Interface similarity improves comparison of DNA-binding proteins: the Homeobox example

Álvaro Sebastián[1], Carlos P. Cantalapiedra[1], and Bruno Contreras-Moreira[1,2]

[1] Laboratorio de Biología Computacional,
Estación Experimental de Aula Dei/CSIC,
Av. Montañana 1005, Zaragoza, España
[2] Fundación ARAID, Paseo María Agustín 36, Zaragoza, España
http://www.eead.csic.es/compbio
{asebastian,bcontreras}@eead.csic.es

Abstract. The recently published 3D-footprint database contains an up-to-date repository of protein-DNA complexes of known structure that belong to different superfamilies and bind to DNA with distinct specificities. This repository can be scanned by means of sequence alignments in order to look for similar DNA-binding proteins, which might in turn recognize similar DNA motifs. Here we take the complete set of Homeobox proteins from *Drosophila melanogaster* and their preferred DNA motifs, which would fall in the largest 3D-footprint superfamily and were recently characterized by Noyes and collaborators, and annotate their interface residues. We then analyze the observed amino acid substitutions at equivalent interface positions and their effect on recognition. Finally we estimate to what extent interface similarity, computed over the set of residues which mediate DNA recognition, outperforms BLAST expectation values when deciding whether two aligned Homeobox proteins might bind to the same DNA motif.

Keywords: protein-DNA interface, DNA motif, substitution matrices

1 Introduction

3D-footprint [1] (`http://floresta.eead.csic.es/3dfootprint`) is a database that dissects sequence readout in protein-DNA complexes of known structure, extracted from the Protein Data Bank [2], identifying molecular contacts that contribute to specific recognition and inferring structure-based position weight matrices from the atomic coordinates. Currently the database contains over 2700 complexes, which can be assigned to SCOP superfamilies [3]. After removing redundancy, the most populated superfamily turns out to be that of homeodomain-like proteins, including Homeobox transcription factors, which have been the subject of extensive crystallographic and spectroscopic studies due to their key role in developmental processes in multicellular organisms [4].

Furthermore, Homeobox proteins are of special interest since the publication of the work by Noyes and collaborators [5], in which the authors characterized the

binding specificities of 85 *Drosophila melanogaster* homeodomains. This repertoire of homologous transcription factors provides a formidable opportunity to study the correlation between the mutations that naturally occur at the interface of Homeobox proteins and their effect on binding specificity.

In this paper we apply the structural knowledge contained in 3D-footprint to: i) define the set of most commonly used interface residues across Homeobox proteins; and ii) elucidate to what extent interface similarity between pairs of homeodomains correlates with the recognition of similar DNA motifs.

After a cross-validation benchmark we find that interface position-specific substitution matrices (ISUMs), automatically inferred from training sets of homeodomains, perform better than BLOSUM62, and significantly better than BLAST expectation values, in the task of deciding whether two aligned Homeobox proteins bind to the same DNA motif.

2 Material and Methods

2.1 Homeobox protein sequences and DNA motifs

A dataset of 85 *D.melanogaster* Homeobox protein sequences and their 2240 DNA binding sites, first published by Noyes [5], was used to build 85 position weight matrices (PWMs) in TRANSFAC format using both CONSENSUS [6] and MEME [7], choosing in each case the resulting PWM with largest information content.

2.2 Structural alignment of homeodomains and identification of interface residues

A multiple structural alignment of non-redundant homeodomains extracted from 3D-footprint [1] was compiled as previously explained [10], and all identified interface interactions annotated as hydrogen bonds, water-mediated hydrogen bonds or hydrophobic interactions. For simplicity interface residues were numbered following the schema using by Noyes [5]. The list of annotated homeodomains includes 35 Protein Data Bank chains:

`1au7_B_1,1b72_A_1,1b8i_A_1,1b8i_B_1,1e3o_C_1,1fjl_C_1,1h89_C_1,1h89_C_2,`
`1hlv_A_2,1ic8_A_1,1ig7_A_1,1ign_A_1,1ign_A_2,1jgg_B_1,1jt0_C_1,1le8_A_1,`
`1mnm_C_1,1nk3_P_1,1puf_A_1,1puf_B_1,1w0t_B_1,1w0u_B_1,1yz8_P_1,1zq3_P_1,`
`2d5v_B_1,2h1k_B_1,2hdd_A_1,2kdz_A_1,2kdz_A_2,2qhb_A_1,2r5y_A_1,2yvh_C_1,`
`3cmy_A_1,3d1n_I_1,9ant_B_1`

