PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Comparison of DNA binding across protein superfamilies

Bruno Contreras-Moreira,[1,2,3]* Javier Sancho,[3,4] and Vladimir Espinosa Angarica[3,4]

[1] Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas, Av. Montañana 1.005. 50059 Zaragoza, España

[2] Fundación ARAID, Paseo María Agustín 36, Zaragoza, España

[3] Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza. Corona de Aragón, 42. Edificio Cervantes, 50009 Zaragoza, España

[4] Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza. Pedro Cerbuna, 12. 50009 Zaragoza, España

ABSTRACT

Specific protein–DNA interactions are central to a wide group of processes in the cell and have been studied both experimentally and computationally over the years. Despite the increasing collection of protein–DNA complexes, so far only a few studies have aimed at dissecting the structural characteristics of DNA binding among evolutionarily related proteins. Some questions that remain to be answered are: (a) what is the contribution of the different readout mechanisms in members of a given structural superfamily, (b) what is the degree of interface similarity among superfamily members and how this affects binding specificity, (c) how DNA-binding protein superfamilies distribute across taxa, and (d) is there a general or family-specific code for the recognition of DNA. We have recently developed a straightforward method to dissect the interface of protein–DNA complexes at the atomic level and here we apply it to study 175 proteins belonging to nine representative superfamilies. Our results indicate that evolutionarily unrelated DNA-binding domains broadly conserve specificity statistics, such as the ratio of indirect/direct readout and the frequency of atomic interactions, therefore supporting the existence of a set of recognition rules. It is also found that interface conservation follows trends that are superfamily-specific. Finally, this article identifies tendencies in the phylogenetic distribution of transcription factors, which might be related to the evolution of regulatory networks, and postulates that the modular nature of zinc finger proteins can explain its role in large genomes, as it allows for larger binding interfaces in a single protein molecule.

Proteins 2010; 78:52–62.

Key words: protein–DNA interface; direct and indirect readout; superfamily; atomic interactions.

INTRODUCTION

The specific interactions between short DNA sequences and proteins are a central feature of a wide group of processes in cell biology and organism development. Therefore, the study of the mechanisms of specific DNA binding by dedicated proteins has raised most attention. In addition to genetic, biochemical, and molecular biology approaches, it seems clear that a systematic study of the characteristics of the complexes formed between proteins and DNA at the atomic scale will provide a better understanding of the recognition process. To date, several reports have shed some light into the structural and functional characteristics of DNA-binding protein families[1–5] and the sequences recognized by DNA binding domains (DBD).[6–10] These studies have resulted in important contributions describing the interplay between DNA and protein during the recognition process and the structural determinants both at the protein and DNA contact surfaces responsible for specific recognition.

The high scientific relevance of the problem of protein–DNA recognition has contributed to a great increase in the number of high-quality structures of DNA-binding proteins (DBPs) reported in the Protein Data Bank.[11] The structures, especially those of their complexes with DNA, have provided valuable

Abbreviations: DBD, DNA binding domain; GR, glucocorticoid receptor-like; H, homeodomain-like; HE, homing endonucleases; IAS, interface alignment score; LR, lambda repressor-like; P53, p53-like; RE, restriction endonucleases; RHH, ribbon–helix-helix; RMSD, root mean square deviation; TF, transcription factor; WH, winged helix; ZF, C2H2/C2HC zinc fingers.

Additional Supporting Information may be found in the online version of this article.

The authors state no conflict of interest.
Grant sponsor: Consejo Superior de Investigaciones Científicas; Grant number: 200720I038; Grant sponsor: Banco Santander Central Hispano, Fundación Carolina and Universidad de Zaragoza; Grant sponsor: Consejo Superior de Investigaciones Científicas, JAE program; Grant sponsor: Fundación Aragón I+D; Grant sponsor: BFU; Grant number: BFU 2007-61476/BMC.
*Correspondence to: Bruno Contreras-Moreira, Laboratory of Computacional Biology, Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas, Av. Montañana 1.005. 50059 Zaragoza, España. E-mail: bcontreras@eead.csic.es
Received 31 March 2009; Revised 22 June 2009; Accepted 24 June 2009
Published online 6 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22525

insight into the stereochemical principles of binding, including how particular base sequences are recognized and how the DNA structure is often modified on binding.[3] The availability and steady growth of structural data of protein–DNA complexes has constituted the ground for a group of computational studies describing the characteristics of the amino acid–base interactions that determine binding specificity,[12,13] the different types of readout mechanisms involved in DNA recognition[14,15] and the evolutionary conservation of the residues located at contact interfaces.[16–18]

Protein–DNA interaction can be seen as a reaction in which one or more protein domains dock to the major and/or minor grooves of a DNA double helix. It is known that specificity is determined by the contribution of direct readout,[12] that is, associated to direct atomic contacts formed between atoms from amino acid side chains and nitrogen bases, indirect readout,[14,15] that is, mediated by the conformational changes undergone by DNA and the contribution of residues that are not in direct contact, and desolvation of the contact interface upon binding. The diverse studies carried out so far have suggested that the relative contributions of each one of the mechanisms related to specificity are different for each DNA-binding protein. It has also been claimed that the combination of the inter- and intramolecular readout energies leads to an enhanced specificity of recognition. The existence of a "universal" or "generic" protein–DNA recognition code at the atomic level has been proposed based on the strength of contact preferences.[12] Nevertheless, many amino acids form favorable contacts with different bases, making it necessary to generalize a deterministic recognition code to a probabilistic binding profile maximizing the likelihood of observed protein–DNA contacts.[19] However, several other reports question the existence of such kind of generic code for protein–DNA interaction,[20] whereas others argue that a family-specific code might exist.[3]

In a wide-ranging study of all the available structures of DBDs,[21] Siggers *et al.* were able to cluster these domains according to the geometric conservation of the contact interface with DNA. They found that, with few exceptions, proteins within a structural family form definite clusters. Another remarkable conclusion from this work is that, although proteins with similar folds tend to dock in similar ways, important differences are observed that seem to correlate with the level of sequence conservation at the docking interface.[22] Siggers *et al.* also proved that homologous interfaces tend to maintain certain contacts, even if this requires a distortion of the DNA. However, this study was mainly focused on the geometric properties at the protein side of contact interfaces, and thus was unable to address some interesting questions such as: (a) what is the relative contribution of specific amino acids and bases at the interface to specific binding, and (b) what nitrogen bases contribute the most to indirect readout recognition and how this contribution affects the specificity of recognition. We have recently developed a simple methodology to generate atomistic representations of protein–DNA interfaces,[23] which has been previously used with fairly good results to generate structure-based models of transcription factor binding sites. This computational protocol, named DNAPROT, permits a detailed structural dissection of the interfaces accounting for direct and indirect readout of DNA, including the contribution of the interactions mediated by water molecules and allowing sampling and optimization of amino acid side-chains rotamers[24] and has successfully been applied to a representative set of crystallographic structures and homology models.

