HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2016 May 19.

Published in final edited form as:

Nat Biotechnol. 2015 December ; 33(12): 1242–1249. doi:10.1038/nbt.3343.

Affinity regression predicts the recognition code of nucleic acid binding proteins

Raphael Pelossof[1], **Irtisha Singh**[1,2], **Julie L. Yang**[1,2], **Matthew T. Weirauch**[3,4,5], **Timothy R. Hughes**[5], and **Christina S. Leslie**[1,*]

[1] Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY.

[2] Tri-I Program in Computational Biomedicine, Weill Cornell Graduate College, New York, NY.

[3] Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

[4] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH.

[5] Donnelly Centre, University of Toronto, Toronto, ON.

Abstract

Predicting the affinity profiles of nucleic acid-binding proteins directly from the protein sequence is a major unsolved problem. We present a statistical approach for learning the recognition code of a family of transcription factors (TFs) or RNA-binding proteins (RBPs) from high-throughput binding assays. Our method, called affinity regression, trains on protein binding microarray (PBM) or RNA compete experiments to learn an interaction model between proteins and nucleic acids, using only protein domain and probe sequences as inputs. By training on mouse homeodomain PBM profiles, our model correctly identifies residues that confer DNA-binding specificity and accurately predicts binding motifs for an independent set of divergent homeodomains. Similarly, learning from RNA compete profiles for diverse RBPs, our model can predict the binding affinities of held-out proteins and identify key RNA-binding residues. More broadly, we envision applying our method to model and predict biological interactions in any setting where there is a high-throughput 'affinity' readout.

* Corresponding author: Christina S. Leslie, ; Email: cleslie@cbio.mskcc.org.

Ed sum:

DNA and RNA-binding specificities can be predicted from the sequence of members of protein families using a machine learning approach called affinity regression trained on experimental data from the same family

Author Contributions

RP developed the affinity regression algorithm, carried out all the model training and statistical analyses, and helped to write the manuscript. IS performed the protein domain structural analyses and PBM motif analyses. JLY assisted with processing RNA compete data sets and performed the RBP motif analyses. MW and TRH provided independent PBM-derived homeodomain motif data and advised on PBM and RNA compete motif analyses and on the manuscript writing. CSL advised on algorithm development and statistical analyses, supervised the research, and wrote the manuscript.

Competing Financial Interests Statement

The authors have no competing financial interests.

A long-term goal in the study of gene regulation is to understand the evolution of transcription factor (TF) and RNA-binding protein (RBP) families, namely how changes in protein domain sequence lead to differences in DNA- or RNA-binding preference[1, 2]. To be generally applicable, such analyses require data sets with a large number and diversity of training examples. Recent technological advances have enabled the assessment of the relative preferences of proteins to DNA and RNA on an unprecedented scale[1, 3–8]. Much of the newly available TF binding data comes from protein binding microarray (PBM) experiments, where the DNA-binding preferences of an individual fluorescently tagged TF are measured using a universal array of >40K double-stranded DNA probes[3]. The largest existing compendium of *in vitro* binding data for diverse RBPs uses the RNA compete assay, which measures the binding affinity of an RBP against >200K single-stranded RNA probes[7, 8]. We asked whether exploiting these data with sophisticated multivariate statistical techniques might allow us to learn *family-level* models of the DNA or RNA preferences of large classes of TFs and RBPs.

To this end, we developed a machine learning approach called *affinity regression* to learn the nucleic acid recognition code for TF or RBP families directly from the protein sequence and probe-level binding data from PBM or RNA compete experiments. Unlike previous methods[9, 10], our approach requires neither a summarization of binding data as motifs, nor an alignment of protein domain sequences, but instead works directly from amino acid and nucleotide *k*-mer features and allows us to accurately predict the binding profile – and generate a high-quality binding motif – for a TF or RBP not seen in training directly from its protein sequence alone. Moreover, by using the trained interaction model to map binding data back onto features of the protein sequence, we can identify key residues that contribute to the binding specificities of individual proteins.

Results

Training a “recommender system” to model biological interaction data

We propose a general statistical framework for any problem where the observed data can be explained as interactions between two kinds of inputs. While this problem setting is ubiquitous in computational biology, most algorithmic work comes from recommender systems such as Netflix, where users select movies that they like and the recommender algorithm tries to suggest appropriate movies for a new user. By describing each movie by a set of features (e.g. {“comedy”, “horror”, length, actors}) and each user by personal features ({age, gender, geographic location, marital status, Facebook likes}), the recommender seeks to learn relationship rules between the feature spaces of users and movies (e.g. “30-year-old British men like comedy movies with Mr. Bean”).

Here we model high-throughput binding data, such as PBM data for a large family of TFs, using a recommender system formulation. Rather than learn rules for movie preferences of users, we learn rules for binding preferences of TFs for DNA probes. Given a family of structurally-related TF binding domains and their PBM binding profiles, we introduce an algorithm called *affinity regression* to learn a model that explains the binding data as interactions between amino acid *K*-mer features of the protein domain sequences and nucleotide *k*-mer features of the DNA probes (**Fig. 1a**). The algorithm learns a weighting on

all interactions between TF K-mer features and DNA k-mer features that accurately explains one input's preference for the other given the observed binding data. For example, we may learn the rule that in the homeodomain family, the sequence of protein residues 'FQNR' contributes to binding ('likes') the DNA sequence 'TAATTA'.

Formally, we set up a bilinear regression problem to learn an interaction matrix W between TFs, represented by the input matrix P, and DNA probes, represented by the input matrix D, that reconstructs the output matrix Y of observed binding profiles (**Fig. 1a**). Each TF protein sequence is represented by its K-mer count features as a row in P, and each DNA probe sequence by its k-mer count features as a row in D; columns in Y represent the binding profiles of different TFs across probes. The affinity regression interaction model is formulated as:

$$DWP^T \approx Y$$

where D, P, Y are known and W is unknown.

Here the number of probes is very large (10,000s) while the number of TFs is much smaller (a few 100). To obtain a better conditioned system of equations, we multiply both sides of the equation on the left by Y^T (**Fig. 1b** and Methods); the outputs then become pairwise similarities between binding profiles rather than the binding profiles themselves. We then apply a series of transformations to obtain an optimization problem that is tractable with modern solvers (see Methods, Supplementary Note). We use singular value decomposition to cut down the rank of the input matrices and thus reduce the dimensions of the interaction matrix W to be learned. We then convert from a bilinear to a regular regression problem by taking a tensor product of the input matrices (analogous to tensor kernel methods in the dual space[11, 12]) and solve for W with ridge regression. In our experiments, we used $K = 4$ for amino acid K-mer features of TF and RBP protein sequences, $k = 6$ for DNA probe features, and $k = 5$ for RNA probe features, motivated by parameter choices in existing string kernel literature[13, 14] (Supplementary Note).

We can interpret the affinity regression model through mappings to its feature spaces[15]. For example, to predict the binding preferences of an unknown TF, we can right-multiply its protein sequence feature vector through the trained DNA-binding model to predict the similarity of its binding profile to those of the training TFs (**Fig. 1c**). To reconstruct the binding profile of a test TF from the predicted similarities, we assume that the test binding profile is in the linear span of the training profiles and apply a simple linear reconstruction (Supplementary Note, **Fig. 1c**). Finally, to identify the residues that are most important for determining the DNA-binding specificity, we can left-multiply a TF's predicted or actual binding profile through the model to obtain a weighting over protein sequence features, inducing a weighting over residues. We call these right- and left-multiplication operations "mappings" onto the DNA probe space and the protein space, respectively.

Affinity regression outperforms nearest neighbor on homeodomains

We trained an affinity regression model on PBM profiles for 178 mouse homeodomains from a previous study from Berger et al.[1] We transformed the probe intensity distributions to emphasize the right tail of the intensity distribution, containing the highest affinity probes (see Supplementary Note), and used pairwise similarities of transformed profiles as outputs. Our task was to learn a model for homeodomain to DNA probe binding interactions that would generalize to held-out protein sequences, so that for example we could predict the binding motif for a test homeodomain from its amino acid sequence alone.