Furthermore, these homeodomains were sampled to calculate pairwise interface alignments. This procedure starts by reducing their protein-DNA interfaces to two-dimensional matrices, which we call *interface matrices*, that are expected to capture most details of their binding mode. Then, a pair of such matrices can be aligned by i) matching interface amino acid residues whose contact patterns overlap, and ii) by penalizing pairs of residues with distinct contact maps. As a by-product of these interface alignments we also obtain structure-based alignments of their bound DNA sequences, as shown in Figure 1.

2.3 Annotation of interface residues in Homeobox proteins

The interface positions of all 85 *D.melanogaster* Homeobox protein sequences were assigned by means of local BLASTP [8] alignments to 3D-footprint entries.

2.4 DNA motif alignment and similarity scoring

All 85 *D.melanogaster* position weight matrices (PWMs, see Supplementary Material), which were generated with the DNA binding sites described in section 2.1, were aligned against each other with the STAMP software[9], using an ungapped Smith-Waterman algorithm and taking the Pearson Correlation Coefficient as the similarity score. This *similarity score* takes values in the range [-L , L], where L is the length of the PWM.

2.5 Cross-validation parameters

The original dataset was split into training and validation subsets of 68 and 17 homeodomains, respectively. This process was repeated for 10 rounds with different random training and validation sets. Training sets were used to compute ISUMs for each interface position, while validation sets were used to benchmark the DNA motif predictions made by applying the previously calculated ISUMs.

2.6 Generating interface substitution matrices (ISUMs)

Homeobox domain sequences were globally aligned with MUSCLE [11] and their interface positions labelled. For each of the 8 interface positions, in bold in Table 1, the 4 most abundant amino acids were selected. All the 2^{10} possible binary score variations with repetitions among pairs of these 4 residues were computed. Obviously non binary scores are possible and probably more realistic, but at the cost of increasing the search space. For example, the four residues most frequently found in interface position 2 (G,R,Q,K) could be assigned the following 10 substitution scores, which represent the chance of mutating one residue to another while preserving the ability to recognize the same DNA motif: $GG \rightarrow 1, RR \rightarrow 1, GQ \rightarrow 0, KK \rightarrow 1, QR \rightarrow 0, KR \rightarrow 1, QQ \rightarrow 1, GK \rightarrow 0, KQ \rightarrow 0, GR \rightarrow 0$.

For each interface position the best score variations, those that maximized the Pearson correlation between interface scores and the corresponding DNA motif alignment scores, were selected and used to build symmetric interface substitution matrices (ISUMs). Ten sets of ISUMs were generated (one per training set) and used independently to perform DNA motif predictions within each of the corresponding validation sets. The final ISUMs in Tables 2 and 3 are the average of 10 cross-validations rounds.

2.7 Pairwise alignments of Homeobox domains

Pairs of protein sequences from the validation set where aligned with the BLASTP program [8]. Resulting E-values where normalized by taking the natural logarithm and dividing with respect to self-match E-values, as shown for sequences A and B in Equation 1:

$$normEval(A,B) = \begin{cases} 0 & if\ Eval(A,B) > 0; \\ 1 & if\ Eval(A,B) < AbsMinEv; \\ \frac{ln(Eval(A,B))}{ln(minEval))} & if\ minEval \geq AbsMinEv; \\ \frac{ln(Eval(A,B))}{ln(AbsMinEv)} & if\ minEval < AbsMinEv. \end{cases} \quad (1)$$

where $minEval = min(Eval(A,A), Eval(B,B))$ and $AbsMinEv$ is the absolute smallest non-zero E-value returned by BLAST ($1e^{-180}$). Normalized E-values take values in the range 0-1, with small values corresponding to low sequence similarity.

