Published online 4 March 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac008

A deterministic code for transcription factor-DNA recognition through computation of binding interfaces

Marco Trerotola[1,2], Laura Antolini[3], Laura Beni[1], Emanuela Guerra[1,2], Mariano Spadaccini[4], Damiano Verzulli[4], Antonino Moschella[5] and Saverio Alberti[1,5,*]

[1]Laboratory of Cancer Pathology, Center for Advanced Studies and Technology (CAST), University "G. D' Annunzio", Via L. Polacchi 11, 66100 Chieti, Italy, [2]Department of Medical, Oral and Biotechnological Sciences, University "G. d'Annunzio", 66100 Chieti, Italy, [3]Center for Biostatistics, Department of Clinical Medicine, Prevention and Biotechnology, University of Milano-Bicocca, 20052 Monza, Italy, [4]Unit of Informatics, University "G. d'Annunzio", 66100 Chieti, Italy and [5]Unit of Medical Genetics, Department of Biomedical Sciences - BIOMORF, University of Messina, via Consolare Valeria, 98125 Messina, Italy

Received September 02, 2021; Revised December 05, 2021; Editorial Decision January 10, 2022; Accepted February 28, 2022

ABSTRACT

The recognition code between transcription factor (TF) amino acids and DNA bases remains poorly understood. Here, the determinants of TF amino acid-DNA base binding selectivity were identified through the analysis of crystals of TF-DNA complexes. Selective, high-frequency interactions were identified for the vast majority of amino acid side chains ('structural code'). DNA binding specificities were then independently assessed by meta-analysis of random-mutagenesis studies of Zn finger-target DNA sequences. Selective, high-frequency interactions were identified for the majority of mutagenized residues ('mutagenesis code'). The structural code and the mutagenesis code were shown to match to a striking level of accuracy ($P = 3.1 \times 10^{-33}$), suggesting the identification of fundamental rules of TF binding to DNA bases. Additional insight was gained by showing a geometry-dictated choice among DNA-binding TF residues with overlapping specificity. These findings indicate the existence of a DNA recognition mode whereby the physical-chemical characteristics of the interacting residues play a deterministic role. The discovery of this DNA recognition code advances our knowledge on fundamental features of regulation of gene expression and is expected to pave the way for integration with higher-order complexity approaches.

INTRODUCTION

The basic determinants of the binding of transcription factor (TF) amino acid (AA) side chains to target DNA sequences remain poorly understood (1–7). Multiple mechanisms of DNA sequence recognition have been recognized, which include cooperative binding, cofactor recruitment, DNA kinking and bending, flanking DNA sequence recognition (8), DNA methylation (5), allosteric effects (1–7,9), binding synergy through repeat symmetry (10), indirect TF-DNA interactions (11), induced folding of intrinsically disordered regions, dynamic recognition of TF target DNA sequences (12) and higher order assembly of enhancer sequences (13).

Taken together, these findings have led to the concept that TF bind DNA through intricate, interactive contacts, which are not amenable to simple ruling (14–19). Discrepancies between TF AA-DNA base interaction models (1,20–23) and their weak predictive power (15,16,19,24,25) supported this skeptical point of view. Consistent, genome-wide analyses have rarely identified statistically over-represented motif syntax rules, questioning whether they indeed exist and whether they impose evolutionary constraints on regulatory elements function (13,26).

Never the less, pivotal studies, including chromatin immunoprecipitation coupled to sequencing (ChIP-seq), succeeded in identifying discrete elements of TF-DNA recognition processes (8,27). Single-base resolution was then achieved by *in vitro* TF-DNA arrays (28,29), SELEX (30) and SNP-SELEX-based approaches (27), suggesting shared underlying principles in TF recognition of target DNA sequences. Still, these remained incompletely understood, and the existence of a deterministic TF-DNA recognition code remained unclear (31,32).

In this work we challenged a model of deterministic binding between TF-AA and DNA bases. To explore such TF-DNA recognition code, we devised converging strategies for identifying determinants of TF AA-DNA base binding selectivity. The TF that uses an α-helix for major groove DNA recognition (for the sake of simplicity these will be called

*To whom correspondence should be addressed. Tel: +39 347 3445123; Email: salberti@unime.it

'α-helix TF') comprises the largest TF subset (21,25,33) and has been suggested to play a pivotal evolutionary role in the shaping of the genetic code (see (34) for structural and thermodynamic arguments in favor of this model). Several α-helix TF-DNA complex structures have been solved to high resolution (https://www.rcsb.org/), allowing to perform a broad analysis of crystals of TF-DNA complexes. Computation of the TF-DNA binding interfaces was thus utilized to identify protein-DNA non-bonded contacts (hydrogen bonds, Van der Waals interactions, hydrophobic bonds and salt bridges). Selective, high-frequency interactions were identified for the majority of AA side chains ('structural code'). An entirely independent strategy was devised to challenge this TF-DNA interaction code, via meta-analysis of random-mutagenesis studies of Zn fingers and DNA target sequences. This led to identify TF-DNA interaction determinants *in vitro,* through the recognition of high-affinity-binding AA-DNA base pairs ('mutagenesis code'). The structural code and the mutagenesis code were here shown to match to a striking level of accuracy.

These findings indicated the existence of a DNA recognition mode whereby the physicochemical characteristics of the interacting residues play a key deterministic role. This TF AA-DNA base-binding alphabet was shown to possess a broad, TF class-independent validity, providing the fundaments for defining the mechanisms of DNA recognition by TF.

MATERIALS AND METHODS

Protein-DNA crystal structures

Protein-DNA complexes of eukaryotic α-helix TF that had been crystallized and solved with a resolution of ≤3 Å (http://www.rcsb.org/) were analyzed (Figure 1 and Supplementary Table S1A). Duplicate structures were excluded. In the case of multimeric structures comprising identical cells, only a single cell was taken into account, to prevent bond-counting bias, i.e. repetitive counting of identical bonds occurring in identical structures. Subsets of higher resolution TF-DNA crystals were separately analyzed, using resolution thresholds of ≤2.5 Å or of ≤2 Å, respectively (Supporting Information, Supplementary Tables S2–S4). Co-crystal structures were displayed with Swiss-PdbViewer (https://spdbv.vital-it.ch/) or RasMol (http://rasmol.org/). RasMol and Raster3D (http://www.bmsc.washington.edu/raster3d/) (35) were utilized for computer graphics. Structure analysis was performed on a Silicon Graphics Octane r12000 workstation.

Identification of protein-DNA contacts

Non-bonded interactions in protein-DNA complexes were identified in TF-DNA interfaces (22) (http://www.rcsb.org/) (84 structures, training dataset). A validation dataset was utilized that contained independent PDB entries (44 structures) (Supplementary Table S1A).

The HBPLUS program (https://www.ebi.ac.uk/thornton-srv/software/HBPLUS/) (22,36) was used to locate specific proximal donor and acceptor atom pairs and to calculate theoretical parameters fitting distinct bond types.

Hydrogen-bond donor and acceptor atom pairs were identified as described (22). Donor-acceptor atoms angle values were accepted between 180°, which is the optimal hydrogen-bond geometry, and 90°, which is the minimal threshold for hydrogen-bond occurrence. Maximum-distance thresholds were adopted of 2.5 Å between hydrogen-acceptor atoms and of 3.25 Å for donor-acceptor atoms.

Van der Waals forces vary with the inverse of the sixth power of the distance between the interacting atoms. Following the procedures described in the Results section, Van der Waals contacts were defined as all contacts between atoms not involved in hydrogen bonds that were <4 Å apart (Supplementary Table S2A).

Electrostatic interactions vary inversely with the square of the distance, and are difficult to calculate at a whole-structure level. Juxtaposed opposite charges (e.g. Arg-$NH2^{(+)}$ and $^{(-)}$P groups) were identified as clear interactions of this type.

Hydrophobic interactions depend on the increase in entropy gained by the removal of a hydrophobic surface from ordered solvating water, and as such they are difficult to be explicitly computed (37). Clear hydrophobic interactions were identified as contacts between hydrophobic side chains and the methyl group of thymines.

HBPLUS generated *.nb2 files (non-bonded contacts) and *.hb2 files (hydrogen bonds). Graphical 2D representations of the HBPLUS *.hb2 and *.nb2 outputs were generated with NUCPLOT (36) (Supplementary Table S2). Bond length thresholds were implemented in the NUCPLOT analyses. NUCPLOT 2D graphics were generated for all TF AA-DNA co-crystals (https://github.com/marcotrerotola/TF-finder/branches).

