BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 26–33
doi:10.1093/bioinformatics/btm539

Sequence analysis

Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control

A.D.J. van Dijk[1], C.J.F. ter Braak[2], R.G. Immink[3], G.C. Angenent[3] and R.C.H.J. van Ham[1,*]

[1]Applied Bioinformatics, PRI, Wageningen UR, Droevendaalsesteeg 1, [2]Biometris, Wageningen UR, Bornsesteeg 47 and [3]Bioscience, PRI, Wageningen UR, Droevendaalsesteeg 1, Wageningen, The Netherlands

Received on July 12, 2007; revised on October 13, 2007; accepted on October 19, 2007

Advance Access publication November 17, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Transcription factor interactions are the cornerstone of combinatorial control, which is a crucial aspect of the gene regulatory system. Understanding and predicting transcription factor interactions based on their sequence alone is difficult since they are often part of families of factors sharing high sequence identity. Given the scarcity of experimental data on interactions compared to available sequence data, however, it would be most useful to have accurate methods for the prediction of such interactions.

Results: We present a method consisting of a Random Forest-based feature-selection procedure that selects relevant motifs out of a set found using a correlated motif search algorithm. Prediction accuracy for several transcription factor families (bZIP, MADS, homeobox and forkhead) reaches 60–90%. In addition, we identified those parts of the sequence that are important for the interaction specificity, and show that these are in agreement with available data. We also used the predictors to perform genome-wide scans for interaction partners and recovered both known and putative new interaction partners.

Contact: roeland.vanham@wur.nl

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

1 INTRODUCTION

Transcription factor networks are extensively studied (Davidson and Erwin, 2006; Walhout, 2006). In these networks, combinatorial control plays an important role: specific combinations of *cis* elements are present in the promoter regions of genes, and only specific combinations of transcription factors (TFs) bind to these elements. In the latter case, interactions between TFs are essential for functioning of the network; this makes the study of TF–TF interactions of paramount importance.

Large-scale investigations of protein–protein interactions have resulted in enormous amounts of interaction data (Gandhi *et al.*, 2006; Mathivanan *et al.*, 2006; Salwinski *et al.*, 2004; von Mering *et al.*, 2007). Nevertheless, these data currently represent only a small fraction of the real 'interactome', and therefore new approaches are needed to obtain accurate and predictive information about TF–TF interaction capacities. Here, we propose to analyze the sequence determinants of specificity in interacting TFs.

The analysis of TF–TF interactions based on sequence data is difficult, since a transcription factor family typically consists of a number of proteins related to each other as a result of duplications (Babu *et al.*, 2004; Teichmann and Babu, 2004). Hence there is often high sequence similarity, and standard conservation analysis does not provide insight into the determinants of interaction specificity. Moreover, existing interaction prediction methods are not applicable in the context of these families.

Many existing interaction prediction methods are not based on sequence data only but use auxiliary data such as localization data, expression data, structural data or information about interactions of orthologs (Shoemaker and Panchenko, 2007). Those that do rely on sequence data have limited applicability for the interactions among transcription factor family members. InterPro-based predefined 'sequence signatures' (Sprinzak and Margalit, 2001; Sprinzak *et al.*, 2006) or Pfam domain assignments (Chen and Liu, 2005) are useful for whole genome-based analysis but are not applicable for TF families since these signatures will hardly vary. Other approaches are either aimed at groups of all-vs-all interacting proteins (Li *et al.*, 2006) and as such, are not useful to explain interaction specificity in TF–TF interactions, or use structural information (Li and Li, 2005). A promising approach has been the mining of motifs in sets of proteins that share a common interaction partner (Fang *et al.*, 2005; Neduva *et al.*, 2005). However, the resulting motifs have not been applied in the prediction of interactions, and it is not a priori clear how to combine single sequence-based motifs into features for an interaction predictor. Finally, support vector machines have been applied using either amino acid triplet occurrences (Shen *et al.*, 2007) or more general 'sequence signature products' (Martin *et al.*, 2005); the disadvantage in this approach is that there is not a direct way to obtain information about which parts of the sequence are important for the interactions.

*To whom correspondence should be addressed.

Here, we used sequence information to develop predictors for interaction specificity as well as obtain information about those parts of the sequence that determine this specificity. The latter makes our method amenable to targeted experimental verification of predictions. The main reason to use sequence information is that it is readily available for many organisms and that the ultimate means to understand those interactions, experimental three-dimensional structures, are far out of reach as of today, in spite of tremendous progress obtained by large-scale structural genomics approaches (Levitt, 2007).

Our approach consists of training a classification algorithm using experimental interaction data. As input for the algorithm, we use short sequence motifs overrepresented in the sequences of the interacting proteins. Currently we used one approach [D-STAR, (Tan *et al*., 2006)], which finds motifs defined over pairs of sequences. In the context of interaction prediction, this offers a clear advantage over other motif search algorithms. Note, however, that our approach is not depending on this method and that other motif search approaches, e.g. MEME (Bailey and Elkan, 1995), can be used. Essential in our approach is that we select relevant motif pairs via feature selection, by using Random Forest (Breiman, 2001), which had the best overall performance in a recent evaluation of interaction prediction methods (Qi *et al*., 2006). To validate the approach, we apply it to simulated datasets as well as to several biological datasets.

