Structural (and sequence-based) analysis of transcriptional regulation

Bruno Contreras-Moreira,[1] Irma Lozada Chávez,[2] and Vladimir Espinosa Angarica[3]

[1] *Estación Experimental de Aula Dei, CSIC, Zaragoza, España*
[2] *Centro de Ciencias Genómicas, UNAM, México*
[3] *Depto de Bioquímica y Biología Molecular y Celular, Universidad de Zaragoza, España*

Most computational approaches to transcriptional regulation use sequence-based methodologies, that aim to discover regulatory motifs in genomic segments. Here we argue that the current content of the Protein Data Bank (PDB) can provide invaluable data that drive the prediction of regulatory interactions within genomes. First, we dissect protein-DNA interfaces and find atomic interactions that contribute to sequence-specific recognition, mainly hydrogen bonds and Van der Waals contacts. These specificity determinants can be expressed in terms of atomic weight matrices, that are shown to be robust in bootstrap experiments and yield scores that correlate with approximate measures of binding specificity. Second, using example transcription factors from *Escherichia coli* we find that some protein-DNA interfaces have sequence-dependent DNA geometries that constitute indirect readout mechanisms, in agreement with previous reports. Third, we are able to build structure-based position weight matrices that capture both types of recognition mechanisms and test them in genomic experiments, with results comparable to sequence-based methodologies. We conclude that the PDB can be further exploited in exploring transcriptional regulation and other biological processes mediated by protein-DNA interactions.

MOTIVATION:

Specific interactions between protein and DNA molecules lie at the core of fundamental cellular processes such as transcriptional regulation. Although many of these interactions have been experimentally described at atomic scale, most predictive computational methods employed when exploring transcriptional regulation use sequence-based methods. This work aims to prove that the repertoire of protein-DNA complexes annotated at the Protein PDB can be effectively used in characterizing regulatory networks.

BACKGROUND:

The task of finding associations between transcription factors and their targets has been so far addressed by building statistical models that encode DNA binding preferences, that can then be used to scan genomes with the aim of identifying putative binding sites. These models, position weight matrices (PWMs) obtained after aligning previously known binding sites, reduce the recognition process to the sequence level, ignoring the tri-dimensional contacts at the protein-DNA interface.

The alternative approach proposed here considers the structural mode of binding of transcription factors in order to search for sequences that are preferred for particular geometries, taking advantage of seminal papers[1,2] that describe the direct and indirect readout components and reported ways to compute them.

Direct readout is associated to contacts established among amino acid residues and nitrogenous base pairs at the interface, separating Van der Waals contacts, which generally do not confer sequence specificity, with the exception of the hydrophobic interactions involving the C7 of thymine, and hydrogen bonds. The different types of atomic interactions at this interface have been thoroughly studied [3]. However, the scarcity of PDB complexes limited these studies to the level of residues. The increasing number of high-resolution protein-DNA complexes available, 279 50%

non-redundant PDBs, can now be used to re-approach this question at the atomic level, paving the way for more detailed studies.

Indirect readout does not involve direct atomic contacts, it rather is the result of DNA shape and deformability being dependent of the sequence of nucleotides.

In contrast with attempts to apply atomic force fields [4,5], this work proposes a fairly simple method that exploits protein-DNA complexes at the atomic scale and combines interaction preferences with empirical estimations of the DNA deformation energy in order to build structure-based PWMs that can be taken for the prediction of transcription factor binding sites in genomic sequences.

Water mediated bridges are ignored int his paper, as they have been found to be mainly related to gap-filling functions at interfaces, with very little overall contributions to binding specificity.

For this task we used the 3DNA software, the harmonic function and the set of experimentally-derived parameters described by Olson. We then run an in silico ”saturation mutagenesis” analysis of the native DNA chain replacing at the structural level the nucleotide found in a given position by the other three possible bases, preserving the sequence of the other positions in the binding site. Each mutated sequence generated by this procedure was threaded into the initial structure in the same way described above and the direct and indirect readout scores were computed for the complete sequence.

METHODS AND RESULTS:

Validation of protein-DNA atomic interaction matrices

The catalogue of hydrogen bonds and Van der Waals contacts observed in a non-redundant selected set of 279 crystallographic protein-DNA complexes culled from the PDB was used to calculate atomic weight matrices. Close inspection of these matrices unveils some previously reported trends, such as the preference of arginine (by means of its NH1 and NH2 groups) to be bonding with guanine (with acceptor atoms N7 and O6). With respect to Van der Waals interactions, the main discriminatory group was found to be C7 methyl group from thymine. However, in order to validate the information captured in these matrices a more rigorous exercise was performed, by doing 1000 bootstrap replicates in which a random third of the interface collection was scored using weight matrices derived from the remaining two thirds. This experiment (Figure 1) quantifies the overtraining of our matrices and their predictive value, as well as any biases that could arise from the PDB composition.

