all of the BEEML-PBM PWMs before converting them to PFMs.

Mahoney *et al*. developed a multiple PFM alignment program, STAMP, that produces reliable motif alignments (Mahony *et al*., 2007a; Mahony *et al*., 2007b). They further showed a mutual information analysis between the motif alignments and the protein alignments could be used to determine some of the key interacting residues for various TF families (Mahony *et al*., 2007b). We implemented a similar multiple alignment program in MATLAB to test additional scoring metrics and methods of guide tree construction. After developing metrics which incorporated the per column IC, we found that the best metric was SSDs (sum of squared differences), also one of the two best metrics reported by Mahony *et al*. We also found that ungapped local alignments yielded the best PFM alignments and that it was important to perform motif core alignment, as outlined by Mahoney *et al*. when aligning the relatively short HD motifs to generate a multiple motif alignment (MMA). Motif cores are defined as consecutive positions in a motif having an IC above 0.3 bits, or if there were not at least four consecutive columns with IC >0.3 bits, then the four consecutive columns with the highest total IC were used. If the motif was <4 base pairs long, then the entire motif was designated as the motif core. The guide tree we used for progressive motif alignments was based on the Euclidian distances between PFMs, which is simpler than the *p*-value-based alignment trees used by STAMP and produces similar results.

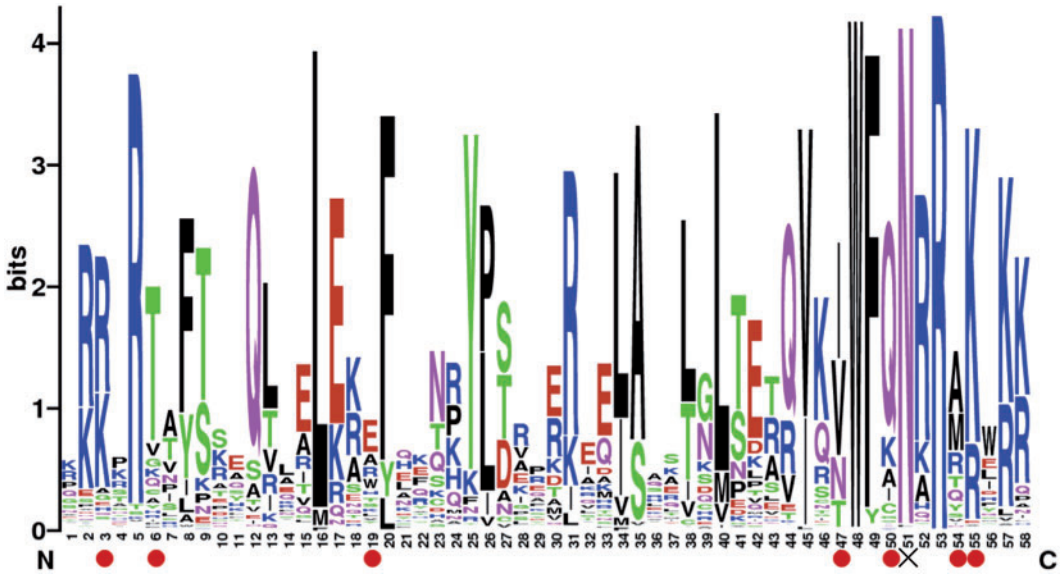

Fig. 1. Sequence logo of the MAFFT-generated HD multiple sequence alignment used for training the recognition models. The circles denote positions identified by our feature selection method (positions 3, 6, 19, 47, 50, 54, 55). Most HD proteins contain Asn51 ('X' symbol), which is a critical residue in recognition that binds Adenine with high specificity. HDs lacking Asn51, such as Lag1, tend to have very divergent recognition motifs

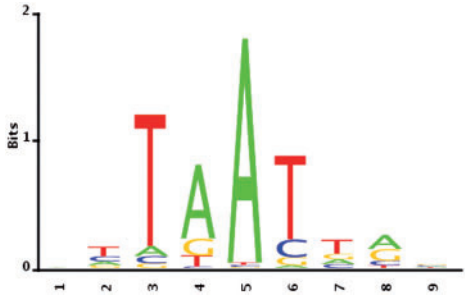

Fig. 2. Average PFM for the trimmed HD multiple motif alignment

We trimmed flanking positions from the MMA that had a mean IC of <0.05. The final trimmed MMA consisted of nine positions. Figure 2 shows the logo for the average PFM for the entire HD dataset. The A at position 5 in the average HD PFM is almost completely conserved. This preference reflects the strong conservation of Asn at position 51, which effectively specifies the conserved A (Ades and Sauer, 1995; Fig. 1)

