BIOINFORMATICS

Vol. 22 no. 14 2006, pages e74–e80
doi:10.1093/bioinformatics/btl215

Comparative footprinting of DNA-binding proteins

Bruno Contreras-Moreira[1,*] and Julio Collado-Vides[1]

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av.Universidad, s/n, 62210 Cuernavaca, Morelos, México

ABSTRACT

Motivation: Comparative modelling is a computational method used to tackle a variety of problems in molecular biology and biotechnology. Traditionally it has been applied to model the structure of proteins on their own or bound to small ligands, although more recently it has also been used to model protein-protein interfaces. This work is the first to systematically analyze whether comparative models of protein-DNA complexes could be built and be useful for predicting DNA binding sites.

Results: First, we describe the structural and evolutionary conservation of protein-DNA interfaces, and the limits they impose on modelling accuracy. Second, we find that side-chains from contacting residues can be reasonably modeled and therefore used to identify contacting nucleotides. Third, the DNASITE protocol is implemented and different parameters are benchmarked on a set of 85 regulators from *Escherichia coli*. Results show that comparative footprinting can make useful predictions based solely on structural data, depending primarily on the interface identity with respect to the template used.

Availability: DNASITE code available on request from the authors

Contact: contrera@ccg.unam.mx

Supplementary information: http://www.ccg.unam.mx/Computational_Genomics/supplementary/ismb2006

1 INTRODUCTION

Comparative modelling is now a mature technology that predicts the three-dimensional arrangement of a protein sequence given an alignment to one or more template proteins of known structure. The use of protein models may range from site-directed mutagenesis and molecular replacement to molecular docking and protein design and engineering (Baker and Sali, 2001; Contreras-Moreira *et al*., 2002). The actual use of a protein model will depend on its expected accuracy, dictated primarily by the sequence similarity to the templates used (Contreras-Moreira *et al*., 2005; Chothia and Lesk, 1986). Together with sequence alignment errors, this is a main factor affecting model quality (Tramontano *et al*., 2001). This factor has also been found to be critical when reconstructing protein-protein interfaces (Aloy *et al*., 2003); the more similar the sequences, the more predictable the details of the interface.

*To whom correspondence should be addressed.

In this paper we ask these questions to a different system, the interface between proteins and nucleic acids. There has been great interest in understanding these interactions, given the biological relevance of genetic regulation (Sarai and Kono, 2005). For this reason a good amount of experimental work has been dedicated to this problem, most of it now part of the Protein Data Bank (PDB) (Berman *et al*., 2000). This work takes all this experimental data, i.e. crystallographic and NMR structures, in order to:

(1) determine if there are any evolutionary trends which might explain the divergence of protein-nucleic acid interfaces and therefore support comparative modelling of these complexes

(2) assess if footprinting predictions can be made by comparative modelling of protein-DNA complexes

The motivation for this analysis stems from a variety of approaches recently tested on experimentally determined complexes, that isolate and characterize the preferred recognised sequences of transcription factors by using physical (Aloy *et al*., 1998; Gromiha *et al*., 2005; Kono and Sarai 1999; Luscombe *et al*., 2001; Morozov *et al*., 2005; Nadassy *et al*., 1999; Pabo and Nekludova 2000; Paillard and Lavery 2004; Selvaraj *et al*., 2002; Siggers *et al*., 2005; Steffen *et al*., 2002) and evolutionary metrics (Kaplan *et al*., 2005; Raviscioni *et al*., 2005). Here we demonstrate that comparative modelling can help explain or predict the repertoire of known binding sites of a given regulator, annotated in resources such as RegulonDB (Salgado *et al*., 2006), for proteins for which no structural description is available, provided that we know the structure of homologous proteins.

This work presents the first systematic benchmark of comparative modelling protein-DNA complexes with the aim of predicting DNA operator sites. First we compile a non-redundant set of protein-DNA complexes to assess the conservation of their interfaces. The results show that comparative modelling of these complexes is possible with one restriction: as sequence similarity diminishes protein-DNA interfaces diverge exponentially. Second we implement a protocol that we call DNASITE that builds comparative models of protein-DNA interfaces using tools and datasets widely used by the structural bioinformatics community. Finally we choose the appropriate parameters and test the performance of DNASITE on a set of 85 *Escherichia coli* regulator proteins for which RegulonDB contains known binding-sites with experimental evidence.

2 METHODS

Collecting protein-DNA complexes

We retrieved all PDB entries (as of August 9, 2005) containing both protein and DNA coordinates, and selected all protein chains less than 12Å away from any DNA segment. This list of chains was pruned using a 95% sequence identity cut-off to get a non-redundant set, using the web server PISCES (Wang and Dunbrack, 2003). We then put every selected chain together with the contacting nucleic acid molecules and called that a PN complex, where P stands for protein and N for nucleic acid. The resulting library contained 273 crystallographic and NMR structures and is available as supplementary material.