This article addresses some of the aforementioned issues and systematically explores the conservation of structural features of binding interfaces, centering the study both at the protein and DNA sides of docked complexes. In particular, the DNAPROT methodology is applied to nine superfamilies from the Structural Classification of Proteins (SCOP)[25] and the results demonstrate that evolutionarily unrelated DNA-binding domains conserve important specificity statistics, such as the ratio of indirect/direct readout and the frequency of atomic interactions, but also unveil patterns that are superfamily-specific. Although it had already been described that prokaryotic genomes have a dominant proportion of Winged helix (WH) transcription factors, known to be functional as dimers, and that larger metazoan genomes are enriched in zinc finger (ZF) proteins, here we propose that this evolutionary trend is related to the modular nature of ZFs, which can be concatenated to ensure enough binding specificity, by means of larger interfaces, in a single protein molecule.

MATERIALS AND METHODS

Set of protein–DNA complexes and SCOP annotation

The subset of protein–DNA complexes in the Protein Data Bank[11] (release May 15th, 2009) was downloaded and the accompanying list of clusters of protein chains with 95% of sequence identity (under the derived_data directory) was parsed to define a nonredundant set of 175 monomeric complexes. For each cluster, the chain with best resolution was taken, considering NMR structures only when no crystallographic structures were available. For homodimeric complexes, only one chain was taken, the one appearing first in the corresponding nonredundant list. All nonredundant chains found as part of heterodimers or higher order complexes were considered. The protein sequence of each nonredundant complex was searched against SUPERFAMILY (version 1.69) using the superfamily.pl script with default parameters[26] in order to be assigned to SCOP superfamilies, which are expected to share evolutionary history,[25] and

Table I
Structural Descriptors of 9 DNA-Binding Superfamilies

Superfamily (SCOP domains)	Domains/ protein	Multiple alignment length (columns)	Mean RMSD (Å, residues)	Mean % sequence identity	Mean IAS	Interface (columns)	% Core (columns)	% Multiuse (columns)	% Different rotamers (columns)
HE (14)	1.4	216	3.23 (83)	19	3.65	33	67	68	89
ZF (35)	2.7	35	1.66 (26)	38	3.46	7	86	83	100
H (47)	1.1	191	2.87 (44)	16	2.53	37	54	35	100
GR (22)	1	111	2.06 (61)	41	4.21	16	44	43	100
WH (40)	1.1	232	3.55 (44)	10	2.38	33	61	70	94
RE (15)*	1	469	3.58 (39)	6	3.27	57 [66]	21 [9]	33 [67]	90 [80]
LR (13)	1	109	2.77 (47)	19	3.34	17	41	43	100
RHH (6)	1	104	2.3 (41)	15	2.44	4	75	0	100
P53 (18)*	1	363	3.42 (69)	13	3.05	26 [34]	50 [21]	23 [29]	83 [100]

SCOP domains with boundaries defined by SUPERFAMILY[26] were superposed with MAMMOTHmult.[27] The mean number of domains extracted per complex is shown in the second column, while the length of the resulting multiple alignments is given in the third column. Restriction enzymes and p53-like domains are marked with asterisks to indicate that their multiple rigid superpositions include domains that cannot be fit in frame, which affects the calculated statistics. The *mean RMSD* column shows the mean core size (in residues) and the resulting root mean square deviation of all pairwise domain structural alignments generated in the course of the progressive multiple alignment. The next column gives the mean sequence identity of these pairwise alignments. The *mean IAS* columns present the mean protein-DNA interface similarity score of Siggers *et al.*[21] The remaining columns give more detailed information about the interfaces, as shown in Figure 3, and are all calculated in the frame of reference defined by the structural superpositions of domains. Bracketed statistics for restriction enzymes and p53-like domains correspond to hidden Markov multiple alignments computed with SUPERFAMILY, calculated to demonstrate that different alignments still unveil large differences among domains. First, the *interface* column states how many columns of the original multiple alignment include interface residues, those that establish atomic interactions with nitrogen bases. Then, the *core* column shows the fraction of interface columns shared by at least two protein-DNA complexes in the same superfamily. Next, the *multiuse* column shows the fraction of core columns that includes residues that form atomic interactions of different types in different complexes. Finally, the *different rotamer* column states the fraction of core columns that include interface residues whose side chains belong to different rotameric states.

to precisely define domain boundaries. To minimize sampling problems, only those superfamilies containing more than five complexes were further considered in this work. The full list of 175 complexes is available as Supporting Information and their mean sequence identity percentages are shown in Table I. Complexes 1au7_B, 1e3o_C, 1ic8_A, 2d5v_B, 2h8r_A, 1fok_A and 2o61_A were not considered as they contain two DNA-binding domains from different SCOP superfamilies.