Non-bonded contact assessment

Non-bonded contacts in each structure were quantified upon normalization, whereby each DNA-binding AA received a score of 1.00. If a single AA residue bound more than one structure, a fractional score was attributed to each interaction (e.g., if an AA bound one base, one sugar moiety and one phosphate group a score of 0.33 was assigned to each one of the three contacts). TF AA contacts to the target DNA bases in 5′, mid, 3′ positions were separately compiled. Contacts of ZnCoS AA in position 2 were compiled together with those in position 3 (38). A similar convention was used for HTH. Zipper-type TF interact with DNA over four turns of an α-helix versus the three of ZnCoS and HTH. For ease of comparison, the two middle DNA bases and the corresponding AA contacts in leucine zippers were grouped into the 'mid' position (Supplementary Tables S3 and S4).

Zn finger mutagenesis studies meta-analysis

A meta-analysis of the listed Zn finger mutagenesis studies (38–44) was performed (training dataset). The largest single study (45) was not included in the meta-analysis and was used as a test dataset. A normalized score system was devised that followed the same rationale as for the TF-DNA

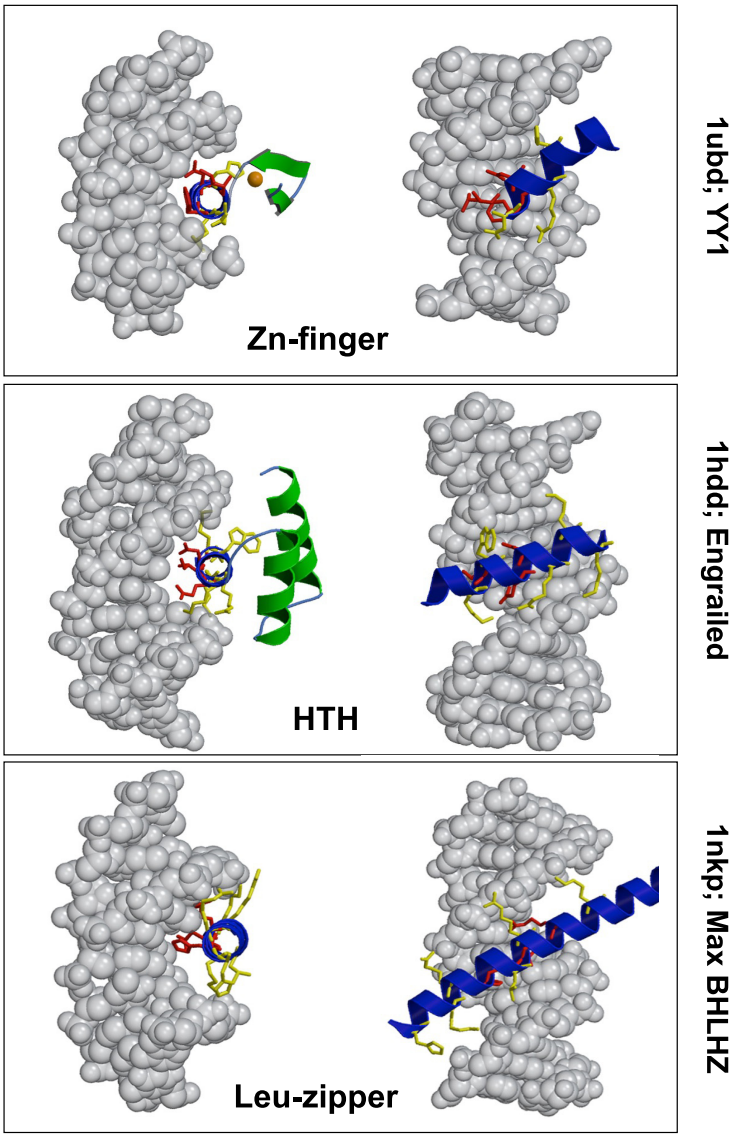

Figure 1. Structure of the α-helix TF. Beta-beta-alpha Zn finger (1ubd; YY1 Zn finger domain) (top); HTH (1hdd; Engrailed homeodomain) (mid); leucine zipper (1nkp; Max BHLHZ region) (bottom). Full-length DNA-binding α-helices (in blue) are shown as ribbon diagrams. Side views of the DNA-binding α-helix are shown in the right panels. Straight-on views (left) demonstrate contacts to the DNA bases (red) and to the backbone (yellow). Orange sphere (top, left): coordinating Zn atom. Figures were generated with Raster 3D - Molscript.

structures. Each Zn finger mutant residue that bound (corresponded to) a single DNA base received a score of 1.00. If an AA bound more than one base, a fractional score was computed (e.g. if an AA bound three bases, each interaction received a score of 0.33). If a quantitative analysis was available, e.g. if the relative strength of binding to different residues was quantified, the score unit was subdivided accordingly (e.g. if an AA bound two different bases, one of them with a 4-fold higher affinity, this received a score of 0.8; 0.2 was attributed to the lower affinity contact). Matrices of TF AA-DNA base contacts were constructed for each independent mutagenesis study; all-inclusive matrices of position-specific and overall contacts were subsequently compiled (Supplementary Table S4).

DNA bases nearest-neighbor analysis

The observed distribution of bases in TF target sequences was described according to absolute occurrence and relative frequency of individual bases (Supplementary Tables S5 and S6). Expected distributions were computed as the product of observed frequencies of two neighboring bases, under the assumption of random association between pairs of adjacent bases. Comparative observed distributions of (unordered) pairs of bases in the TF target sequences are presented in Supplementary Table S6Di. Overall observed/expected distributions were computed by averaging Zn finger, HTH, leucine zipper frequencies, under the assumption of equal contribution by each TF-DNA structure datasets (Supplementary Table S6 Dii 'overall').

Statistical analysis

If contacts between individual AA and the four DNA bases were random, one would expect 25% of AA contacts to each DNA base. This was assumed as the null hypothesis in Fisher exact test analysis of TF AA-DNA base binding distributions in TF-DNA structures and in Zn finger mutants. The probability of distribution homogeneity was assessed by χ^2 analysis. Normalized scores were rounded-up to full-digits for comparisons. The top-scoring contact pair frequencies in the different TF classes were compared using a one-tailed Fisher exact test. Combinatorial analysis was utilized to quantify the probabilities of random occurrence of series of top-scoring contact pairs across the different TF classes.

RESULTS AND DISCUSSION

Position weight matrix models of TF AA-DNA base contacts

DNA-binding motifs in TF are typically identified according to position weight matrix (PWM) models of over-represented sequence/position-matching AA (4). Whenever such PWM are utilized for TF-binding specificity description (46), analysts implicitly assume independent binding of individual AA to DNA bases, and additive effects on TF-DNA recognition (4).

We attempted this approach on TF-DNA structures for which co-crystals were available (Supplementary Table S1A). PWM allowed to identify distinct features, e.g. the well-known high frequency binding of Arg to G bases and the prevalence of DNA-binding sites at positions -1, +2–3, +6–7 of the DNA binding α-helix, together with degenerate binding to neighbouring bases (Supplementary Table S1B). However, limited additional information could be gathered, consistent with previous analyses (27).

Comparison with actual bonds, as obtained from the analysis of TF-DNA structures, revealed massive divergence of PWM from accurate descriptions. As an example, PWM analysis of the 1aay TF-DNA structure (Supplementary Table S1B) failed to reveal 9 instances of actual bonds (T156 to C3, F144 to T5, I128 to G6, R103 to G8, F118 to G8, E121 to C9, D120 to C53, D176 to C59, K179 to C57). Further, PWM predicted R180 to bind G2 but not G60 and R124 to bind G7 rather than G8. Corresponding findings were obtained for the other TF-DNA structures analyzed (Supplementary Table S1B). These findings indicated that co-crystal structure analysis provided much more powerful/higher resolution data than PWM modeling.

TF-DNA contacts in 3D structures of TF-DNA complexes

Consistent with previous findings (1,7,28,47–49), our results indicated reliable predictions of TF-DNA bonds using physico-chemical features of TF contacts with DNA. Hence, we went on to analyze the 3D structure of TF-DNA co-crystals for geometric parameters associated to non-covalent bonds (16,22,23,39–45,50–52).