Comparing complexes by means of protein structural alignments

The next step of our procedure was to compare the protein chains of all complexes using structural alignments, as a way of minimizing possible alignment errors. For this we used the program MAMMOTH (Ortiz *et al*., 2002) and considered only pairs of complexes that yielded –ln(E) values over 4.5 and had at least 10% of sequence identity, to eliminate non statistically significant matches. From more than 37000 comparisons, 442 passed this filter and were used to plot the conservation of protein-nucleic acid interfaces as sequence similarity changed. Each of these pairs resulted in a structural superposition with an associated sequence alignment. Eight folds from the Structural Classification of Proteins (SCOP) (Murzin *et al*., 1995) dominate this dataset, as shown in Results.

Calculating interface agreement between superposed complexes

For each complex pair (A,B) we calculated three numbers: the sequence identity (ID_{ab}) between protein chains P_a and P_b; the structural agreement of the amino acid residues participating in the interface (P-$RMSD_{ab}$); and the structural agreement of the interface nucleotides (N-$RMSD_{ab}$). Calculating ID_{ab} is simple, matches in the sequence alignment divided by the total number of aligned residues. The other two numbers are calculated from the structural superposition of PN_a over PN_b in six steps:

(1) P_a residues contacting N_a nucleotides are put in set P_{ac}.

(2) P_b residues aligned to those in P_{ac} are put in P_{bc}.

(3) Residues in P_{ac} and P_{bc} are taken in pairs to calculate their root-median-square deviation. We call this number P-$RMSD_{ab}$.

(4) For each residue in P_{ac}: closest nucleotide in N_a is put in set N_{ac}.

(5) For each residue in P_{bc}: closest nucleotide in N_b is put in set N_{bc}.

(6) Nucleotides in N_{ac} and N_{bc} are taken in pairs to calculate their root-median-square deviation. We call this number N-$RMSD_{ab}$.

Protein residues were represented by their C_α atoms, while for nucleotide bases we took N9 (purines) and N1 (pyrimidines) atoms. For step 1, a protein-nucleic acid contact is defined as a pair of atoms placed less than 12Å away from each other, following the work of Aloy *et al.* (Aloy *et al*., 1998). For step 2 we require aligned protein residues to be within 4Å from each other after superposition.

Calculating side-chain modelling accuracy

1477 H-bonding residues from our library of superposed complexes were modelled with the program SCWRL2.7 (Dunbrack and Karplus, 1993) and RMSD values were calculated for each model-experimental pair of side-chains. For each pair(A,B), first A was used as template to predict B side-chains and then B was chosen as template.

Table 1. Protein-DNA recognition matrix compiled by the authors (CM parameter set) from a set of 273 95% non-redundant complexes. Contacts were identified using a distance threshold of 4Å (from any side-chain atom to any atom in the purine/pyrimidine ring). Each value is a log-odd calculated as in (Mandel-Gutfreund, *et al*., 2001)

	C	G	A	T
D	+0.26	−0.49	−1.79	−1.11
P	−1.31	−1.81	−0.73	−0.15
I	−1.06	−1.64	−0.53	−0.99
K	−0.54	+1.05	−0.75	+0.35
W	+0.44	+0.34	−0.47	+0.07
C	−0.74	−1.83	−0.85	−0.36
G	−2.57	−2.57	−2.57	−2.57
F	−0.76	+0.01	+0.06	+0.30
Q	+0.21	+0.49	+0.63	+0.25
S	−0.40	+0.42	−0.50	+0.62
N	+0.41	+0.46	+0.98	+0.65
L	−1.76	−1.29	−1.03	−0.65
V	−0.97	−2.57	−0.43	−0.06
E	+0.53	−1.65	−1.62	−1.09
Y	+0.55	+0.60	+0.36	+0.88
R	+0.76	+1.96	+0.56	+1.09
T	+0.26	−0.35	−0.41	+0.44
M	−0.40	+0.31	+0.10	+0.39
A	−1.10	−1.31	−1.21	−0.27
H	−0.39	+1.01	−0.49	+0.54

Implementation of DNASITE

The DNASITE protocol was programmed in Perl and C and is conceptually very simple. The input is a protein sequence and these are the steps that follow:

(1) Search for homologous protein-DNA complexes with three iterations of PSI-BLAST (Altschul *et al*., 1997), using a sequence library made of the proteins in our non-redundant set of complexes plus the sequences in SWISSPROT (Sep, 2005) (Bairoch and Apweiler, 2000).