Estimating binding specificity from interface scores

Atomic interaction matrices can be used to score protein-nucleic acid interfaces, by simply adding the weights associated to the observed atomic interactions. We performed this calculation with the structures of 11 *E.coli* transcription factors (TrpR, Rob, PurR, PhoB, NarL, MetJ, MarA, FadR, DnaA, CRP and LacR) after threading the set of annotated binding sites in RegulonDB, and made a scatter-plot of the resulting score range against the number of sites scored. Figure 2 shows that relatively unspecific transcription factors, such as CRP, yield greater score ranges, suggesting that atomic matrices provide biologically meaningful scores.

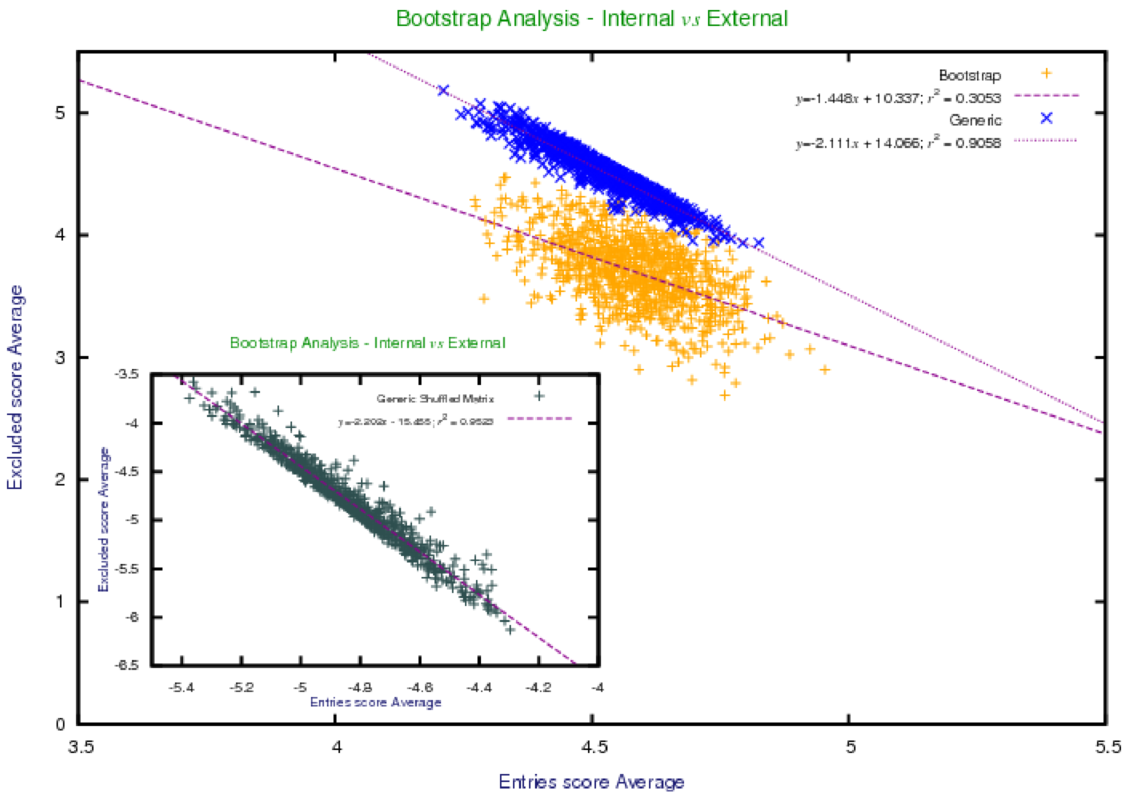


FIG. 1: Scatter plot and regression analysis of the distribution of the mean score values for the bootstrap and excluded sets, the generic and the shuffled matrices. The discrimination power of generic and bootstrap matrices is very similar, and significantly different from that of reshuffled matrices, in the sub-chart, that display no discrimination competency.