Selection of protein-DNA co-crystals

Protein-DNA complexes of eukaryotic α-helix TF (2) (Figure 1 and Supplementary Movie S1), that were solved at a resolution of better than 3 Å, were selected (Protein Data Bank, https://www.rcsb.org/). To prevent statistical bias from inclusion of multiple identical structures, the study design required a minimum of one AA/DNA base contact-pair difference among TF-DNA complexes that were included in the analysis (Supplementary online material). Besides improving the robustness of cumulative binding data, by avoiding repetitive counting of identical bonds in identical structures, this 'one-AA difference' allowed to assess the existence of cooperative effects in DNA recognition by neighboring AA, e.g. via di-nucleotide alphabets (8,53) (see below).

Identification of DNA-protein contacts

TF-DNA bonds, such as hydrogen bonds, Van der Waals, electrostatic and hydrophobic interactions, were systematically identified in the selected TF-DNA co-crystal structures (22,36).

Hydrogen-bond donor and acceptor atom pairs were identified first, utilizing the set of parameters described in the Experimental Procedures. Distance thresholds were adopted of 2.5 Å between hydrogen-acceptor atoms and of 3.25 Å for donor-acceptor atoms. Donor-acceptor atom angle values were accepted when comprised between 90° and 180°.

Van der Waals forces, as dipole-dipole interactions, dipole-induced dipole interactions, induced dipole-induced dipole interactions, can act between any two molecules or atoms. This type of force is short-range and varies according to the inverse sixth power of the distance between interacting molecules. Van der Waals bond force has been redefined by low-temperature atomic force microscopy (54). Such force was found to depend on the radius of interacting atoms and on their molecular context. Thus, for most rigorous parameter settings, we did set to identify Van der Waals bonded contacts over gradients of progressively wider distances between potential interacting atoms, under the expectation that the number of actual contacts would plateau above the highest-energy distance, which was expected to be ≈ 3.8 Å (Supplementary Table S2A). This was verified across the entire TF-DNA dataset. We thus set to utilize an operational threshold of 4 Å for Van der Waals bond identification.

Electrostatic interactions are long-range forces (55), that follow the Coulomb's law, whereby the potential bond energy varies inversely versus the distance between opposite charges. Hydrophobic interactions can occur between any residue depending on the increase in entropy gained by the removal of hydrophobic surface area from ordered solvating water. Both electrostatic and hydrophobic interactions are thus inadequately accounted-for by static analyses of TF-DNA crystals, and were restrictively identified as pairs of juxtaposed opposite charges (electrostatic bonds) and as contacts between hydrophobic side chains and thymine methyl groups (hydrophobic contacts) (22,36) (Figure 2).

TF-DNA contacts were searched for in 84 structures in the training dataset and in 44 structures in the test dataset (Supplementary Table S1A). Invididual NUCPLOT analyses allowed to identify all non-bonded contacts that followed the 'structural rules' defined above, in each of

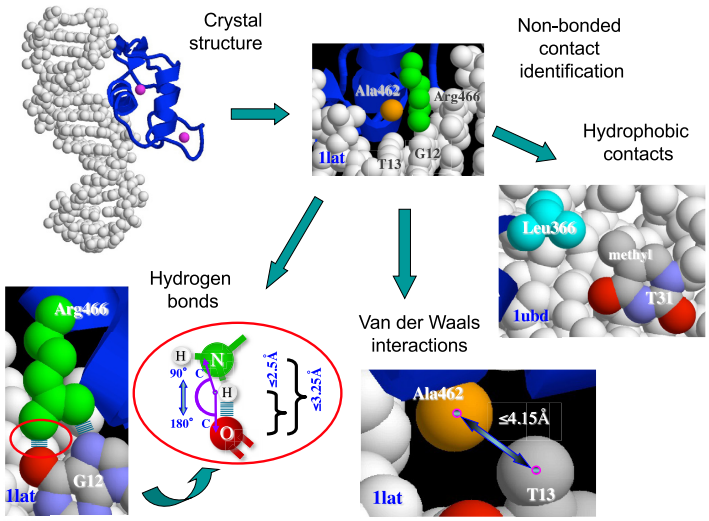

Figure 2. Strategy for decoding TF-DNA binding rules. The structure of co-crystals of TF and target DNA was displayed in ribbons and spacefill diagrams using RasMol. The 1lat, glucocorticoid receptor, and 1ubd, YY1 Zn finger, were utilized to exemplify the different types of non-bonded contacts. Hydrogen bonds are in blue-gray. Target DNA bases are in CPK (Corey, Pauling, Koltun; cyan: N, red: O, gray: C; light gray: H). Zn atoms are in magenta. Red oval: parameter ranges for hydrogen bond identification.

the 128 crystal structures analyzed. Individual NUCPLOT graphic outputs (https://github.com/marco-trerotola/TF-finder/branches; an example is presented in Figure 3) were then parsed. TF AA contacts to target DNA bases (5′, 3′ and one/two mid positions) and to ribose and phosphate groups were listed. Each contact was counted as a unit. Fractional scores were attributed to cases of interactions with multiple moieties (bases, ribose, phosphate). A Perl-based software was developed (TF-finder.pl, Supporting Data) (https://github.com/marco-trerotola/TF-finder; Ocean code DOI: 10.24433/CO.6477754.v1) that listed individual bonds as identified in NUCPLOT and collected them in bond classes, according to AA type, target DNA base or ribose or phosphate, and position in the α-helix, for each TF class (Supplementary Tables S3A–C). This allowed to obtain a first comprehensive view of TF-DNA bonds in a large, high-resolution crystal structure dataset.

DNA contacts of Zn-coordinating structures

Zn-coordinating structures (ZnCoS) comprise beta-beta-alpha Zn fingers, hormone-nuclear receptors, loop-sheet-helix and GAL4-type TFs. ZnCoS are the most abundant subset of α-helix TF (21,25). The analysis of 29 high-resolution ZnCoS-DNA complexes demonstrated high TF AA-DNA base specificities. In spite of the different frequency of use of specific AA at different positions in the binding helix, no detectable variation in binding specificity according to position was recorded (Supplementary Table S3A). It should be noted that invariant AA-DNA base specificity was predicted in the case of deterministic recognition of DNA bases by TF AA. Correspondingly predicted was a differential use of specific AA, according to the geometry of the TF-DNA binding interface (see below). Hence, these findings provided a first, key support for the correctness of our model.

Validation of the ZnCoS binding rules in Helix-Turn-Helix (HTH) and leucine zippers

A fundamental issue was if and to what extent the ZnCoS binding rules were valid for other TF classes. Thus, ZnCoS binding specificities were compared with those of HTH and leucine zippers. The large number of analyzed structures that fell within our selection criteria for each TF class (Supplementary Table S1A) allowed to reach high statistical reliability throughout this work.

In spite of the different frequency of utilization of specific AA in DNA contacts across the different TF classes (Supplementary Tables S3A–C), a striking similarity of TF AA-DNA base-binding specificity was observed. Matched top-scoring TF AA-DNA base specificity was observed for 17 out of 19 base-binding AA. The combinatorial by-chance probability for this result was a low 2.1×10^{-22}.

In addition, in spite of the large number of variables at play ($n = 240$; 20AA*4bases*3 α-helix positions), the binding specificity was equally conserved across different positions in the DNA-binding helix, throughout TF classes (Supplementary Tables S3A–C). The combinatorial probability of a by-chance occurrence of this result was 3.4×10^{-21}, consistent with a model of deterministic recognition of DNA bases by specific TF AA.

Validation of the TF-DNA binding rules in a test structure dataset

The performance of the TF-DNA binding code was assessed on a test dataset of crystallized TF-DNA complexes. This dataset only included crystal structures that were solved subsequently to those analyzed in the training dataset (21) (Supplementary Tables S3 and S4). Strikingly, all evaluable ZnCoS AA demonstrated conserved binding specificity in the two independent datasets (Supplementary Table S4C), consistent with the identification of fundamental rules of interaction TF with DNA.