2 METHODS

2.1 Motif search

As a first step in our approach, we used the correlated motif search algorithm D-STAR (Tan *et al*., 2006). This finds correlated motifs that are overrepresented in pairs of interacting protein sequences. We refer to the D-STAR motif pairs as 'correlated motif pairs' and to the constituent motifs as 'motif' or 'motif part'. In preliminary experiments, we varied D-STAR parameter settings and analyzed their influence on the performance of the method. Based on parameter settings used in the D-STAR publication (Tan *et al*., 2006), we used motif lengths between 6 and 10 amino acids and numbers of allowed mismatches between 1 and 3. We found that the best performance was obtained with the following settings: motif length = 8, number of allowed mismatches = 3 and maximum number of motifs to retrieve = 500, although the differences in performance were in most cases small (data not shown). D-STAR returns motif instances; multiple instances of the same motif were assembled into a consensus motif by combining all variable characters at each position. To introduce ambiguity in motif definition, the three most variable positions in the motifs were replaced by 'wildcards'; this number was chosen to mimic the D-STAR setting of three mismatches (if multiple positions in the motif had the same ambiguity, we introduced 4 or 5 'wildcards'). To find motif occurrences in new sequences, ps_scan (de Castro *et al*., 2006) was used. Since we found in preliminary investigations that using those 'ambiguous motifs' gave many hits when applied to all proteins in the genome (see below), we used the consensus motifs without wildcards in that case.

2.2 Random forest and variable selection

Correlated motif pairs as found by the motif search algorithm are defined over pairs of sequences; this means that for each protein–protein pair we can define absence or presence of each correlated motif pair. Using experimental interaction data (Table 1), we can now train any classification algorithm to predict 'interaction' or 'no interaction'.

Table 1. Input datasets

TF family (organism)	Experimental data	N_p[a]	N_{int}[b]	ρ_{int}[c]
BZIP (human)	Protein array	57	349	0.21
MADS (*Arabidopsis*)	Y2H	29	111	0.26
Homebox (Mouse)	STRING	16	16	0.12
Homebox (*Arabidopsis*)	STRING	21	79	0.34
Forkhead (*Caenorhabditis elegans*)	STRING	12	23	0.29
Forkhead (Human)	STRING	32	316	0.60

[a]N_p: number of proteins.
[b]N_{int}: number of interactions.
[c]Interaction density: fraction interactions out of all pairwise combinations.

To obtain insight into which features (motif pairs) are important, we used a Random Forest (Breiman, 2001) classification algorithm in combination with a feature-selection procedure (Diaz-Uriate and de Andres, 2006) described previously in the context of gene selection in microarray experiments; both are available as packages in R (www.R-project.org). A Random Forest consists of an ensemble of classification trees whose output is combined. The resulting classification is determined by the class cutoff, i.e. the fraction of tree votes needed for a particular class. We tested several class cutoffs and found that cutoff(interaction) = 0.4 and cutoff(non-interaction) = 0.6 worked best in most cases, although the difference with the default 0.5/0.5 was small. Only for the bZIP data (for description of datasets see below) performance was stronger depending on the value for this cutoff; optimal results were obtained for cutoff(interaction) = 0.2 and cutoff(non-interaction) = 0.8. Note that the class cutoff can help to deal with unbalanced datasets, which is often the case for protein interaction data, where the number of interacting pairs typically is smaller than the number of non-interacting pairs. Other parameter settings in the Random Forest were: number of trees, 5000; mtry, number of variables randomly sampled as candidates at each split = sqrt(number of variables) and replace = TRUE, i.e. sampling with replacement.

Variable selection was performed as described (Diaz-Uriate and de Andres, 2006). Briefly, after gradually removing variables based on an importance measure, the out-of-bag error rates from all the fitted random forests are examined. The solution with the smallest number of variables whose error rate is within u standard errors of the minimum error rate of all forests is chosen. We used the default setting $u = 1$; this strategy can lead to solutions with fewer variables than selecting the solution with the smallest error rate, while achieving an error rate that is not different, within sampling error, from the 'best solution'. Variable selection was performed in two rounds. In the first round, 20% of the features was removed in each selection step [vars.drop.frac = 0.2; (Diaz-Uriate and de Andres, 2006)]. Starting with the features selected that way, in the second round features were selected one-by-one (vars.drop.frac = NULL and vars.drop.num = 1).

To obtain an unbiased estimate of the performance of our method, we performed leave-one-out cross-validation. Here, we removed subsequently each of the sequences and its associated interactions from the input data for D-STAR and the Random Forest, and probed the prediction accuracy of the resulting predictor on this sequence and its interactions.

