PLOS ONE

RESEARCH ARTICLE

Sequence-Specific Recognition of DNA by Proteins: Binding Motifs Discovered Using a Novel Statistical/Computational Analysis

David Jakubec[1,2], Roman A. Laskowski[3], Jiri Vondrasek[1]*

1 Institute of Organic Chemistry and Biochemistry, Prague 6, Czech Republic, **2** Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University in Prague, Prague 2, Czech Republic, **3** EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

* jiri.vondrasek@uochb.cas.cz

Citation: Jakubec D, Laskowski RA, Vondrasek J (2016) Sequence-Specific Recognition of DNA by Proteins: Binding Motifs Discovered Using a Novel Statistical/Computational Analysis. PLoS ONE 11(7): e0158704. doi:10.1371/journal.pone.0158704

Editor: Narayanaswamy Srinivasan, Indian Institute of Science, INDIA

Received: April 11, 2016

Accepted: June 21, 2016

Published: July 6, 2016

Data Availability Statement: The lists of PDB entries considered at the various maximum sequence identity levels, the 3D coordinates and the interaction energies of the amino acid side chain - DNA residue pairs, and the corresponding interaction energy profiles can be found on figshare with DOI 10.6084/m9.figshare.3462785.

Funding: This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic (KONTAKT II programme LH11020, http://www.msmt.cz/), and the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic (research project No. Z40550506, http://

Abstract

Decades of intensive experimental studies of the recognition of DNA sequences by proteins have provided us with a view of a diverse and complicated world in which few to no features are shared between individual DNA-binding protein families. The originally conceived direct readout of DNA residue sequences by amino acid side chains offers very limited capacity for sequence recognition, while the effects of the dynamic properties of the interacting partners remain difficult to quantify and almost impossible to generalise. In this work we investigated the energetic characteristics of all DNA residue—amino acid side chain combinations in the conformations found at the interaction interface in a very large set of protein—DNA complexes by the means of empirical potential-based calculations. General specificity-defining criteria were derived and utilised to look beyond the binding motifs considered in previous studies. Linking energetic favourability to the observed geometrical preferences, our approach reveals several additional amino acid motifs which can distinguish between individual DNA bases. Our results remained valid in environments with various dielectric properties.

Introduction

Interactions of deoxyribonucleic acid (DNA) with proteins form the basis of several essential processes of cellular physiology. These interactions display various levels of specificity towards the designated DNA base sequences. For example, the interactions of DNA with repair enzymes must display low sequence preferences if genome integrity is to be maintained [1–3], and histone proteins involved in nucleosomes have been shown to promote non-specific association with nucleic acids [4–6]. On the other hand, for processess such as regulation of gene expression by transcription factor proteins, DNA sequence recognition with high specificity is critical.

Experimental studies utilising X-ray diffraction or nuclear magnetic resonance spectroscopy have been actively used to explore atomic-level details of proteins, nucleic acids, and their

www.uochb.cz/). Access to computing and storage facilities was provided by ELIXIR CZ and the National Grid Infrastructure MetaCentrum, administered under the programme "Projects of Large Infrastructure for Research, Development, and Innovations." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

complexes for more than half a century. The RSCB Protein Data Bank (PDB) currently hold over 3,000 structures of protein—DNA complexes obtained from a variety of organisms by a range of experimental methods [7, 8]. Structural and biochemical studies of proteins and their cognate DNA sequences were recently performed for some whole organisms [9, 10].

Years of analyses of experimental structures have revealed two principal contributions to the process of specific DNA sequence recognition. The base readout mechanism involves local interactions between the protein DNA-binding domain and the target base sequence, predominantly in the form of a matching pattern of bidentate hydrogen bond donor and acceptor groups [11]. Asparagine and glutamine are capable of distinguishing between the Hoogsteen edge of adenine and the other DNA bases, while specific recognition of guanine by these amino acids is possible from the sugar edge. In addition, arginine can recognise the Hoogsteen edge of guanine [12].