Atomic dissection of interfaces

The DNAPROT algorithm was applied to each monomeric complex to analyze the binding interface in terms of direct, that is, atomic interactions: hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic interactions, and indirect, that is, sequence-specific DNA geometry deformations—readout. It is important to note that DNAPROT considers only those atomic interactions that are sequence-specific, those that involve amino acid side-chains and purine/pyrimidine rings. Among hydrophobic interactions, only thymine C7 interactions are considered. Electrostatic interactions also play a major role in protein–DNA binding, but only a small fraction are expected to contribute to specific recognition, and these correspond to the interface hydrogen bonds mentioned earlier. Using this methodology, we obtained a structure-based position weight matrix for each interface, in which the direct/indirect relative contribution of each nitrogen base is assessed. Briefly, the saturating mutation strategy implemented in DNAPROT iteratively evaluates the interaction potential of a given protein–DNA complex, whereas each base at the crystallographic site is mutated by the other three bases.[23] Each single mutant is processed to obtain the contacts and deformation contribution of the given base and the analysis of all possible mutants renders a matrix in which the direct and indirect readout contributions are linearly combined by means of a deformation weight. Indirectly recognized base pairs are defined as those columns in the aforementioned indirect readout position specific matrix in which at least one nucleotide has a frequency >40 in a 0–100 scale; the prior frequency for all four nucleotides is by default 25. The DNA motif bound by any protein in this dataset is defined as the shortest oligonucleotide that encompasses all directly read bases plus all indirectly read bases that are less than three nucleotides away. This was necessary to adequately handle multimeric PDB structures that include several protein chains bound to the same DNA duplex. For the calculation of the indirectly readout fraction, motifs shorter than four nucleotides were not considered. This cut-off was chosen empirically based on the knowledge of the characteristics of TF-binding sites, which usually correspond to regions of 3–6 nucleotides.[28–30] Motifs of three nucleotides were excluded to avoid overestimating indirect readout fractions in the case of interfaces with few dissected atomic interactions.

Multiple alignment of DNA-binding superfamilies

The first step was to split all domains contained in the set of complexes of each superfamily, as many proteins contained more than one domain, for example, the set of 13 C2H2/C2HC ZF proteins yielded 35 domains. The domain boundaries reported by SUPERFAMILY, as

explained earlier, were generally followed, but were manually corrected for some concatenated ZF proteins. These domains were then structurally aligned by MAMMOTH-mult,[27] with the aim of putting all binding interfaces in the same frame of reference, and were further analyzed by using the Interface Alignment software,[21] which reports interface alignment scores (IAS) for pairs of complexes. Further multiple alignments were calculated using SUPERFAMILY hidden Markov models. DANGLE 0.63 (http://kinemage.biochem.duke.edu/software) was used to calculate side-chain torsion angles. To assign single interaction roles to interface residues, which can be easily displayed in a multiple alignment, the following rules were applied in this order of priority:

1. if the residue forms one or more hydrogen bonds it is called a hydrogen bond residue,
2. if the residue forms one or more hydrophobic interactions it is called a hydrophobic residue, and
3. if the residue forms one or more water-mediated hydrogen bonds it is called a water-mediated residue.

Rotamers were clustered using the CPAN Algorithm::Cluster module (http://search.cpan.org/dist/Algorithm-Cluster/) by requiring two side chains to be in the same cluster if both their χ_1 and χ_2 angles were $<40^\circ$ away.[31] Multiple alignments and superposition PDB files are available at http://www.eead.csic.es/compbio/suppl/dna_families/mammoth.zip and http://www.eead.csic.es/compbio/suppl/dna_families/alignments.zip, respectively.

Distribution of DNA-binding proteins at genomic scale

We downloaded the genomic annotations for predicted transcriptions factors from release 2.0 of DBD (http://www.transcriptionfactor.org),[32] completing this information with the genomic assignments for restriction and homing endonucleases (HE), which were kindly provided to us by curator Derek Wilson. This set was filtered in the following steps:

1. Genomes in which the subset of DBPs that belong to our defined set of nine superfamilies accounts for $<50\%$ of the annotated DBPs were discarded. This was done to avoid species in which the repertoire of DBPs is not adequately represented by these nine superfamilies.
2. Genomes with <100 annotated ORFs were discarded.
3. Genomes with <10 annotated TFs were also rejected.
4. Phyla with <4 genomes were ignored, with the exception of nematoda, for which only three genomes are available.

The remaining 490 annotated proteomes were used to generate Figure 4.

RESULTS AND DISCUSSION

This section presents the results of several analyses carried out with a nonredundant set of 175 protein–DNA complexes obtained as explained in section Materials and Methods. This set contains nine protein superfamilies, which are now listed together with their abbreviated names and the number of complexes included in each superfamily: C2H2/C2HC ZFs 13, HE 11, RE 16, lambda repressor-like (LR, 12), homeodomain-like (H, 38), p53-like (P53, 17), WH 38, glucocorticoid receptor-like (GR, 22), and Ribbon–helix-helix (RHH, 8).

Contribution of indirect readout to DNA recognition across superfamilies

The first question that this article aims to address is: how important indirect readout mechanisms are in different DNA-binding superfamilies? For this purpose, we applied the DNAPROT[23] algorithm to all members of the set described earlier and estimated how many base pairs of the bound DNA motif are, on average, recognized by means of sequence-specific deformations of the DNA duplex. These indirect readout estimations correspond to energetic potentials of deformation associated to DNA base steps.

The results are displayed in Figure 1, with two observations to be made: (a) the fraction of base pairs that are indirectly read within DNA motifs—that is, the median values of the distributions represented by the box plots—is typically small, with an average of 20% and (b) restriction endonucleases (RE) have a substantially larger proportion of indirectly readout bases [depicted as filled hexagons in the interface of Fig. 1(B)]. Overall, the contribution of indirect readout mechanisms, as dissected by our methodology, is rather small. However, it is important to note that these are superfamily generalizations, that is, individual complexes may depart from the superfamily-shared behavior as implied by the range of variation in each superfamily. For instance, among members of the C2H2/C2HC ZF superfamily, the Wilms tumor suppressor was found to substantially distort DNA upon binding, and hence more than half of its DNA motif is subject to indirect readout mechanisms.[34]

Two of the superfamilies included in this study correspond to the proteins with enzymatic activity, the restriction, and HE. As can be seen in Figure 1, for these two cases, we obtained the highest and lowest indirect readout contributions to binding specificity, which is in agreement with the experimental data obtained for some members of these superfamilies. A recent study of DNA recognition by RE underscored the relevance of this indirect type of reading of DNA, as they proved that a mutant enzyme deficient of direct contacts showed no loss of sequence specificity.[35] There are also other examples, such as the reports of the molecular structure of