2.3 Simulations

To validate our method and obtain insight into its robustness towards experimental noise, we used simulated datasets, generated by implanting motifs into random sequences and subsequently defining

presence of interaction based on motif occurrences. First, 25 random sequences were generated (sequence length 300). At each position in the sequence, one amino acid was randomly chosen as the 'most occurring amino acid', and randomly assigned a probability between 0.0 and 1.0. The 19 other amino acids were assigned the remaining probability. This procedure resulted in sequences with average sequence identity of 34%, which is comparable to the sequence identities found in the six transcription factor families that we used, which range from 23% to 34%. Subsequently, motifs were implanted in these sequences with each motif part having a number of occurrences drawn from a uniform distribution between 0 and the number of proteins. We used experimental short motifs from http://lmd.embl.de/lmd/Yeast_SinglesTA.html; we randomly pooled together the following motif pairs: (PxxxRxLS, TxxLF), (LxxQQ, PxxxIxY), (GxxxYxxL, DxxDxxxD), (YLxxLxxL, KASxxxQ) and (LxxLxK, LxDLxK). For various settings of network parameters (see below) we generated sequences, implanted one or more motif parts and defined interactions between sequences based on motifs.

In the experimental datasets, we found in most cases that specific combinations of motifs are needed for interactions. Therefore, presence of a single motif pair in a pair of sequences did not simply define interaction in our simulations. Rather, interaction was defined by first sorting all pairs of sequences based on the number of motifs present. Then, starting with the pairs with most motif pairs, sequence pairs were defined to be interacting, until the fraction of interacting pairs would reach the desired limit. The value for ρ_{int} (fraction of interacting pairs over all pairwise combinations) was 0.25. Note that because of the way we define interactions (based on motif occurrences) the interaction fraction will in general be slightly lower than this value (see Results section), depending on the distribution of the motifs in the sequences. Also note that the absence or presence of each motif pair is independent of the absence or presence of other motif pairs. Subsequently, false positive interactions and false negative interactions were generated by adding or deleting interactions; this was done for a fraction of 0.0, 0.125 and 0.25, resulting in nine different noise levels, from 0.0 false positives and 0.0 false negatives up to 0.25 false positives and 0.25 false negatives. At each parameter setting, five independent replicas were used. Performance was assessed using cross-validation as described above, but since in the case of simulated data both 'real' (data without noise) and 'observed' interactions (data with noise) are available, we report the performance with respect to both.

2.4 Support vector machine approach

To be better able to place the performance of our prediction method into perspective, we implemented a recently published method (Shen *et al*., 2007) that uses Support Vector Machine (SVM). Briefly, we applied the SVM^{light} package (Joachims, 1999) and the same cross-validation approach as described above for our Random Forest-based method. The SVM parameters C and γ were optimized using a grid of values as described in Shen *et al*. (2007). For more details, see Supplementary Table 2.

2.5 Input data

As biological datasets, we wanted to use trusted experimental data sources as well as data from a large-scale interaction database, to illustrate potential large-scale applications of our method. As experimental datasets, data for the human bZIP (Newman and Keating, 2003) and *Arabidopsis* MADS (de Folter *et al*., 2005) transcription factor families were used (see Table 1).

In addition, four datasets were selected from the STRING database that integrates known and predicted interactions (von Mering *et al*., 2007). The STRING data was first clustered into transcription factor-related KOGs (Tatusov *et al*., 2000), based on names for KOGs provided in ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kog. Subsequently, selection of the datasets was guided by using a cutoff of at least 10% for the interaction density ρ_{int}, i.e. the fraction of protein pairs that are interacting. Note that the majority of the transcription factor-related KOGs in STRING has a lower interaction density, which probably reflects incompleteness in the STRING data. Two datasets from two different organisms for each family were chosen, since differences between the same families in different organisms might illustrate characteristics of our approach. A final requirement was that we wanted structural information for these families, in order to validate our analysis. For the homeobox and forkhead domains, structural information is indeed available [PDB 1fjl, (Wilson *et al*., 1995) and PDB 2007, (Stroud *et al*., 2006), respectively], so we selected these two families. The KOG identifiers for the selected homeobox and forkhead families are 0773 and 2294, respectively.

Table 2. Number of motifs and prediction performance (based on leave-one-out cross-validation) per dataset

TF family (organism)	N[a]	Spec[b]	Cov[b]	Acc[b]	SVM[c]
BZIP (human)	8	0.6 (0.4)	0.3 (0.3)	0.8 (0.1)	0.7 (0.2)
MADS (*Arabidopsis*)	7	0.5 (0.4)	0.4 (0.3)	0.8 (0.2)	0.8 (0.2)
Homeobox (Mouse)	2	0.7 (0.2)	0.3 (0.4)	0.9 (0.1)	0.9 (0.1)
Homeobox (*Arabidopsis*)	2	0.5 (0.3)	0.3 (0.3)	0.6 (0.1)	0.6 (0.1)
Forkhead (*C.elegans*)	2	0.2 (0.3)	0.4 (0.5)	0.6 (0.3)	0.7 (0.2)
Forkhead (Human)	6	0.6 (0.4)	0.5 (0.4)	0.8 (0.2)	0.7 (0.2)

[a]Number of motifs selected by the variable selection procedure.
[b]Performance is assessed by specificity (Spec), coverage (Cov) and accuracy (Acc), where specificity is the fraction of predicted interactions that are correct, coverage is the number of interactions that is indeed predicted and accuracy is the accuracy over all predictions (both interactions and non-interactions).
[c]Accuracy of SVM-based prediction. For detailed performance of SVM, see Supplementary Table 2.