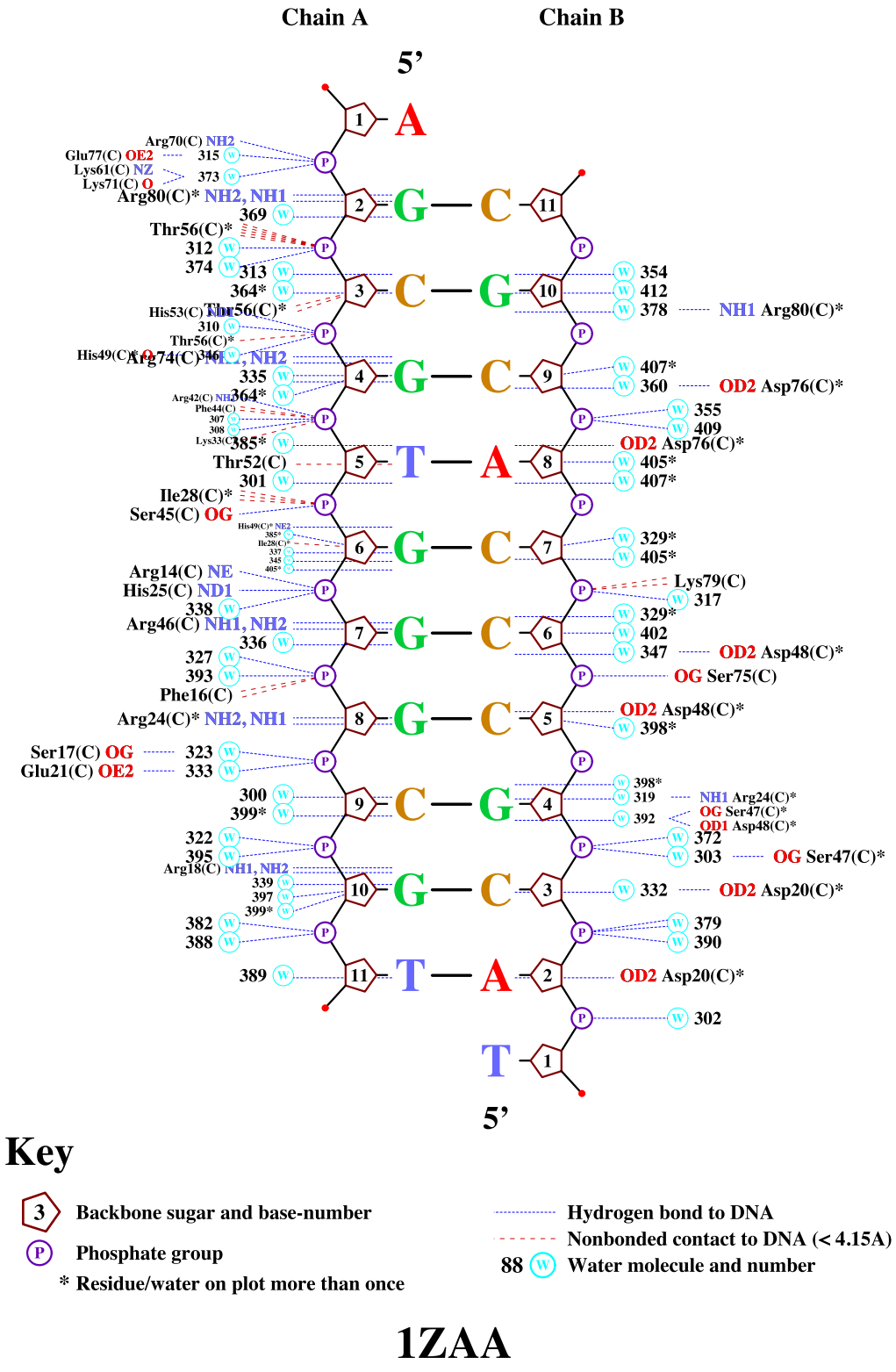

Figure 3. DNA-TF interaction diagrams. The example is provided of analysis of the 1zaa zinc finger, the murine ZIF268 immediate early gene (KROX-24) TF-DNA co-crystal (2). Two-dimensional diagram schematics of non-bonded contacts at the contacting surfaces of TF-DNA complexes. The HBPLUS program (https://www.ebi.ac.uk/thornton-srv/software/HBPLUS/) (22,36) was used to locate specific proximal donor and acceptor atom pairs. Rendering of HBPLUS non-covalent surface bond analysis was generated with NUCPLOT 2D graphics (36). Chains A (bases 1–11) and B (bases 1–11): target DNA; Chain C (AA 3–87): Zn finger. Backbone sugar and base number, AA number and water molecule numbering are as indicated in the bottom inset.

High-resolution TF-DNA structure analysis

The resolution of TF-DNA crystals affects the accuracy with which residues at the TF-DNA interface are located in space, and this may in turn affect the accuracy of identification of non-bonded contacts (21). Thus, higher resolution structures (<2.5 Å or <2.0 Å) were separately analyzed (Supplementary Table S4A), to assess the performance of the recognition code draft.

The 'all α-helix' code performed remarkably well in the high-resolution dataset, with a correspondence of 17 of 18 top scores ($P \leq 5.8 \times 10^{-11}$) and of 7 of 8 ancillary specificities (Supplementary Table S4A), together with an even sharper binding specificity profile. The same trend was maintained in structures with a resolution <2 Å. In spite of the considerably smaller sample size, the binding mode of the latter was found to correspond to the 'all α-helix' code in 14 of 15 evaluable cases (top scores; $P \leq 1.5 \times 10^{-8}$) (Supplementary Table S4A), confirming the accuracy of the TF-DNA recognition rules we had identified.

Experimental validation—meta-analysis of Zn finger mutagenesis studies

A formal possibility existed that a systematic bias could be associated with the structure of physiologically relevant TF-DNA pairs. In other words, conserved TF AA-DNA base pairs could be more frequently identified, not because of intrinsic binding propensity, but rather because of selection for a conserved function of specific DNA and TF sequences. We thus devised a totally independent search strategy for assessing selectivity in TF-DNA contacts, through a meta-analysis of large-scale Zn finger mutagenesis studies (Figure 4). The rationale of this approach was that random mutagenesis of both specific DNA and TF sequences would have selected *in vitro* AA-DNA base matching pairs, that would not depend on the *in vivo* evolutionary history of the corresponding TF and target DNA. A clear-cut set of TF AA-DNA base contact preferences was expected from the global assessment of mutagenesis studies. This binding profile ('Zn finger mutagenesis code') had, then, to be used to assess/ validate the 'TF AA-DNA base structural code'.

Mutagenized Zn finger binding preferences

Large-scale analysis of Zn finger-DNA binding specificities was conducted in several independent studies (16,23,38–45,50–52) (Figure 4). Inclusion criteria in our investigation were defined, according to those adopted in meta-analysis of clinical trial data (56). Main inclusion measures were (i) conduction of random mutagenesis of Zn finger DNA-binding AA, (ii) mutant selection by direct binding to randomized target DNA sequences and (iii) large numbers of selected mutants in each of the studies (≥100) (38–44) (Supplementary Table S5B). This strategy allowed to considerably extend previous attempts (Supplementary Table S5C–E), and to compile a comprehensive binding-specificity matrix of all available mutants ($n = 846$).

Sharp binding preferences of mutagenized residues were identified for 678 mutants (80.1%) (Supplementary Table S5B). This 'Zn finger mutagenesis code' was assessed by comparison with additional, independent mutagenesis studies, i.e. (i) a large single study of Zn finger mutagenesis (45) (Supplementary Table S5C) and (ii) a comprehensive dataset of smaller, additional mutagenesis investigations (Supplementary Table S5E). Notably, 13 of 14 evaluable binding specificities were found to correspond to those predicted by the global mutagenesis meta-analysis, thus robustly validating our analytical strategy.

Comparison of the 'Zn-finger-mutagenesis' and 'crystal-structure' rules

The 'structure' code and the 'mutagenesis' code were compared and overall fitness measures were computed. A first fitness measure was computed by contrasting the top binding preferences in the 'structure' versus 'mutagenesis' datasets. A second fitness measure was computed by a systematic comparison of observed versus expected binding profiles across the 'structure' and 'mutagenesis' datasets.