(2) Use local PSI-BLAST alignments to build the protein backbone of the modelled complex, using the template's coordinates. Accept only models that align residues known to be contacting nucleotides in the template.

(3) Add SCWRL side-chains keeping the template DNA in frame. We can choose to model only mutated side-chains.

(4) Identify binding residues as those less than 4.5Å away from any atom in the purine/pyrimidine ring, a similar distance to that used previously by Mandel-Gutfreund (Mandel-Gutfreund and Margalit, 1998). These residues are used to calculate the % interface identity (IID).

(5) Thread DNA sequences into the modelled complex and evaluate the matching using logarithmical protein-DNA 20x4 recognition matrices, such as those derived by Mandel-Gutfreund (Mandel-Gutfreund *et al*., 2001). The scoring function (Equation 1) is additive, assuming that each residue in the interface contributes equally to the matching score. A family-specific correction might be applied, calculating a correction term derived from the background substitution frequencies contained in the PSI-BLAST position-specific scoring matrices (PSSM) and the protein-DNA matrix used, as described in Equation 2. The idea is that amino acid substitutions might be indicating which nucleotide bases are preferred at each position, somehow capturing context-dependent preferences. DNA deformation for each

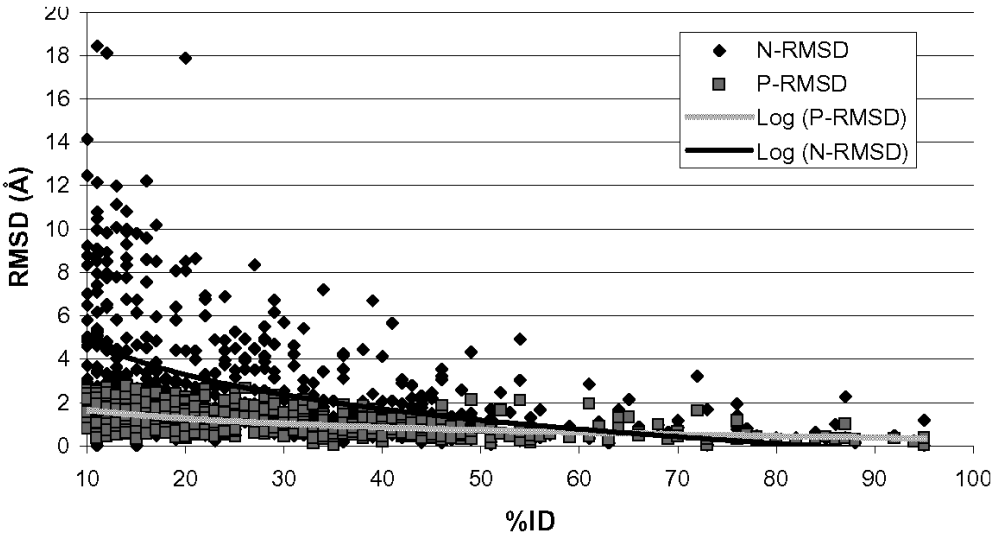

Fig. 1. Interface conservation in terms of P-RMSD and N-RMSD. 442 pairs of protein-nucleic acid complexes were superposed and the conservation of their interfaces plotted against their protein sequence identity. Two measures are reported: P-RMSD, the median deviation of the protein residues taking part in the interface; N-RMSD, the median deviation of the nucleotides of the interface. Logarithmical regression lines are added to assist in the interpretation.

threaded sequence is approximately estimated using the X3DNA package (Lu and Olson, 2003), in order to consider also indirect readout mechanisms (Gromiha *et al*., 2005). Briefly, DNA parameters (step, shift, slide, rise, tilt, roll, twist) are calculated from the template DNA molecule and then used to approximate deformation energies based on sequence-dependent parameters (Olson *et al*., 1998) (Marc Parisien, personal communication). The native DNA molecule is used as a reference and an arbitrary cut-off is set to skip sequences with large deformation energies. To ensure fast computation times, shortcuts are applied when the number of possible DNA sequences is greater than 4^9. Only the top fraction of sequences is selected to build a footprinting matrix. If the number of selected sequences is less that 50 the DNA sequence of the template complex is added.