Affinity regression followed by linear reconstruction enabled accurate prediction of probe-level binding intensities from homeodomain sequence (Supplementary Note). For example, **Fig. 1d** plots the predicted versus experimental probe intensities for Cart1, using a model trained on 90% of the homeodomains where Cart1 was one of the held-out examples. In particular, probes containing the three 8-mers that are most enriched at the top of the intensity distribution are correctly predicted by probe reconstruction to have high affinities to Cart 1 (**Fig. 1c**). Moreover, the correlation between predicted and experimental probe intensities was similar to the correlation between experimental probe intensities from replicate Cart1 PBM experiments (replicate-replicate correlation 0.63, replicate-prediction correlation 0.62, **Fig. 1e**; see **Supplementary Fig. 1** for other TFs).

In 10-fold cross validation on held-out homeodomains, affinity regression strongly outperformed prediction based on the BLOSUM nearest neighbor, where the training domain that is most similar to each test example based on global sequence alignment with BLOSUM substitution scores is considered the nearest neighbor, and this neighbor's binding profile is used for prediction (**Fig. 1f**; **Supplementary Fig. 2**). Indeed, not only did affinity regression outperform nearest neighbor methods in 10-fold cross validation when evaluated either on correlation with experimental binding intensities across all probes ($p < 8.0e-6$, one sided KS test) or on detection of the 1% highest affinity probes ($p < 5.6e-4$, one sided KS test), it also performed almost as well as an ‘oracle’ method, where we chose the optimal training example binding profile as the prediction (**Fig. 1g**). These results demonstrate the strong statistical performance of the family-level TF-DNA binding model learned with affinity regression.

Interaction model identifies DNA binding specificity residues

Since the affinity regression model captures interaction information between K-mer features of the TF amino acid sequences and DNA k-mers, we next asked whether the trained model could identify which residues in the homeodomain sequences determine DNA binding specificity. To achieve this, we trained a model W on all the homeodomain PBM data, and we ‘mapped’ each TF's PBM binding profile Y through the probe k-mer matrix and the interaction model, $Y^T DW$, to get a weighting over amino acid K-mers. Using this weighting, we obtained a mapping score for each K-mer in the TF domain sequence as well as a positional importance score for each residue by summing weights of the K-mer windows containing it (Supplementary Note, **Fig. 2a**). A heatmap of these positional importance scores for a subset of the training data, including the Hox proteins and PYP-containing TALE domains, is shown in **Fig. 2b** (see also **Supplementary Fig. 3**). The DNA-contacting

residues receive the highest scores in this heatmap, producing a bright band of important residues towards the end of the multiple sequence alignment. In addition, other regions are highlighted for specific classes of homeodomains, and importantly, these residues are not found among those conserved across all homeodomains (top of heatmap, **Fig. 2b**).

To assess the statistical significance of the mapping scores at each *K*-mer in the domain sequence, we trained 10,000 affinity regression models for different randomizations of the *K*-mer features in each input sequence, used the empirical null distribution of scores at each *K*-mer position to define a nominal *p*-value, and corrected for multiple non-independent tests using the Benjamini-Hochberg-Yekutieli procedure (see Supplementary Note, **Supplementary Fig. 4**). For example, **Fig. 2c** shows the positional importance profile for two distinct homeodomains, Hoxa9 and Pknox1, with significant positional *K*-mers (FDR < 0.05) shown in bold face on the sequences at the bottom. The Hoxa9 profile shows the largest significant peak over the third helix α_3, corresponding to the DNA contacting residues. Structural alignment of Hoxa9 with Hesx-1 suggests that two glutamic acids in alpha helix α_1 interact with arginines in α_2 and α_3, forming salt bridges that stabilize the binding configuration[16, 17]. Our positional *K*-mer analysis finds a significant peak over α_1 containing both glutamic acids (LEKE), and the major peak over α_3 also contains the arginine residue of a salt bridge; there is a third peak over α_2 (which does not pass FDR < 0.05) that contains the arginine for the other salt bridge. The residues corresponding to the DNA contacts (red) and the identified components of the salt bridges (cyan) are shown on the Hoxa9 co-crystal structure in **Fig. 2d** (highlighted residues defined in Methods.)

By contrast, Pknox1 is a three-amino acid loop extension (TALE) homeodomain, and the positional importance profile derived from the affinity regression model indeed identifies a peak corresponding to the TALE residues PYP[18] in between alpha helices α_1 and α_2 (**Fig. 2c**), which has been reported to be involved in the Knox homeodomain-DNA target interaction in an analysis of the plant homeodomain OSH15[19]. In addition, sequence alignment of OSH15 and Pknox1 suggests that the hydrophobic residues WL in the significant peak over helix α_1 may contribute to a hydrophobic core that stabilizes the homeodomain[19]. **Fig. 2e** shows the structure for human PKNOX1 aligned to the previous co-crystal structure with the core DNA contacting residues and TALE residues as identified by significant positional *K*-mers annotated in red; significant residues in green may contribute to the hydrophobic core, while residues in orange are identified as significant by the model but to our knowledge are not directly supported in the literature.

Predicted binding profiles yield accurate mouse homeodomain motifs

We next sought to confirm that the predicted binding profile can be used to generate a reliable DNA binding motif. Summarizing a PBM binding profile as a single position-specific scoring matrix (PSSM) can be problematic, as there are numerous motif discovery algorithms (summarized and benchmarked in Weirauch et al.[20]) that produce different results from each other and often return multiple motifs. Despite these caveats, we decided to compare the results of applying the same motif discovery algorithm to *predicted* binding profiles and to actual PBM experimental data, to see if similar motifs were obtained.

For the mouse homeodomains, we used affinity regression to predict binding profiles using 10-fold cross-validation. For each held-out domain, we applied the motif discovery algorithm Seed-and-Wobble[3] to its predicted binding profile as well as to the PBM binding profile of its nearest neighbor in the training set. For both affinity regression and nearest neighbor, we retained the algorithm's top three motifs. To define ground truth motifs, we generated three Seed-and-Wobble motifs for each PBM profile and selected a 'target' motif by comparison to the UniPROBE database (see Methods). We then used Kullback-Leibler divergence (D_{KL}) to compare the predicted motifs for each test homeodomain to the target motif and reported the best match for each method.

Fig. 3a shows the comparison of affinity regression versus nearest neighbor for the task of generating a motif close to the target motif; here we transformed the log(D_{KL}) scores by subtracting the minimum log(D_{KL}) score over the set, so that all values are positive and small values correspond to well-predicted motifs. For guidance on what is a good or poor score, we identified homeodomains for which we have replicate experiments and computed the log(D_{KL}) of the best matching motif from the replicate PBM experiment to the target motif (Supplementary Note); we took the median of these scores as our threshold for strong motif prediction performance. Regions where the performance of affinity regression or nearest neighbor is as good or better than this "median replicate" score are shown in gray in **Fig. 3a**. Overall, similar numbers of homeodomains are better predicted by affinity regression as nearest neighbor (90 versus 87, with one tie), and there is no significant difference in performance based on log(D_{KL}) scores between the two methods (using $p < 0.05$ threshold, Wilcoxon signed rank test). Several examples where affinity regression and nearest neighbor both succeed, both fail, or diverge in performance are shown in **Fig. 3b**.

Affinity regression gives accurate motifs for diverse homeodomains

We next turned to a newly generated data set of 218 homeodomains from diverse species for which PBM experiments and motif analyses have been carried out[21]. Before predicting and evaluating motifs, we assessed how well affinity regression, trained on the mouse homeodomain set alone, could predict binding data for these diverse homeodomains. The PBM data in the Weirauch et al. study used a different probe design than the original mouse homeodomain data set; however, 8-mer Z-scores[1] summarized from PBMs with different probe designs can be compared. Therefore, we trained a modified version of affinity regression where we represented every 8-mer by constituent k-mers of length $k = 1, \ldots, 7$ and regressed against the 8-mer Z-scores on the mouse homeodomain data set (see Supplementary Note). For the Z-score model, we trained on a subset of 75 non-redundant mouse homeodomains defined by Alleyne et al.[9], who previously tried to predict Z-scores from homeodomain sequence by training independent regression models for each 8-mer. Alleyne et al. found that their regression models could not outperform a nearest neighbor approach based on a 15 amino acid representation of the homeodomains in leave-one-out-cross-validation; by contrast, the Z-score affinity regression model outperformed their best reported result (**Supplementary Table 1**).