2.8 Validation of the calculated ISUMs

All possible pairwise alignments from the validation set, extracted from the global multiple alignment mentioned in section 2.6, were sampled and their interfaces scored with the trained ISUMs and with the generic substitution matrix BLOSUM62 [12]. In addition, all sequence pairs from the validation set were also re-aligned with BLASTP in order to calculate normalized E-values, which we used as a measure of overall sequence similarity. Data pairs of interface and motif alignment scores were scatter-plotted, linear regression estimated by least-squares fitting and statistical parameters calculated with statistical software R [13]. A receiver operating characteristic (ROC) curve was also plotted taking as *truth test* a motif *similarity score* ≥ 5 . This threshold was tuned after benchmarking the content of TRANSFAC database v9.3 [14], in order to obtain a sensitivity (True Positive Ratio) of 0.7 and a specificity (1-False Positive Ratio) of 0.9.

3 Results

3.1 Comparison and clustering of Homeobox interface architectures

A non-redundant set of homeodomains included in 3D-footprint [1] was analyzed and their protein-DNA interfaces reduced to two-dimensional matrices. Subsequently these 2D interface matrices were compared to each other, end the corresponding DNA motif alignments extracted, as illustrated in Figure 1. As a result of this structural analysis we found that Homeobox DNA motifs usually fit one of 7 subtypes, shown in Figure 2, which approximately encompass

the 11 groups originally proposed by Noyes [5]. These clusters show the structural equivalence between different subtypes of Homeobox DNA motifs, which we consider in section 3.6 in order to call incorrect DNA alignments.

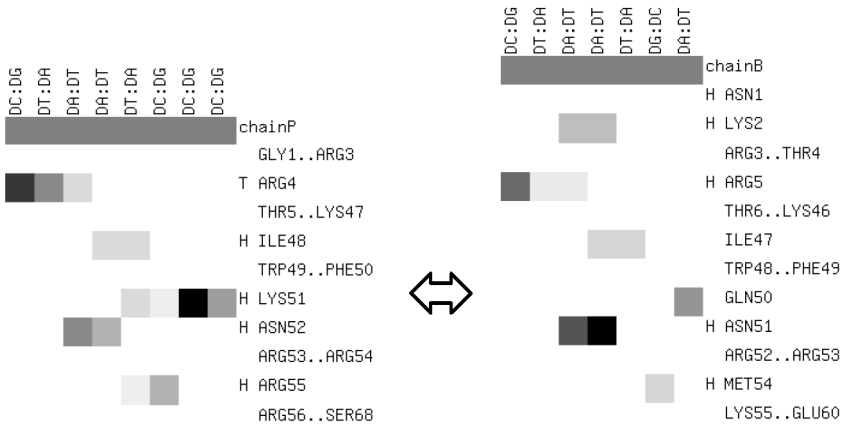

Fig. 1. Structure-based interface alignment of homeodomains 1zq3_P (left) and 2h1k_B (right). Each complex features interface residues in the Y axis and DNA nucleotides in the X axis (grey bar on top). Interactions are depicted as filled squares with density proportional to the number of atomic contacts. Aligned (equivalent) interface residues are placed in the same row. The resulting DNA motif alignment is: CTAATCCC / CTAATGA- .

3.2 Defining a consensus Homeobox protein-DNA interface

Homeobox proteins (and in general homeodomain-like proteins) usually bind to DNA with a conserved architecture. This observation can be used to infer interface conservation directly from sequence alignments between transcription factors. However, a consensus set of residues which i) faithfully represent the interface architecture, and ii) minimize the loss of information, must be defined beforehand.

In the case of Homeobox proteins we used structural data collected in the 3D-footprint database [1] to annotate the critical interface residues involved in DNA recognition, which are shown in Table 1. Out of the 18 surveyed interface positions, which vary in terms of number of contacts and in frequency across homeodomain-like sequences, we shortlisted the same 8 positions proposed by Noyes [5] in order to facilitate the comparison of our results. These positions include 47, 50 and 54, which have been previously reported to be the key determinant positions for DNA recognition in mouse Homeobox transcription factors [15].

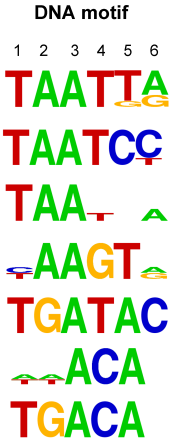

Fig. 2. Multiple alignment of Homeobox DNA motif subtypes, inferred from pairwise structure-based interface comparisons.