However, the local, linear model of DNA sequence recognition by a complementary pattern of hydrogen bond donor and acceptor groups was found to be incomplete. In some protein—DNA complexes, the readout of the nucleic acid shape can be equally, or even more, important [11]. The DNA often adapts non-canonical forms in the interaction region, and its propensity to form various local distortions is dependent on a larger base sequence context [13–18]. GC-rich regions of the genome have a predisposition to adopt A-like conformations, while high AT content results in a narrowing of the minor groove, creating more negative electrostatic potential, a feature universally recognised by arginine side chains [17, 19].

Base readout and DNA shape recognition are both utilised to some extent in the majority of protein—DNA complexes. The formation of hydrogen bonds between the protein and the nucleic acid can induce a sequence-dependent distortion, which may, in turn, enable the formation of a new set of specific contacts. Therefore, the two modes cannot be separated if a complete description of the recognition process is to be obtained [13, 16, 20]. Structural data suggest that while the amino acid composition of the interface often provides sufficient information to distinguish between individual families of transcription factors, subtle differences in the dynamic properties of the cognate DNA region can guide the higher-resolution recognition by specific members of a single protein family [11, 16, 21].

In spite of these insights, no recognition code applicable to all protein families has been described to date, although recognition codes for a few genome editing enzymes are known [11, 22]. Current knowledge of the non-local properties of large blocks of DNA residues is lacking, while the dynamic aspects of the protein—nucleic acid interaction remain difficult to investigate both theoretically and experimentally. On the other hand, studies of amino acid—DNA base interactions, which probe atomic-level details of the direct readout mechanism, can readily be performed. While limited experimental data on these interactions are available [23, 24], computer technology has enabled analyses that simultaneously investigate the binding mechanism in thousands of protein—DNA complexes. Indeed, a substantial part of our understanding of the interactions involving these elementary biomolecular building blocks has been derived from studies utilising bioinformatics and other computational approaches.

The pioneering work of Berg and von Hippel combining the experimental results available at that time with a statistical mechanical framework offered one of the first rigorous theoretical treatments of specificity in protein—DNA interactions [25].

Mandel-Gutfreund and Margalit were among the first to utilise a data set of three-dimensional structures to derive contact potentials for prediction of protein—DNA interaction sites. They found that amino acids that carry bidentate hydrogen bond donor and acceptor groups and therefore enable DNA base recognition in a one-to-one manner, are strongly favoured at the interface [26]. Luscombe *et al.* also observed significant correlations between the populations of the same amino acid side chains and their cognate DNA bases. In addition, some other

contacts that did not feature bidentate hydrogen bonding motifs were dubbed "context-specific." Although the amino acids involved could not by themselves distinguish between individual DNA bases, their presence at the interface was deemed essential for the stabilisation of the respective complexes [27].

Dror *et al.* have recently performed a detailed analysis of the binding mechanisms *via* which homeodomains recognise their DNA binding sites. By combining protein and DNA sequence and shape covariation analysis with binding data obtained from high-throughput methods, specific positions containing amino acids facilitating DNA shape recognition were uncovered in the N-terminal tail of the homeodomain [28]. This study thus highlighted the importance of DNA geometry in binding site recognition. Further effort was also made in the study of DNA shape readout by other transcription factor protein families [29].

Many online databases focusing on protein—DNA complexes have been established over the past decade, their functionality ranging from simple repositories to sites offering complex analytical tools [27, 30–33].

The aforementioned studies of base readout have been based on statistical analysis and decomposition of the interfacial region of existing experimental structures. This treatment does not explicitly consider the physico-chemical characteristics of the interacting molecules. A different approach, utilising the methods of computational chemistry, is possible. Indeed, theoretical studies to calculate the interaction properties of amino acid—DNA base dimers have been conducted at both the quantum mechanical (QM) and empirical potential levels [34–40].

In this work, we combine an approach based on statistical analysis of existing experimental structures of protein—DNA complexes with molecular mechanical (MM) calculations. We have recently shown the very satisfactory performance of these computational methods when calculating the interaction energies of dimers of amino acids with DNA bases *in vacuo* in comparison with mid-level DFT calculations [41], and a similar level of agreement has been observed in comparison with high-level correlated QM results [42]. Here, we probe the explicit contribution of the charged phosphate group to the recognition of DNA bases on a physical basis for the first time. We weighted the importance of various interaction motifs based on their relative abundance in the structures of real protein—DNA complexes. This allowed us to view the interaction specificity as a function of the energetic favourability and geometrical conservation of the binding motifs. Finally, we tested the validity of the observed interaction preferences in different dielectric environments in an effort to effectively capture the intricate electric properties of the protein—nucleic acid interface.