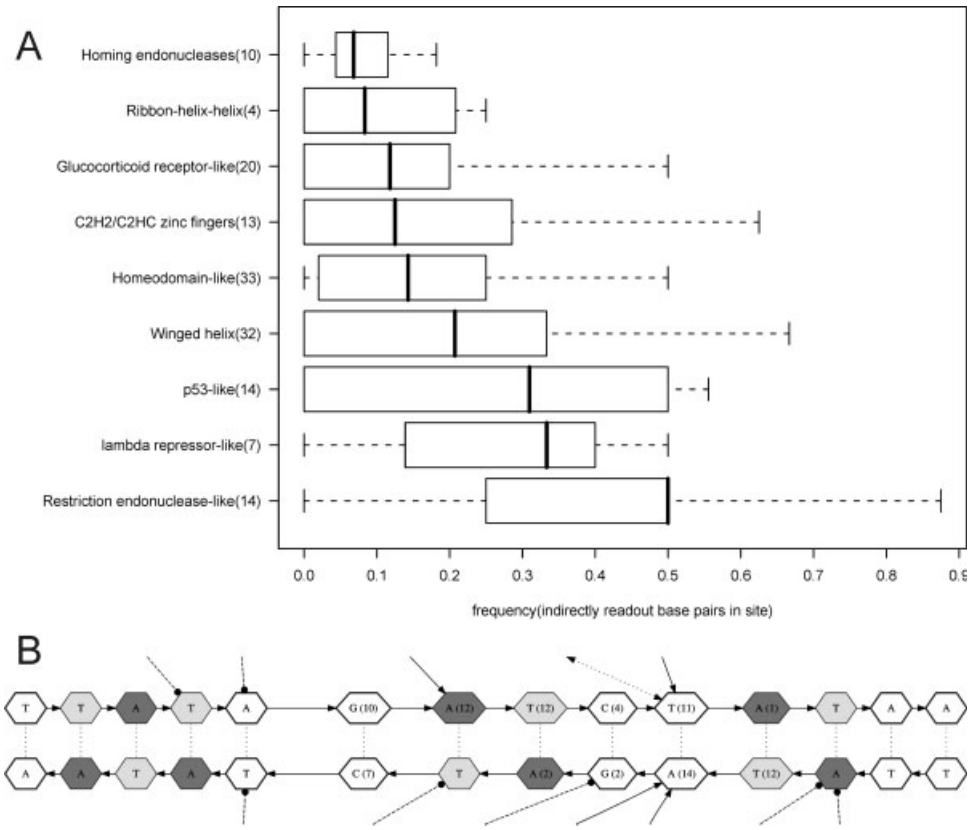

Figure 1

Relevance of indirect readout across representative SCOP superfamilies. (**A**) Box plots of the distribution of frequencies of indirectly recognized motif positions in nine different DNA-binding SCOP superfamilies, calculated from the corresponding number of monomeric complexes in parenthesis, shown next to its abbreviation. The boxes correspond to the interquartile range, the median is represented with a bold line and the minimum and maximum values are marked with whiskers connected with a dashed line. Outliers of the distribution are represented with open dots. (**B**) Atomic interface graph of restriction enzyme BgIII[33] (PDB: 1dfm_B), with seven base pairs in the bound DNA motif involved in indirect readout mechanisms, shown as filled bases. The arrows represent atomic interactions at the interface—that is, solid arrows = hydrogen bonds, dashed arrows = water-mediated bonds, dotted arrows = hydrophobic interactions—and are displayed to illustrate the fact that motif positions are often recognized by means of both indirect and direct readout.

BgIII[33] and NaeI,[36] which found large DNA rearrangements that occur upon specific binding. Therefore, the structural evidence at hand indicates that restriction enzymes extensively utilize indirect readout to bind DNA sequences very specifically. Enzyme MunI (PDB: 1d02) turns out to be a special case, as it induces a kink on its target oligonucleotide,[37] but this deformation is not translated into a significant increase in binding specificity as estimated by DNAPROT (see Supporting Information Fig. 2).

For the HE, those that yield the lowest fraction of indirect readout, our findings are in accordance with the biological function of these enzymes, engaged in processes of specific chromosome cleavage.[38] Proteins within this family contain a high number of direct contacts, as can be seen in Figure 2, with a relatively low contribution of indirect reading of DNA.[39] Figure 1 also shows that the median indirect readout contributions of the others superfamilies, most of which are transcription factors, are somewhere in-between the values obtained for the enzymes. Although the contribution of indirect reading to specificity is not fully understood yet, the analysis of our results and previous reports suggests that the generic function of the superfamily, for example whether the members recognize very specific sequences, a large number of different sequences, a restricted number of similar but not identical sequences or being completely unspecific—might be related to the ratio of indirect readout. Although some reports propose that for some specific transcription factors[40] and RE,[34] the indirect reading of DNA is case-dependent, the evidence presented here implies that the structural constraints of superfamilies impose limits on the indirect readout fractions of individual DBDs. In the case of RE, proteins with a common core fold[41] but with highly different DNA-binding regions,[3] that must bind well defined sequences with

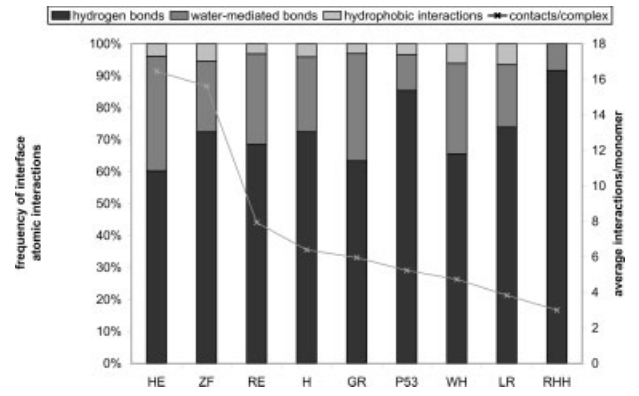

Figure 2

Direct readout explained in terms of atomic interactions at the protein–DNA interface. The bar plot depicts the mean contribution of hydrogen bonds, water-mediated bonds, and hydrophobic interactions to direct readout across the set of nine SCOP superfamilies (see abbreviations in Fig. 1). A line plot is overlaid showing the average number of interface atomic interactions for each superfamily, according to the secondary *Y*-axis at the right end of the diagram.