Table 1. Survey of interface residues in homeodomains, numbered as in the work of Noyes [5]. The 8 most relevant residues for Homeobox proteins are in bold.

residue number	**2**	**3**	4	**5**	6	29	41	42	43	44	46	**47**	**50**	**51**	53	**54**	**55**	58
Hbonds	3	3	2	21	1	1	1	2	7	1	8	17	9	21	1	5	5	1
water-Hbonds	4								1		4	1	6	1				1
hydrophobic									1			8	2	1				

3.3 Derivation of Interface Substitution Matrices (ISUMs)

A 10-round cross-validation experiment was set up in order to calculate Interface Substitution Matrices (ISUMs) from randomly sampled training subsets of 68 Homeobox sequences. The resulting average matrices are shown in Table 2. ISUMs showed to be very similar from one training subset to another. In particular, inspection of the total 690 amino acid substitution scores generated (out of a theoretical maximum of 800: 10 per interface position x 8 positions x 10 rounds), only 160 scores differed from one training set to another and 80 belonged to positions 5 and 51. These later positions indeed contribute little to increase the correlation between the interface and the DNA recognized, as already envisaged by Noyes [5], and this might explain their variability.

As a control experiment, we also generated ISUMs for two randomly chosen non-interface positions (36 and 66, Table 3), which display an even higher variability in its amino acid composition and a lower correlation with the DNA motifs recognized. In general, any other sequence positions without interface roles contribute very little, if anything, to the DNA motif correlation. As a consequence, it is possible to increase the list of interface residues if necessary, as only relevant interface residues will have a correlation impact. In fact we ob-

served this behaviour when using an enlarged list of 14 interface positions (data not shown).

Table 2. Average ISUMs for the 8 most important residues (in bold) of the Homeobox binding interface. The values in the matrix are average scores that evaluate the effect of mutating interface residues, measured in terms of DNA motif similarity score. Substitution scores take values in the range [0 , 1], A score close to 0 means that a substitution does not contribute to increase the correlation between interface similarity and DNA motif similarity across pairs of proteins in the training set.

2	A	E	K	R
A	0.9	0	0	0
E	0	0.8	0	0
K	0	0	1	1
R	0	0	1	1

3	A	H	K	R
A	1	0	0	0
H	0	1	0.9	1
K	0	0.9	0	0
R	0	1	0	1

5	Q	R	S	T
Q	0.5	0.9	0.15	0.2
R	0.9	1	0.05	0.1
S	0.15	0.05	0.5	0.85
T	0.2	0.1	0.85	0.5

47	I	N	T	V
I	1	0	1	1
N	0	0.1	0	0
T	1	0	1	1
V	1	0	1	1

50	A	I	K	Q
A	1	0	0	0
I	0	0.85	0	0
K	0	0	0.2	0
Q	0	0	0	1

51	L	N
L	0.5	0.1
N	0.1	0.9

54	A	M	R	T
A	1	1	0	1
M	1	1	0	1
R	0	0	0	0
T	1	1	0	1

55	K	Q	R
K	1	0.1	0
Q	0.1	0.5	0.1
R	0	0.1	0

Table 3. Average ISUMs for two random non-interface positions (in bold) of the Homeobox binding interface. Note that these matrices contain more than 4 residues, as different cross-validation rounds often find a different set of frequent residues for these positions. Substitution scores take values in the range [0 , 1].

36	A	H	K	M	N	Q	S
A	0.1	0.4	0	0.1	0	0	0.1
H	0.4	0.7	0	0.1	0.1	0.1	0.1
K	0	0	0	0	0	0	0
M	0.1	0.1	0	0.2	0	0	0
N	0	0.1	0	0	0	0	0
Q	0	0.1	0	0	0	0	0
S	0.1	0.1	0	0	0	0	0.3

66	A	G	K	P	Q	S
A	0.7	0	0	0.6	0.2	0.3
G	0	0	0	0	0	0
K	0	0	0	0	0	0
P	0.6	0	0	0.1	0	0.3
Q	0.2	0	0	0	0	0
S	0.3	0	0	0.3	0	0.3

3.4 Evaluation of ISUMs

In each cross-validation round the derived ISUMs were evaluated on the collection of pairwise alignments of the remaining 17 protein sequences (the evaluation subset), annotating the 8 critical interface residues defined earlier. The correlations between interface similarity scores and the corresponding DNA motif similarities were calculated using the generated ISUMs and compared to those obtained using BLOSUM62 (for the comparison of interface residues) and the

normalized E-value (for the complete protein sequences). Results are shown in Table 4.