Fig. 3c shows an example of predicted versus experimental 8-mer Z-scores for an Oikopleura dioica homeodomain assayed by Weirauch et al. The overall rank correlation of

predicted and experimental Z-scores is high (ρ = .765), and 48% of the top 100 8-mers based on predicted Z-scores overlap with the top 100 8-mers determined from experimental Z-scores. Moreover, running the PWM-Align-Z algorithm[21] on top 100 predicted 8-mers produces a motif similar to the one obtained from the top experimental 8-mers (**Fig. 3c**). Overall, the Z-score affinity regression model strongly outperformed BLOSUM nearest neighbor for prediction of Z-scores on the diverse Weirauch et al. homeodomains based on Spearman correlation or AUPR for discriminating the top 1% of 8-mers from the bottom 50% ($p < 1e-16$ and $p < 6.91e-9$, signed rank test, respectively; **Fig. 3d**, **Supplementary Fig. 5a,b**). Only on the difficult task of discriminating between the top 1% and bottom 99% of 8-mers does affinity regression statistically tie BLOSUM nearest neighbor.

We then asked whether we could derive accurate motifs for these diverse homeodomains from the Z-scores or binding profiles predicted by affinity regression, using models trained on mouse homeodomains only. The previous study used four separate motif discovery algorithms[21] – BEEML[22], Feature-REDUCE[20], PWM-Align, and PWM-Align-Z – and used cross-validation on replicate experiments for each TF to select among algorithms and parameter settings to produce the final reported motif. However, as previously observed[20], the motifs generated by the different algorithms have very different statistical properties, with BEEML and FeatureREDUCE producing low information content/degenerate motifs and PWM-Align and PWM-Align-Z giving higher information content motifs (**Supplementary Fig. 6**). Therefore, motifs derived from predicted versus experimental Z-scores/binding intensities can only be compared when generated by the same algorithm. We chose PWM-Align-Z, which takes as input the top 8-mers ranked by Z-score, and BEEML, which uses probe-level binding data, as motif algorithms for our analysis.

We first used the Z-score affinity regression model to predict 8-mer Z-scores for each Weirauch et al. homeodomain and derived PWM-Align-Z motifs from the top 100 predicted 8-mers. We compared performance to nearest neighbor motifs on the data set of 75 non-redundant mouse homeodomains, where training set motifs were again generated by PWM-Align-Z and assessed performance by $\log(D_{KL}) - \min \log(D_{KL})$ relative to PWM-Align-Z motifs generated directly from the experimental data. We found that the motifs predicted by affinity regression were significantly closer to ground truth motifs than nearest neighbor motifs ($p < 0.014$, Wilcoxon signed rank test; **Supplementary Fig. 7**; Supplementary Note). By examining the bimodal motif score distributions (**Supplementary Fig. 7**) and visually inspecting motifs, we concluded that motifs satisfying a score threshold of 5 were generally close to ground truth. **Fig. 3e** shows the D_{KL}-based score for each predicted motif versus the ground truth motif for the Weirauch data set, plotted against phylogenetic distance for the corresponding homeodomain from the nearest training set homeodomain (Supplementary Note, **Supplementary Fig. 8**); specific examples are highlighted in red, with experimental and predicted motifs shown in **Fig. 3f**. Whereas the motif score is positively correlated with phylogenetic distance ($R \sim 0.482$), there are still many motifs at high phylogenetic distance that satisfy the motif quality threshold.

As a second motif assessment, we used BEEML to extract motifs from binding profiles predicted by affinity regression and compared to previously reported ground truth BEEML motifs[21]. Since BEEML can converge to a suboptimal motif or fail to converge, we ran

BEEML 3-4 times per homeodomain on predicted and true binding profiles (Supplementary Note) and reported the motif that was closest to the ground truth BEEML motif for both affinity regression and nearest neighbor. To obtain motifs with higher information content, we scaled BEEML energy matrices as previously described[10] (Supplementary Note). We were able to compare performance for 181 (out of 218) test homeodomains for which at least one BEEML run converged for each method and found that affinity regression significantly outperformed nearest neighbor ($p < 1.3e-3$, Wilcoxon signed rank test; **Supplementary Fig. 9**; Supplementary Note). Finally, we compared the accuracy of the best affinity regression motif to those produced by the PreMoTF method[10], which trains a random forest model to predict scaled BEEML motifs from homeodomain amino acid features. We again found that the best affinity regression BEEML motif significantly outperformed PreMoTF ($p < 1.31e-4$, Wilcoxon signed rank test; **Supplementary Fig. 9**, Supplementary Note).

Affinity regression learns a model of RBP-RNA interactions

To demonstrate that our approach is not limited to TFs and PBM data, we turned to a recent study that performed 231 RNA compete binding experiments to assay the binding preferences of 207 RBPs[8]. This diverse data set comprises seven structural classes of RBPs from multiple organisms, with good representation of two larger classes RBPs – the RNA-recognition motif (RRM) proteins and the KH domains. We carried out a filtering process to identify a subset of 130 RBPs that shared similar 4-mers (Supplementary Note), containing many RRM proteins as well as some KH domains, and asked whether affinity regression model could learn general principles of RBP-RNA interactions for these examples.

We used 10-fold cross-validation on these 130 RNA compete experiments to assess performance of affinity regression for the prediction of RNA binding affinities from RBP amino acid sequence. **Fig. 4a** shows that affinity regression systematically outperforms nearest neighbor for the binding profile prediction task ($p < 1.74e-4$ vs. NN, $p < 3e-6$ vs. BLOSUM NN, one-sided KS test; **Supplementary Fig. 10**), here evaluated based on Spearman correlation of the predicted and experimentally measured binding intensities across over 200K probes. Indeed, we also significantly outperform nearest neighbor and BLOSUM nearest neighbor when evaluated by detection of the top 1% brightest probes in the experimental binding data ($p < 1e-4$ vs. NN, $p < 1e-4$ vs. BLOSUM NN, one-sided KS test; **Fig. 4b**, **Supplementary Table 2**). Using BLOSUM substitution scores to compute the nearest neighbor performed worse than simply using similarity in the 4-mer space, possibly because the protein sequences are less sequence similar than in the homeodomain case and many have multiple RBP domains. Affinity regression also did not come as close to ‘oracle’ performance, i.e. prediction based on the optimal nearest neighbor for the scoring metric, as in the homeodomain case, perhaps due to the diversity of RBP sequences.

Next we asked whether we could identify residues contributing to RNA-binding specificity, as we did for DNA-binding specificity in mouse homeodomains. To do this, we first split the RBP sequences into their constituent RNA-binding domains and trained a domain-level affinity regression model (Supplementary Note). We then mapped the predicted binding profile through the probe matrix and the trained model ($Y^T DW$) to obtain positional K-mer

and residue scores over individual domain sequences, as before. **Fig. 4c** shows a subset of the resulting heatmap of positional importance scores derived from the model (see **Supplementary Fig. 11** for all training domains). Similar to before, we used an empirical null model to assess the significance of high-scoring positional *K*-mer scores and identified *K*-mers that satisfied an FDR < 0.15 threshold (Supplementary Note; **Supplementary Fig. 12**). For example, one of the significant regions for RBFOX1, an RRM RBP in the heatmap, is the subsequence GFGFVT, which belongs to a beta sheet that contacts the RNA and contains both phenylalanines that are known to be critical for RNA binding[23] (**Fig. 4d**; see **Supplementary Fig. 13** for additional examples).

Finally, to assess how well we could predict binding motifs for RBPs, we trained a Z-score affinity regression model using data for all 207 RBPs without filtering in a 10-fold cross-validation setting (Supplementary Note). Here, we trained on 7-mer Z-scores as reported in the website cisBP-RNA, and we represented each 7-mer by *k*-mers of length $k = 1, \ldots, 6$. We used the top 100 7-mers predicted by affinity regression as input to PWM-Align-Z to generate binding motifs and compared to ground truth motifs generated by the same algorithm on the experimental binding data. **Fig. 4e** shows a subset of affinity regression predicted motifs and ground truth motifs for the RNA compete data (see **Supplementary Fig. 14** for all motifs). We found that the motifs generated by the Z-score affinity regression model strongly outperformed nearest neighbor motifs ($p < 7.66e-10$, Wilcoxon signed rank test; **Supplementary Fig. 15**), demonstrating the power and generalizability of our approach.

