Affinity regression predicts the recognition code of nucleic acid–binding proteins

Raphael Pelossof¹, Irtisha Singh^{1,2}, Julie L Yang^{1,2}, Matthew T Weirauch^{3–5}, Timothy R Hughes⁵ & Christina S Leslie¹

Predicting the affinity profiles of nucleic acid–binding proteins directly from the protein sequence is a challenging problem. We present a statistical approach for learning the recognition code of a family of transcription factors or RNA-binding proteins (RBPs) from high-throughput binding data. Our method, called affinity regression, trains on protein binding microarray (PBM) or RNAcompete data to learn an interaction model between proteins and nucleic acids using only protein domain and probe sequences as inputs. When trained 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, when trained on RNAcompete profiles for diverse RBPs, our model correctly predicts the binding affinities of held-out proteins and identifies key RNA-binding residues, despite the high level of sequence divergence across RBPs. We expect that the method will be broadly applicable to modeling and predicting paired macromolecular interactions in settings where high-throughput affinity data are available.

A long-term goal in the study of gene regulation is to understand the evolution of transcription factor (TF) and 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 PBM experiments, where the DNA-binding preferences of a fluorescence-tagged TF are measured by a universal array of >40,000 double-stranded DNA probes³. The largest compendium of *in vitro* binding data for diverse RBPs uses the RNAcompete assay, which measures the binding affinity of an RBP against >200,000 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.

We developed affinity regression, a machine-learning approach to predict the nucleic acid recognition code for TF or RBP families directly from protein sequence and probe-level binding data from PBM or RNAcompete 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, 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 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. Although this problem setting is ubiquitous in computational biology, most algorithmic work comes from ‘recommender systems’ such as those used by Netflix, where the recommender algorithm tries to suggest appropriate movies for a new user on the basis of other users’ reported preferences. By describing each movie by a set of features (such as genre, length and actors) and each user by personal features (age, gender, geographic location, marital status, Facebook ‘likes’), the algorithm seeks to learn relationship rules between the feature spaces of users and movies (for example, “30-year-old British men like comedies with Mr. Bean”).

Here we use a recommender system formulation to model high-throughput binding data, such as PBM data for a large family of TFs, by learning 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.

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

¹Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. ²Tri-I Program in Computational Biology and Medicine, Weill Cornell Graduate College, New York, New York, USA. ³Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio, USA. ⁴Division of Biomedical Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio, USA. ⁵Donnelly Centre, University of Toronto, Toronto, ON, Canada. Correspondence should be addressed to C.S.L. (cleslie@cbo.mskcc.org).

Received 15 May; accepted 30 July; published online 16 November 2015;
doi:10.1038/nbt.3343

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 (tens of thousands) is much larger than the number of TFs (a few hundred). To obtain a better-conditioned system of equations, we multiply both sides of the equation on the left by Y^T (Fig. 1b and Online 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 (Online Methods and 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 the problem from a bilinear one to a regular regression one 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 spaces¹⁵. 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 and Fig. 1c). Finally, to identify the residues that are most important in 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', on the DNA probe space and on 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 study¹. We transformed the probe intensity distributions to emphasize the right tail of the intensity distribution, containing the highest-affinity probes (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 we could, for example, predict the binding motif for a test homeodomain from its amino acid sequence.

Affinity regression followed by linear reconstruction enabled accurate prediction of probe-level binding intensities from homeodomain sequence (Supplementary Note). For example, Figure 1d plots the predicted and experimental probe intensities for the homeodomain Cart1 using a model trained on 90% of the homeodomains in which 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 were correctly predicted by probe reconstruction to have high affinities to Cart1 (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 and Supplementary Fig. 1).

In tenfold cross-validation on held-out homeodomains, affinity regression strongly outperformed prediction based on BLOSUM nearest neighbor, where the training domain that is most similar to each test example on the basis of global sequence alignment with BLOSUM substitution scores is considered the nearest neighbor, and this neighbor's binding profile is used for prediction (Fig. 1f and Supplementary Fig. 2). We also compared against a simple nearest-neighbor approach, using Euclidean distance in the K -mer vector space to identify the nearest neighbor. Indeed, not only did affinity regression outperform both nearest-neighbor methods in tenfold cross-validation when evaluated on correlation with experimental binding intensities across all probes ($P < 8.0 \times 10^{-6}$, one-sided Kolmogorov-Smirnov (KS) test) and on detection of the 1% highest-affinity probes ($P < 5.6 \times 10^{-4}$, one-sided KS test), it also performed almost as well as an 'oracle' method, in which 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.