Table 4. Pearson correlation coefficients between Homeobox DNA motif and interface similarity using different scoring schemes in ten cross-validation rounds.

round	ISUMs	BLOSUM62	normalized E-value
1	0.83	0.78	0.71
2	0.73*	0.59	0.45
3	0.71*	0.58	0.55
4	0.78	0.72	0.65
5	0.45	0.45	0.40
6	0.77	0.76	0.57
7	0.80*	0.71	0.64
8	0.66	0.74	0.42
9	0.82*	0.69	0.49
10	0.86	0.79	0.74

In 4 of the 10 repetitions, the correlation obtained using ISUMs matrices showed to be at least 10% better than any other scoring, and in all but one repetitions it was at least 10% better than the normalized BLAST E-value. In addition, we measured the predictive power of interface comparison by means of a Receiver Operating Characteristic (ROC), plotted in Figure 3. The ROC curve shows a significant improvement in the sensitivity and specificity when using ISUMs matrices in the range of specificity [0.4 , 1] in comparison with the other measures, in particular when compared with BLAST expectation values.

3.5 Limitations of BLASTP alignments when predicting Homeobox DNA motifs

Table 5 highlights 10 homeodomain pairs which display high DNA motif and interface similarities but low overall protein sequence similarity, and hence small normalized E-values. These alignments illustrate that often protein domain alignments, such as local alignments produced by BLASTP, might fail to explain the binding of similar DNA motifs. What is the frequency of these events? Among all the unique validation alignments (1138) there were 155 (13%) where normalized E-value was less than 0.2 and DNA motif similarity more than 0.6 . Out of these 155, 124 had an ISUM score higher than 0.6 and therefore demonstrate that there are a substantial amount of cases where DNA motifs can only be properly predicted taking into account interface similarity. Moreover, all of these 155 alignments have overall ISUMs scores higher than 0.2, so ISUMs clearly have a lower false negative rate than E-values.

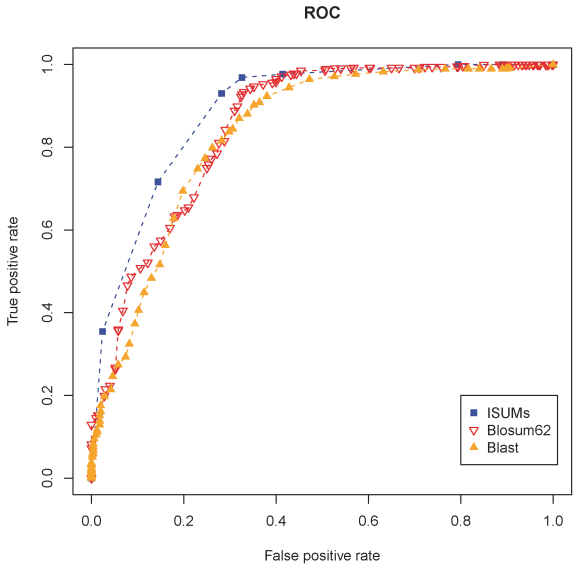

Fig. 3. ROC curve of Homeobox DNA motif predictions with ISUMs, BLOSUM62 and normalized BLAST E-values.

3.6 Limitations of sequence-based DNA motif alignments

The set of pairwise alignments of DNA motifs, generated during the validation of ISUMs, were compared to the set of structure-based interface alignments of homeodomains (section 3.1, Figure 2). It was observed that sequence-based dynamic programming, as carried out by the STAMP software, occasionally yielded incorrect alignments, which failed to represent the underlying common interface architecture. These alignment errors occurred mainly when aligning DNA motifs of subtypes TAATnn and TGATAC. As can be seen in Supplementary Material, these difficult alignments were associated to an insignificant average STAMP E-value of 0.90 .

4 Discussion and conclusions

The results presented in this work support the hypothesis that the residues directly involved in DNA recognition can explain and capture binding preferences better than the complete protein sequence as aligned by BLASTP.