Remarkably, correspondence of top binding preferences between structure contacts datasets and mutagenized Zn finger data were observed for 14 of 16 evaluable AA ($P \leq 3.7 \times 10^{-9}$) (Table 1, highlighted; Supplementary Table S4C). The two apparent exceptions (Ile and Val) appeared of limited value, as Ile generated just 1.5 total contacts in the mutagenesis studies and the Val binding specificity in mutants corresponds to the second ranking one in α-helix structures. A comparison of binding specificities across subgroups of α-helix structures revealed that the binding preferences of Ser in HTH and leucine zipper structures better fitted those in mutagenized Zn fingers than those in Zn-CoS structures (Table 1 and Supplementary Table S4C). As ZnCoS share the same structure as the Zn finger mutants, the better fit of the latter with the HTH and leucine zipper data was far from trivial. Corresponding findings were obtained for Ala, Gly, Leu and Val, indicating shared recognition modes across structures and a global robustness of the derived recognition code.

The ZnCoS 'structure' rules were then used to predict the frequency of occurrence of specific TF AA-DNA base pairs in Zn finger mutants (Supplementary Table S5A). Observed versus expected ratios (O/E) indicated the matching of predictions of the mutagenesis code versus those from the structural code, according to the contact distribution profiles in the two datasets. O/E were correctly predicted by the ZnCoS-DNA recognition code, with an average O/E across different AA of 78.1 ± 8.1%. A similar analysis was performed using the 'all α-helix' interaction rules. Average O/E of the 'all-α-helix' code across all AA was 95.7 ± 7.5%, consistent with a better predictive power of the largest dataset-based recognition code (Supplementary Table S5B).

As predicted by a deterministic model of TF AA-DNA base recognition, the binding specificities of Zn finger mutants proved essentially invariant versus position in the α-helix (Supplementary Table S4B). The combinatorial probability of a by-chance occurrence of this match was 8.3×10^{-25}.

The binding specificities of the structural code and the mutagenesis code were then compared over the entire datasets. This showed matching of the binding preference profiles of the structural code and the mutagenesis code to a striking level of accuracy ($P = 3.1 \times 10^{-33}$), consistent

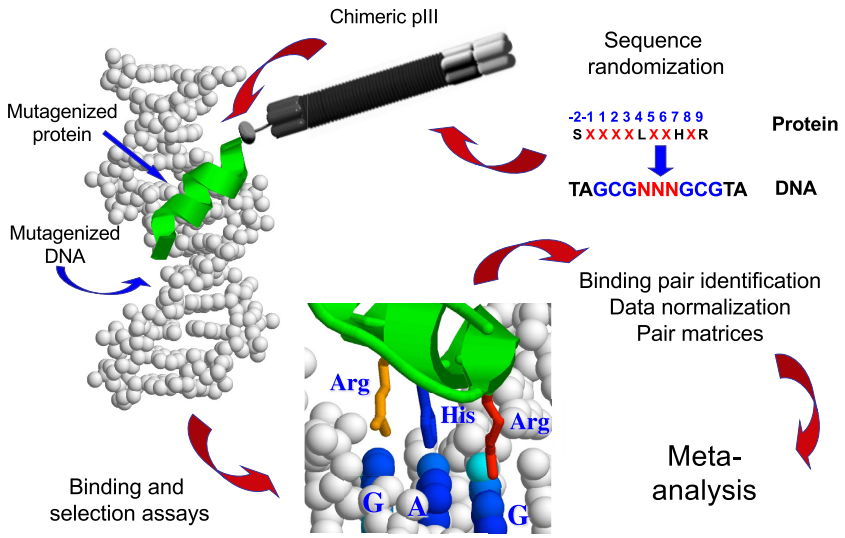

Figure 4. Strategy for the analysis of the DNA mutagenesis studies. (From upper right corner) Scheme of the TF and target DNA mutagenesis (residues are numbered from the first helical position (+1); DNA: Zif268 operator sequence, as adapted from (45)) → chimeric phage construction and selection → AA-DNA base-pair identification, quantitative analysis and normalization → individual-study quantitative analysis → comprehensive meta-analysis. Figures were generated with RasMol. TF are in ribbons, stick and spacefill diagrams.

with correspondence to fundamental rules of TF binding to DNA bases.

A DNA recognition code

The conserved binding specificities observed across the co-crystallized TF-DNA structures and the contact residue mutagenesis datasets allowed the pooling of all of the corresponding binding data, for an additional refinement of the computed frequencies of TF AA-DNA base-binding pairs (Table 1 and Supplementary Table S4). This AA-DNA base recognition code is graphically presented in Figure 5.

Comparison between the presently derived code with previous seminal studies is presented in Supplementary Table S5C–E. We refer the reader to each instance of comparison for detailed analysis. Briefly, (i) frequent instances, such as G recognition by Arg and C recognition by acidic residues, were identified by essentially all codes; (ii) rare instances of DNA base-TF AA recognition appeared only identifiable in the largest of the datasets analyzed; (iii) binding preferences, i.e. quantitative analysis of ≥ 1 DNA base recognition, were only identifiable in the full, merged datasets of crystal structures and mutagenesis studies we had generated.

Fundamental features of TF binding to DNA

The TF-DNA recognition code we have identified is deterministic in nature. For the sake of clarity, it is worth remarking that recognition of four (4) DNA bases by twenty (20) AA implies that multiple AA can recognize individual DNA bases. Guanine is bound by positively charged Arg, His, Lys, Cytosine is selected for by aspartic and glutamic acid. Adenine is recognized by amide and phenolic groups. Thymine is selected for by multiple AA classes, main among them those of bulky hydrophobic AA, together with those the low-space occupancy (Ala, Gly) and of alcoholic side chains (Ser, Thr, Cys) AA (Supplementary Tables S4 and S5; Figure 5). The methyl group at carbon 5 (C5M) of Thymine plays a key role in AA binding, and 95% of Ala contacts to Thymine occur as hydrophobic contacts to C5M. Reciprocally, of the 51 bonded contacts to the Thymine C5M, two thirds were shown to occur versus hydrophobic AA; 94% of the latter were Val or Ile. The converse was also found to be true, as 100% of Ile and 79% of Val bound Thymine C5M.

A special case is that of Trp, which frequently contacts DNA in HTH. However, in 27 out of 28 instances Trp only contacts the DNA backbone (Supplementary Table S3B). Hence, it appears to have limited DNA base discrimination capacity, rather working as a non-base-specific enhancer of binding of HTH to DNA. Trp can exert contacts to DNA in the bacterial trp repressor/operator complex (57) and in TFIIIA (58). In both cases, Trp only contacts the DNA backbone.

Nearest-neighbor analysis of target DNA bases

Early findings indicated that the vast majority of interactions that occur between a TF and individual DNA bases were independent from each other, and generated additive binding (1,4). However, alternative models of di-nucleotide recognition were proposed (8,53). We thus performed a nearest-neighbor analysis to detect whether preferential association between adjacent DNA bases occurred at TF target sites (28,59).

The observed distribution of the 1110 computed single-base contacts was 239 (21.5%) for A, 301 (27.1%) for T, 339 for G (30.5%), 231 for C (20.8%) (Supplementary Table S6). Predicted occurrences were 25% contacts for each base. Hence, albeit observed values were rather close to the expected value, a larger utilization of purines is suggested to be of functional value.

Expected values of di-nucleotides in DNA target sequences were then estimated under the assumption of random pairwise association. Observed values of contacted