Given a PN complex, with L interface nucleotides contacting C protein residues and a scoring matrix, the scoring function is calculated as follows:

$$Score(PN) = \sum_{i=1}^{L}\sum_{j=1}^{C} match\ (P_i, N_j, matrix) \qquad (1)$$

To calculate the family correction for a given residue P_j in contact with nucleotide base N_i, each of the 20 possible aminoacid (aa) substitution frequencies in a PSSM are considered:

$$Corr(P_j, N_i) = \sum_{x=1}^{20} freq(aa(x))match(aa(x), N_i, matrix) \qquad (2)$$

DNASITE benchmark

The set of known and putative regulator proteins in *E.coli* was taken as a test set, including 3 SCOP folds. Each of those sequences was used as input for DNASITE and 85 comparative models were obtained (IHF was excluded from this test as it was considered to be non-sequence specific). Each of these 85 models was built using different parameters that will be referred to using these codes:

- Def: default parameters, using a 2001 Mandel-Gutfreund matrix, up to three contacts per residue and a DNA deformation cut-off of 1.6 kcal/mol.
- CM: uses a matrix built by the authors from the non-redundant set of complexes, based only on distance cut-offs (see Table 1).
- Sc3: uses SCWRL3.0 (Canutescu *et al*., 2003), instead of version 2.7, to compare the performance.
- Df1: uses a DNA deformation energy cut-off of 1 kcal/mol.
- Df2: uses a DNA deformation energy cut-off of 2 kcal/mol.
- Df3: uses a DNA deformation energy cut-off of 3 kcal/mol.
- C1: only one contact per residue is considered, the closest one.
- M: conservative, models only mutated side-chains, the rest are taken as in the template complex.
- F: uses family-specific correction.
- P: P-value cut-off for selecting threaded sequences.

The footprint matrices generated by DNASITE were aligned against the corresponding set of known binding sites extracted from RegulonDB (Jan, 2006) using the program PATSER (Hertz and Stormo, 1999). Each site is flanked by segments of 10 nucleotides. Alignments yielding significant scores, over the cut-off estimated by PATSER for each matrix, were considered as recovered sites and for those the average ln(P-value) was calculated. Finally, the aligned sites were used to build a sequence logo with WebLogo (Crooks *et al*., 2004).

3 RESULTS

3.1 Protein-DNA interface conservation

Figure 1 shows N-RMSD and P-RMSD values obtained from a total of 442 non-redundant complex superpositions plotted against %ID. Individual N-RMSD and P-RMSD data points are depicted and logarithmic regression lines are added to help interpretation. Note that interface nucleotides accumulate larger deviations when superposed than their contacting residues. Furthermore, both N-RMSD and P-RMSD are significantly correlated to %ID, with correlation coefficients of −0.43 and −0.52 respectively. Nucleotide median deviations for complexes with at least 30% of sequence identity tend to be close to 2Å, more precisely within the 1.4 ± 1.2Å interval.

As mentioned earlier, 8 SCOP folds are over-represented in our dataset, the most common being the DNA/RNA binding 3-helical