Approximating DNA-deformation

The indirect readout contribution was estimated by approximating the deformation costs of threading a nucleotide sequence into a DNA backbone with fixed geometric step parameters (step, shift, slide, rise, tilt, roll, twist) taken from the native complex , using 3DNA software, the harmonic function and the set of experimentally-derived parameters described by Olson[2].

By iteratively mutating all positions in the original DNA template we compute the cost associated to every mutation, unveiling positions along the molecule in which not all nitrogenous bases are equally compatible. Positions 4,5,6 in the CRP logo of Figure 4 correspond to indirect readout nucleotides, with no direct contacts.

Benchmark of structure-based PWMs with genomic-sized sequences

At this point it is possible to build structure-based PWMs by mutating to saturation the template DNA sequence, calculating both the deformation costs and the sum of interface interaction weights associated to every mutation. The probability associated to nucleotides in each position is calculated as $p(n) = e^{(-((1-D)*direct(n)+D*indirect(n)))}$, where D is a linear weight given to the deformation-based indirect readout.

We can then test the potentiality of these PWMs to discriminate cognate sequences of transcription factors, defined as known binding sites with experimental evidence in RegulonDB, from

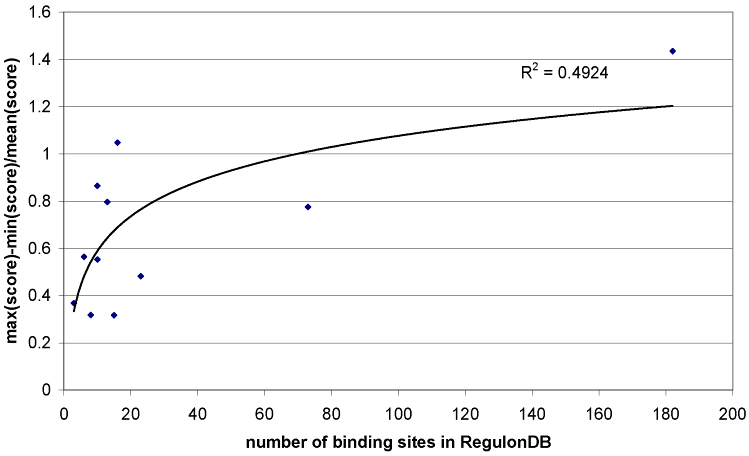


FIG. 2: Scatter plot of score variability and number of regulatory targets, with a correlation coefficient when log-transformed of 0.70 (pairs=11, R2=0.49, p=0.015)

large sets of same %GC random sequences where few potential binding sites should be found.

This was accomplished by calculating Receiver Operating Characteristic (ROC) curves for the eleven *E.coli* transcription factors mentioned earlier, and the area under those curves, that measures the discrimination ability in a single number. Figure 3 shows the ROC curves for CRP, calculated using different D indirect readout contributions.

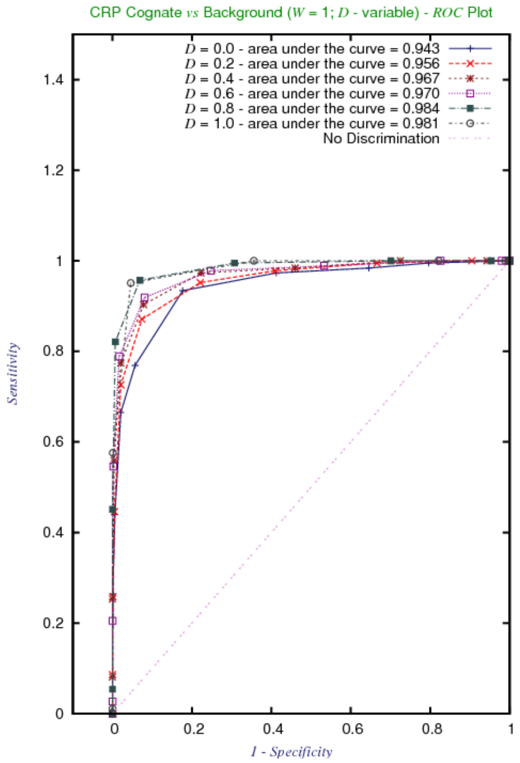


FIG. 3: ROC curve of the structure-based PWM for CRP (pdbcode 1CGP) after scanning 10^6 random sequences. This complex is an example in which the indirect readout component is actually more important for specificity than the interface contacts