Table 1. DNA bases recognized by transcription factor residues

	G	A	T	C	sum	P values
Ala						
ZnCoS [a]	1 [b]	0	2 [c]	0	3	0.299781 [d]
HTH	1.2	0.3	9.8	0.2	11.5	0.000011
Leu-zip.	0.3	2.6	17.3	0	20.2	1.295197e-008
α-helix	2.5	2.9	29.1	0.2	34.7	2.431388e-014
Mutant	3	2.5	30.5	6	42	3.670886e-011
All	5.5	5.4	59.6	6.2	76.7	0
Arg						
ZnCoS	48.9	1	4.7	3.2	57.8	0
HTH	69.5	5.8	13	1	89.3	0
Leu-zip.	26	5.6	1.3	5	37.9	1.243052e-008
α-helix	144.3	12.4	19	9.2	184.9	0
Mutant	263	4.5	5.5	3	276	0
All	407.3	16.9	24.5	12.2	460.9	0
Asn						
ZnCoS	0.5	6.8	0.5	0.8	8.6	0.000707
HTH	3.6	38.2	9.8	4.8	56.4	1.003164e-011
Leu-zip.	0.2	7.9	9	4.9	22	0.03052431
α-helix	4.3	52.9	19.3	10.5	87	2.031708e-014
Mutant	2.5	49.5	6.5	2	60.5	0
All	6.8	102.4	25.8	12.5	147.5	0
Asp						
ZnCoS	0.7	2.6	0.8	12.3	16.4	0.000218
HTH	0	0	0	3	3	0.007383
Leu-zip.	0	0	0	0	0	
α-helix	0.7	2.6	0.8	15.3	19.4	5.345477e-006
Mutant	5.5	1	2.5	71.5	80.5	0
All	6.2	3.6	3.3	86.8	99.9	0
Cys						
ZnCoS	0	0	1	1	2	0.572407
HTH	1.5	1.5	2.5	0	5.5	0.572407
Leu-zip.	0	0	0	0	0	
α-helix	1.5	1.5	3.5	1	7.5	0.549668
Mutant	0	0	0	0	0	
All	1.5	1.5	3.5	1	7.5	0.549668
Gln						
ZnCoS	1	11	0.5	0.5	13	2.631283e-006
HTH	8.6	26.2	14.8	8.8	58.4	0.004492
Leu-zip.	0	0.5	2.5	0	3	0.11161
α-helix	9.6	37.7	17.8	9.3	74.4	2.297967e-006
Mutant	1	48.5	11.5	1.5	62.5	0
All	10.6	86.2	29.3	10.8	136.9	0
Glu						
ZnCoS	0.3	3.6	1.3	23.6	28.8	2.029343e-011
HTH	0.5	0.7	3.9	11.2	16.3	0.000347
Leu-zip.	0.8	2.1	1	11.8	15.7	0.000083
α-helix	1.6	6.4	6.2	46.6	60.8	0
Mutant	4	5	4	14.5	27.5	0.024104
All	5.6	11.4	10.2	61.1	88.3	0
Gly						
ZnCoS	1	0	4	0	5	0.03511
HTH	4.4	1	7.6	1.5	14.5	0.053427
Leu-zip.	0	0	0	0	0	
α-helix	5.4	1	11.6	1.5	19.5	0.001996
Mutant	6.5	1.5	6.5	1.5	16	0.261464
All	11.9	2.5	18.1	3	35.5	0.000172
His						
ZnCoS	5	1.5	1	0.5	8	0.071898
HTH	6	0.3	2.2	0.3	8.8	0.007383
Leu-zip.	12.1	0.5	2.1	1.1	15.8	0.000018
α-helix	23.1	2.3	5.3	1.9	32.6	2.502147e-008
Mutant	29.5	2	5	8	44.5	2.160181e-009
All	52.6	4.3	10.3	9.9	77.1	0
Ile						
ZnCoS	0	0	1	0	1	0.391625
HTH	0	6.2	7.5	0.4	14.1	0.002222
Leu-zip.	0	0	0	0	0	
α-helix	0	6.2	8.5	0.4	15.1	0.002222
Mutant	0	1	0.5	0	1.5	0.391625
All	0	7.2	9	0.4	16.6	0.000895

	G	A	T	C	sum	P values
Leu						
ZnCoS	0	1	1	0	2	0.572407
HTH	0.3	0	2.5	0.5	3.3	0.11161
Leu-zip.	0	0	0	0	0	
α-helix	0.3	1	3.5	0.5	5.3	0.03511
Mutant	0	0.5	6	3.5	10	0.012858
All	0.3	1.5	9.5	4	15.3	0.002905
Lys						
ZnCoS	40.9	1.8	4.5	2.5	49.7	0
HTH	18.6	3	6	0.3	27.9	1.380057e-006
Leu-zip.	0	2	0.8	0	2.8	0.299781
α-helix	59.5	6.8	11.3	2.8	80.4	0
Mutant	16.5	3	12	0	31.5	0.000073
All	76	9.8	23.3	2.8	111.9	0
Met						
ZnCoS	0	0	0	0	0	
HTH	0	1.9	0.3	1	3.2	0.299781
Leu-zip.	0	0	0	0	0	
α-helix	0	1.9	0.3	1	3.2	0.299781
Mutant	0	1	1	0	2	0.572407
All	*0*	2.9	*1.3*	*1*	5.2	0.283886
Phe						
ZnCoS	0	0.2	1.2	0.5	1.9	0.391625
HTH	1	0	0.8	0	1.8	0.572407
Leu-zip.	0	0	0.5	0	0.5	0.391625
α-helix	1	0.2	2.5	0.5	4.2	0.299781
Mutant	0	0	0	0	0	
All	1	0.2	2.5	0.5	4.2	0.299781
Pro						
ZnCoS	0	0	1	0	1	0.391625
HTH	0	3.1	4.9	0	8	0.029291
Leu-zip.	0	0	0	0	0	
α-helix	0	3.1	5.9	0	9	0.011726
Mutant	0	0	0	0	0	
All	0	3.1	5.9	0	9	0.011726
Ser						
ZnCoS	4	2	2.7	1.8	10.5	0.801252
HTH	9.7	3.7	12	6.3	31.7	0.171797
Leu-zip.	0	0	1	0	1	0.391625
α-helix	13.7	5.7	15.7	8.1	43.2	0.103092
Mutant	9	4	37.5	9	59.5	1.994801e-010
All	22.7	9.7	53.2	17.1	102.7	4.560347e-009
Thr						
ZnCoS	0	0	3.5	3.5	7	0.046012
HTH	0.5	4.2	4	3	11.7	0.271449
Leu-zip.	0	0	0	0	0	
α-helix	0.5	4.2	7.5	6.5	18.7	0.050835
Mutant	11.5	5.5	60	33.5	110.5	7.127632e-014
All	12	9.7	67.5	40	129.2	6.883383e-015
Trp						
ZnCoS	0	0	0	0	*0*	
HTH	0	0	0	0	*0*	
Leu-zip.	0	0	0	0	*0*	
α-helix	0	0	0	0	*0*	
Mutant	0	0	0	0	*0*	
All	*0*	0	0	0	*0*	
Tyr						
ZnCoS	0.2	0	0	0.2	0.4	
HTH	1	3.2	1.1	1.8	7.1	0.665885
Leu-zip.	0	0	0	0	0	
α-helix	1.2	3.2	1.1	2	7.5	0.665885
Mutant	0	2	0	0	2	0.11161
All	1.2	5.2	1.1	2	9.5	0.188811
Val						
ZnCoS	0.5	0	3.5	0	4	0.007383
HTH	1.5	8.1	12.7	1	23.3	0.001327
Leu-zip.	0.7	2.5	6.7	1.2	11.1	0.029291
α-helix	2.7	10.6	22.9	2.2	38.4	2.239429e-006
Mutant	2	0.5	7	10	19.5	0.004203
All	4.7	11.1	29.9	12.2	57.9	0.000024

Table 1 reports the absolute numbers of TF AA-DNA base contacts identified in comprehensive analyses of TF-DNA crystals (lines 1-4), mutagenesis studies (line 5) and in the all-comprehensive meta-analysis (line 6).

[a]: ZnCOS structures, HTH structures, Leu-zip.: leucine zipper structures; α-helix: sum of ZnCoS, HTH and leucine zipper contacts; Mutant: mutagenized Zn fingers; All: sum of α-helix and mutagenized Zn finger AA-DNA base contacts.

[b]: Table 1 reports the absolute values of TF AA-DNA base contacts.

[c]: Top-scoring contacts across different TF classes are highlighted in yellow; light-yellow highlights indicate weaker correlations.

[d]: Probability of distribution homogeneity, χ^2 analysis.