Affinity regression identifies DNA binding-specificity residues

As 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 '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 and Fig. 2a). We also created heat maps of these positional-importance scores for a subset of the training data, including the Hox proteins and proline-tyrosine-proline (PYP)-containing TALE domains (Fig. 2b and Supplementary Fig. 3). The DNA-contacting residues received the highest scores in this heat map, producing a bright band toward the end of the multiple sequence alignment. In addition, other regions were highlighted for specific classes of homeodomains; notably, these residues are not found among those conserved across all homeodomains (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 nonindependent tests using the Benjamini-Hochberg-Yekutieli procedure (Supplementary Note and Supplementary Fig. 4). For example, Figure 2c shows the positional importance profile for two distinct homeodomains, Hoxa9 and Pknox1, with significant positional K -mers (false discovery rate (FDR) < 0.05) highlighted. 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 Hox-1 suggests that two glutamic acids in 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 found a significant peak over α_1 containing both glutamic

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. The model is represented by the interaction matrix W ; 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 plotted against experimental probe intensities 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) plotted against Cart1 experimental probe intensities. **(f)** Probe correlation performance on held-out homeodomains for affinity regression versus BLOSUM nearest neighbor. Each point is the Spearman correlation between the predicted and actual probe intensities, reporting results on held-out TFs using tenfold cross-validation. **(g)** Prediction performance measured by Spearman correlation of probe intensities (left) and area under precision-recall curve (AUPR) for detection of the top 1% of probes (right) for affinity regression, BLOSUM nearest neighbor, nearest neighbor and an oracle method. 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 mean $\pm$ s.e.m. on correlation with experimental binding intensities across all probes ($P < 8.0 \times 10^{-6}$, one-sided Kolmogorov–Smirnov (KS) test) and on detection of the 1% highest-affinity probes ($P < 5.6 \times 10^{-4}$, one-sided KS test). Affinity regression performed significantly better than both nearest-neighbor methods ($P < 8.0 \times 10^{-6}$, all probe correlation; $P < 5.6 \times 10^{-4}$, top 1% AUPR, one-sided KS test). *** $P < 0.001$.

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 did not pass FDR < 0.05) that contains the arginine for the other salt bridge (Fig. 2d; highlighted residues are defined in Online Methods).

By contrast, Pknox1 is a 3-amino acid (3-aa) loop extension (TALE) homeodomain, and the positional importance profile derived from the affinity regression model indeed identified a peak corresponding to the TALE residues PYP¹⁸ between 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 (ref. 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 homeodomain¹⁹. Figure 2e shows the structure for human PKNOX1 aligned to the published co-crystal structure and highlights the core DNA-contacting residues and TALE residues (as identified by significant positional K -mers).

Affinity regression yields accurate mouse homeodomain motifs

We next sought to confirm that the predicted binding profile could be used to generate a reliable DNA binding motif. Summarizing a PBM binding profile as a position-specific scoring matrix (PSSM) can be problematic, as there are numerous motif discovery algorithms²⁰ that produce different results and often return multiple motifs. Despite these issues, we applied the same motif-discovery algorithm to predicted binding profiles and to actual PBM experimental data to see whether the motifs obtained were similar.

For the mouse homeodomains, we used affinity regression to predict binding profiles with tenfold cross-validation. For each held-out domain, we applied the motif-discovery algorithm Seed-and-Wobble³ to its predicted binding profile as well as to the PBM binding profile of its nearest neighbor in the training set (based on Euclidean distance of K -mer vectors). 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

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) and the predicted binding importance profiles across members of the homeodomain family (bottom). The brightest band of columns corresponds to the core DNA-contacting residues. Binding-specificity features particular to groups of homeodomains were also correctly identified, such as the PYP sequence corresponding to the TALE domain. Red boxes indicate 4-mers with positional importance score satisfying a significance threshold of $FDR = 0.05$ (Supplementary Fig. 4).

(c) Actual mapped amino acid positional importance scores for human PKNOX1 (TALE homeodomain) and mouse Hoxa9. Significant ($FDR < 0.05$, Benjamini-Hochberg-Yekutieli procedure) positional 4-mers are shown in bold (bottom). (d,e) Significant ($FDR < 0.05$, Benjamini-Hochberg-Yekutieli procedure) 4-mers from the positional importance maps highlighted on PDB structures for Hoxa9 (d; PDB 1PUF) (d) and PKNOX1 (e; PDB 1X2N). Residues predicted to contact DNA are shown in red; components of two salt bridges that stabilize the binding conformation are shown in blue (d); a significant region of PKNOX1 potentially contributing to the hydrophobic core is shown in green (e); predicted residues without a known role in binding specificity are shown in orange (residues are defined in Online Methods).

for each PBM profile and selected a target motif by comparison to the UniPROBE database (Online 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.