2.6 Genome-wide prediction

Using the motifs selected by the Random Forest for the experimental datasets (see Results section), all peptide sequences containing these motifs in the respective genomes were found using ps_scan. Genome data were obtained via TAIR (*Arabidopsis*) and Ensembl (mouse, human and *C.elegans*). All possible pairs of motif-containing proteins were then input to the predictor.

3 RESULTS AND DISCUSSION

Our approach towards predicting and understanding interaction specificity at a sequence level consists of a prediction algorithm that is trained using interaction data. After training, the method can predict interactions based on sequence data only. The presence of correlated motif pairs, which are defined over pairs of sequences, was used as input features for the predictor. To find these motif pairs we used an existing correlated motif search algorithm, D-STAR (Tan *et al*., 2006).

An important step is the selection of motifs, since the search returns many motifs and one does not know in advance which and how many of these are 'real'. A Random Forest classification algorithm (Breiman, 2001) was used, together with a recently published feature-selection method (Diaz-Uriate and de Andres, 2006). Based on input interaction data, the classifier selects those features that it needs to build a reliable

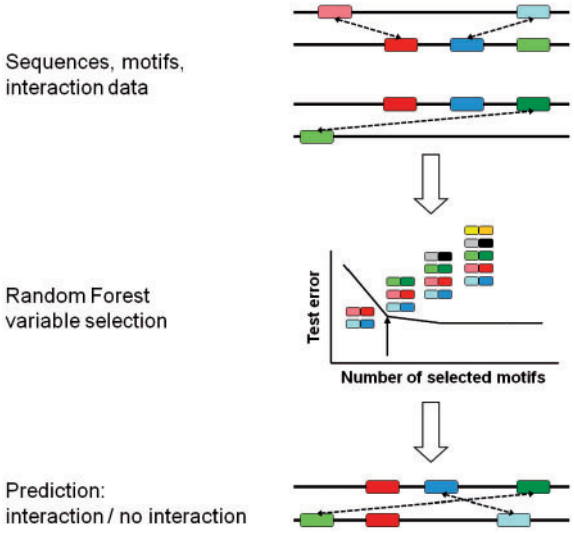

Fig. 1. Motif selection and interaction prediction. Interaction data (top) are used as input to the motif search and to train the Random Forest (middle), which selects the important motif pairs based on the rise in test error when too many motifs are removed. The resulting Random Forest can predict interactions in new input sequences (bottom). Lines with colored boxes indicate protein sequences and motif parts, respectively; complementary motif parts have matching colors. Arrows between proteins indicate interactions based on the presence of a motif pair in the protein pair (i.e. the proteins have complementary motif parts).

predictor. Subsequently, interaction predictions can be made using the presence of the selected motif pairs in pairs of sequences. In Figure 1 this procedure is schematically illustrated. Note that with the datasets we used, the running time of both D-STAR and the Random Forest motif selection is typically between several minutes and a few hours.

3.1 Method validation using simulated datasets

Prior to its application to various biologically relevant datasets (see below), we applied our method on simulated datasets to obtain insight in the robustness of the performance with respect to noise. The datasets were generated by implanting short sequence motifs into random sequences and subsequently defining presence of interaction based on motif occurrences. Specific combinations of motif pairs (see Methods section for description of those) were set to be required for interaction (in a separate set of simulations we defined the presence of one motif pair to be sufficient for determining interaction; since no systematic differences were observed between both sets of simulations, we will discuss only the first set of simulations). The interaction density ρ_{int} (fraction of interacting pairs out of all possible pairwise combinations) was 0.18 ± 0.06, and we added various levels of noise (0%, 12.5% and 25%, both as false positives and false negatives, resulting in nine different noise levels). The performance of the method was assessed using leave-one-out cross-validation.

In Supplementary Table 1, the results for all simulations are provided. Figure 2 shows the average performance of our method, assessed by specificity, coverage and accuracy. The

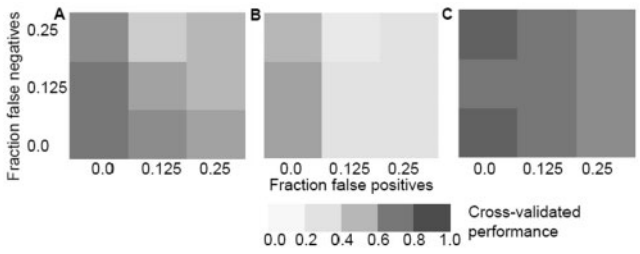

Fig. 2. Influence of noise in interaction data on performance of the algorithm. (**A**) specificity (fraction of predicted interactions that are indeed true), (**B**) coverage (fraction of true interactions that are predicted) and (**C**) accuracy as a function of fraction false positives and fraction false negatives. Error rates are based on leave-one-out cross-validation. Data shown here are results of simulations with fraction interacting combinations 0.25 and are results of five replicas (SDs not shown). Full results of all simulations are shown in Supplementary Table 1.

latter is defined based on both positive and negative predictions and interaction data; specificity and coverage are defined based on positive predictions relative to the available positive interaction data. Specificity is the fraction of predicted interactions that are true positives, and coverage is the fraction of experimental interactions that are indeed predicted.