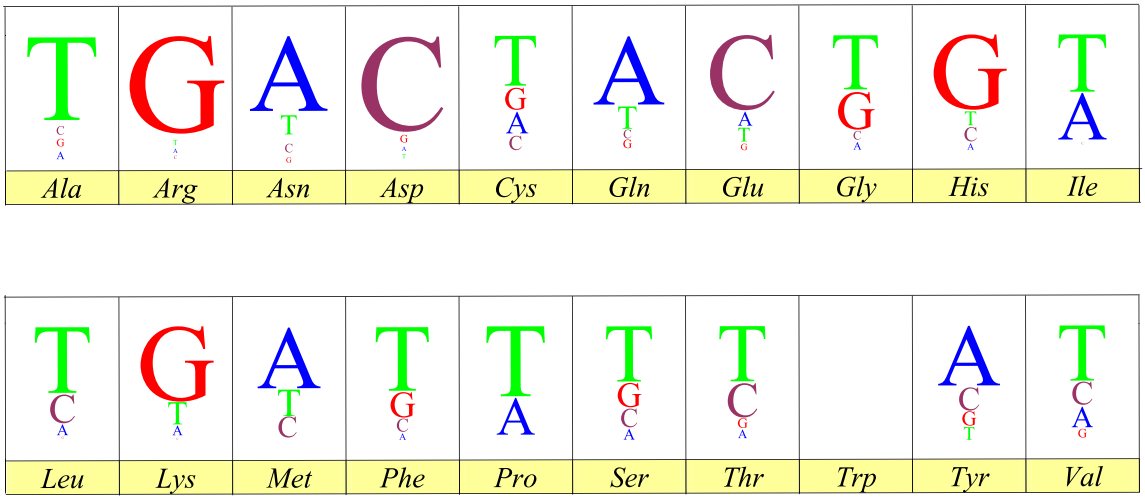

Figure 5. TF-DNA binding code. Weblogo (http://weblogo.berkeley.edu/) (60) of the binding frequencies of amino acids (bottom) to DNA bases (top, in color). The font size of the target DNA base is proportional to the corresponding binding frequency.

DNA di-nucleotides were subsequently assessed. The distribution of the 104 (unordered) pairs of contacted DNA bases observed in leucine zippers is shown in Supplementary Table S6A. Those for HTH and ZnCoS are presented in Supplementary Tables S6B and S6C, respectively.

Observed versus expected distributions of adjacent di-nucleotides are contrasted in Supplementary Table S6D. This indicated vast correspondence of observed versus expected di-nucleotides as predicted from single-base frequencies (Supplementary Table S6Dii, 'Overall'). These findings strongly supported TF recognition of single-nucleotides, as independent target bases, rather than of di-nucleotides (8,53). However, we noticed a few instances of deviation from expected frequencies. The largest ones were shown by the adenine/cytosine pair of adjacent bases, which showed an increase in observed frequency by 27% versus the expected value, and by the Adenine/Guanine pair, which showed a 31% decrease versus the expected value (Supplementary Table S6Dii, 'Overall').

TF AA-DNA base binding geometry

The nearest-neighbor analysis supported a model of deterministic TF AA recognition of target DNA via a one-base alphabet. However, the few instances of deviation from expected frequencies suggested that additional factors were at work. As noted above, the binding specificity of TF AA did not vary according to position in the binding helix, whether in co-crystallized structures (Supplementary Table S4A) or in Zn finger mutants (Supplementary Table S4B). On the other hand, different AA recognizing the same base were used at vastly different frequency at different positions in the DNA-binding α-helix (Supplementary Tables S4A–C). As selection of a specific DNA base can be achieved by more than one AA, and individual AA present with distinct side chain length and volume, this suggested mechanisms of 'choice' between iso-binding AA. The broadest instance appeared that of recognition of Thymine, which is achieved by a large set of different AA. We thus hypothesized that Thymine could be recognized by best-fitting AA, as allowed by the geometry of the other residues in the DNA-binding triplet.

We assessed the predictive power of this model. Among Thymine-binding AA, Ala possesses the shortest side chain. Hence, Ala was expected to be preferentially placed in the center of TF DNA-binding triplets, and to be flanked by AA with longer side chains (Supplementary Table S7B). This was shown to be true in 9 of 12 evaluable cases (75%), versus the expected 33%, if Ala occurrence was dictated by chance alone.

Zn finger mutagenesis studies provided an unbiased test of the 'geometry-driven' model of 'choice' of specific AA at specific positions in the DNA-binding AA triplet. We thus screened Zn finger mutagenesis studies for DNA sequences, that were expected to require the binding of AA side chains of different length. A distinct experimental set was provided by recognition of target GXG DNA sequences. Analysis of GCG versus GTG DNA target sequences in mutagenized structures. led to identify Arg-Glu-Arg for GCG and Arg-Ala-Ser for GTG contacts (45). Both sets of contacts fulfil the DNA binding code rules. However, the short side chain of Ala brings the TF much closer to Thymine in the GTG sequence and leads to recognition of the 3′ G by the shorter Ser side chain (Supplementary Table S7B), suggesting a wedge-like shape of the DNA-binding TF surface.

We went on to validate this model in co-crystals of GCG- or GTG-binding TF (Figure 6A) (60–62). A Zif 268 variant (1a1g) and the interferon regulatory factor 2 (2irf) were identified, that bound GCG or GTG, respectively. In the case of the Zif 268 variant, the long side chain of the central Glu (Figure 6A) allows to host Arg both upstream and downstream, and to have the Arg side chains in extended conformation (Supplementary Table S7B). On the other hand, a central Ala residue in the IRF 2 HTH induces a much closer contact of the TF α-helix to the DNA (Figure 6B). This constrains the upstream Arg to a less extended conformation and obliges to the presence of a flexible down-

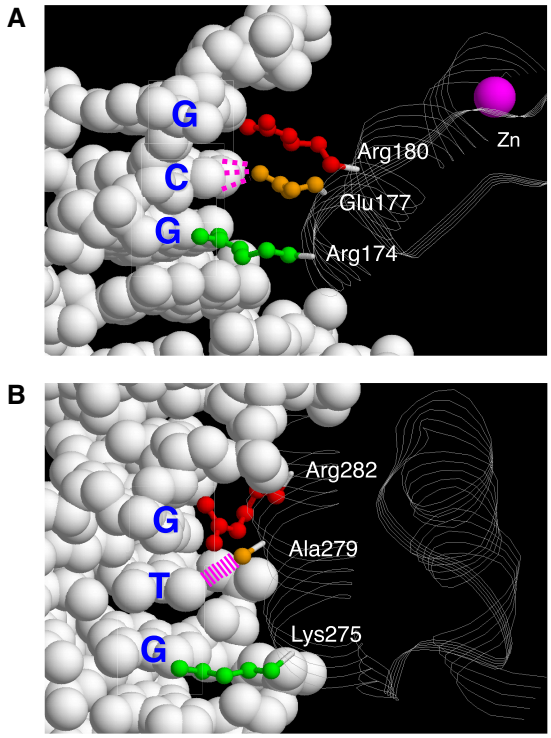

Figure 6. Steric constraints on TF binding to DNA. (**A**) Zn finger Zif 268 variant / target DNA complex (1a1g) (61). Magenta dotted line: hydrogen bond between Glu177 and C. Zn: coordinating Zn atom (magenta). (**B**) HTH Interferon regulatory factor 2 / target DNA complex (2irf) (62). Magenta dashed line: hydrophobic contact between Ala279 and T. Figures were generated with RasMol. TF are in Strands diagrams.

stream residue like Lys that can wrap around its target G (Figure 6B).

Hence, both the mutagenized TF and crystallized structure data support the notion that best-fitting side chain geometries of DNA binding-AA play a role in dictating preferences among 'iso-binding' AA.

CONCLUSIONS

Complex modes of DNA sequence recognition have been proposed, that include cooperative binding, flanking DNA sequence recognition (8,10), dependence on DNA structure (1–7,9), induced folding of TF domains and dynamic recognition of target DNA sequences (12). However, fundamental knowledge on the basic affinity of individual AA for target DNA bases remained missing.

Our findings fill-in this gap, through large-scale, unbiased analyses AA-DNA bases contacts in TF/DNA co-crystals, with orthogonal validation through meta-analysis of random mutagenesis studies of TF AA and target DNA sequences.

These findings provide novel insight into the fundamental rules of binding of individual TF AA and target DNA bases, and identify a deterministic code of TF-DNA recognition. The broad, TF class-independent validity of this TF AA-DNA base recognition code, supports the identification of fundamental mechanisms of DNA recognition by TF. Novel understanding of processes of regulation of gene expression may potentially be reached from using the fundamental rules we have identified, together with integration with higher-order complexity approaches for recognition of target DNA sequences.

DATA AVAILABILITY

The TF-finder software is available at: https://github.com/marco-trerotola/TF-finder Ocean code DOI: 10.24433/CO.6477754.v1.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank thank N. Luscombe, S. Jones, P. Simeone and F. Ciccarelli for discussion and support and S. Maerkl and S. Quake for making available primary pre-publication data.