high specificity, indirect readout can be used as a "prescreening" mechanism during target site location by reducing the search space according to the deformation propensity of the sites.[35] For HE, which include a group of proteins with marked structural differences among them,[39] indirect readout may not be such determinant because direct contacts alone may account for an efficient recognition. In the case of transcription factors, indirect readout contributions correspond to intermediate values possibly related to multiple recognition scenarios—that is, large regulons for global regulators or a compact group of similar sequences for local regulators[42]—and are also likely influenced by the limitations that the DBD imposes on the number of sequences than can be recognized, as described for ZFs.[20]

Contribution of direct readout (atomic interactions) to DNA recognition across superfamilies

This section analyzes direct readout, probably a more tractable mechanism than indirect readout, because its energetic contribution is made of pairs of interfacial atoms of both the protein and DNA molecules that interact. The results depicted in Figure 2 correspond to the frequency of sequence-specific atomic interactions within protein–DNA interfaces. Perhaps the most important observation is that there is an almost constant contribution of hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic interactions to the binding interface across superfamilies. However, it seems clear that hydrogen bonds are the main source of specific interactions (on average they account for 72% of interactions), whereas the relevance of hydrophobic interactions is minor. A similar scenario is discovered when atomic pair potentials are used, providing further evidence in the same direction (see Supporting Information). It should be noted that only thymine C7 hydrophobic interactions were considered, as only these contacts were found to confer specificity in previous work,[43] and this explains this reduced contribution.

These results suggest that the molecular basis of direct readout follows general principles, that are shared by unrelated DBPs, and therefore support the existence of a set of recognition rules, a code, at the atomic level, in agreement with the observations of Luscombe *et al.*[12] This set of rules can be associated to a general set of atomic interactions among chemical groups at the interface, involving the same amino acids and bases in somehow similar chemical contexts from one DBD to another. However, interface residues often contact several nitrogen base groups simultaneously, and the interface architecture imposes geometric restrictions that favor some particular contacts over others. For these reasons, it is not generally possible to translate these atomic preferences into a one-to-one residue-base code, which would necessarily be affected by superfamily-specific conditions, as previously claimed.[44,45] For instance, interactions between asparagine and adenine account for >14% in Hs but only for 4% among WH transcription factors. The same applies for the interactions between lysine and guanine at the interfaces of GR, H, and WH proteins, where the summed relative contribution of water-mediated and direct hydrogen bonds accounts for 26%, 7%, and just 1%, respectively.

As in the previous section, it is worth noting that there are individual complexes that show an array of interactions clearly different from that of their superfamily. Among LRs, the interface of repressor protein P22 c2 (PDB: 2r1j) shows only one direct hydrogen bond per monomer, and recognition seems to be substantially driven by a combination of indirect readout, a key hydrophobic contact, and several water-mediated bonds.[46]

In addition to this, Figure 2 also shows the average number of atomic interactions in each superfamily (see the right vertical axis). In this respect, it would appear that there are two types of superfamilies: (a) those in which the average number of contacts per monomer is between three and seven and (b) C2H2/C2HC ZFs and HE, with >14 interactions per monomer. The first type includes superfamilies, such as Hs and Winged helices, which are known in many cases to be only functional as dimers. Thus, it can be assumed that these proteins will often be binding to DNA targets with an average number of interactions around 12. C2H2/C2HC ZFs can achieve a similar number of contacts without necessarily requiring the formation of dimers, but instead are usually made of several domains in tandem in the same polypeptide. For instance, the Wilms tumor suppressor

mentioned earlier contains four canonical ZFs in a row.[34] HE, usually embedded in introns or inteins, appear to have the largest number of interactions at the interface, and this is consistent with the fact that these double-stranded DNases bind to extraordinarily large recognition sites, from 12 to 40 nucleotides long.[47] If data from Figures 1 and 2 are combined, it could be argued that restriction enzymes (RE) belong to a third type of DBPs, with an intermediate number of contacts at the interface and a very important indirect readout component.

Structural comparison of DNA-binding interfaces

The next step in our analysis is to compare structural determinants within each superfamily, with the aim of updating and extending the pioneering work of Pabo and Nekludova.[45] As a prerequisite each superfamily has to be put in a common frame of reference, and a natural way of achieving this is by means of structure fitting. The degree of similarity of these proteins can be calculated at the domain scale, by means of average root mean square deviations (RMSD), or by focusing on the subset of interface residues, those that mediate direct readout, in which case the IAS of Siggers *et al.*[21] can be used. In addition, we ask if aligned interface residues play a similar interface role across members of the superfamily or whether they have a similar spatial arrangement, which can be measured in terms of side chain torsion angles. Figure 3 shows the superposition and corresponding multiple alignment computed by MAMMOTHmult[27] for 35 C2H2/C2HC ZF domains, extracted from the 13 members of that superfamily, and serves as a guide to Table I, which summarizes the structural analyses performed on the nine superfamilies subject of this study. There are seven interface columns in this superfamily, of which six (86%) are shared among several domains. Among these core interface positions, there is one dedicated exclusively to hydrophobic interactions in two domains (column 18), whereas the rest have mixed uses, dominated by residues that make specific hydrogen bond interactions with nitrogen bases. Figure 3(C) summarizes the rotamer clusters found for interface residues aligned in column 20. Although there are four clusters, the first cluster is the largest and includes different amino acids extracted from different complexes.