Materials and Methods

Data set preparation

A set of 1,584 structures of protein—DNA complexes solved by X-ray crystallography to a resolution better than 2.5 Å and with an R-factor no worse than 0.25 was obtained from the PDB in April 2014 and refined using the PISCES sequence culling server [7, 43]. Only the entries containing at least one double stranded DNA region consisting of at least 4 base pairs were considered. When multiple identical polypeptide chains were included in the PDB structures, such as in the complexes of homomultimeric transcription factors, only a single (alphabetically first) protein chain in complex with DNA was retained. The structures of hetero-*N*-meric protein complexes were separated into *N* independent entries, each one containing a single protein chain interacting with a replica of the recognised DNA double helix. These proteins were further analysed independently, *i.e.*, during the sequence homology assessment (see below). In total, 1,737 unique polypeptide chains in complex with their cognate DNA sequences comprised the data set.

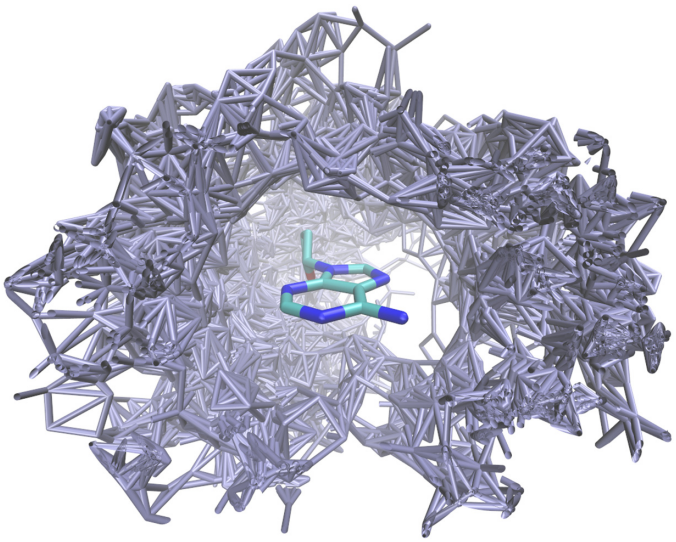

Fig 1. Distribution of asparagine side chains (gray) around an adenosine nucleoside. All molecular graphics were created using using VMD-1.9.2 [46].

doi:10.1371/journal.pone.0158704.g001

From each of these complexes, all interacting amino acid—2'-deoxyribonucleoside 5'-monophosphate (dNMP) dimers were extracted as follows. The interaction between the residues was defined based on the distance-based criteria utilised in the construction of the Atlas of Protein Side Chain Interactions [44]: when the distance between any amino acid reference atom and any DNA residue heavy atom was less than 1.0 Å plus the sum of the atoms' van der Waals radii, a contact was defined [45]. Three reference atoms were selected for each amino acid and coincide with the residue's characteristic side chain group atoms [45]. Pairs in which the nucleotide would originate from the 5' end of the DNA strand were discarded, as these residues naturally lack the 5' phosphate group. No other distinctions were made between the different nucleotide conformers. A total of 47,480 amino acid—dNMP dimers were obtained this way.

When one geometrically transforms all dimers containing a certain amino acid—dNMP combination (for example, all deoxyadenosine 5'-monophosphate [dAMP]—asparagine contacts) into an appropriately chosen common frame of reference, a three-dimensional distribution of the amino acid residues around the nucleotide is obtained (Fig 1). These transformations were performed by minimising the root mean square deviation (RMSD) of the DNA base heavy atoms between all dimers of a particular type.