2.3 Feature selection for machine learning

Machine learning methods like SVMs and RFs are relatively good at capturing both the non-linear and linear relationships between independent variables that give rise to the observed dependent variable. In our case the independent variables are the amino acids in the DBD of each protein. The dependent variables are the elements of the PFM for each protein, 27 free parameters for the 9 positions of the HD motifs (Stormo, 2011). The SVM and RF methods can both perform feature selection as part of the learning process, but we found that ranking features with an adjusted form of mutual information (MIp) (Dunn *et al*., 2008) led to faster training and resulted in more accurate models. To make each PFM matrix discrete, we used a profile alphabet (Wang and Stormo, 2005). Nineteen different multinomial probability distributions over the four bases A, C, G and T were defined and assigned to labels in an alphabet. This allowed us to convert each column of a PFM to a single discrete label. We began with an initial set of seed profiles chosen to cover the space of possible PFM columns and then iteratively refined those profiles. The Euclidean distance between every column from every PFM in the MMA and all 19 profile vectors was calculated. Each PFM column was assigned to the nearest profile. The mean of the set of column vectors assigned to each profile then became the new profile vector. This process was repeated until convergence.

MIp was used to rank the positions in the DBDs. The adjustment to the mutual information proposed by Dunn *et al*. (2008) is based on the idea that the average entropy of each position in an alignment gives each position a particular propensity toward mutual information. A simple correction to the MI, called the average product correction (APC), takes into account the average MI across all of the positions. The final corrected mutual information is $\mathrm{MIp(a,b)} = \mathrm{MI(a,b)} - \mathrm{APC(a,b)}$.

The MIp score was calculated for every possible protein and motif position pair. Using the maximum MIp score for each position in the protein alignment, the set of protein positions, or potential features, was then sorted according to max(MIp) resulting in a sorted set of features. The RF, SVM and *K*-nearest neighbor (KNN) methods (described below) were each used to train a set of models, one model per element of the PFM. WYK encoding (Stormo, 2011) was used so that only free parameters were predicted by the models, since each position of a PFM only has three free parameters due to the constraint that each PFM column sums to one. To avoid imposing an arbitrary MIp score cutoff to determine the feature set to use for training the recognition models, the sorted set of features was used to construct progressively larger feature sets. The first feature set contained just the single DBD position with the highest MIp score. The last feature set in the series contained all of the residues in the DBD.

2.4 Machine learning algorithms

KNNs is the best published approach for predicting the specificity of HD proteins (Alleyne *et al*., 2009; Noyes *et al*., 2008). We included that in our set of methods so that each approach is trained on exactly the same datasets. Methods we tested included SVMs, RFs, neural nets (NNs) and partial least squares regression (PLSR). In pilot studies we found that NN and PLSR did not perform well compared with RF and SVM and NN was computationally expensive to run, so comprehensive tests were only performed to compare the KNN, SVM and RF methods.

2.4.1 k-Nearest neighbors The KNNs algorithm is a very simple method based on the principle that similar inputs generally yield similar outputs. The Hamming distance between every pair of aligned proteins was calculated. For every query protein, the closest proteins in the training set served as the reference proteins. The average of reference proteins' PFMs were then used as the prediction for the query protein. If the k parameter is set to 1, only the closest protein is included in the reference set. If it is set to 2, then the second closest protein is also included, etc. In the case of a tie all of the corresponding PFMs were averaged together. We used the knnflex R package for the KNN analysis. Surprisingly, in preliminary studies involving the HD dataset, we found that tuning the k parameter did not increase performance, so we fixed the $k = 1$ for all subsequent analysis.