The method has a reasonable accuracy; however, coverage seems to be generally lower than specificity. This was not the case when we assessed specificity and coverage of the method when using all data to define the motifs in the motif search phase (so without cross-validation). Apparently, this indicates that the generalization of motifs to unknown sequences (as probed by the cross-validation) is a key and difficult step in the algorithm.

Interaction density does not seem to have a large influence on the prediction performance (data not shown), which is an important point since many classification algorithms suffer from imbalance in the dataset. In general, one would expect interaction densities below 0.5 (the chance that two proteins interact is lower than the chance that they do not interact) and indeed in most of our datasets this is the case. In addition, we analyzed interaction data for all KOGs in the human protein interaction database HPRD (Mishra *et al.*, 2006), which according to a recent survey (Mathivanan *et al.*, 2006) contains the largest number of proteins and interactions compared to other human protein interaction databases; here we found the average interaction density per KOG to be 20% (with a SD of 35%). This means that the fact that our approach can deal with imbalance in the data is an important strength.

3.2 Application to TF families: interaction prediction

We further applied our method to six different transcription factor families, covering different organisms and experimental methods (see Table 1). Results are given in Table 2. Accuracy is 60% or higher, and the average of specificity and coverage over the various datasets is 50% and 40%, respectively (these values are based on leave-one-out cross-validation). It is not easy to compare these values to existing approaches, since most of these are not applicable to the special case of interactions within

transcription factors that we study here. Only for the bZIP family we found an existing prediction approach, consisting of sequence-based rules (Newman and Keating 2003).

Our method (specificity 56%, coverage 34%), performs better than these rules. For example, when these rules are defined in such a way that they identify one-third of the interactions, the number of false positives equals the number of true positives (specificity 50%, coverage 33%). Our performance is somewhat worse than that of an existing approach that was developed specifically to predict bZIP interactions (Fong *et al*., 2004), which has specificity 82% and coverage 54%. The latter approach uses knowledge about which residue–residue interactions take place in coiled coils, based on structural features of the interhelical interface, as well as experimental knowledge on determinants of specificity. This knowledge is generally not available, which is exactly why we aim for a method that uses only sequence information.

To be better able to place the performance of our prediction method into perspective, we also implemented a recently published method (Shen *et al*., 2007) that uses sequence information to predict interactions using SVM; see Supplementary Table 2 for a further discussion. The average accuracy of both approaches is comparable. In three out of six cases, our method has a much better specificity and coverage; in two cases the SVM approach has a better specificity and coverage (although the difference in these two cases is small) and in one case performance is comparable. It is interesting that the three sets where our method is clearly better, bZIP, MADS and human forkhead, are the datasets where our predictor selects the highest number of motifs (see Table 2). This could indicate that the triads used in the SVM are less suited to deal with these datasets where combinatorial effects likely play a role. In addition, the SVM approach does not provide information about which parts of the sequence are important in mediating the interaction, which is an explicit goal of our work (see below).

In general, the performance of our sequence motif-based predictor is satisfactory, especially taking into account the difficulty of interaction prediction based on sequence data only. We also tested that using randomly swapped class indicators, the performance drops significantly, indicating a non-trivial performance of our method.

3.3 Application to TF families: analysis of sequence motifs

In addition to being able to predict interactions, the main output of our method is a set of motifs that it predicts to be sequence level determinants of interaction specificity. In Supplementary Table 3, all the selected motifs are presented. To investigate the meaning of those motifs, we analyzed the homeobox and forkhead datasets, for which dimer structures are available. For these, motifs are found in the respective domains, which are also found at the interface in the experimental structure, as shown in Figure 3 (in Supplementary Table 3, their position within the sequence logo of the respective domains is indicated). This indicates that our method is successful in returning motifs that are likely to be functional. Other motifs are found that are outside the domains with known structure; for these we cannot use the structure to judge whether they could play a role in the interaction or not. However, using InterProScan we found for several of these that they overlap with motifs known or suspected to be involved in mediating interactions (see Supplementary Table 3). For example, in the Arabidopsis homeodomain proteins we find a motif overlapping with the KNOX2 (Pfam Pf03790) motif, which is known to be involved in interactions. This again underlines our success in finding relevant motifs.