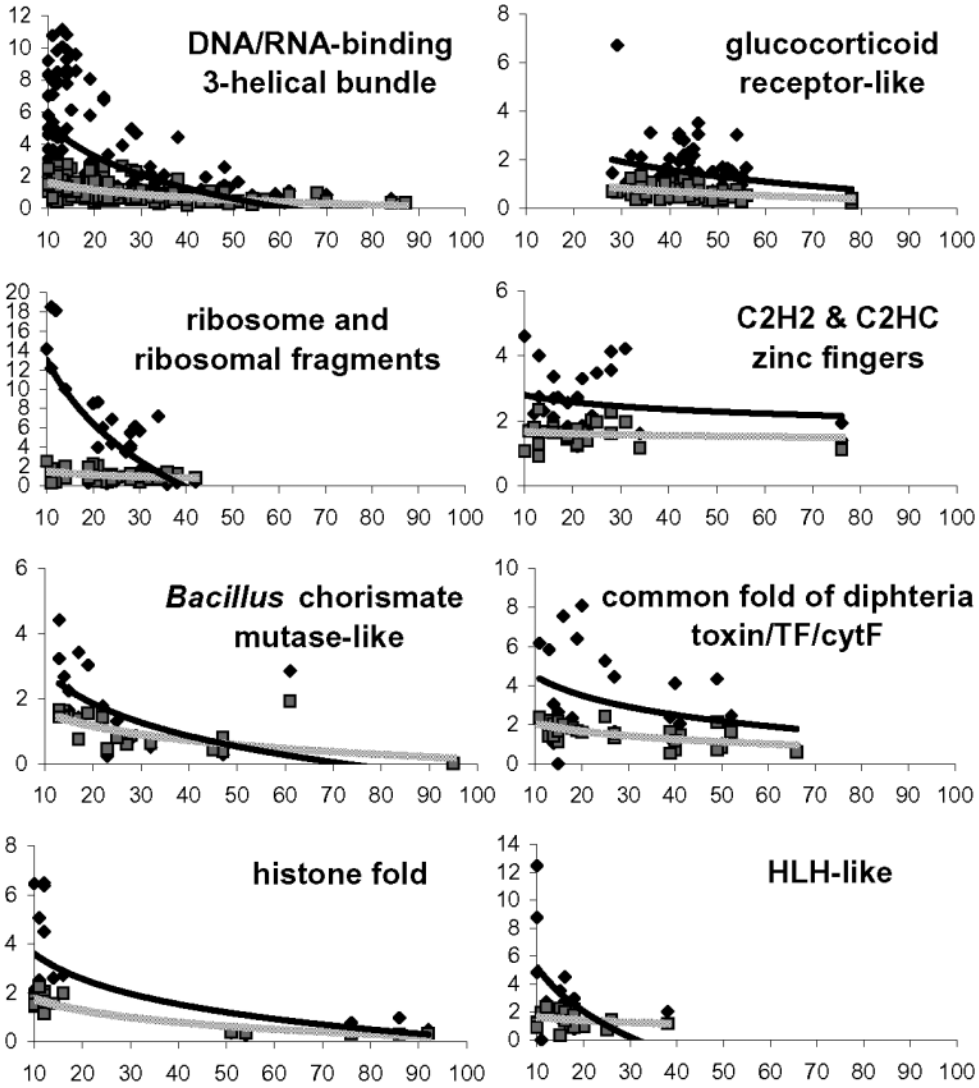

Fig. 2. Interface conservation for 8 representative SCOP folds. Same analysis as in Figure 1, splitting the data corresponding to the most abundant SCOP folds in our dataset. For all panels X-axis is %ID and Y-axis is RMSD measured in Å, with N-RMSD plotted in black and P-RMSD in grey. A majority of *E.coli* transcription factors contain helix-turn-helix motifs and can be classified as DNA/RNA-binding 3-helical bundle folds.

bundle. Figure 2 shows the same analysis performed on these most abundant SCOP folds, showing more specific trends, as also noticed by Siggers (Siggers *et al*., 2005).

These results are encouraging as they indicate that interfaces are structurally and evolutionary related and their sequence similarity is a reasonable estimator of the degree of conservation. However, before we can build comparative models of these complexes we need to previously identify which modelled amino acid residues are contacting DNA bases.

3.2 Side-chain modelling accuracy

In order to identify which residues are contacting nucleotides in a complex we first need to model the residue side-chains. As explained in Materials and Methods, we used the program SCWRL2.7 for this task and found that 77% of H-bonding modelled side chains deviate less than 2.0Å in average with respect to the experimental coordinates, excluding pairs of complexes with less than 30% sequence identity. We concluded that we can reasonably predict side-chain rotamers and therefore which residues are likely contacting nucleotides.

3.3 Footprinting of comparative protein-DNA complexes

Table 2 shows the performance of the DNASITE protocol using our test set of 85 *E.coli* regulators, comprising three folds: DNA/RNA-binding 3-helical bundles, lambda repressors and Met repressors. Three measurements are taken for each run: the percentage of recovered sites, the mean alignment score and the mean significance of alignment scores. This benchmark highlights some parameters settings, those that perform well in recovering RegulonDB sites with significant scores. Three of them were selected, P0.0001, MF and FP0.0001, and a few representative examples of footprinting predictions are shown in Figure 3. What do these parameters