2.4.2 Random forest regression RF is an ensemble method that makes use of a collection of weak decision trees (Breiman, 2001). There are only two main user specified parameters, the total number of trees in the ensemble (ntree) and the number of randomly selected features to use to determine the best split at every node in each tree (mtry), so the method is simple to tune. In practice, it works well even without tuning, although we did tune the mtry parameter for every RF model. We iteratively tried increasing values of mtry, from 2 to 50. For all of these iterations, we set ntree to 50 for increased speed. Next, the mtry versus mean squared error (MSE) curve was smoothed and the mtry parameter that yielded the best MSE value was determined. Then, we used the optimal mtry parameter, and we set ntree to 500, which is the default value. We used the R randomForest package (Liaw and Wiener, 2002) to generate and make predictions with RFs.

2.4.3 Support vector machine regression SVMs are a popular machine learning binary classifier that has been adapted to perform regression. SVMs use a kernel function to map a set of training vectors into a higher dimensional space. They find the linear separating hyperplane that maximizes the margin of separation in this higher dimensional space. A non-negative cost parameter, C, is set by the user and determines the weight of the error term in the minimization (Chang and Lin, 2011). The e1071 R package was used to train and tune the SVM models. This package is based on the libsvm program (Chang and Lin, 2011). The dependent variables were each centered and scaled as recommended. Of the available kernel functions, the authors of libsvm recommend the radial basis function as the simplest to tune and the most generally applicable. The radial basis function includes a gamma parameter, which we tuned along with the cost parameter using the grid search function implemented in the e1071 package. We searched over the range $2^{-15} \leq \gamma \leq 2^3$ and $2^{-5} \leq C \leq 2^3$ using a step size of 2^2. In preliminary studies involving the HD dataset, we tried using all of the different kernel functions available in libsvm. However, we found that the radial basis function performed better than, or as well as, the other kernel functions and it was much less expensive to tune in most cases.

3 RESULTS

We aligned the HD amino acid sequences and DNA binding specificities (described as PFMs) for a set of 263 HD proteins

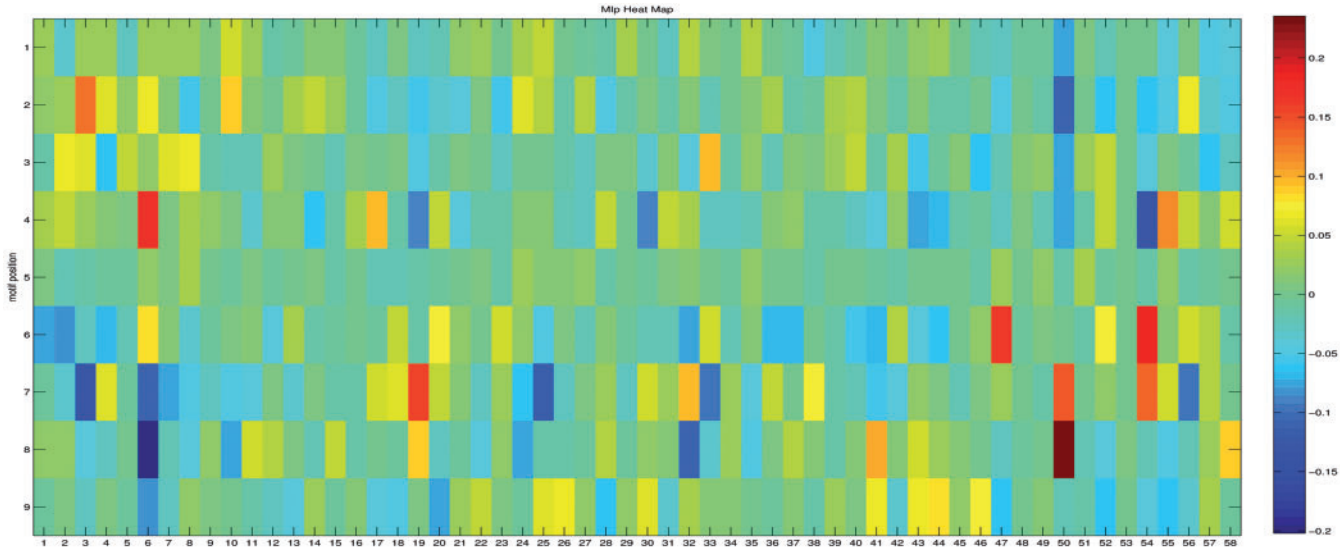

Fig. 3. Heat map showing the protein alignment (horizontal axis) versus motif alignment (vertical axis) MIp matrix