Discussion

Numerous methods have been developed for learning the binding preferences of a single TF from PBM probe data, including rank statistics for scoring preferred 8-mer patterns[3], PSSMs learning methods[3, 24], and more general support vector regression models based on *k*-mer string kernels[25], among others (reviewed and benchmarked previously[20]). Likewise, RNA compete binding data for a single RBP can be summarized by a standard PSSM or *k*-mer enrichment statistics or used to learn binding motifs that incorporate predicted target RNA secondary structure[26]. By contrast, there has been relatively little work on learning the DNA recognition code for a family of TFs from PBM data and, to the best of our knowledge, learning family-level models of RBP binding preferences has not been attempted before. Several studies[9, 10] have tried to learn a family-level DNA-binding model from the mouse homeodomain PBM compendium. These methods used a simplified representation of the input space of protein domain sequences (e.g. DNA-contacting residues, position-specific residues in a multiple alignment) and a reduced output representation of binding motifs (individual Z-scores or PSSMs) and deployed standard machine learning algorithms to learn the mapping from input to output. By contrast, our approach does not involve any reduced representation of the space of protein sequences or binding profiles and outperformed these previous approaches. In the mouse homeodomain setting, using affinity regression with position-specific residues relative to a multiple alignment also gives good prediction of probe intensities, though slightly weaker than with the 4-mer representation ($p < 2.46e-3$ based on Spearman correlation, Wilcoxon signed rank test; **Supplementary Table 3**).

However, learning directly from *K*-mers rather than using a multiple sequence alignment was critical for training on RNA compete profiles for a diverse set of RBPs.

Likewise, the ability to retain richer binding information in the form of probe-level intensities – rather than first compressing the binding profile to a PSSM – is a key feature of our approach. In particular, mapping binding profiles through the model onto the protein *K*-mer space revealed key binding specificity residues in individual TFs and RBPs. There is some debate as to whether PSSMs or richer models are better for representing TF binding information, with some arguing that standard PSSMs are adequate in most cases[27]. We indeed could extract accurate motifs from Z-scores or binding profiles predicted by affinity regression, based on a systematic evaluation of predicted versus ground truth motifs from two different algorithms. However, the performance advantage of the extracted motifs over nearest neighbor was generally more modest than the advantage at the Z-score/binding profile level. We therefore reason that PSSMs, while familiar and interpretable, are a lossy compression of PBM/RNA compete binding data, and that richer representations such as those that use *k*-mers may provide higher accuracy for predicting target sites[28].

Various authors have used predicted secondary structure in the modeling of RBP binding preferences[29–31]. Following Foat and Stormo[30], we used occurrences of 5-mers in the unpaired region of predicted stem loops as separate features from simple 5-mer occurrences (Supplementary Note). We found that the 5-mers in stem loops gave no advantage over simple 5-mers (**Supplementary Table 4**), likely because the current version of the RNA compete assay is designed to avoid probes with secondary structure. However, several newer assays to measure *in vitro* protein-RNA interactions do generate rich statistics for structured RNA probe sequences, including the RNA Bind-n-Seq assay[32] and a method that uses *in situ* transcription to synthesize RNA probes tethered to DNA with a repurposed sequencing instrument[33]. As data from these newer assays becomes available across families of RBPs, it will become important to extend our affinity regression approach to suitably incorporate RNA secondary structure in the feature representation.

Our results show that affinity regression is highly effective for learning and interpreting family-level models of protein-nucleic acid interactions from high-throughput binding compendia. More broadly, affinity regression can be used to train a bilinear interaction model for any macromolecular or cellular interactions where interactors are described by features and where a high-throughput ‘affinity’ readout is available. As one example, we can apply affinity regression to link upstream signaling pathways with downstream transcriptional response in tumors samples, pairing phosphoproteomic measurements with motif hits in gene promoters to predict transcriptional output[34]. High-throughput screening data with quantitative readouts, cell co-culture systems with quantitative phenotypes, and T cell epitope binding data are all potential applications of our approach. We therefore envision our method as a general strategy to model and interpret biological interaction data.

Methods

Additional details on PBM and RNA compete data sets and probe-level data normalization, mathematical development of the algorithm, affinity regression model selection, statistical

significance of amino acid *K*-mer scores, and motif analyses are provided in the Supplementary Note.

Training the affinity regression model

We define affinity regression as the following regularized bilinear regression problem. Let $Y \in \mathbb{R}^{N \times M}$ be a matrix which defines the binding intensities over probes $i = 1, ..., N$ for TFs $j = 1, ... , M$, so that each column of Y corresponds to a PBM experiment. Let $D \in \mathbb{R}^{N \times Q}$ be a matrix that defines the *k*-mer features (in the alphabet of bases) of each probe *i*. Let $P \in \mathbb{R}^{M \times S}$ be a matrix that defines the *K*-mer features (in the alphabet of amino acids) of each TF protein sequence *j*. We set up a bilinear regression problem to learn the weight matrix $W \in \mathbb{R}^{Q \times S}$ on combinations of pairs of TF-probe features:

$$DWP^T \approx Y. \quad (1)$$

To solve this regression problem, we formulate an L_2-regularized optimization problem:

$$argmin_W \|Y - DWP^T\|_2^2 + \lambda \|W\|_2^2$$

where D, P and Y are known **(Fig. 1a)**. We can transform the system to an equivalent system of equation by reformulating the matrix products as Kronecker products[35, 36]:

$$DWP^T \approx Y \iff P \otimes D \quad vec(W) \approx vec(Y) \quad (2)$$

where $\otimes$ is a Kronecker product, and vec(·) is a vectorizing operator that stacks a matrix and outputs the corresponding stacked vector.

Since the number of probes N is very large and the number of TFs is typically small ($M << N$), we may represent the system as a smaller system of equations by using a kernel-like transformation in the output space, namely we left-multiply both sides of Equation (1) by Y^T before the tensor product transformation (Equation (2)) so that our new outputs are the similarities between the original output vectors (see Supplementary Note for error term handling):

$$\left(P \otimes Y^T D\right) vec(W) \approx vec\left(Y^T Y\right). \quad (3)$$

Again this system of equations can be solved using L_2-regularized regression **(Fig. 1b)**. Due to the enormous size of the space of pairs of features (in our case, in the millions), we employ additional compression techniques to solve the system of equations of the affinity regression problem so that it can be solved on a standard desktop computer (see Supplementary Note).

Homeodomain analysis

Motif prediction—We used three motif algorithms in our analysis: Seed-and-Wobble on predicted and experimental binding profiles in the mouse homeodomain data set, and PWM-Align-Z and BEEML on predicted and experimental Z-scores and binding profiles, respectively, on homeodomains from Weirauch et al. For all methods, we determined a high information content core of each 'ground truth' motif obtained by the motif discovery algorithm on experimental data, and we used this core to define the length of the PSSM for motif comparisons based on symmetrized Kullback-Leibler divergence, D_{KL} (see Supplementary Note).

Determination of target ('ground truth') motifs—For ground truth motifs for 178 mouse homeodomains, we applied Seed-and-Wobble to the experimental PBM data, considered the top three motifs for each homeodomain, and chose the motif closest to 'primary' PSSM posted on the UniPROBE database, as measured by the Kullback-Leibler divergence (D_{KL}), as the 'target' motif. The three predicted Seed-and-Wobble PSSMs for affinity regression (respectively, nearest neighbor) were then compared to the target PSSM, and the PSSM with minimum D_{KL} was selected for performance evaluation. For the test set of 218 divergent homeodomains, the target motif was taken to be the PSSM generated by PWM-Align-Z or BEEML, as previously reported[21].