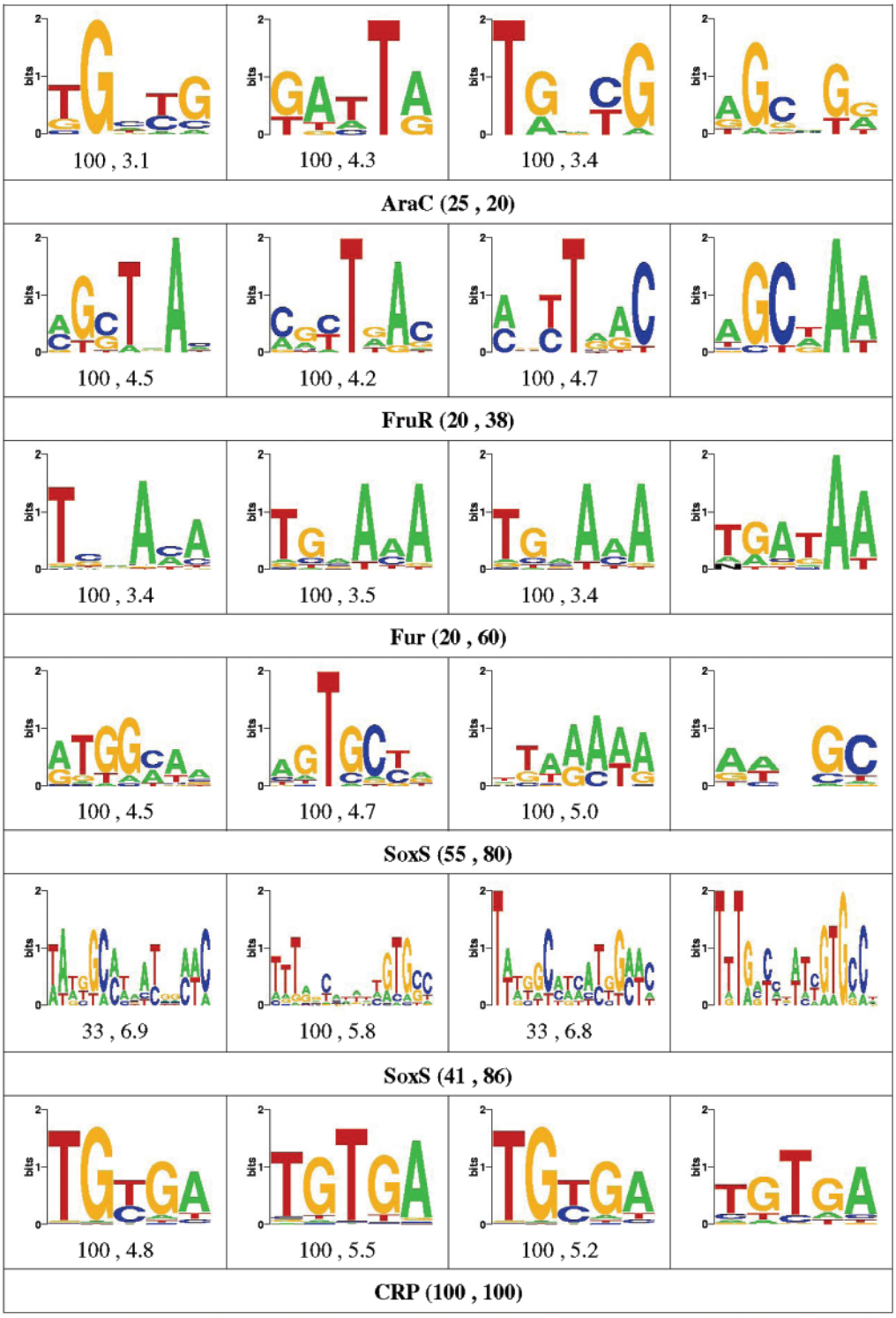

Fig. 3. Representative examples of footprint predictions using the DNASITE protocol. Binding site predictions based on comparative models for 5 *E.coli* regulators. Each row shows the results for a protein-DNA complex and the numbers in parenthesis indicate the corresponding %ID and %IID. The first three columns show the results for the P0.0001, MF and FP0.0001 parameter sets, including the % of recovered sites and the average alignment site score; the fourth shows the consensus matrix calculated by CONSENSUS/WCONSENSUS (Hertz and Stormo, 1999) on the RegulonDB sequences, as an independent control. Two independent predictions for SoxS are displayed here, using two different template complexes, one of them (55, 80) spanning only one of the DNA-contacting domains. The FP0.0001 (55, 80) prediction recovers 100% of sites, but includes false positives, as can be seen in the logo. Note that the MF (55, 80) correct prediction is also included into the (41, 86), whilst P0.0001 and FP0.0001 (41, 86) predictions do not recover all known binding sites and obtain incorrect sequence logos. SoxS is an example of split site, composed of two subsites. Our current benchmark methodology often cannot recover split sites.

Table 2. Performance of different DNASITE parameter sets tested on a total of 85 *E.coli* DNA-binding proteins with mean % sequence identity of 35 and % interface identity of 46. The first column labels each parameter set, encoded as mentioned in Materials and Methods. The second column shows the mean % of RegulonDB sites aligned with a significant score by PATSER. The third column shows the mean -ln(P) score for each DNA-binding protein, as reported by PATSER. The last column shows the mean significance of recovered sites, calculated as ln(P) – significance threshold

Parameter set	% Sites recovered	Mean –ln(P)	Mean significance
Def	94	4.7	1.5
CM	90	4.5	1.3
Sc3	94	4.6	1.7
Df1	95	4.7	1.9
Df2	94	4.6	1.5
Df3	94	4.6	1.4
C1	98	4.3	2.1
M	97	4.6	2.4
F	93	4.8	1.8
P0.01	93	4.5	1.6
P0.001	94	4.4	2.0
P0.0001	94	4.2	2.5
MF	96	4.6	2.5
FP0.001	93	4.5	2.2
FP0.0001	97	4.4	2.9