We also compared the results for the two forkhead and the two homeodomain datasets. In the latter case, there is no

Table 3. Genome-wide prediction of interaction partners[a]

TF family (organism)	Number of predicted partners		
	Total	Pred–pred[b]	Pred-known[c]
BZIP (human)	33	33	24
MADS (*Arabidopsis*)	1	1	0
Homeobox (Mouse)	33[d]	33[d]	33[d]
Homeobox (*Arabidopsis*)	10[e]	10[e]	0
Forkhead (*C.elegans*)	2[f]	2[f]	0
Forkhead (Human)	1741/19[g]	1741/19[g]	1741/19[g]

[a]Motifs selected by the Random Forest procedure were used in genome-wide scan; proteins that contained one or more motif parts were subsequently used in the Random Forest predictor to predict interactions with the TF family proteins.
[b]Number of additional proteins that are predicted to have at least one interaction with one of the other additional proteins.
[c]Number of additional proteins that are predicted to have at least one interaction with one of the proteins of the TF family.
[d]Note that among these proteins, one actually is a homeobox protein that was not included in the original STRING dataset.
[e]Note that four of these proteins are in fact homeobox proteins that were not included in the original STRING dataset.
[f]Note that these two proteins are in fact forkhead proteins that were not included in the original STRING dataset.
[g]Here 1741 additional proteins are found with the predictor trained on all STRING data, and 19 with the predictor trained on the STRING data with score above 0.25 (see text for details).

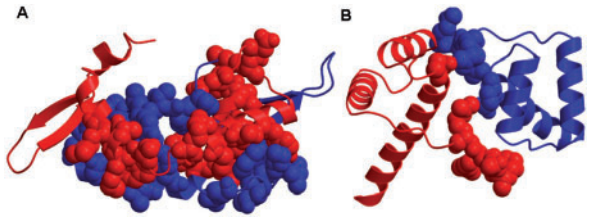

Fig. 3. Motif validation using structural information. (**A**) Human forkhead dimer structure (Foxp2, Protein Data Bank id 2a07). Spacefill indicates interface residues that are part of sequence motifs found in human (Arg543, Thr547, Trp548, Lys549, Ala551, Val552, Arg553, Pro506, Phe507, Thr508, Tyr509, Ala510, Ile513) and *C.elegans* (Ile517, Met518, Glu519, Asp522, Arg523, Tyr531, Phe534, Thr535, Phe538) datasets. (**B**) Drosophila homeodomain dimer (Paired Protein, PDB id 1fjl). Spacefill indicates interface residues that are part of sequence motifs found in mouse (Lys0, Gln1, Arg2, Arg3) and *Arabidopsis* (Glu42). In both (A) and (B), the two monomers that form the dimer are shown in blue and red, respectively.

sequence homology outside the homeodomain itself (the Arabidopsis homeobox proteins fall in another subclass than the mouse homeobox proteins); however, for the forkhead proteins some of the motifs found in one species are present in the sequences of the other species, even if there is no overlap between the motifs themselves. Specifically, for each of the two motif pairs found in *C.elegans*, one of the motif parts is also found in some of the human sequences. Conversely, for three of the human motif pairs both motif parts are found in *C.elegans* sequences. These three motifs are not enough to predict interaction according to the 'human predictor'. However, we find that the presence of these motifs in *C.elegans* is really indicative of interaction, since out of the 33 protein–protein pairs with at least one of these motif pairs, 15 are indeed interacting [there are in total 23 interactions, giving specificity (coverage) of 0.45 (0.65)]. This illustrates the potential of using the selected motifs to predict interactions in other genomes.

To understand the role of the selected sequence motifs, we analyzed in the various datasets whether only a single motif is enough for the model to predict interaction or that specific combinations of motifs are required. To answer this question, we grouped all interacting pairs into 'interaction classes', based on the specific combinations of motifs that they contain. As shown in Figure 4, there are differences between the various families in the number of interaction classes and the number of motifs that define these interaction classes. The mouse homeobox contains only one interaction class, which is defined by two motifs, meaning that both motifs that were found for this dataset have to be present in order to predict 'interaction'. On the contrary, for the *Arabidopsis* homeobox, there is one interaction class consisting of two motifs, but also two interaction classes consisting of each of the two motifs separately.

Interestingly, the *Arabidopsis* homeobox and human forkhead, which have higher interaction densities (ρ_{int}, see Table 1) than respectively mouse homeobox and *C.elegans* forkhead, also have a higher number of interaction classes. Intuitively, one would expect that when there are more interactions there are also more interaction classes needed in order to maintain specificity, and the same effect is indeed seen in the simulated datasets (data not shown).

A high number of motifs do not necessarily mean that extended combinations of motifs are needed for interaction; e.g. in the bZIP transcription factor family, many interactions are predicted based on the occurrence of a single motif pair or a combination of only two motif pairs. We refer to the combination of multiple motif pairs being needed for interaction prediction as 'motif synergy'. Figure 4B illustrates this concept showing the single classification tree that best fits the selected motifs for the human forkhead dataset. This tree is not directly comparable to the Random Forest classifier, but again it shows a motif class with only one motif needed for interaction (left side) and motif classes with multiple motifs ('motif synergy', right side).