Phylogenetic tree construction—We pooled 75 non-redundant training mouse homeodomain sequences with an additional 218 more divergent homeodomains from Weirauch et al.[21] Multiple sequence alignment was performed using ClustalX, and this alignment was used to generate the phylogenetic tree (Jalview) based on average distance using percent identity. Every branch was assigned a score by averaging the log(D_{KL}) scores of the subbranches.

Protein Structures—PyMOL was used to visualize the PDB protein structures. Highlighted residues are as follows: 1PUF (Hoxa9): red, A/206-209, A/248-259 (DNA binding residues), cyan, A/220-223, 256 (salt bridge residues). 1X2N (PKNOX1): red, A/52-65 (DNA binding residues) and A/32-35 (TALE), green A/25-29, and orange A/46-49.

RNA binding protein analysis

RNA motif prediction—We used PWM-Align-Z to produce a PSSM for each RBP RNA compete experiment using $k = 7$ as the width of the k-mers and $N = 100$ top k-mers for the alignment (see Supplementary Note).

Protein Structure—Highlighted residues for PDB structure 2ERR (RBFOX1) are: red, A/147-150 (EIIF) and A/157-162 (GFGFVT), both RNA-proximal regions.

RNA motif visualization—We visualized the PSSMs from 207 RBPs, including both RRM and KH subfamilies using the motifStack (version 1.4.0) R package and plotted them in a circularized phylogenetic tree.

Software availability

Source code that implements the main affinity regression algorithm and runs the simulation experiments described in the Supplemental Note is available as a Supplementary File. A full implementation of the affinity regression algorithm, scripts used to generate the analyses in the study, and processed PBM and RNA compete data can be obtained from https://bitbucket.org/leslielab/affreg.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Kate Cook for providing her PWM-Align-Z script and Quaid Morris for suggestions on RBP analysis. This work was supported in part by CIHR grant MOP-111007 (T.R.H.) and NIH grants HG006798 and CA143840 (C.S.L.).

References

1. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. Jun 27; 2008 133(7):1266–76. PubMed PMID: 18585359. Pubmed Central PMCID: PMC2531161. [PubMed: 18585359]
2. Liu J, Stormo GD. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. Bioinformatics. Sep 1; 2008 24(17):1850–7. PubMed PMID: 18586699. Pubmed Central PMCID: PMC2732218. [PubMed: 18586699]
3. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology. Nov; 2006 24(11):1429–35. PubMed PMID: 16998473. Pubmed Central PMCID: PMC4419707.
4. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome research. Jun; 2010 20(6):861–73. PubMed PMID: 20378718. Pubmed Central PMCID: PMC2877582. [PubMed: 20378718]
5. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. Jan 17; 2013 152(1-2):327–39. PubMed PMID: 23332764. [PubMed: 23332764]
6. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell. Jun 27; 2008 133(7):1277–89. PubMed PMID: 18585360. Pubmed Central PMCID: PMC2478728. [PubMed: 18585360]
7. Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature biotechnology. Jul; 2009 27(7):667–70. PubMed PMID: 19561594.
8. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. Jul 11; 2013 499(7457):172–7. PubMed PMID: 23846655. Pubmed Central PMCID: PMC3929597. [PubMed: 23846655]
9. Alleyne TM, Pena-Castillo L, Badis G, Talukder S, Berger MF, Gehrke AR, et al. Predicting the binding preference of transcription factors to individual DNA k-mers. Bioinformatics. Apr 15; 2009 25(8):1012–8. PubMed PMID: 19088121. Pubmed Central PMCID: PMC2666811. [PubMed: 19088121]
10. Christensen RG, Enuameh MS, Noyes MB, Brodsky MH, Wolfe SA, Stormo GD. Recognition models to predict DNA-binding specificities of homeodomain proteins. Bioinformatics. Jun 15;

2012 28(12):i84–9. PubMed PMID: 22689783. Pubmed Central PMCID: PMC3371834. [PubMed: 22689783]

11. Brunner C, Fischer A, Luig K, Thies T. Pairwise support vector machines and their application to large scale problems. J Mach Learn Res. 2012; 13(1):2279–92.
12. Vert JP, Qiu J, Noble WS. A new pairwise kernel for biological network inference with support vector machines. BMC bioinformatics. 2007; 8(Suppl 10):S8. PubMed PMID: 18269702. Pubmed Central PMCID: PMC2230501. [PubMed: 18269702]
13. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome research. Sep; 2012 22(9):1723–34. PubMed PMID: 22955984. Pubmed Central PMCID: PMC3431489. [PubMed: 22955984]
14. Leslie C, Eskin E, Noble WS. The spectrum kernel: a string kernel for SVM protein classification. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2002:564–75. PubMed PMID: 11928508. [PubMed: 11928508]
15. Tenenbaum JB, Freeman WT. Separating Style and Content with Bilinear Models. Neural Comput. 2000; 12(6):1247–83. [PubMed: 10935711]
16. Hirsch JA, Aggarwal AK. Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. The EMBO journal. Dec 15; 1995 14(24): 6280–91. PubMed PMID: 8557047. Pubmed Central PMCID: PMC394752. [PubMed: 8557047]
17. Torrado M, Revuelta J, Gonzalez C, Corzana F, Bastida A, Asensio JL. Role of conserved salt bridges in homeodomain stability and DNA binding. The Journal of biological chemistry. Aug 28; 2009 284(35):23765–79. PubMed PMID: 19561080. Pubmed Central PMCID: PMC2749150. [PubMed: 19561080]
18. Burglin TR. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic acids research. Nov 1; 1997 25(21):4173–80. PubMed PMID: 9336443. Pubmed Central PMCID: PMC147054. [PubMed: 9336443]
19. Nagasaki H, Sakamoto T, Sato Y, Matsuoka M. Functional analysis of the conserved domains of a rice KNOX homeodomain protein, OSH15. The Plant cell. Sep; 2001 13(9):2085–98. PubMed PMID: 11549765. Pubmed Central PMCID: PMC139453. [PubMed: 11549765]
20. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nature biotechnology. Feb; 2013 31(2):126–34. PubMed PMID: 23354101. Pubmed Central PMCID: PMC3687085.
21. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. Sep 11; 2014 158(6): 1431–43. PubMed PMID: 25215497. Pubmed Central PMCID: PMC4163041. [PubMed: 25215497]
22. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. PLoS computational biology. Dec.2009 5(12):e1000590. PubMed PMID: 19997485. Pubmed Central PMCID: PMC2777355. [PubMed: 19997485]
23. Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, et al. Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. The EMBO journal. Jan 11; 2006 25(1):163–73. PubMed PMID: 16362037. Pubmed Central PMCID: PMC1356361. [PubMed: 16362037]
24. Chen X, Hughes TR, Morris Q. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics. Jul 1; 2007 23(13):i72–9. PubMed PMID: 17646348. [PubMed: 17646348]
25. Agius P, Arvey A, Chang W, Noble WS, Leslie C. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. PLoS computational biology. 2010; 6(9) PubMed PMID: 20838582. Pubmed Central PMCID: PMC2936517.
26. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS computational biology. 2010; 6:e1000832. PubMed PMID: 20617199. Pubmed Central PMCID: PMC2895634. [PubMed: 20617199]

27. Zhao Y, Stormo GD. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature biotechnology. Jun; 2011 29(6):480–3. PubMed PMID: 21654662. Pubmed Central PMCID: PMC3111930.

28. Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell reports. Apr 25; 2013 3(4):1093–104. PubMed PMID: 23562153. Pubmed Central PMCID: PMC3640701. [PubMed: 23562153]

29. Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. Nature methods. Jun; 2011 8(6):444–5. PubMed PMID: 21623348. [PubMed: 21623348]

30. Foat BC, Stormo GD. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. Molecular systems biology. 2009; 5:268. PubMed PMID: 19401680. Pubmed Central PMCID: PMC2683727. [PubMed: 19401680]

31. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome biology. 2014; 15(1):R17. PubMed PMID: 24451197. Pubmed Central PMCID: PMC4053806. [PubMed: 24451197]

32. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Molecular cell. Jun 5; 2014 54(5):887–900. PubMed PMID: 24837674. Pubmed Central PMCID: PMC4142047. [PubMed: 24837674]

33. Buenrostro JD, Araya CL, Chircus LM, Layton CJ, Chang HY, Snyder MP, et al. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nature biotechnology. Jun; 2014 32(6):562–8. PubMed PMID: 24727714. Pubmed Central PMCID: PMC4414031.