A quantitative evaluation of the structural statistics of binding interfaces in all nine superfamilies can be found in Table I. Overall, structural analysis finds that ZFs are the smallest DNA-binding domains, which are very similar between them under both RMSD and IAS metrics, and have a very compact set of core interface residues, as described in the literature.[5] In addition, ZF proteins contain on average 2.7 domains per protein chain, in contrast with most other superfamilies, which contain on average only one DNA-binding domain per chain. There are, however, other observations to be made to this table. For instance, in terms of RMSD, p53-like, and Restriction endonuclease (RE) domains are found to be the most divergent, as expected for their low percentages of sequence identity,[22] and indeed these superfamilies include domains which cannot be superposed in frame. Multiple alignments based on SUPERFAMILY hidden Markov models confirm the observed structural divergence (see Table I and Supporting Information). WH domains are also divergent, but their superposition is in frame. In terms of IAS, proteins from the GR superfamily are found to have very similar interfaces, in contrast with WHs, that appear to have a very flexible way of binding to the major groove of DNA. With respect to interface size, RHH and ZF proteins stand out for being compact DNA binders, which tend to use a small subset of residues to drive specific recognition. On the other end of the scale are REs, which have a large number of interface positions. Despite their different interface sizes, proteins from these superfamilies have an average of 55% of interacting residues in the core. These are positions observed in several members of the superfamily, which suggest that a common binding architecture might exist. However, REs display a different trend, as they seem to have almost unique arrays of interface residues when compared with each other. We now analyze the interaction roles of interface residues, which can display different recognition uses. H and RHH domains tend to have conserved roles for their interface residues, whereas the remaining superfamilies are able to use the same interface position to have, say, a hydrogen bond in one complex and a hydrophobic interaction in another complex.

Finally, Table I also indicates that most (83–100%) interface positions across superfamilies can be expected to have at least two side chain rotamer conformations in the same superfamily.

The data in Table I can also be used as a guide for comparative modeling exercises. For instance, the selection of protein templates, which should then be aligned to a query sequence in order to build a three-dimensional molecule,[48] will be affected by the degree of conservation of binding interfaces, as suggested by previous reports.[49,50] In addition, the variable size of the core interface across superfamilies is expected to affect the outcome of methods that predict interface residues.[51] Finally, the finding that interface side-chains cluster in a few groups can be exploited in molecular modeling exercises to drive conformational searches.

The structural analysis of evolutionarily related DNA-binding proteins yields the following conclusions. As with indirect readout, it is observed that the DBD of proteins from a given superfamily share structural features, but there is variability within superfamilies. It seems that DBDs with a common ancestor have evolved

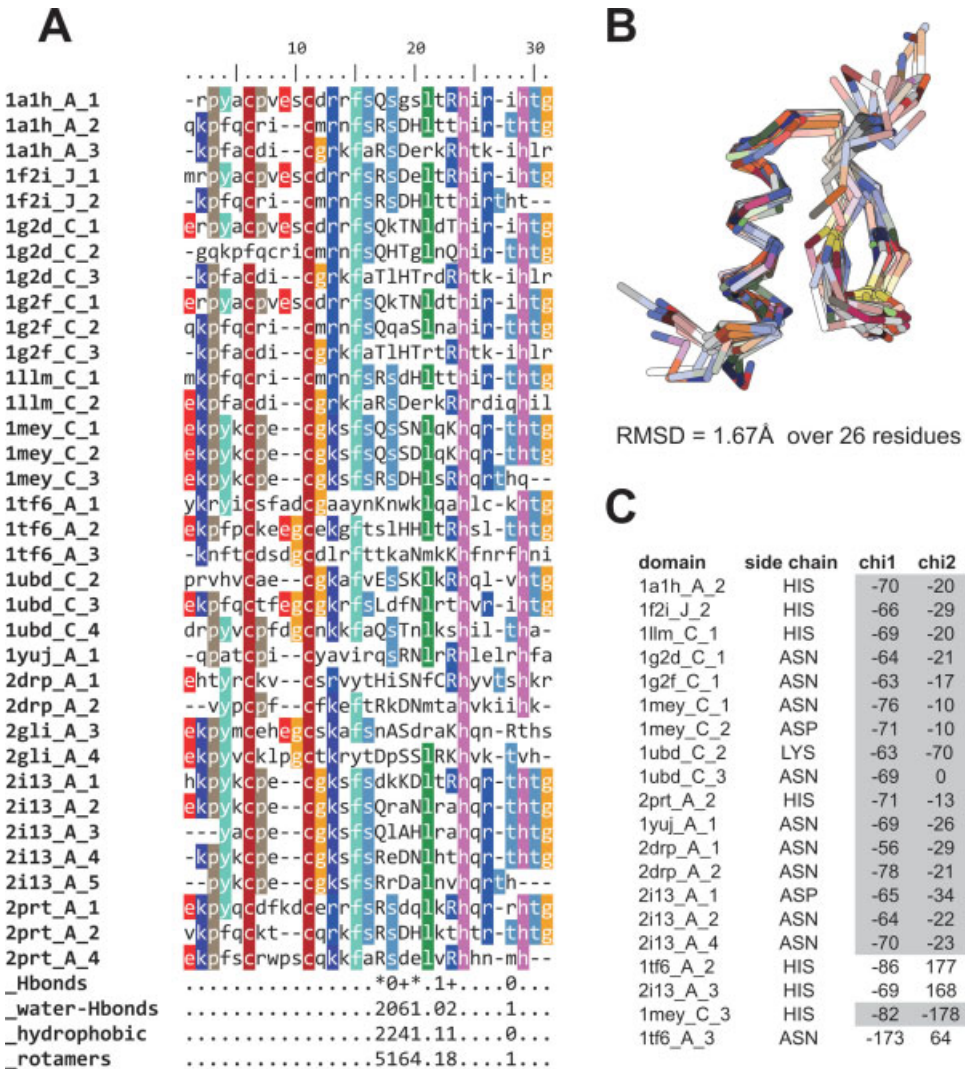

Figure 3

Alignment and superposition of 35 C2H2/C2HC zinc finger domains. (**A**) Structure-based multiple sequence alignment, with ruler on top and interface summaries at the bottom. Column 17, an example of multiuse interface position, contains 23 residues that form hydrogen bonds, whereas two residues take part in water-mediated bonds and another two form hydrophobic interactions. These 27 equivalent interface residues can be grouped in five side chain rotamer clusters. Column 18 is an example of interface residue dedicated exclusively to hydrophobic contacts, observed in only two fingers (alanine in 2i13_A_3 and histidine in 1g2d_C_2) that belong to the same rotamer cluster. In this panel, '*' means >20 residues and "+" means >10 residues. (**B**) Backbone cartoon of the corresponding structural superposition. (**C**) Summary of four side chain rotamer clusters found for interface residues aligned in column 20 of the multiple alignment.