To compare affinity regression to nearest neighbor for the task of generating a motif close to the target motif (Fig. 3a), we transformed the $\log(D_{KL})$ scores by subtracting the minimum $\log(D_{KL})$ score over the set, so that all values were positive and small values corresponded 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 ('median replicate' score) as the threshold for strong motif prediction performance (Fig. 3a). Overall, similar numbers of homeodomains were better predicted by affinity regression than nearest neighbor (90 versus 87, with one tie), and there was no significant difference in performance based on $\log(D_{KL})$ scores between the methods ($P < 0.05$, Wilcoxon signed rank test) (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 out²¹. 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 that study, by Weirauch *et al.*²¹, used a different probe design from the mouse homeodomain data set; however,

8-mer Z-scores¹ summarized from PBMs with different probe designs can be compared. Therefore, we trained a modified version of affinity regression in which every 8-mer is represented by constituent *k*-mers of length $k = 1, \dots, 7$ and regressed against the 8-mer Z-scores on the mouse homeodomain data set (Supplementary Note). For the Z-score model, we trained on a subset of 75 published nonredundant mouse homeodomains⁹, from a study that tried to predict Z-scores from homeodomain sequence by training independent regression models for each 8-mer. The authors' regression models could not outperform a nearest-neighbor approach⁹ based on a 15-aa 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).

Figure 3c shows an example of predicted and experimental 8-mer Z-scores for an *Oikopleura dioica* homeodomain assayed by Weirauch *et al.*²¹. The overall rank correlation of predicted and experimental Z-scores was high ($\rho = 0.765$), 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 algorithm²¹ on the top 100 predicted 8-mers produced 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.*²¹ homeodomains, as evaluated by Spearman correlation and area under precision-recall curve for discriminating the top 1% of 8-mers from the bottom 50% ($P < 1 \times 10^{-16}$ and $P < 6.91 \times 10^{-9}$, signed rank test, respectively) (Fig. 3d and Supplementary Fig. 5a,b). Only in the task of discriminating

Figure 3 DNA binding profiles predicted by affinity regression generate accurate binding motifs for diverse homeodomains. **(a)** In tenfold cross-validation, for each test TF we predicted probe intensities, generated PSSMs by Seed-and-Wobble and compared these predicted motifs to PSSMs estimated directly from the experimental data. 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 with similar score ranges, with no statistical significance between their performances ($P > 0.05$, one-sided KS test). **(b)** Examples of predicted PSSMs, 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 nonredundant mouse homeodomains, versus experimental Z-scores for SNAPOd2T00005194001, one of the diverse homeodomains assayed by Weirauch *et al.*²¹. Binding motifs generated by PWM-Align-Z on the basis of the top 100 8-mers predicted by affinity regression and the top 100 8-mers on the basis of actual Z-scores. **(d)** Performance comparison of the Z-score affinity regression model versus oracle nearest neighbor, BLOSUM nearest neighbor and nearest neighbor in 4-mer space. Error bars represent mean $\pm$ s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; ns, $P > 0.05$. **(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.*²¹ homeodomains (Supplementary Fig. 8). Motif accuracy is reported as $\log(D_{KL})$ – minimum $\log(D_{KL})$ relative to ground-truth motifs generated by PWM-Align-Z; green region indicates motif score < 5 . **(f)** Examples of predicted and ground-truth motifs based on PWM-Align-Z motif extraction. AR, affinity regression; NN, nearest neighbor; min, minimum.

between the top 1% and bottom 99% of 8-mers was affinity regression statistically tied with 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 Weirauch *et al.* study²¹ used four separate motif discovery algorithms²¹—BEEML²², Feature-REDUCE²⁰, 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 observed²⁰, the motifs generated by different algorithms have very different statistical properties, with BEEML and FeatureREDUCE producing low-information content or degenerate motifs and PWM-Align and PWM-Align-Z giving higher-information content motifs (Supplementary Fig. 6). Therefore, motifs derived from predicted and experimental Z-scores or binding intensities can be compared only 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.*²¹ 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 nonredundant mouse homeodomains, where training set motifs were again generated by PWM-Align-Z and assessed performance by $\log(D_{KL})$ – minimum $\log(D_{KL})$ 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 and 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. Figure 3e shows the D_{KL} -based score for each predicted motif versus the ground-truth motif for the Weirauch *et al.*²¹ data set, plotted against phylogenetic distance for the corresponding homeodomain from the nearest training set homeodomain (Supplementary Note and Supplementary Fig. 8); experimental and predicted motifs are

shown in **Figure 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

them to previously reported ground-truth BEEML motifs²¹. Because 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