34. Osmanbeyoglu HU, Pelossof R, Bromberg JF, Leslie CS. Linking signaling pathways to transcriptional programs in breast cancer. Genome research. Nov; 2014 24(11):1869–80. PubMed PMID: 25183703. Pubmed Central PMCID: PMC4216927. [PubMed: 25183703]

35. Golub, GH.; Van Loan, CF. Matrix computations. Fourth edition.. Vol. xxi. The Johns Hopkins University Press; Baltimore: 2013. p. 756

36. Penrose R. A generalized inverse for matrices. Mathematical Proceedings of the Cambridge Philosophical Society. 1955; 51(03):406–13.

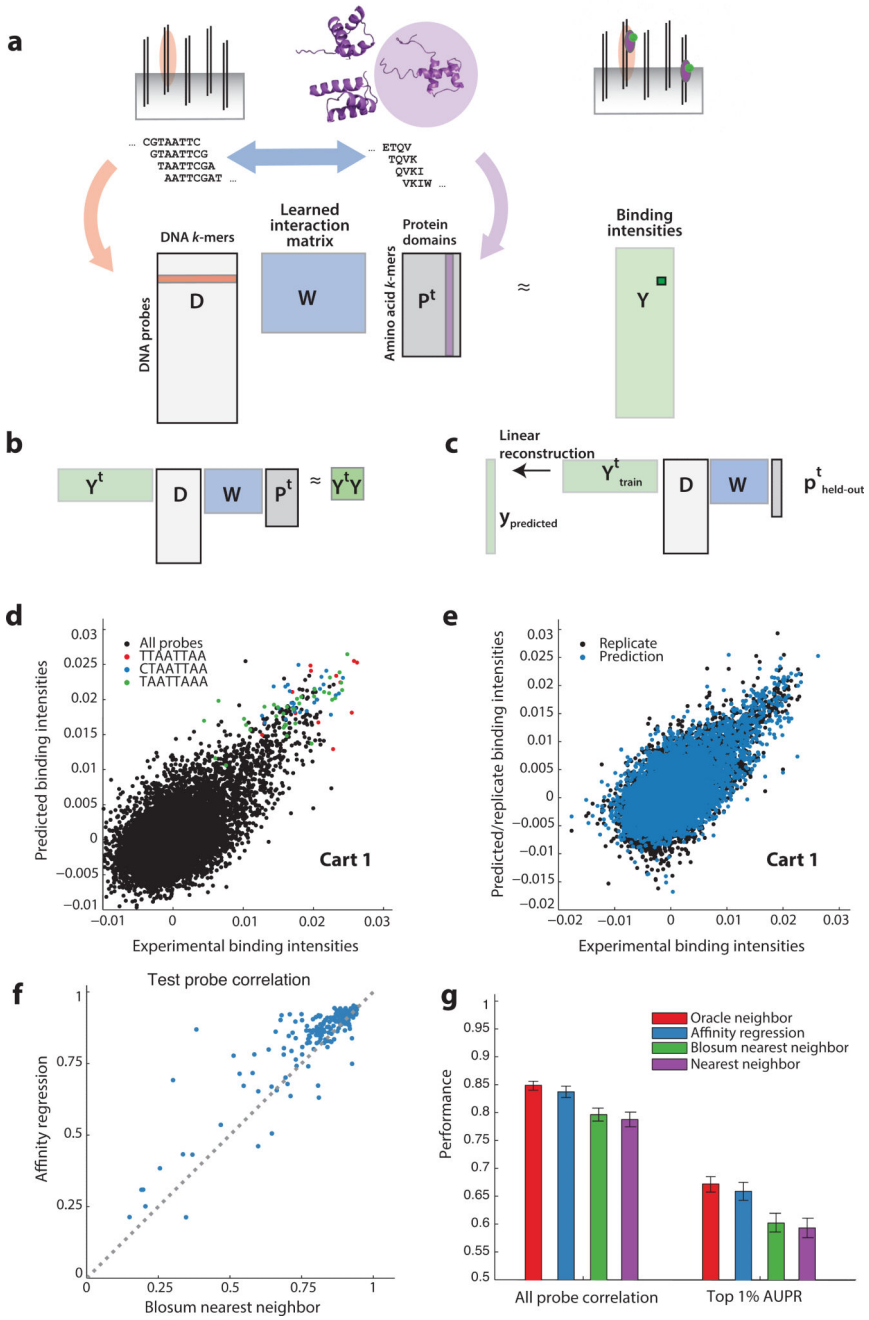

Figure 1. Affinity regression learns highly accurate models of transcription factor-DNA binding interactions from protein binding microarray experiments

a) Affinity regression decomposes the binding intensity for each TF and DNA probe as a weighted interaction between the *k*-mer features of the probe and the *K*-mer features of the TF amino acid sequence. Training the interaction model involves solving a regularized bilinear regression to minimize errors in reconstructing the probe intensity data across all TFs and probes. The model is represented by the interaction matrix W, whereas P and D represent the *K*-mer features of protein sequences and the *k*-mer features of DNA probes, respectively. **b)** Lowering the number of equations by left multiplication with Y^T makes the

problem computationally feasible on a standard computer, and the matrix $Y^T D$ is amenable to low rank approximation. **c)** Full-dimensional probe intensity profile prediction is achieved by mapping the lower dimensional solution back into the span of the training probe intensity profiles. **d)** Predicted probe intensities (y-axis) are plotted against experimental probe intensities (x-axis) for the homeodomain Cart1, using a model trained on 90% of the mouse homeodomain PBM data set with Cart1 among the held-out proteins. Probes containing the three most enriched 8-mers are correctly predicted to have high intensities. **e)** Replicate experimental probe intensities (black) and predicted probe intensities (blue) are both plotted against Cart1 experimental probe intensities, showing that the prediction method has a similar level of variation as replicate noise. **f)** Probe correlation performance on held-out homeodomains for affinity regression (y-axis) versus BLOSUM nearest neighbor (x-axis). Each point is the Spearman correlation between the predicted and actual probe intensities, reporting results on held-out TFs using 10-fold cross-validation. **g)** The bar plots show prediction performance measured by Spearman correlation of probe intensities (left) and AUPR (area under precision-recall curve) for detection of the top 1% of probes (right) for affinity regression, BLOSUM nearest neighbor, nearest neighbor, and an ‘oracle’ method that chooses the training example with optimal performance for the evaluation metric (best possible neighbor). ‘BLOSUM nearest neighbor’ uses local alignment scores with the BLOSUM50 substitution matrix to compute the nearest neighbor; ‘nearest neighbor’ uses Euclidean distance in the *k*-mer vector space to identify the nearest neighbor. Error bars represent the standard error of the mean across 10 folds. Affinity regression is significantly better than both nearest neighbor methods, and there is no significant difference between affinity regression and the ‘oracle’ method.