3.4 Genome-wide prediction

To investigate the wider applicability of the interaction motifs found in the different datasets, we analyzed their genome-wide occurrence. Proteins that contained at least one motif part were

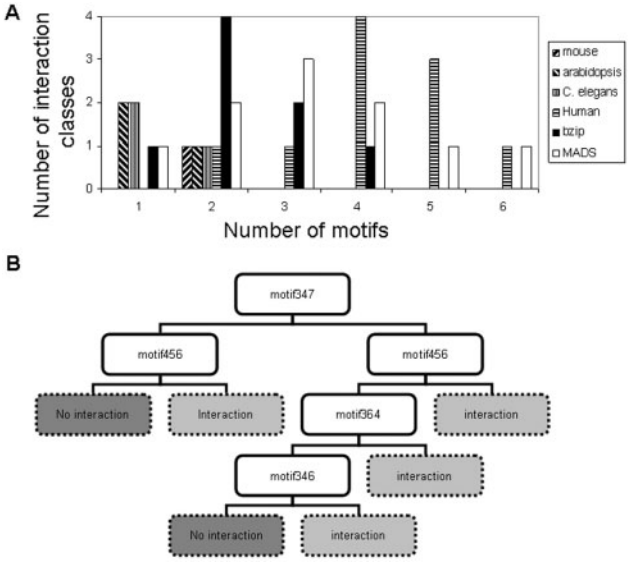

Fig. 4. Interacting protein pairs are clustered into 'interaction classes' based on their sequence motifs (for details see text). (**A**) Histogram of number of interaction classes with the indicated number of motifs for the various interaction datasets. For example, for the mouse homeobox dataset, only when both motifs are present, the model predicts 'interaction'. For most other datasets, interaction classes with a single motif as well as with specific combinations of motifs do occur. (**B**) Classification tree for the human forkhead dataset. The tree starts at the top and at each node, if a motif is present the path goes right and if it is not present the path goes left. Note that this tree is not directly comparable to the Random Forest classifier. This tree shows that either one specific motif is enough to predict 'interaction' (motif456, at the left side of the tree) or specific motif combinations are needed in this dataset to predict 'interaction'.

submitted to the predictor. In several cases, additional predicted interactions were found (see Table 3), and many of those could indeed be verified in literature. For the human bZIP and human forkhead data, we used the human protein interaction database HPRD (Mishra *et al.*, 2006) to assess the predicted additional interactions via an association test based on the hypergeometric distribution.

3.4.1 Human bZIP Approximately half of the additional predicted interactors contain coiled coils [as assessed using SMART; (Letunic *et al.*, 2006)], so it is not unlikely that they could interact with the coiled coil containing bZIP proteins. Among them are some phosphatases and kinases (e.g. CDC42-binding protein kinase beta, and Rho-associated, coiled-coil containing protein kinase 2), and it is indeed known that the bZIP transcription factors are activated and deactivated by phosphorylation/dephosphorylation. The overlap with HPRD is significant in this case ($P \approx 10^{-9}$). Of course one should also keep in mind that HPRD does not exhaustively sample interaction space.

3.4.2 Arabidopsis MADS Here only one additional interactor was predicted, for which we could find no evidence.

3.4.3 Mouse homeobox The predicted interactors include several transcription factors, for which we could find literature evidence for their interactions with homeobox transcription factors: GATA6 (Decker *et al*., 2006), SOX3 (Dailey and Basilico, 2001), SP9 [Sahara *et al*., 2007, Neural Development 2007, online] and SRF (Ju *et al*., 2006). For the latter the interaction site on the homeobox domain has been mapped, and it matches one of the selected motifs (data not shown). In addition, some chloride channels were predicted as interactors. These are known target genes (Ando *et al*., 2005; Costantini *et al*., 2005), so this interaction might indicate the presence of a feedback mechanism between the product of the target gene and the transcription factor.

3.4.4 Arabidopsis homeobox The majority of the predicted interactors of the *Arabidopsis* homeobox TFs was of unknown function. However, one interaction was predicted for UBC26, a ubiquitin conjugating enzyme; supporting evidence for this interaction is provided by the known interaction between the related zebrafish UBC9 and VSX-1, a homeobox TF (Kurtzman and Schechter, 2001). Two other predicted interactors with known function were a hydrolase and a peptidase, for which however no supporting evidence is available.

3.4.5 C.elegans forkhead Two new interactors were predicted, which are in fact forkhead proteins as well (but were not included in the original dataset).

3.4.6 Human forkhead In this case the predictions are questionable, since over 1700 additional interactors are predicted. Comparison of the overlap with HPRD gave a *P*-value of 0.73, indicating that there is no significant association. The likely reason here is noise in the input data; this dataset has the highest interaction density, which might well mean that many of the interactions are not correct. Indeed, of the input STRING data, only two interactions were found in HPRD as well, and these are the two with the highest values of STRING score. When applying our method on a subset of the STRING data using a STRING score cutoff of 0.25, the predictor finds 19 additional proteins that are in HPRD, with 3 interactions and 84 non-interactions validated by HPRD, 103 predicted interactions not found in HPRD and no HPRD-based interactions missing. The *P*-value now is 0.26, which is better than before (note again that HPRD is not exhaustive). These 19 proteins include some with annotation 'similar to forkhead protein' as well as ATRX (a transcriptional regulator) and a SET domain containing protein, which has histone methylation activity. These are all related to transcriptional activity and interactions with forkhead transcription factors seem plausible. Note that we only use some short sequence motifs so it is remarkable to get an overrepresentation of transcription-related proteins. On the basis of these results, our method seems promising for genome-wide interaction prediction.