Figure 4 Affinity regression learns a predictive model of RBP-RNA interactions from RNAcompete experiments. **(a)** Test probe correlation comparison between BLOSUM nearest neighbor and affinity regression for 130 RBPs, with tenfold cross-validation and showing performance for held-out proteins. Each point represents the Spearman correlation between the predicted and actual RNAcompete probe intensities. **(b)** Performance on held-out RBPs using tenfold cross-validation for affinity regression, nearest-neighbor methods and an oracle that returns the optimal training example as neighbor. Error bars represent mean $\pm$ s.e.m. Affinity regression performed significantly better than both BLOSUM nearest neighbor and nearest neighbor ($P < 10^{-4}$, one-sided Kolmogorov–Smirnov test), and there was no significant difference between affinity regression and the oracle neighbor for probe-intensity Spearman correlation and top 1% probe-prediction area under the receiver operating curve (AUROC). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. **(c)** Predicted binding importance profiles across a subset of RRM proteins (see **Supplementary Note** for KH domains), computed by mapping K -mer weights y^TDW onto each RRM. RBPs that have multiple RRM binding domains are represented as multiple rows. The learned model finds several amino acid K -mers correlated with binding. Red boxes indicate amino acids 4-mers with positional importance score satisfying an FDR threshold of 5% (**Supplementary Fig. 12**). **(d)** Co-crystal structure of human splicing factor RBOX1 (PDB 2ERR) in complex with the RNA sequence UGCAUGU; significant positional K -mers corresponding to the sequence GFGFVT, containing two phenylalanines critical for RNA-binding within a β -sheet contacting the RNA, as well as the RNA-proximal K -mer (EIIF) are shown in red. **(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 RNAcompete experiment; the outer wheel shows the affinity regression (AR) predicted motif on unseen TFs in a tenfold cross-validation setting.

regression and nearest neighbor. To obtain motifs with higher information content, we scaled BEEML energy matrices as previously described¹⁰ (**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 ($P < 1.3 \times 10^{-3}$, Wilcoxon signed rank test) (**Supplementary Fig. 9** and **Supplementary Note**). Finally, we compared the accuracy of the best affinity regression motif to those produced by the PreMoTF method¹⁰, 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 ($P < 1.31 \times 10^{-4}$, Wilcoxon signed rank test; **Supplementary Fig. 9** and **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 RNAcompete binding experiments to assay the binding preferences of 207 RBPs⁸. This diverse data set comprises seven structural classes of RBPs from multiple organisms, with good representation of two larger classes of 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 share similar 4-mers (**Supplementary Note**), which contained many RRM proteins and some KH domains, and asked whether the affinity regression model could learn general principles of RBP-RNA interactions for these examples.

We used tenfold cross-validation on these 130 RNAcompete experiments to assess performance of affinity regression for the prediction of RNA binding affinities from RBP amino acid sequence. Affinity regression systematically outperformed nearest neighbor for the binding profile–prediction task (**Fig. 4a**) ($P < 1.74 \times 10^{-4}$ vs. nearest neighbor; $P < 3 \times 10^{-6}$ vs. BLOSUM nearest neighbor, one-sided KS test; **Supplementary Fig. 10**), here evaluated on the basis of Spearman correlation of the predicted and experimentally measured binding intensities across more than 200,000 probes. Affinity regression also significantly outperformed nearest neighbor and BLOSUM nearest neighbor when evaluated by detection of the top 1% brightest probes in the experimental binding data ($P < 1 \times 10^{-4}$ and $P < 1 \times 10^{-4}$, respectively, one-sided KS test) (**Fig. 4b** and **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 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 owing 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 described above (**Fig. 4c** and **Supplementary Fig. 11**). We again 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** and **Supplementary Fig. 12**). For example, one of the significant regions for the RRM RBP RBFOX1 was the

subsequence GFGFVT, which belongs to a β -sheet that contacts the RNA and contains both phenylalanines critical for RNA binding²³ (**Fig. 4d** and **Supplementary Fig. 13**).