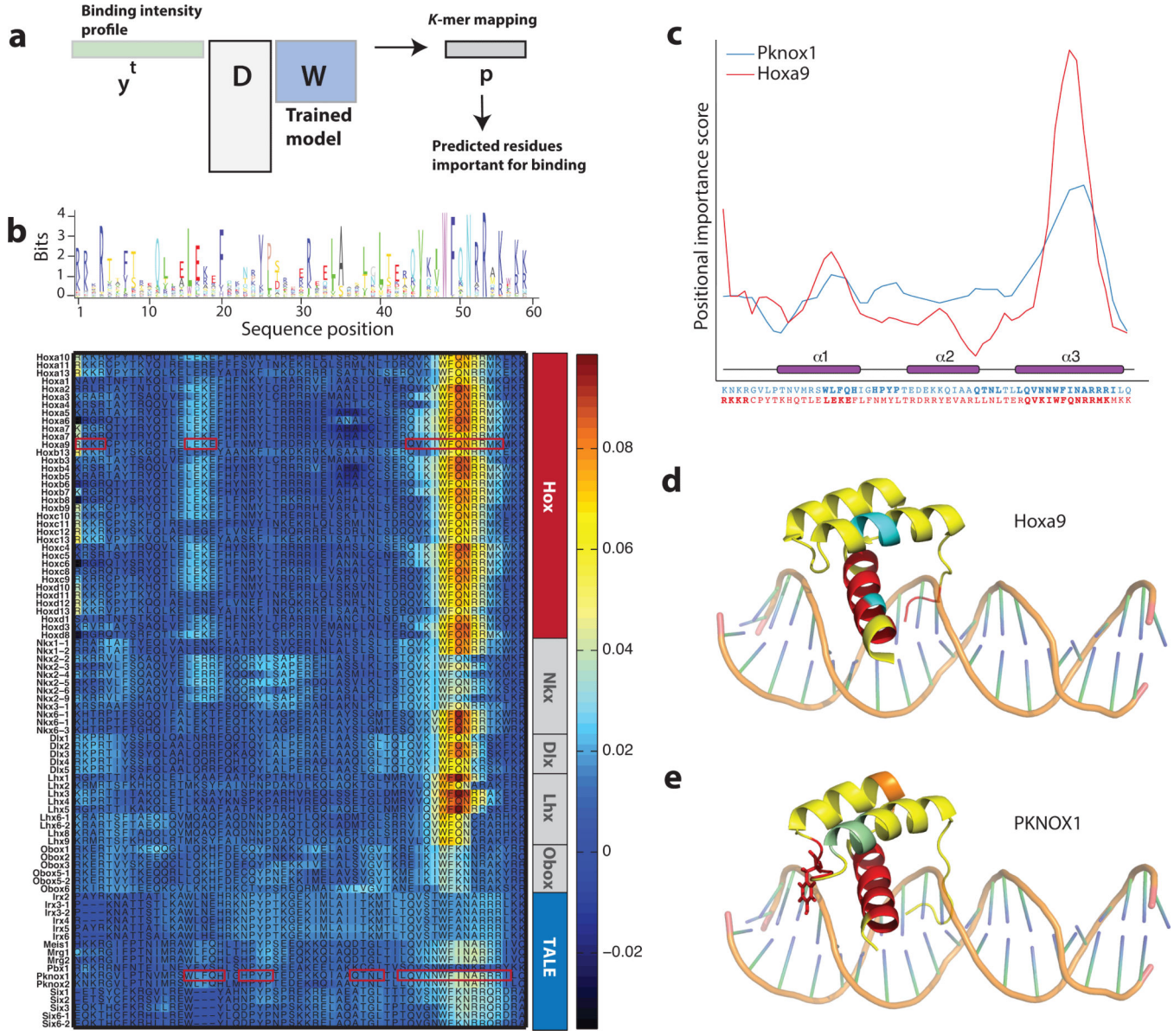

Figure 2. Affinity regression identifies key residues that contribute to homeodomain-DNA binding specificity

a) Mapping the experimental or predicted PBM intensity profile through the model produces a weighting over amino acid *K*-mers, which is used to compute a positional importance profile over residues of the TF sequence binding. **b)** Sequence conservation of the homeodomain family (top track) and the predicted binding importance profiles across members of the homeodomain family (bottom map) are shown. Binding importance profiles are computed from *K*-mer weights via $_{y}{}^{T}DW$ and mapped to each TF sequence. The brightest band of columns corresponds to the core DNA-contacting residues. Binding-specificity features particular to groups of homeodomains are also correctly identified, such as the PYP sequence corresponding to the TALE domain. For Hoxa9 and Pknox1, 4-mers with positional importance score satisfying a 5% FDR threshold are shown with red boxes (see **Supplementary Fig. 4** for all mouse homeodomains). **c)** Actual mapped amino acid

positional importance scores are shown for human PKNOX1 (TALE homeodomain) and mouse Hoxa9. A local peak can be seen for PKNOX1 at the TALE domain (PYP) that does not appear for Hoxa9. Statistically significant positional 4-mers are shown in boldface on the sequences at the bottom of the panel. **d,e)** Statistically significant 4-mers from the positional importance maps for Hoxa9 and Pknox1 are highlighted on known structures from PDB. For Hoxa9, the PDB co-crystal structure is shown; for PKNOX1, the homeodomain structure is aligned to the previous co-crystal structure. The protein is shown in yellow, and the predicted residues that contact DNA are in red. In Hoxa9, identified components of two salt bridges that stabilize the binding conformation are in cyan; in PKNOX1, a significant region potentially contributing to the hydrophobic core is shown in green; predicted residues without a known role in binding specificity are indicated in orange. See methods and materials for highlighted residues.

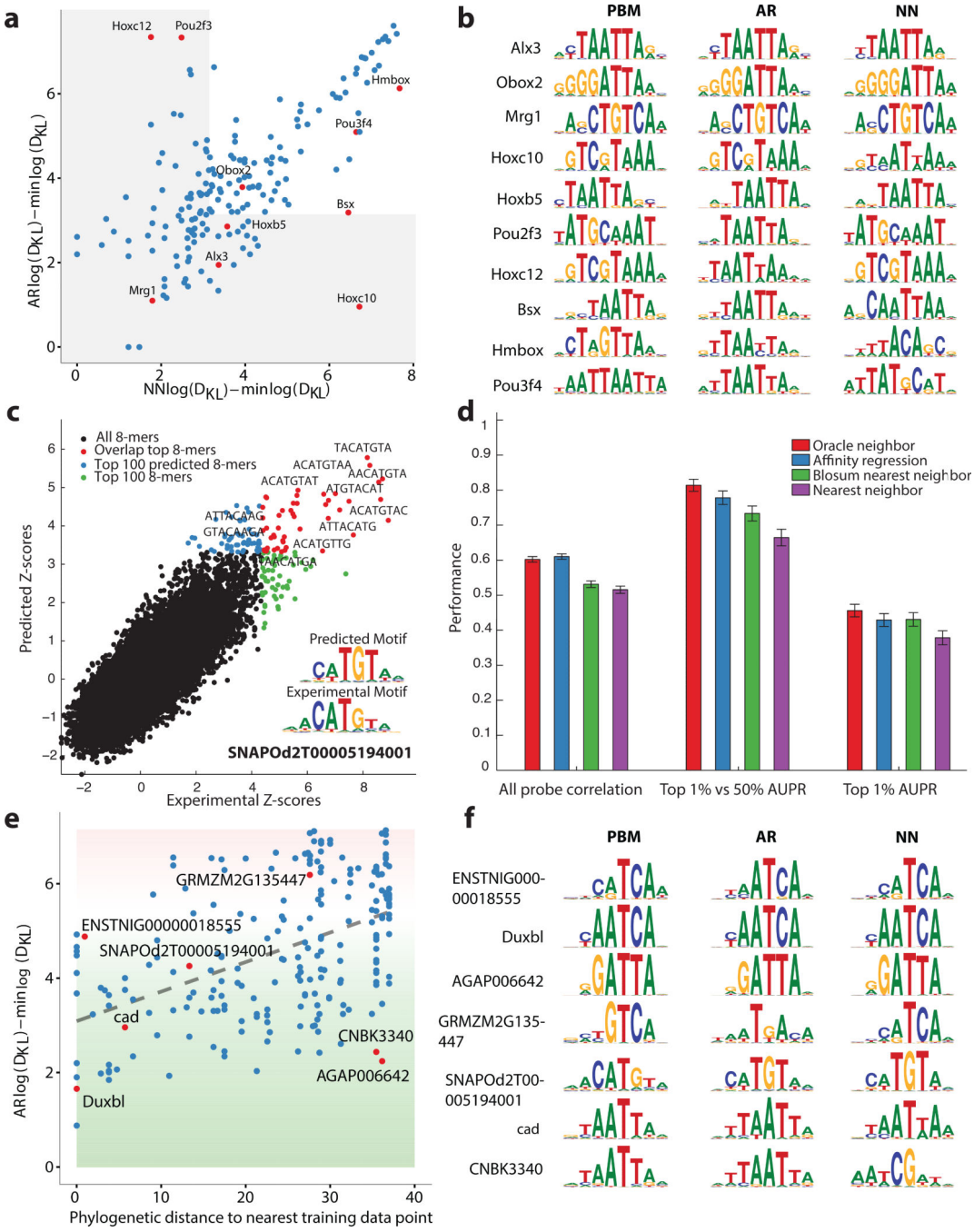

Figure 3. DNA binding profiles predicted by affinity regression generate accurate binding motifs for diverse homeodomains

a) In 10-fold cross-validation, for each test TF we predicted probe intensities, generated PSSMs using Seed-and-Wobble, and compared these predicted motifs to PSSMs estimated directly from the experimental data. We used the $\log_2$ Kullback-Leibler divergence (D_{KL}) to compare motifs; these scores are shifted by adding the min D_{KL} to all values, so that the adjusted scores are all positive and small values correspond to good detection of the target motif. The gray regions correspond to motif detection that is as good or better than the