FUNDING

The financial support of Telethon - Italy (#GGP02353), Foundation Compagnia di San Paolo – Torino, Italy (#2489IT), Italian Association for Cancer Research, Italy (#2114_1096), Italian Ministry for the University and Research (PS-DM29174), ABO Project SpA, Italy (CH01D0081) and Marie Curie-TOK Fellowship – EC VI Framework Programme (#014541) is gratefully acknowledged.
Conflict of interest statement. None Declared.

REFERENCES

1. Marabotti,A., Spyrakis,F., Facchiano,A., Cozzini,P., Alberti,S., Kellogg,G.E. and Mozzarelli,A. (2008) Energy-based prediction of amino acid-nucleotide base recognition. *J. Comput. Chem.*, **29**, 1955–1969.
2. Pavletich,N.P. and Pabo,C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, **252**, 809–817.
3. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
4. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
5. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, 500–503.
6. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
7. Schneider,B., Cerny,J., Svozil,D., Cech,P., Gelly,J.C. and de Brevern,A.G. (2014) Bioinformatic analysis of the protein/DNA interface. *Nucleic Acids Res.*, **42**, 3381–3394.
8. Le,D.D., Shimko,T.C., Aditham,A.K., Keys,A.M., Longwell,S.A., Orenstein,Y. and Fordyce,P.M. (2018) Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. USA*, **115**, E3702–E3711.

9. Ng,A.H.M., Khoshakhlagh,P., Rojo Arias,J.E., Pasquini,G., Wang,K., Swiersy,A., Shipman,S.L., Appleton,E., Kiaee,K., Kohman,R.E. *et al.* (2021) A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.*, **39**, 510–519.
10. Afek,A., Schipper,J.L., Horton,J., Gordan,R. and Lukatsky,D.B. (2014) Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. USA*, **111**, 17140–17145.
11. Gordan,R., Hartemink,A.J. and Bulyk,M.L. (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.*, **19**, 2090–2100.
12. Jana,T., Brodsky,S. and Barkai,N. (2021) Speed-Specificity trade-offs in the transcription factors search for their genomic binding sites. *Trends Genet.*, **37**, 421–432.
13. Avsec,Ž., Weilert,M., Shrikumar,A., Krueger,S., Alexandari,A., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A. *et al.* (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, **53**, 354–366.
14. Gupta,A., Christensen,R.G., Bell,H.A., Goodwin,M., Patel,R.Y., Pandey,M., Enuameh,M.S., Rayla,A.L., Zhu,C., Thibodeau-Beganny,S. *et al.* (2014) An improved predictive recognition model for cys(2)-his(2) zinc finger proteins. *Nucleic Acids Res.*, **42**, 4800–4812.
15. 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.
16. Choo,Y. and Klug,A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.
17. Matthews,B.W. (1988) Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
18. Brennan,R.G. and Matthews,B.W. (1989) Structural basis of DNA-protein recognition. *Trends Biochem. Sci*, **14**, 286–290.
19. Wolfe,S.A., Grant,R.A., Elrod-Erickson,M. and Pabo,C.O. (2001) Beyond the "recognition code": structures of two cys2his2 zinc finger/TATA box complexes. *Structure (Camb.)*, **9**, 717–723.
20. Suzuki,M. and Yagi,N. (1994) DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl. Acad. Sci. USA*, **91**, 12357–12361.
21. 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.
22. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, reviews001.1.
23. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
24. Beerli,R.R. and Barbas,C.F. 3rd (2002) Engineering polydactyl zinc-finger transcription factors. *Nat. Biotechnol.*, **20**, 135–141.
25. Jamieson,A.C., Miller,J.C. and Pabo,C.O. (2003) Drug discovery with engineered zinc-finger proteins. *Nat. Rev. Drug Discov*, **2**, 361–368.
26. Zeng,J., Yan,J., Wang,T., Mosbrook-Davis,D., Dolan,K.T., Christensen,R., Stormo,G.D., Haussler,D., Lathrop,R.H., Brachmann,R.K. *et al.* (2008) Genome wide screens in yeast to identify potential binding sites and target genes of DNA-binding proteins. *Nucleic Acids Res.*, **36**, e8.
27. Yan,J., Qiu,Y., Ribeiro dos Santos,A.M., Yin,Y., Li,Y.E., Vinckier,N., Nariai,N., Benaglio,P., Raman,A., Li,X. *et al.* (2021) Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**, 147–151.
28. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
29. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
30. Liu,J. and Stormo,G.D. (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res.*, **33**, e141.
31. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
32. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays*, **24**, 466–475.
33. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
34. Alberti,S. (1997) The origin of the genetic code and protein synthesis. *J. Mol. Evol.*, **45**, 352–358.
35. Merritt,E.A. and Bacon,D.J. (1997) Raster3D - photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.
36. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (1997) NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res.*, **25**, 4940–4945.
37. Cozzini,P., Fornabaio,M., Marabotti,A., Abraham,D.J., Kellogg,G.E. and Mozzarelli,A. (2002) Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 1. Models without explicit constrained water. *J. Med. Chem.*, **45**, 2469–2483.
38. Choo,Y. and Klug,A. (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. USA*, **91**, 11168–11172.
39. Jamieson,A.C., Wang,H. and Kim,S.H. (1996) A zinc finger directory for high-affinity DNA recognition. *Proc. Natl. Acad. Sci. USA*, **93**, 12834–12839.
40. Desjarlais,J.R. and Berg,J.M. (1994) Length-encoded multiplex binding site determination: application to zinc finger proteins. *Proc. Natl. Acad. Sci. USA*, **91**, 11099–11103.
41. Isalan,M., Klug,A. and Choo,Y. (1998) Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry*, **37**, 12026–12033.
42. Segal,D.J., Dreier,B., Beerli,R.R. and Barbas,C.F. 3rd (1999) Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc. Natl. Acad. Sci. USA*, **96**, 2758–2763.
43. Jamieson,A.C., Kim,S.H. and Wells,J.A. (1994) In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, **33**, 5689–5695.
44. Liu,Q., Xia,Z. and Case,C.C. (2002) Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *J. Biol. Chem.*, **277**, 3850–3856.
45. Choo,Y. and Klug,A. (1994) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl. Acad. Sci. USA*, **91**, 11163–11167.
46. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
47. Donald,J.E. and Shakhnovich,E.I. (2005) Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Res.*, **33**, 4455–4465.
48. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
49. Maienschein-Cline,M., Dinner,A.R., Hlavacek,W.S. and Mu,F. (2012) Improved predictions of transcription factor binding sites using physicochemical features of DNA. *Nucleic Acids Res.*, **40**, e175.
50. Desjarlais,J.R. and Berg,J.M. (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl. Acad. Sci. USA.*, **89**, 7345–7349.
51. Greisman,H.A. and Pabo,C.O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657–661.
52. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA*, **98**, 7158–7163.
53. Klug,A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Ann. Rev. Biochem.*, **79**, 213–231.
54. Kawai,S., Foster,A.S., Bjorkman,T., Nowakowska,S., Bjork,J., Canova,F.F., Gade,L.H., Jung,T.A. and Meyer,E. (2016) Van der waals interactions and the limits of isolated atom models at interfaces. *Nat. Commun.*, **7**, 11559.
55. Prelesnik,J.L., Alberstein,R.G., Zhang,S., Pyles,H., Baker,D., Pfaendtner,J., De Yoreo,J.J., Tezcan,F.A., Remsing,R.C. and Mundy,C.J. (2021) Ion-dependent protein–surface interactions from intrinsic solvent response. *Proc. Natl. Acad. Sci.*, **118**, e2025121118.
56. Yusuf,S., Peto,R., Lewis,J., Collins,R. and Sleight,P. (1985) Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardiovasc. Dis.*, **27**, 335–371.

57. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
58. Nolte,R.T., Conlin,R.M., Harrison,S.C. and Brown,R.S. (1998) Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc. Natl. Acad. Sci.*, **95**, 2938.
59. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-Binding specificities of human transcription factors. *Cell*, **152**, 327–339.
60. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
61. Elrod-Erickson,M., Benson,T.E. and Pabo,C.O. (1998) High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure*, **6**, 451–464.
62. Fujii,Y., Shimizu,T., Kusumoto,M., Kyogoku,Y., Taniguchi,T. and Hakoshima,T. (1999) Crystal structure of an IRF-DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. *EMBO J.*, **18**, 5028–5041.