Detection of the clusters. As illustrated in Fig 2, the directional nature of some non-covalent interaction modes, notably hydrogen bonds, leads to the clustering of amino acid residues relative to the nucleotide in 3D [27, 41]. The rigorous identification of these clusters has previously been described in detail [41]. In brief, after all amino acid—dNMP dimers of a certain type had been transformed to superpose the DNA bases as described above, we picked out each amino acid in turn and calculated the RMSD between its reference atoms and the reference atoms of all other amino acids in the respective distribution. The amino acid for which the number of contacts with RMSD less than 1.5 Å was the largest was then recognised as a cluster representative and, together with its RMSD-derived neighbours (the cluster), taken out of the distribution. This process was repeated until 6 clusters had been identified in each distribution, or until the last cluster isolated was too sparsely populated to be considered significant. A total of 12,935 dimers were found within some of the 469 clusters [41]. It should be noted that the

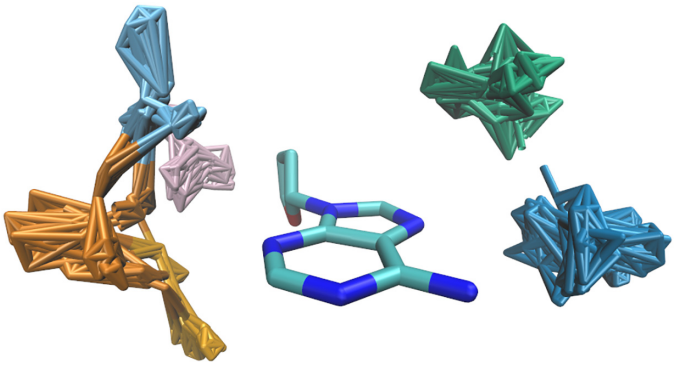

Fig 2. Side chain clusters identified in the distribution shown in Fig 1. The clusters are shown in colours (dark blue, vermillion, light blue, pink, green, orange).

doi:10.1371/journal.pone.0158704.g002

absolute number of contacts found in the clusters varied greatly between the various amino acid—DNA residue dimer types. While only few tens of contacts formed the clusters in the majority of distributions involving non-polar amino acids, up to several hundreds of contacts comprised the clusters in dimers which are known to form motifs significant for the process of direct sequence recognition (*i.e.*, arginine—guanine).

Treating the data set bias. While the redundant polypeptide chains corresponding to identical protein units within individual PDB files had already been discarded (see the treatment of homomultimeric proteins above), we had not at this point investigated the sequence identity of entries originating from different PDB structures. The high sequence similarity across different PDB entries would introduce bias into the data set, as the contacts originating from homologous protein structures would appear overpopulated compared to the contacts extracted from protein families for which few structures are currently available.

The bias was treated as follows. First, a global sequence alignment utilising the Needleman-Wunsch algorithm [47] and using the BLOSUM62 substitution matrix [48] was performed for all unique pairs of protein chains. The package *needle*, available in the EMBOSS-6.4.0.0 molecular biology suite, was used to perform these alignments [49]. Having obtained the sequence identity score for all pairs of protein chains, we constructed a set in which the sequence identity of any two proteins was less than 100%, *i.e.*, we removed the entries containing identical proteins.

Furthermore, we generated three additional sets of protein—DNA complexes, ones in which the sequence identity of any pair of protein chains was less than 30%, 90%, or 95%. These sets were generated as follows. For each of the 1,737 polypeptide chains, a list of proteins having sequence identity greater or equal to X% (X = 30, 90, 95) was compiled. These lists were then merged for each X to create a total list of homologous structures at the particular sequence identity level. The complements of these total lists are the sets of structures for which the sequence identity of any pair of protein chains is less than X%. The sets should be viewed as a more-or-less random selection of protein—DNA complexes satisfying the maximum sequence identity criteria; the randomness is limited by the fact that from each subset of homologous proteins, the alphanumerical order of the PDB identificators determined our selection. In fact, the same "randomness" was used in the construction of the set containing non-identical protein chains; however, in this case, the only difference between the PDB entries is the unlikely variation in the DNA sequences.

Table 1 presents the numbers of amino acid—dNMP dimers found for individual DNA base types. The populations of the sets constructed at the various maximum sequence identity levels are shown. The removal of identical protein entries discards over half of the available

Table 1. Number of contacts involving individual DNA base types after redundant entries had been discarded.