to develop flexible binding interfaces that allow equivalent positions to bind to different nitrogen bases and accommodating variants of rotamer conformations, whereas retaining some key core positions, such as glutamine 50 in Hs.[45] Despite the overall conservation of geometry within superfamilies, the number of interface changes required to distinguishing one DBD from the other can be very small, as illustrated by CRP and FNR, two homologous TFs in *Escherichia coli* that can be interconverted with just a couple of point mutations.[52]

Distribution of DNA-binding proteins across genomes

In this section, we wish to focus on the distribution of DBPs across completely sequenced genomes, with the aim of searching for evolutionary trends that might complement our structure-based analysis. Figure 4 summarizes the proportion of annotated DBPs that belong to any of the considered superfamilies among 490 proteomes grouped in 15 phyla, which have been sorted by mean proteome size. First inspection indicates that the relative proportion of endonucleases is rather low in

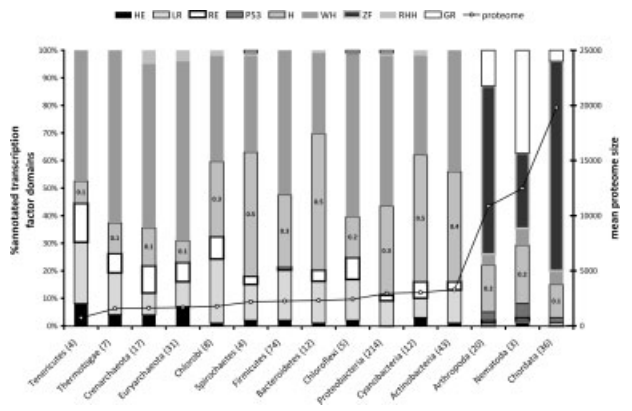

Figure 4

Genomic distribution of DNA-binding superfamilies. Bar plot of the distribution of nine DBD superfamilies across the annotated genomes of 15 phyla. The data for each phyla was derived from the number of genome sequences shown in parenthesis. The fraction annotated as homeodomains is shown in number, as this superfamily corresponds to at least 10% of the annotated TFs in all phyla. A line plot is overlaid that shows the mean proteome size of each phyla, as the secondary *Y*-axis describes. The arthropoda, nematoda, and chordata phyla are summarized in the text as metazoans. This figure was prepared with data from the DBD database (http://www.transcriptionfactor.org).[32]

most genomes, with only a little presence of these domains in higher organisms. Our data show that HE are more abundant in tenericutes and extremophiles of the phylum euryarcheaota than in other bacteria, presumably because their opportunities for homing are limited.[38] HEs are also scarce among metazoans, possibly because their segregated germ lines impedes horizontal transmission.[38] With respect to RE, these domains are more frequent in tenericutes and crenarcheota, although are frequently found in all bacterial genomes, as shown in Figure 4.

The bar plot also shows that Hs are the only superfamily of TFs which have an important (10% or more) presence across phyla, from prokaryotes to archaea and eukaryotes. WH TFs are also present in all phyla, but when they are the dominant proteins in prokaryotes[53] and archeaea (with proteome sizes below 5000 bases), they represent just a minor (5%) fraction of TFs in metazoan genomes. Lambda receptor proteins represent 7–23% of TFs in organisms with small proteomes, but are $<$1% of metazoan TFs. The opposite case are GR TFs, which are almost absent in small genomes but account for 13 and 37% of TFs in arthropoda and nematoda, respectively, with a minor fraction in chordata. Finally, though C2H2/C2HC ZFs are almost absent in small genomes,[54] they seem to be the preferred TFs of metazoan organisms, as they account for 27–76% of annotated TFs.

This data, together with the observations in Figure 2, suggests that transcriptional regulation has followed two distinct evolutionary paths in small and large genomes. In prokaryotic genomes, such as *E. coli*, regulation is dominated by WH TFs, that only accumulate a sufficient number of interface contacts after dimerization, a reaction which is usually dependent on some effector signal.[55] Instead, metazoan organisms, with genomes that can be several orders of magnitude larger, have chosen modular C2H2/C2HC ZF domains, that can be easily concatenated in evolution and ensure enough binding specificity, that is, enough atomic interactions spanning a longer oligonucleotide, in a single protein molecule. Furthermore, it is well documented that ZFs can still interact with other proteins and form multimeric complexes.[5]

CONCLUSION

After analyzing a comprehensive collection of protein–DNA complexes, we estimate that the average contribution of indirect readout to specific binding is approximately of one every five DNA bases, with the notable exception of restriction enzymes, which double its contribution. Furthermore, proteins from the same superfamily often display uneven indirect readout behaviors. With respect to direct readout, hydrogen bonds dominate DNA recognition, with a minor fraction of hydrophobic interactions. The constant contribution of atomic interactions across superfamilies supports the existence of a general set of recognition rules at the atomic level. It also appears that most superfamilies have a number of atomic interactions per monomer near six, except HE and C2H2/C2HC ZF, which have $>$14 contacts on average.

Comparison of proteins from the same superfamily by means of structural fits indicates that some superfamilies show larger interface variability than others, particularly restriction enzymes and p53-like proteins. Aligned interface residues of H and RHH domains are less likely to switch their interaction type than the rest, and, in general, their side chains cluster in rotamer groups.

In summary, though direct readout uses a generic code of recognition at the atomic level, the architecture of individual DBPs show subtle superfamily deviations that determine a more case-specific way of DNA recognition. This corresponds to an scenario in which the same set of rules—that is, the interactions between specific groups from amino acids and nitrogen bases—can be tuned at the structural level, involving deformation of DNA, the geometry of the binding domain and rotamer variants of amino acid side-chains, to ensure a specific binding and versatility of DNA recognition.