(Table 1). MIp analysis of these proteins successfully ranked the most important key residues with only one false positive in the case of the HD dataset (Fig. 3, Table 2). Position 19 had a high MIp score but to our knowledge it has never been observed to make base-specific interactions as it is found on the opposite side of the domain from the DNA-recognition surface (Fraenkel *et al*., 1998). The other top six positions identified by our MIp-based feature selection analysis are key residues known to interact with the bases directly in at least one structure (Fraenkel *et al*., 1998; Passner *et al*., 1999; Wolberger *et al*., 1991) or influence specificity by substitution between HDs (Damante *et al*., 1996). While other groups have used overlapping, but somewhat different, sets of key residues (Alleyne *et al*., 2009), we find that including additional features beyond the set selected using MIp actually decreased performance for the SVM and KNN models, and only increased performance slightly for RF (see below).

Assessments of each method were based on 10-fold cross validation. The entire set of 263 HD proteins and their motifs were randomly divided into 10 subsets and in each of 10 training runs, 9 subsets were used for training the model and the remaining 1 was used to assess the accuracy of the model. Accuracy was measured as the MSE for each parameter of the predicted PFM compared with the observed PFM for each protein in the test set.

Figure 4 shows the performance for the KNN, SVM and RF methods as increasing numbers of features (protein positions) are included. All methods increase in performance (decrease in MSE) for the first seven features, after which they plateau or even increase MSE. Table 2 lists the positions added as features, the order of which was determined by the MIp ranking. Those features that improved the MSE appreciably for all methods are highlighted in yellow. For the KNN and SVM methods, performance tended to decrease when eight or more features were employed. The performance of the RF model did increase slightly in general as more features were added. Apparently the RF method does a slightly better job of internal feature selection here than even the SVM method. Based on the literature (Damante *et al*., 1996; Ekker *et al*., 1994; Kissinger *et al*., 1990; Noyes *et al*., 2008), six of these seven residues are thought to be important residues in sequence specific DNA recognition in at least some contexts.

Our previously published KNN-based method, flyhd (Noyes *et al*., 2008), was used to make predictions for all 154 mouse proteins considered in this study. Flyhd made predictions for 130 of these. A RF-based model was trained using the same training data employed by flyhd, which consists of, on average, 22 binding sites selected by B1H per fly HD protein. The MSE for predicting the mouse PFMs with flyhd was 0.0159, whereas the MSE for the RF model was 0.0113, 29% lower. Training an RF model using the latest

Table 2. MIp-ranked features for HD protein and motif alignments

Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Residue number	50	54	6	47	1	3	55	41	33	17	32	58	10	44	20
Features	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Residue number	52	46	38	26	43	2	8	56	25	4	18	7	24	30	11

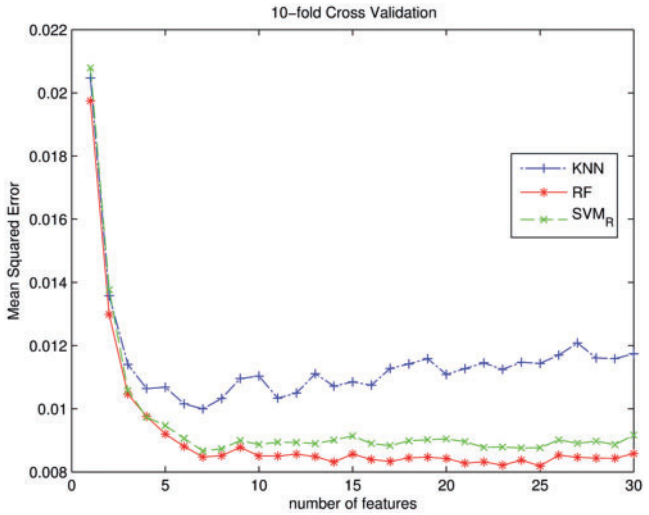

Fig. 4. Plot of the number of features used to train the KNN, RF and SVM models versus the 10-fold cross validation MSE values. After seven features are included the MSE stoped decreasing for KNN and SVM and did not decrease much for RF. Only the top 30 features were considered

SOLEXA motifs available from flyFactorSurvey lowered the MSE by an additional 8.6% to 0.0099.