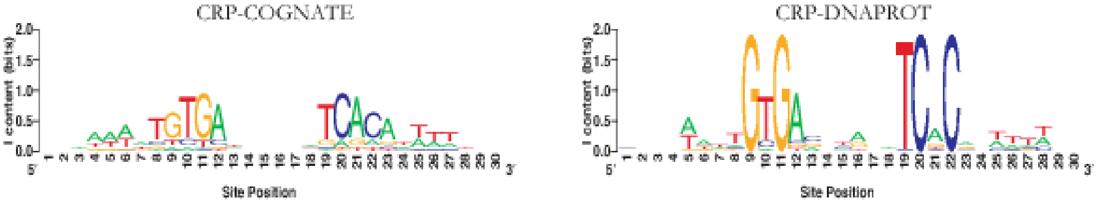
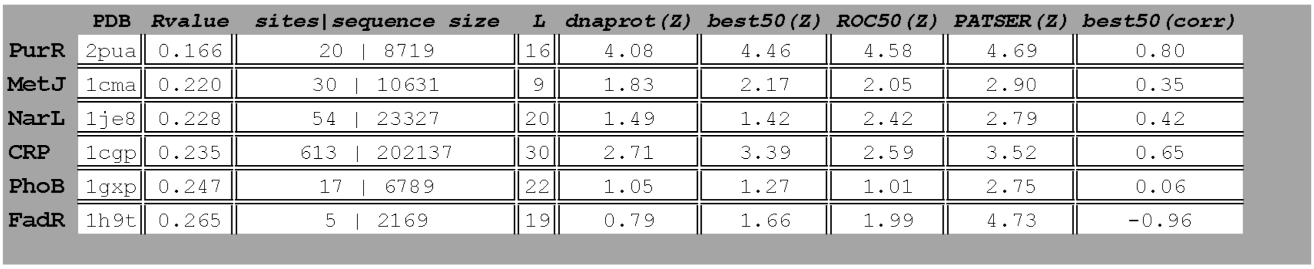


FIG. 4: Compariosn of sequence logos compiled from cognate sites (right) and derived from the best 50 scoring sites as evaluated by the structure-based PWM

To further evaluate our method we ran a last round of genomic sequences scans in parallel with CONSENSUS/PATSER[Hertz], a widely used sequence-based method that requires prior knowledge of binding sites, obtaining results that are comparable for 4 out 6 transcriptions factors (CRP,NarL,PurR and MetJ). If these 6 TFs are ranked in terms of the R-factor of their structural models, it turns out that PhoB and FadR, those for which we obtained worst results, are of lower quality. This last round of experiments highlights the importance of structural data, as detailed in Figure 5.



	PDB	Rvalue	sites|sequence size	L	dnaprot(Z)	best50(Z)	ROC50(Z)	PATSER(Z)	best50(corr)
PurR	2pua	0.166	20 | 8719	16	4.08	4.46	4.58	4.69	0.80
MetJ	1cma	0.220	30 | 10631	9	1.83	2.17	2.05	2.90	0.35
NarL	1je8	0.228	54 | 23327	20	1.49	1.42	2.42	2.79	0.42
CRP	1cgp	0.235	613 | 202137	30	2.71	3.39	2.59	3.52	0.65
PhoB	1gxp	0.247	17 | 6789	22	1.05	1.27	1.01	2.75	0.06
FadR	1h9t	0.265	5 | 2169	19	0.79	1.66	1.99	4.73	-0.96

FIG. 5: Performance of our method DNAPROT compared to a standard sequence-based method (PATSER), measured in terms of median Z-scores of true positive sites. Note the correlation coefficients between the scores of two completely different methods for evaluating sites.

CONCLUSIONS:

We find here that starting from a single good quality structural model of a given TF-DNA complex, predictions can be made in genomic sequences by using matrices of atomic interaction propensities and the empirical indirect readout information extracted from the structural model itself, providing a novel way for predicting promoter operator sites for transcription factors with few or no experimentally characterized binding sites.

REFERENCES

1. Michael Gromiha, M., Siebers, J.G., Selvaraj, S., Kono, H. & Sarai, A. (2004). Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. J Mol Biol 337, 285-94.

2. Olson,W.K., Gorin,A.A., Lu,X.J., Hock, L.M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A 95, 11163-8.

3. Luscombe, N.M., Laskowski, R.A. & Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29, 2860-74.

4. Paillard, G. & Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. Structure 12, 113-22.

5. Havranek, J.J., Duarte, C.M. & Baker, D. (2004). A simple physical model for the prediction and design of protein-DNA interactions. J Mol Biol 344, 59-70.