(adjusted) median log(D_{KL}) between motifs from replicate experiments. For most TFs, affinity regression and nearest neighbor produce PSSMs in a similar score range, and these with no statistical significance between their performance ($p > 0.05$, one-sided KS tests). **b)** Examples of predicted PSSMs are presented with corresponding target PSSMs (derived from experimental PBM data). **c)** Example of predicted Z-scores from the Z-score affinity regression model, trained on 75 non-redundant mouse homeodomains, versus experimental Z-scores for SNAPOd2T00005194001, one of the diverse homeodomains assayed by Weirauch et al. Binding motifs generated by PWM-Align-Z based on the top 100 8-mers predicted by affinity regression and the top 100 8-mers based on actual Z-scores are shown. **d)** Performance comparison of the Z-score affinity regression model versus the 'oracle' nearest neighbor, BLOSUM nearest neighbor, and nearest neighbor in 4-mer space. Error bars represent the standard error of the mean across 10 folds. **e)** Motif accuracy of affinity regression predicted motifs, generated by running PWM-Align-Z on the top 100 predicted 8-mers, versus phylogenetic distance from the nearest training set homeodomain for all 218 Weirauch et al. homeodomains, based on the phylogenetic tree shown in **Supplementary Fig. 8**. Motif accuracy is reported as log(D_{KL}) – min log(D_{KL}) relative ground truth motifs generated by PWM-Align-Z; motif scores < 5 are shown in the green region and indicate accurate motifs, while those above this threshold are in the red region. **f)** Examples of predicted and ground truth motifs based on PWM-Align-Z motif extraction.

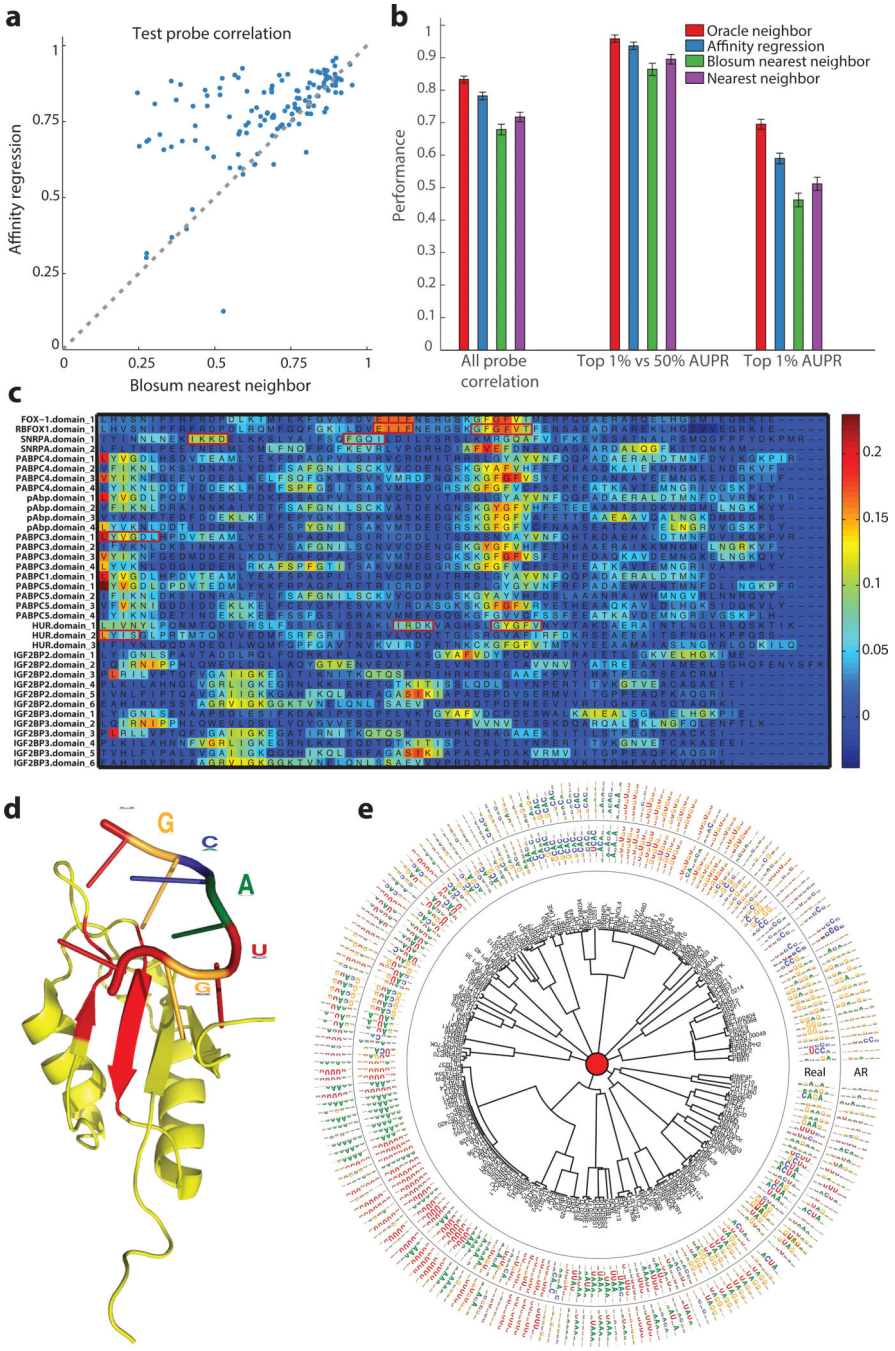

Figure 4. Affinity regression learns a predictive model of RBP-RNA interactions from RNA compete experiments

a) Test probe correlation comparison between BLOSUM nearest neighbor and affinity regression for 130 RBPs, using 10-fold cross-validation and showing performance for held-out proteins. Each point is the Spearman correlation between the predicted and actual RNA compete probe intensities. **b)** The bar plots show performance on held-out RBPs using 10-fold cross-validation for affinity regression, nearest neighbor methods, and an oracle that returns the optimal training example as neighbor. Error bars represent the standard error of the mean across 10 folds. Affinity regression performs significantly better than both

BLOSUM nearest neighbor and nearest neighbor, and the is no significant difference in comparison to the 'oracle' neighbor for probe intensity Spearman correlation and top 1% probe prediction AUROC. **c)** Predicted binding importance profiles across a subset of RRM proteins (see Supplementary Note for KH domains), computed by mapping *K*-mer weights ${}_{Y}{}^{T}DW$ onto each RRM. RBPs that have multiple RRM binding domains are represented as multiple rows. The learned model finds several amino acid *K*-mers that are correlated with binding. For specific RBPs, amino acids 4-mers with positional importance score satisfying a 5% FDR threshold are shown with red boxes (see **Supplementary Fig. 12** for all RBPs). **d)** The co-crystal structure shows human splicing factor RBFOX1, one of the RRM RBPs in the heatmap, in complex with the RNA sequence UGCAUGU; identified in red are significant positional *K*-mers corresponding to the sequence GFGFVT, containing two phenylalanines critical for RNA-binding within a beta sheet contacting the RNA, as well as the RNA-proximal *K*-mer (EIIF). **e)** Predicted PSSMs for protein subfamilies with the RRM and KH domains. The inner PSSM wheel shows the PWM-Align-Z PSSM for the actual RNA compete experiment, while the outer wheel shows the affinity regression predicted motif on unseen TFs in a 10-fold cross-validation setting.