Finally, a survey of the frequency of transcription factor superfamilies across 15 phyla finds clear patterns of distribution, confirming that prokaryotic genomes are preferentially regulated by Winged helices, whereas metazoans are rich in ZF TFs. Here we propose that the modular nature of ZF domains, which can be concatenated to

ensure enough binding specificity in a single protein molecule, can explain this evolutionary trend.

ACKNOWLEDGMENTS

The authors thank the organizers of the International Workshop in memoriam of Ángel Ramírez Ortiz. This article is dedicated to the memory of Ángel. They also thank Derek Wilson for his help in assigning endonuclease domains across genomes.

REFERENCES

1. Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wuthrich K. Homeodomain-DNA recognition. Cell 1994;78:211–223.
2. Harrison SC, Aggarwal AK. DNA recognition by proteins with the helix-turn-helix motif. Annu Rev Biochem 1990;59:933–969.
3. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biol 2000;1:REVIEWS001.
4. Pabo CO, Sauer RT. Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem 1992;61:1053–1095.
5. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 2000;29:183–212.
6. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Trends Biochem Sci 1988;13:207–211.
7. Choo Y, Klug A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc Natl Acad Sci USA 1994;91:11168–11172.
8. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol 1995;253:370–382.
9. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J Mol Biol 1986;188:415–431.
10. Takeda Y, Sarai A, Rivera VM. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc Natl Acad Sci USA 1989;86:439–443.
11. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007;35(Database issue):D301–D303.
12. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
13. Mandel-Gutfreund Y, Baron A, Margalit H. A structure-based approach for prediction of protein binding sites in gene upstream regions. Pac Symp Biocomput 2001;6:139–150.
14. Selvaraj S, Kono H, Sarai A. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. J Mol Biol 2002;322:907–915.
15. Michael Gromiha M, Siebers JG, Selvaraj S, Kono H, Sarai A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. J Mol Biol 2004;337:285–294.
16. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 2002;320:991–1009.
17. Mirny LA, Gelfand MS. Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res 2002;30:1704–1711.
18. Raviscioni M, Gu P, Sattar M, Cooney AJ, Lichtarge O. Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. J Mol Biol 2005;350:402–415.
19. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res 2005;33:5781–5798.
20. Wolfe SA, Grant RA, Elrod-Erickson M, Pabo CO. Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. Structure 2001;9:717–723.
21. Siggers TW, Silkov A, Honig B. Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. J Mol Biol 2005;345:1027–1045.
22. Contreras-Moreira B, Collado-Vides J. Comparative footprinting of DNA-binding proteins. Bioinformatics 2006;22:e74–e80.
23. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B. Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics 2008;9:436.
24. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
25. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
26. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919.
27. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 2005;21:3255–3263.
28. Defrance M, Janky R, Sand O, van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat Protoc 2008;3:1589–1603.
29. van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 2000;28:1808–1818.
30. van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 1998;281:827–842.
31. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.
32. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 2008;36(Database issue):D88–D92.
33. Lukacs CM, Kucera R, Schildkraut I, Aggarwal AK. Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 A resolution. Nat Struct Biol 2000;7:134–140.
34. Stoll R, Lee BM, Debler EW, Laity JH, Wilson IA, Dyson HJ, Wright PE. Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA. J Mol Biol 2007;372:1227–1245.
35. Little EJ, Babic AC, Horton NC. Early interrogation and recognition of DNA sequence by indirect readout. Structure 2008;16:1828–1837.
36. Huai Q, Colandene JD, Topal MD, Ke H. Structure of NaeI-DNA complex reveals dual-mode DNA recognition and complete dimer rearrangement. Nat Struct Biol 2001;8:665–669.
37. Deibert M, Grazulis S, Janulaitis A, Siksnys V, Huber R. Crystal structure of MunI restriction endonuclease in complex with cognate DNA at 1.7 A resolution. EMBO J 1999;18:5805–5816.
38. Burt A, Koufopanou V. Homing endonuclease genes: the rise and fall and rise again of a selfish element. Curr Opin Genet Dev 2004;14:609–615.
39. Stoddard BL. Homing endonuclease structure and function. Q Rev Biophys 2005;38:49–95.

40. Locasale JW, Napoli AA, Chen S, Berman HM, Lawson CL. Signatures of protein-DNA recognition in free DNA binding sites. J Mol Biol 2009;386:1054–1065.
41. Niv MY, Ripoll DR, Vila JA, Liwo A, Vanamee ES, Aggarwal AK, Weinstein H, Scheraga HA. Topology of Type II REases revisited; structural classes and the common conserved core. Nucleic Acids Res 2007;35:2227–2237.
42. Lozada-Chavez I, Angarica VE, Collado-Vides J, Contreras-Moreira B. The role of DNA-binding specificity in the evolution of bacterial regulatory networks. J Mol Biol 2008;379: 627–643.
43. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins 1999;35:114–131.
44. Pabo CO, Sauer RT. Protein-DNA recognition. Annu Rev Biochem 1984;53:293–321.
45. Pabo CO, Nekludova L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol 2000;301:597–624.
46. Watkins D, Hsiao C, Woods KK, Koudelka GB, Williams LD. P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. Biochemistry 2008;47:2325–2338.
47. Belfort M, Roberts RJ. Homing endonucleases: keeping the house in order. Nucleic Acids Res 1997;25:3379–3388.
48. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
49. Siggers TW, Honig B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Res 2007;35:1085–1097.
50. Contreras-Moreira B, Branger PA, Collado-Vides J. TFmodeller: comparative modelling of protein-DNA complexes. Bioinformatics 2007;23:1694–1696.
51. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004;20:477–486.
52. Spiro S, Gaston KL, Bell AI, Roberts RE, Busby SJ, Guest JR. Interconversion of the DNA-binding specificities of two related transcription regulators, CRP and FNR. Mol Microbiol 1990;4:1831–1838.
53. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res 2003;31:1234–1244.
54. Bouhouche N, Syvanen M, Kado CI. The origin of prokaryotic C2H2 zinc finger regulators. Trends Microbiol 2000;8:77–81.
55. Martinez-Antonio A, Janga SC, Salgado H, Collado-Vides J. Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. Trends Microbiol 2006;14:22–27.