Seq. identity†	30%	90%	100%
Adenine	2,137 (7.9%)	3,087 (11.5%)	5,200 (18.5%)
Guanine	2,477 (6.9%)	3,411 (8.2%)	6,237 (16.7%)
Cytosine	2,007 (8.0%)	2,783 (10.0%)	4,899 (19.4%)
Thymine	2,305 (9.0%)	3,224 (11.5%)	5,373 (17.5%)
Total	8,926 (7.9%)	12,505 (10.3%)	21,709 (18.0%)

The numbers in parentheses indicate the proportion of contacts found in the clusters.
†—indicates that the mutual identity of any pair of sequences in the set is less than *X*%.

doi:10.1371/journal.pone.0158704.t001

dimers (only 21,709 dimers remain from the original 47,480 dimers observed before the sequence identity culling was applied), while relatively little difference is seen when comparing the numbers of contacts found at the 30% and 90% maximum sequence identity levels. This suggests that the bias in the original set was caused by several overpopulated protein families, and that the these redundant entries were already discarded at the 90% sequence identity level. The difference between the populations obtained at 90% and 95% identity levels is negligible and thus not shown. Moreover, the relative populations of the clusters were struck harder by the treatment of bias. For example, from the 12,935 dimers that comprised the clusters before the sequence identity culling, only 3,897 remain after discarding structures containing 100% identical proteins. This can be rationalised by the fact that the discarded homologous structures were more likely to provide geometrically similar contacts. The percentages of contacts found in the clusters at each sequence identity level are shown in parentheses in Table 1.

Identification of the contacts with DNA bases. The previously described procedure led to the retrieval of dimers in which the amino acid may be found in proximity to any of the dNMP's base, 2'-deoxyribose, or phosphate moieties. To quantitatively assess the contribution of the DNA backbone groups to the interaction specificity, we isolated a subset of contacts in which the amino acids interact directly with the DNA base moiety. These contacts were again identified based on distance-dependent criteria: if the distance between any amino acid and any DNA base atom was found less than 1.0 Å plus the sum of the atoms' van der Waals radii, the dimer was labelled as containing a base-directed interaction. Table 2 presents for various maximum sequence identity levels the number of base-directed contacts found in the distributions containing the respective DNA base types.

We thus obtained two sets of contacts: the first contains all amino acid—dNMP dimers found in the respective non-redundant sets of protein—DNA complexes, while the second

Table 2. As Table 1, but only including contacts in which the amino acid interacts directly with the DNA base.

Seq. identity†	30%	90%	100%
Adenine	1,080 (11.6%)	1,548 (17.1%)	2,462 (22.2%)
Guanine	1,313 (10.2%)	1,761 (11.5%)	3,011 (17.2%)
Cytosine	1,000 (9.5%)	1,358 (11.6%)	2,213 (18.9%)
Thymine	1,359 (10.7%)	1,879 (13.6%)	2,886 (17.2%)
Total	4,752 (10.5%)	6,546 (13.4%)	10,572 (18.7%)

The numbers in parentheses indicate the proportion of contacts found in the clusters.
†—indicates that the mutual identity of any pair of sequences in the set is less than *X*%.

doi:10.1371/journal.pone.0158704.t002

constitutes a subset of the former, containing only those dimers in which the amino acid interacts directly with the base moiety. In the framework of the pair-wise additive empirical calculations (see below), it was possible to investigate the exact interaction energy contribution of the interaction with the base moiety to the total interaction energy. To this end, we created two additional sets of contacts by duplicating the former and stripping their sugar-phosphate moieties. The energy of interaction between the amino acid and the sequence-specifying base moiety can be obtained from these dimers. Moreover, direct comparison with the contacts involving dNMPs can be made, revealing the quantitative contribution of the DNA backbone to the recognition process.

System preparation. The procedure atomising the interactions between proteins and DNA into the pairs of interacting residues described above led to the retrieval of amino acid—dNMP dimers. For multiple reasons, we decided to get rid of the atoms constituing the protein backbone groups. The inclusion of the C_α amide and carbonyl groups would introduce charged species into the molecule, greatly complicating the interpretation of the gas phase interaction energies. Second, each peptide bond group would have to be capped, creating intra- and intermolecular interactions that do not exist in nature. Finally, the properties of the atoms constituing the protein backbone are the same in each standard α-amino acid. Therefore, the binding motifs involving the peptide bond groups can hardly be viewed as representative of some preferred interaction mode between a specific amino acid—DNA residue dimers.