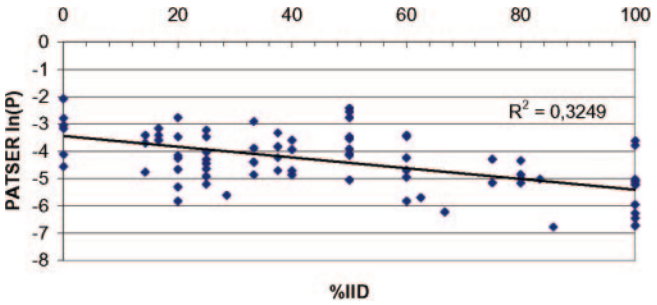

Fig. 4. Interface identity as quality predictor for DNASITE. FP0.0001 scores for 85 modelled complexes are plotted against % interface identity. The observed correlation coefficient is −0.57. This means that high IID values predict better DNASITE footprints.

mean? They suggest that keeping the conserved part of the interface from the template is a good idea (M), in agreement with previous observations (Sandelin and Wasserman, 2004), and that applying family-specific corrections helps in many cases (F). In addition, it seems to be a good choice to select only threaded sequences with low ln(P) values. The different solutions provided by each strategy might not be identical, but perhaps looking for consensus predictions may help discriminate between right and wrong predictions. 73 of these 85 predictions correspond to regulators that have more than 5 annotated binding sites in RegulonDB.

Figure 4 shows that the % interface identity (IID) correlates negatively with the obtained PATSER scores in our benchmark. The correlation coefficient ranges from −0.24 (C1) to −0.57 (FP0.0001). A linear regression line is also plotted, showing a poor R^2 value, due to the large variability of the data. A much weaker correlation is observed when % sequence identity is used instead (data not shown). This suggests that IID is really the important number when comparing different complexes, since mutations in the interface will probably mean changes in the recognised set of nucleotide sequences.

4 DISCUSSION

The assumption behind comparative modelling is that similar sequences will have very similar structures. However, similar protein structures need not have the same biological or molecular function. In our modelling problem two questions need to be answered. The first is whether a homologous protein really binds to DNA. The second is what nucleotide sequences are being recognised by this protein. We might try to answer the first question by calculating the net charge of the suspected binding protein, as suggested by Ahmad (Ahmad and Sarai, 2004), or using any related experimental evidence. However, in this work we focused on the second question.

The reported results suggest that template complexes can be used to estimate the nucleotide preferences of related proteins, as already anticipated (Morozov *et al*., 2005). These results also support the choice of FP0.0001 parameters if score significance is to be maximized. Another lesson learned here is that a conservative approach when predicting footprints is useful, keeping unchanged as much of the template complex as possible (M parameters). This could be saying that we are not very good at predicting preferred DNA sequences from scratch, perhaps because we have only tested generic recognition matrices (Pabo and Nekludova, 2000). Our results also suggest that family-specific DNA preferences can be estimated from protein sequence profiles, improving the observed alignment scores. This might help overcome the limitations of generic recognition matrices, as protein-DNA preferences might be context-specific (Kaplan *et al*., 2005). Besides family corrections, DNASITE could benefit from using tailor-made protein-DNA recognition matrices, were family-specific associations could be encoded. For instance, a homeodomain-like matrix could be derived. Preliminary work suggests that these matrices can significantly improve results but further exploration is needed.

This computational tool can generate different solutions that might be used to build a consensus. If no consensus is reached then probably the wise thing to do is to ignore these predictions. Along with the set of binding sequences selected, DNASITE also produces the motif length, a variable that non-structural footprinting methods need to estimate by other means.

DNASITE can be applied to regulators for which no experimental evidence is available at all, for instance cases where no footprint experiments have been performed. For this reason this tool can potentially be useful for the purpose of curating DNA-binding sites. Furthermore, the algorithm has been implemented using a collection of widely used tools (PSI-BLAST, SCWRL and X3DNA).

This approach makes a simplified use of interface geometry and does not explicitly distinguish H-bond interactions from Van der Waals contacts, allowing fast but perhaps less accurate predictions. Water-mediated H-bonds are also ignored as they don not seem to contribute much to specific protein-DNA recognition (Luscombe *et al*., 2001). Perhaps considering these questions would improve the method, but this remains to be tested.

A weakness of this method is that it depends on the availability of related protein-DNA complexes. For the set of approximately 300 regulators in *E.coli*, less than a third can be studied with this protocol. Probably more regulators could be modelled using more sophisticated protein alignment algorithms, but those cases would need to be benchmarked as well.