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 tenfold cross-validation setting (**Supplementary Note**). Here we trained on 7-mer Z-scores as reported in the CISBP-RNA database (<http://cisbp-rna.ccb.utoronto.ca/>) and represented each 7-mer by k -mers of length $k = 1, \dots, 6$. We used the top 100 7-mers predicted by affinity regression as input to PWM-Align-Z to generate binding motifs and compared these to ground-truth motifs generated by the same algorithm on the experimental binding data (**Fig. 4e** and **Supplementary Fig. 14**). We found that the motifs generated by the Z-score affinity regression model strongly outperformed nearest-neighbor motifs ($P < 7.66 \times 10^{-10}$, 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 patterns³, PSSM learning methods^{3,24} and more general support-vector regression models based on k -mer string kernels²⁵, among others²⁰. Likewise, RNAcompete binding data for a single RBP can be summarized by a standard PSSM or k -mer enrichment statistics or be used to learn binding motifs that incorporate predicted target RNA secondary structure²⁶. By contrast, there has been relatively little work in learning the DNA recognition code for a family of TFs from PBM data, and, to 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 (for example, DNA-contacting residues or 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. Our approach does not involve reduced representation of the space of protein sequences or binding profiles and outperformed these previous approaches. In the mouse homeodomain setting, affinity regression with position-specific residues relative to a multiple alignment gives good prediction of probe intensities, though slightly weaker than with the 4-mer representation ($P < 2.46 \times 10^{-3}$ 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 RNAcompete 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 residues for binding specificity in individual TFs and RBPs. There is 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 cases²⁷. Indeed, we could extract accurate motifs from Z-scores or binding profiles predicted by affinity regression, as suggested by 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 or binding-profile level. We therefore reason that PSSMs, although familiar and interpretable, are a 'lossy compression' of PBM or RNAcompete binding data and that richer representations, such as those that use *k*-mers, may provide higher accuracy for predicting target sites²⁸.

Various studies have used predicted secondary structure in the modeling of RBP binding preferences^{29–31}. Following Foat and Stormo³⁰, 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**), probably because the current version of the RNAcompete assay is designed to avoid probes with secondary structure. However, several newer assays to measure *in vitro* protein–RNA interactions generate rich statistics for structured RNA probe sequences, including the RNA Bind-n-Seq assay³² and a method that uses *in situ* transcription to synthesize RNA probes tethered to DNA with a repurposed sequencing instrument³³. As data from these newer assays become available across families of RBPs, it will become important to extend the 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 interaction in which interactors are described by features and for which a high-throughput 'affinity' readout is available. As one example, affinity regression has been applied to link upstream signaling pathways with downstream transcriptional response in tumor samples, pairing phosphoproteomic measurements with motif hits in gene promoters to predict transcriptional output³⁴. High-throughput screening data with quantitative readouts, cell 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

Methods and any associated references are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank K. Cook for providing the PWM-Align-Z script and Q. Morris for suggestions on RBP analysis. This work was supported in part by Canadian Institute for Health Research grant MOP-111007 (T.R.H.) and US National Institutes of Health grants HG006798 and CA143840 (C.S.L.).