Therefore, we replaced the peptide bond carbonyl and amide groups of the amino acid with hydrogen atoms in each amino acid—DNA residue dimer. This process is consistent with those described in other works on the interactions of amino acids [39, 40, 50]. The result of this geometry culling is called the C_α representation of the amino acid, in which a methyl group caps the C_β atom. In the case of proline, only the carboxyl group was removed and the five-membered heterocycle remained [41].

Due to the way nucleic acid residues are labelled in PDB structures, the extraction of the *N*th DNA nucleotide resulted in the phosphate moieties lacking the O3' oxygen atom belonging to 2'-deoxyribose of the immediately preceeding (N – 1)th residue. This oxygen atom was added at the end of a vector of length 1.610 Å originating at the P atom and perpendicular to the plane containing the atoms OP1, OP2, and O5'. The specific length was chosen because it represents the equilibrium bond length between the two atoms in the Amber94 force field [51].

As the dimers were extracted from X-ray structures only, no hydrogen atoms were originally present. This problem was remedied utilising a custom UCSF Chimera-1.8.1 [52] script, to add the hydrogen atoms to both the amino acid and DNA residues for all contacts. Given the C_α representations, proline was modelled as a neutral tetrahydropyrrole and glycine as methane. Despite the software being able to decide on the correct protonation based on the local environment [52], histidine was protonated on N_ε in all contacts, as the population of N_δ-protonated side chains was insufficient for their separate analysis. Guanine and cytosine were represented by their dominant keto forms, and adenine and thymine by the dominant amino forms. Guanine was set to be protonated on the N1 atom instead of N3. In the two sets of contacts with isolated DNA bases, hydrogen atoms were added to N9 or N1 atoms in purine or pyrimidine bases, respectively. A single hydrogen was added to the O3' atom of the phosphate group.

Interaction energy determination

As noted in the Introduction, the total number of amino acid—DNA residue dimers in all four sets of contacts (over 60,000) and the size of some complexes (over 60 atoms) heavily favour the use of MM techniques over QM calculations. We have already shown the very satisfactory performance of MM methods for the calculation of interaction energies of amino acid side

chain—DNA base dimers [41]. Therefore, we followed the same computational protocol to calculate interaction energies of the extended complexes presented in this work.

Derivation of the missing parameters. The parameters used for the atoms of the C_α representations of amino acids, the atoms of the isolated nitrogenous bases, and atoms of the dNMPs were those derived for the Amber94 or, where applicable, Amber99SB-ILDN force field [51, 53]. The atoms not present in the force field (C_α hydrogen atoms, proline H1 atom, H1 and H9 atoms in isolated pyrimidine and purine bases, respectively, and the phosphate group's O3' and its attached hydrogen atom) had their non-electrostatic parameters assigned from chemically equivalent atom types. The constants of interactions between bonded atoms not present in the original force field were manually added based on the functional similarities of the atoms involved.

Partial atomic charges were assigned to the added hydrogen atoms manually in order to retain an integral total charge of each residue: +1.0 for arginine and lysine; −1.0 for aspartate, glutamate, and dNMPs; and 0.0 for the remaining amino acids and isolated DNA bases. Only the dominant forms of the species at pH = 7 were considered. When multiple hydrogen atoms were added, the charges were split symmetrically.

Computational protocol. The interaction energy calculations were performed using the supermolecular approach. First, the amino acid—dNMP (base) dimer had its hydrogen atoms' positions optimised using conjugate gradient energy minimisation while the heavy (non-hydrogen) atoms were confined to their original positions. Single point energy was then calculated on this optimised dimer geometry. The dimer was then split and a single point energy calculation was immediately calculated for monomer. Hydrogen atom positions were optimised for each monomer by itself, and then a single point energy was calculated. The difference between the single point energy of the monomer after it had been isolated from the complex and after it was optimised by itself is the deformation energy of that monomer. The interaction energy was then calculated as the difference between the single point energy of the optimised complex and the sum of the single point energies of the monomers present in the non-optimised, dimer conformation, plus the sum of the monomers' deformation energies. All MM interaction energy calculations and geometry optimisations were performed using GROMACS-4.5.5 compiled in double precision [54].