4 DISCUSSION

The goal of this study was to find improved recognition models for the HD family of TFs. Although they are a very abundant TF family in nearly all eukaryotic organisms, previous models have focused on simple nearest neighbor type predictions of specificity based on large archives of reference recognition motifs (Alleyne *et al*., 2009; Noyes *et al*., 2008). These previous models have not been informed by the combination of feature selection derived from mutual information and modeling against reliably aligned recognition motifs, which could limit their predictive power. We had previously published a nearest neighbor approach (Noyes *et al*., 2008) for predicting PWMs for new HDs, but we found that an RF-based model trained with the same fly B1H Sanger data had better performance on a mouse test set. Alleyne *et al*. (2009) also used a nearest neighbor approach but they only attempted to predict 8mer enrichment scores for novel HD proteins. They compared various machine learning methods but found nothing that was better than nearest neighbors, but their efforts may have been hampered by an inability to align the motifs they used for training and therefore having to rely on 8mer enrichments. Our results demonstrate that with ample high quality and quantitative training data sophisticated machine learning methods are capable of determining very good recognition models for HD proteins. We believe that these models should be broadly applicable to other families of TFs with the caveat that large deviations in domain docking within subgroups could complicate family analysis (Pabo and Nekludova, 2000; Siggers and Honig, 2007). In support of this view, we have preliminary data demonstrating that this type of approach can provide improved models for DNA recognition by zinc finger proteins even though they have been extensively studied (Benos *et al*., 2001; Liu and Stormo, 2008; Persikov and Singh, 2011).

In the 10-fold cross validation analysis, the average MSE for this RF model was 0.0085 (Fig. 4). Figure 5 compares the observed and predicted motif logos for 12 different HD DBDs that are in the range of the average MSE (0.0080–0.0091) to illustrate the expected accuracy of the predictions. We have produced a web-based prediction tool, PreMoTF (Predicted Motifs for Transcription Factors) (http://stormo.wustl.edu/PreMoTF), for predicting PFMs

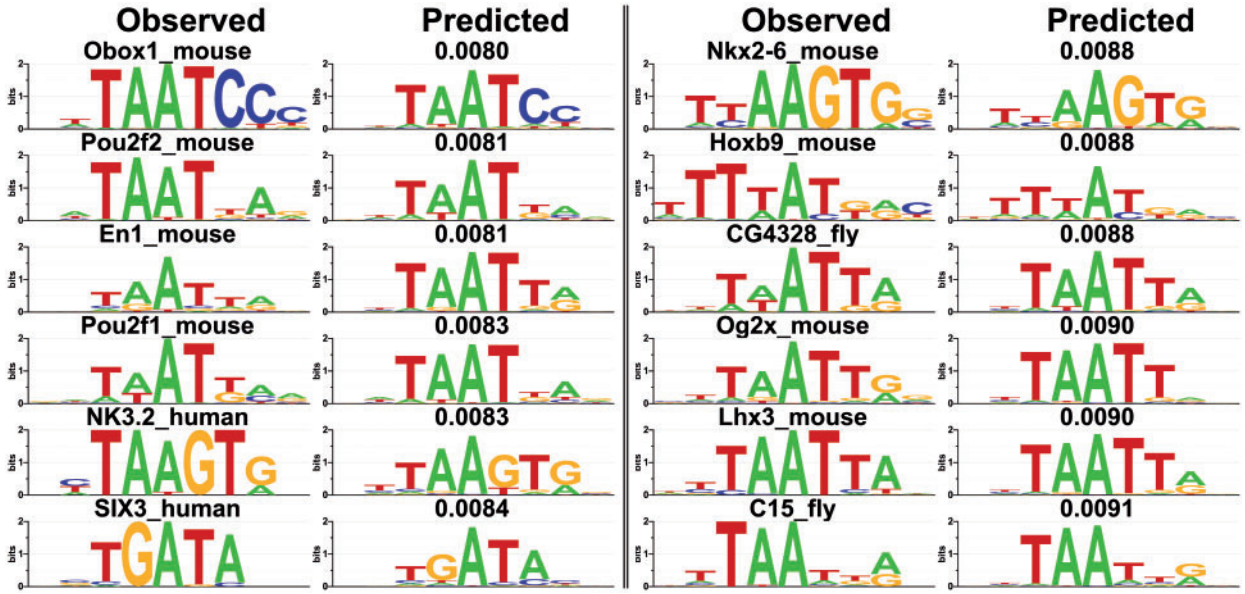

Fig. 5. Comparison of logos for actual and predicted motifs. The predicted motifs are from the 10-fold cross validation analysis RF model and positions 3, 6, 19, 47, 50, 54, 55. The names above each observed motif are the HD domain used for prediction and the MSE between the observed and predicted PFMs are provided above the predicted motifs