Work is under way with other families and so far the results obtained suggest that this holds true for other families, including bZIP and zinc finger transcription factors.

This work presents the performance of simple, binary ISUMs; it remains to be tested whether richer matrices, which would take longer to compute, can

Table 5. Example of Homeobox pairwise alignments with low overall protein similarity and high DNA motif and interface similarities. Motif similarity, ISUM and BLOSUM62 scores are normalized by dividing by motif length.

pair	DNA Motifs	Motif similarity	Interface	norm.E-value	ISUM	BLOSUM62
Lim1 / Slou	kyTaATTr/yaATTAam	0.91	RGRVQNSK/RRRIQNTK	0.1	0.75	0.71
Ro / CG4136	gyTAATTA/yAATTArs	0.85	RRRIQNAK/RHRVQNAK	0	1	0.78
Bsh / CG32105	gyymATTA/yTAATTAaw	0.83	RKRTQNMK/KRRVQNAK	0.1	0.88	0.51
H2.0 / Lim3	vkwTwATwAA/vyTAATTA	0.77	SWRVQNMK/KRRVQNAK	0.05	0.75	0.56
Ap / CG15696	TmATTArs/btTAATTr	0.75	KRRVQNAK/RLRIQNAR	0.06	0.75	0.65

be created and whether different weights can be assigned to different interface positions to improve the observed DNA motif - interface correlation.

This protocol could in principle be used with different proteins families to establish a set of family-specific ISUMs that would help in the prediction of DNA motifs for *orphan* transcription factors. However, the kind of required data, such as the data produced by Noyes and collaborators, is unfortunately not available for most families. Nevertheless, our results provide quantitative evidence supporting the use of standard substitution matrices for evaluation of interface conservation, as previously suggested by other authors [16, 17].

While our results support the general use of interface knowledge when evaluating sequence alignments of transcription factors, they also indicate that annotating interfaces can be particularly important in cases where full domain alignments yield poor scores, as in these cases highly similar interfaces can be masked by overall low similarity alignments.

It is important to recall that our results show that sequence-based alignment methods might fail to produce the correct DNA motif alignment between members of the same family, provided they are sufficiently divergent. This observation justifies the use of structural data for the comparison of transcription factors, whenever available, as done in this paper.

As a result of this work we have now added similarity scores and interface matrices to our weekly updated database 3D-footprint, which will make it easier to annotate and correctly align interfaces in different protein families.

Acknowledgements.

This work was funded by Programa Euroinvestigación 2008 [EUI2008-03612].

References

1. Contreras-Moreira, B.: 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucleic Acids Res. 38(Database issue), D91–97 (2010)
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E.: The Protein Data Bank. Nucleic Acids Res. 28(1), 235–42 (2000)

3. Murzin, A.G., Brenner,S.E., Hubbard,T., Chothia,C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540 (1995)
4. Lewis, E.B.: A gene complex controlling segmentation in Drosophila. Nature 276, 565–570 (1978)
5. Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., Wolfe, S.A.: Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133, 1277–1289 (2008)
6. Hertz, G.Z., Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563–577 (1999)
7. Bailey, T.L., Williams, N., Misleh, C., Li, W.W.: MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 34, W369–373 (2006)
8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990)
9. Mahony, S., Benos, P.V.: STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 35, W253–258 (2007)
10. Contreras-Moreira,B., Sancho,J., Espinosa Angarica,V. : Comparison of DNA binding across protein superfamilies. Proteins, 78(1):52–62 (2009)
11. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004)
12. Henikoff, S., Henikoff, J.G.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. 89, 10915–10919 (1992)
13. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009)
14. Wingender, E., Dietze, P., Karas, H., Knuppel, R.: TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 24, 238–241 (1996)
15. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Peña-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T., Khalid,F., Zhang,W., Newburger,D., Jaeger,S.A., Morris,Q.D., Bulyk,M.L., Hughes,T.R.: Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell 133(7), 1266–1276 (2008)
16. Luscombe,N.M., Thornton,J.M.: Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J. Mol. Biol. 320(5): 991–1009 (2002)
17. Morozov, A.V., Siggia, E.D.: Connecting protein structure with predictions of regulatory sites. Proc. Natl. Acad. Sci. 104, 7068–7073 (2007)