AUTHOR CONTRIBUTIONS

R.P. developed the affinity regression algorithm, carried out all the model training and statistical analyses, and helped to write the manuscript. I.S. performed the protein domain structural analyses and PBM motif analyses. J.L.Y. assisted with processing RNAcompete data sets and performed the RBP motif analyses. M.T.W. and T.R.H. provided independent PBM-derived homeodomain motif data and advised on PBM and RNAcompete motif analyses and on the manuscript writing. C.S.L. advised on algorithm development and statistical analyses, supervised the research, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Berger, M.F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
- Liu, J. & Stormo, G.D. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24**, 1850–1857 (2008).
- Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435 (2006).
- Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
- Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
- Noyes, M.B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**, 1277–1289 (2008).
- Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
- Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
- Alleyne, T.M. *et al.* Predicting the binding preference of transcription factors to individual DNA *k*-mers. *Bioinformatics* **25**, 1012–1018 (2009).
- Christensen, R.G. *et al.* Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* **28**, i84–i89 (2012).
- Brunner, C., Fischer, A., Luig, K. & Thies, T. Pairwise support vector machines and their application to large scale problems. *J. Mach. Learn. Res.* **13**, 2279–2292 (2012).
- Vert, J.P., Qiu, J. & Noble, W.S. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics* **8** (suppl. 10), S8 (2007).
- Arvey, A., Agius, P., Noble, W.S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22**, 1723–1734 (2012).
- Leslie, C., Eskin, E. & Noble, W.S. The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomputing* **2012**, 564–575 (2002).
- Tenenbaum, J.B. & Freeman, W.T. Separating style and content with bilinear models. *Neural Comput.* **12**, 1247–1283 (2000).
- Hirsch, J.A. & Aggarwal, A.K. Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *EMBO J.* **14**, 6280–6291 (1995).
- Torrado, M. *et al.* Role of conserved salt bridges in homeodomain stability and DNA binding. *J. Biol. Chem.* **284**, 23765–23779 (2009).
- Bürglin, T.R. 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 (1997).
- Nagasaki, H., Sakamoto, T., Sato, Y. & Matsuoka, M. Functional analysis of the conserved domains of a rice KNOX homeodomain protein, OSH15. *Plant Cell* **13**, 2085–2098 (2001).
- Weirauch, M.T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* **31**, 126–134 (2013).
- Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
- Zhao, Y., Granäs, D. & Stormo, G.D. Inferring binding energies from selected binding sites. *PLOS Comput. Biol.* **5**, e1000590 (2009).
- Auweter, S.D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25**, 163–173 (2006).
- Chen, X., Hughes, T.R. & Morris, Q. RankMotif: a motif-search algorithm that accounts for relative ranks of *K*-mers in binding transcription factors. *Bioinformatics* **23**, i72–i79 (2007).
- Agius, P., Arvey, A., Chang, W., Noble, W.S. & Leslie, C. High resolution models of transcription factor–DNA affinities improve *in vitro* and *in vivo* binding predictions. *PLoS Comput. Biol.* **6**, e1000916 (2010).
- Kazan, H., Ray, D., Chan, E.T., Hughes, T.R. & Morris, Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.* **6**, e1000832 (2010).
- Zhao, Y. & Stormo, G.D. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* **29**, 480–483 (2011).
- Gordân, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports* **3**, 1093–1104 (2013).
- Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G.G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8**, 444–445 (2011).
- Foat, B.C. & Stormo, G.D. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol. Syst. Biol.* **5**, 268 (2009).
- Maticzka, D., Lange, S.J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15**, R17 (2014).
- Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell* **54**, 887–900 (2014).
- Buenrostro, J.D. *et al.* Quantitative analysis of RNA–protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
- Osmanbeyoglu, H.U., Pelossof, R., Bromberg, J.F. & Leslie, C.S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res.* **24**, 1869–1880 (2014).

ONLINE METHODS

Data and statistics. Additional details on PBM and RNAcompete 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 that defines the binding intensities over probes $i = 1, \dots, N$ for TFs $j = 1, \dots, M$, 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

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

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}

$$DWP^T \approx Y \Leftrightarrow P \otimes D \text{vec}(W) \approx \text{vec}(Y) \quad (2)$$

where $\otimes$ is a Kronecker product, and $\text{vec}(\cdot)$ 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 ($M \ll N$), 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):

$$(P \otimes Y^T D) \text{vec}(W) \approx \text{vec}(Y^T Y) \quad (3)$$

Again this system of equations can be solved using L_2 -regularized regression (**Fig. 1b**). Owing 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 (**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.*²¹. 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} (**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 (<http://thebrain.bwh.harvard.edu/uniprobe/>), 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 reported²¹.

Phylogenetic tree construction. We pooled 75 nonredundant 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 sub-branches.

Protein structures. PyMOL was used to visualize the PDB protein structures. Highlighted residues are as follows: in Hoxa9 (PDB 1PUF) (**Fig. 2d**), red, A/206–209, A/248–259 (DNA binding residues); cyan, A/220–223, 256 (salt bridge residues); in PKNOX1 (PDB 1X2N) (**Fig. 2e**), red, A/52–65 (DNA binding residues) and A/32–35 (TALE); green, A/25–29; orange, A/46–49.

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

Protein structure. Highlighted residues for RBFOX1 (PDB 2ERR) (**Fig. 4d**) are as follows: 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 **Supplementary Note** is available as **Supplementary Code**. A full implementation of the affinity regression algorithm, scripts used to generate the analyses in the study, and processed PBM and RNAcompete data can be obtained from <https://bitbucket.org/leslielab/affreg>.

35. Golub, G.H. & Van Loan, C.F. *Matrix Computations* 4th edn. (The Johns Hopkins University Press, Baltimore, 2013).

36. Penrose, R. A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.* **51**, 406–413 (1955).