based on protein sequence. It currently contains prediction tools for HD proteins and additional protein families will be added as they are developed.

ACKNOWLEDGEMENTS

We gratefully acknowledge Yue Zhao for providing the HD BEEML-PBM PWM models and for helpful advice and discussions. We also thank the other members of the Stormo, Wolfe and Brodsky labs for insightful comments and discussions.

Funding: This work supported by National Institutes of Health grants HG00249 (GDS) and HG004744 (SAW and MHB).

Conflict of Interest: none declared.

REFERENCES

Ades,S.E. and Sauer,R.T. (1995) Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry*, **34**, 14601–14608.

Alleyne,T.M. *et al*. (2009) Predicting the binding preference of transcription factors to individual DNA k-mers. *Bioinformatics*, **25**, 1012–1018.

Bateman,A. *et al*. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.

Benos,P.V. *et al*. (2001) SAMIE: statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.*, **6**, 115–126.

Benos,P.V. *et al*. (2002a) Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays*, **24**, 466–475.

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

Berger,M.F. *et al*. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.

Breiman,L. (2001) Radom forests. *Mach. Learn.*, **45**, 5–32.

Burglin,T.R. (1997) Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. *Nucleic Acids Res.*, **25**, 4173–4180.

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.

Choo,Y. and Klug,A. (1994a) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.

Choo,Y. and Klug,A. (1994b) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.

Choo,Y. and Klug,A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Damante,G. *et al*. (1996) A molecular code dictates sequence-specific DNA recognition by homeodomains. *The EMBO J.*, **15**, 4992–5000.

Dunn,S.D. *et al*. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Ekker,S.C. *et al*. (1994) The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. *EMBO J.*, **13**, 3551–3560.

Finn,R.D. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Fraenkel,E. *et al*. (1998) Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. *J. Mol. Biol.*, **284**, 351–361.

Gehring,W.J. *et al*. (1994) Homeodomain proteins. *Annu. Rev. Biochem.*, **63**, 487–526.

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

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kissinger,C.R. *et al*. (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.

Lewis,E.B. (1978) A gene complex controlling segmentation in Drosophila. *Nature*, **276**, 565–570.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Liu,B. *et al*. (1990) Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed homeodomain/DNA complex. *Biochem. Biophys. Res. Commun.*, **171**, 257–259.

Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.

Mahony,S. *et al*. (2007a) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.

Mahony,S. *et al*. (2007b) Inferring protein DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, i297–i304.

Matthews,B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.

Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.

Noyes,M.B. *et al*. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.

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.

Passner,J.M. *et al*. (1999) Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature*, **397**, 714–719.

Persikov,A.V. *et al*. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.

Persikov,A.V. and Singh,M. (2011) An expanded binding model for Cys(2)His(2) zinc finger protein-DNA interfaces. *Phys. Biol.*, **8**, 035010.

Sato,K. *et al*. (2004) Dissecting the Engrailed homeodomain-DNA interaction by phage-displayed shotgun scanning. *Chem. Biol.*, **11**, 1017–1023.

Seeman,N.C. *et al*. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.

Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.

Siggers,T.W. *et al*. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.

Stormo,G.D. (2011) Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics*, **187**, 1219–1224.

Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.

Stormo,G.D. *et al*. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.*, **10**, 2997–3011.

Tupler,R. *et al*. (2001) Expressing the human genome. *Nature*, **409**, 832–833.

Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.

Wolberger,C. *et al*. (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell*, **67**, 517–528.

Wolfe,S.A. *et al*. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.

Wolfe,S.A. *et al*. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Ann. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.

Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.

Zhu,C. *et al*. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.

Zhu,L.J. *et al*. (2011) FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, D111–D117.