It should be remarked that a more realistic benchmark still needs to be done, using DNASITE footprints to blindly predict binding sites in the context of a genome. It is anticipated that these footprints may have relatively large false positive rates in comparison with more traditional approaches since they tend to be shorter, therefore allowing more random hits to be aligned. Therefore, future users should benefit by combining DNASITE with other structural and non-structural methods.

ACKNOWLEDGEMENTS

B.C.M. thanks Heladia Salgado for her support in using RegulonDB data, Martín Peralta for his help in constructing consensus matrices and Marc Parisien for his advice and code for calculating DNA deformation energies with X3DNA. We acknowledge suggestions by anonymous referees. This work has been supported by a post-doctoral fellowship from Universidad Nacional Autónoma de México awarded to B.C.M. and by NIH grant RO1-GM071962.

ABBREVIATIONS

PDF, Protein Data Bank; RMSD, root-median-square deviation; SCOP, structural classification of proteins; ID, sequence identity; IID, interface sequence identity; PSSM; position-specific scoring matrix.

REFERENCES

Ahmad,S. and Sarai,A. (2004) Moment-based prediction of DNA-binding proteins, *J Mol Biol*, 341, 65–71.

Aloy,P., Ceulemans,H., Stark,A. and Russell,R.B. (2003) The relationship between sequence and interaction divergence in proteins, *J Mol Biol*, 332, 989–998.

Aloy,P., Moont,G., Gabb,H.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1998) Modelling repressor proteins docking to DNA, *Proteins*, 33, 535–549.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 25, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TREMBL in 2000, *Nucleic Acids Res*, 28, 45–48.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics, *Science*, 294, 93–96.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank, *Nucleic Acids Res*, 28, 235–242.

Canutescu,A.A., Shelenkov,A.A. and Dunbrack,R.L.,Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction, *Protein Sci*, 12, 2001–2014.

Contreras-Moreira,B., Ezkurdia,I., Tress,M.L. and Valencia,A. (2005) Empirical limits for template-based protein structure prediction: the CASP5 example, *FEBS Lett*, 579, 1203–1207.

Contreras-Moreira,B., Fitzjohn,P.W. and Bates,P.A. (2002) Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era, *Appl Bioinformatics*, 1, 177–190.

Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator, *Genome Res*, 14, 1188–1190.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins, *Embo J*, 5, 823–826.

Dunbrack,R.L.,Jr and Karplus,M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction, *J Mol Biol*, 230, 543–574.

Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2005) Role of inter and intramolecular interactions in protein-DNA recognition, *Gene*, 364, 108–113.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics*, 15, 563–577.

Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge, *PLoS Comput. Biol.*, 1, e1.

Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins, *Proteins*, 35, 114–131.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Res*, 31, 5108–5121.

Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level, *Nucleic Acids Res*, 29, 2860–2874.

Mandel-Gutfreund,Y., Baron,A. and Margalit,H. (2001) A structure-based approach for prediction of protein binding sites in gene upstream regions, *Pac Symp Biocomput*, 139–150.

Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites, *Nucleic Acids Res*, 26, 2306–2312.

Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models, *Nucleic Acids Res*, 33, 5781–5798.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, 247, 536–540.

Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites, *Biochemistry*, 38, 1999–2017.

Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, *Proc Natl Acad Sci USA*, 95, 11163–11168.

Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, *Protein Sci*, 11, 2606–2621.

Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol*, 301, 597–624.

Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms, *Structure* (Camb), 12, 113–122.

Raviscioni,M., Gu,P., Sattar,M., Cooney,A.J. and Lichtarge,O. (2005) Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity, *J Mol Biol*, 350, 402–415.

Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J., Martinez-Antonio,A. and Collado-Vides,J. (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions, *Nucleic Acids Res*, 34, D394–397

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics, *J. Mol. Biol.*, 338, 207–215.

Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions, *Annu Rev Biophys Biomol Struct*, 34, 379–398.

Selvaraj,S., Kono,H. and Sarai,A. (2002) Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding, *J Mol Biol*, 322, 907–915?

Siggers,T.W., Silkov,A. and Honig,B. (2005) Structural alignment of protein--DNA interfaces: insights into the determinants of binding specificity, *J Mol Biol*, 345, 1027–1045

Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding, *Bioinformatics*, 18 (Suppl 1), S22–30.

Tramontano,A., Leplae,R. and Morea,V. (2001) Analysis and assessment of comparative modeling predictions in CASP4, *Proteins* (Suppl), 22–38.

Wang,G. and Dunbrack,R.L.,Jr. (2003) PISCES: a protein sequence culling server, *Bioinformatics*, 19, 1589–1591.