4 CONCLUSION

We present here a method to analyze transcription factor interactions at the sequence level and to predict these interactions based on sequence information only. By using simulated datasets, we demonstrated its robustness towards experimental noise and its capability to deal with unbalanced data. The latter is an important feature since most interaction datasets have a low interaction density. In various biological datasets, motifs selected by our approach can be interpreted as mediating interactions, and indeed were found at the interface in several cases. The prediction of additional interaction partners should be considered as preliminary work; one limitation is that only proteins that interact in a 'similar' way (using the same sequence motifs) as the TFs among themselves would be found. The core of our results is an interaction predictor for TF families. This predictor gives insight into the sequence level determinants of these interactions. In future work, it will be possible to extend our approach further towards genome-wide interaction prediction. A main application of our algorithm will be towards various sets of TF families, including predictions of interactions in other genomes. In addition, experimental validation through targeted mutagenesis of predicted motifs is currently being performed.

ACKNOWLEDGEMENTS

This work was supported by the BioRange programme (SP 2.3.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI).

Conflict of Interest: none declared.

REFERENCES

Ando,Z. *et al*. (2005) Slc12a2 is a direct target of two closely related homeobox proteins, Six1 and Six4. *FEBS J*., **272**, 3026–3041.

Babu,M.M. *et al*. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol*., **14**, 283–291.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn*., **21**, 51–80.

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

Chen,X.W. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.

Costantini,D.L. *et al*. (2005) The homeodomain transcription factor lrx5 establishes the mouse cardiac ventricular repolarization gradient. *Cell*, **123**, 347–358.

Dailey,L. and Basilico,C. (2001) Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. *J. Cell. Physiol*., **186**, 315–328.

Davidson,E.H. and Erwin,D.H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science*, **311**, 796–800.

de Castro,E. *et al*. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*., **34**, W362–W365.

de Folter,S. *et al*. (2005) Comprehensive interaction map of the Arabidopsis MADS box transcription factors. *Plant Cell*, **17**, 1424–1433.

Decker,K. *et al*. (2006) Gata6 is an important regulator of mouse pancreas development. *Dev. Biol*., **298**, 415–429.

Diaz-Uriate,R. and de Andres,S.A. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Fang,J.W. *et al*. (2005) Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics*, **6**, 277.

Fong,J.H. *et al*. (2004) Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol*., **5**, R11.

Gandhi,T.K.B. *et al*. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet*., **38**, 285–293.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B. *et al*. (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, Cambridge, MA, pp. 169–184.

Ju,J.H. *et al*. (2006) Physical and functional interactions between the prostate suppressor homeoprotein NKX3.1 and serum response factor. *J. Mol. Biol.*, **360**, 989–999.

Kurtzman,A.L. and Schechter,N. (2001) Ubc9 interacts with a nuclear localization signal and mediates nuclear localization of the paired-like homeobox protein Vsx-1 independent of SUMO-1 modification. *Proc. Natl Acad. Sci. USA*, **98**, 5602–5607.

Letunic,I. *et al*. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.

Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.

Li,H.Q. and Li,J.Y. (2005) Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics*, **21**, 314–324.

Li,H.Q. *et al*. (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, **22**, 989–996.

Martin,S. *et al*. (2005) Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

Mathivanan,S. *et al*. (2006) An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **7** (Suppl. 5), S19.

Mishra,G.R. *et al*. (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.*, **34**, D411–D414.

Neduva,V. *et al*. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, 2090–2099.

Newman,J.R.S. and Keating,A.E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*, **300**, 2097–2101.

Qi,Y.J. *et al*. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins Struct. Funct. Bioinformatics*, **63**, 490–500.

Sahara,S. *et al*. (2007) Sp8 exhibits reciprocal induction with Fg and 8 but has an opposing effect on anterior-posterior cortical area patterning. *Neural Develop.*, **2**, 10.

Salwinski,L. *et al*. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Shen,J.W. *et al*. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.

Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Sprinzak,E. *et al*. (2006) Characterization and prediction of protein-protein interactions within and between complexes. *Proc. Natl Acad. Sci. USA*, **103**, 14718–14723.

Stroud,J.C. *et al*. (2006) Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14**, 159–166.

Tan,S.H. *et al*. (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, **7**, 502.

Tatusov,R.L. *et al*. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Teichmann,S.A. and Babu,M.M. (2004) Gene regulatory network growth by duplication. *Nat. Genet.*, **36**, 492–496.

von Mering,C. *et al*. (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.

Walhout,A.J.M. (2006) Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.*, **16**, 1445–1454.

Wilson,D.S. *et al*. (1995) High-resolution crystal-structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.