Solvation effects. To introduce biological relevance to the interaction energy calculations, we included the effects of the surrounding water environment. Molecular dynamics (MD) simulations are usually performed using discrete water models in which each water molecule is treated explicitly. This representation of the solvent is not suitable for the interaction energy calculations presented here. In particular, the requirement of the constrained geometry of the solute would introduce artificial energy gaps between the dimer and monomer conformations of the interacting molecules when solvent relaxation was taken into account.

An alternative approach is to part with the discrete representation of the water molecules and to treat the solvent as a continuous dielectric environment. The electrostatic interaction of the solute with this implicit solvent model is described by the Poisson equation. Given the technical difficulties of solving this differential equation, various approximations to the (linearised) Poisson equation have been derived. Among these, the generalised Born (GB) formalism is the most widely used in simulations of biomolecules [55–57]. Combined with a term accounting for the disruption of the solvent structure due to the presence of the molecule proportional to its surface area (SA), the GB approach can be used to calculate the free energy of solvation of any molecule for which the set of atomic Born radii are known [58, 59].

The GB/SA formalism can be seamlessly integrated with our protocol for interaction energy calculation. The Hawkins-Cramer-Truhlar (HCT) model [60, 61] was used for calculating the effective Born radii. This model was shown to provide the closest agreement with explicit

solvent simulations when calculating the dynamic properties of DNA [56]. While the van der Waals radii and screening constants required for the implicit solvent calculations involving amino acids were already available in GROMACS-4.5.5 [54], the atomic parameters of multiple nucleic acid atoms were missing. These were imported from the freely available source code of the TINKER 7.1 molecular modelling package (http://dasher.wustl.edu/tinker/). The interaction energies were then calculated using the Amber99SB-ILDN force field [51, 53] following the protocol described above. The calculations were performed in environments with relative permitivities of 4, 16, and 80. These values were chosen to approximate the electric properties of the protein interior, protein—nucleic acid interface, and the water environment, respectively.

It should be noted that the used GB/SA model places the dielectric medium everywhere around the molecular cavity, including the regions where neighbouring amino acids or base steps would be naturally present. It can be expected that this treatment affects the interaction energies. Future additions to the model could try to remedy this artificial behaviour by effectively including the cavitation and electrostatic effects of the neighboring residues.

Results and Discussion

Theoretical treatment of specificity

Interaction energy profiles. To establish a link between the relative interaction energies of the various binding motifs and their geometrical conservation revealed by the clustering, we first constructed an interaction energy profile for each distribution. To this end, a histogram of the interaction energies provided by the dimers found in the distribution was created, and the number of bins was calculated from the Freedman-Diaconis formula [62]. The cluster containing the dimers providing on average the lowest (*i.e.*, most stabilising) interaction energies was then identified. A histogram of interaction energies provided by the members of this cluster was made, respecting the bin boundaries calculated for the respective distribution. The two histograms were then overlaid as shown in Fig 3.

Criteria of specificity. The clusters in each distribution represent geometrically conserved arrangements of the interacting partners. This conservation does not, however, automatically imply a role in the direct DNA sequence recognition. For example, contacts featuring single hydrogen bond donor or acceptor groups are naturally sterically constrained because of the directional requirements of hydrogen bonds and are therefore prone to being found in clusters. However, single hydrogen bonds are not sufficient to distinguish between individual DNA bases [12].

Based solely on geometrical criteria, the possibilities of specific base recognition by a single amino acid are limited to the few dimers featuring bidentate hydrogen bonds (see Introduction). Given that it is possible, especially in a high-dielectric environment, to achieve a similar level of stabilisation by utilising a combination of other non-covalent interaction modes [39], it may be desirable to augment this definition of specificity by explicitly considering the interaction energies of the respective dimer conformations.

We have already suggested that the presented spatial distributions effectively reflect the limited accessibility of the nucleotide in the DNA double helix to the approaching amino acids. Assuming that these distributions are configurationally saturated (in the same sense that an ergodic MD simulation would saturate the conformational space), the following requirements can be made for an interaction to be viewed as significant for the sequence recognition process:

1. The orientation of the amino acid relative to the DNA base (dNMP) must be found within one of the geometrical clusters. This condition implies that the respective interaction mode

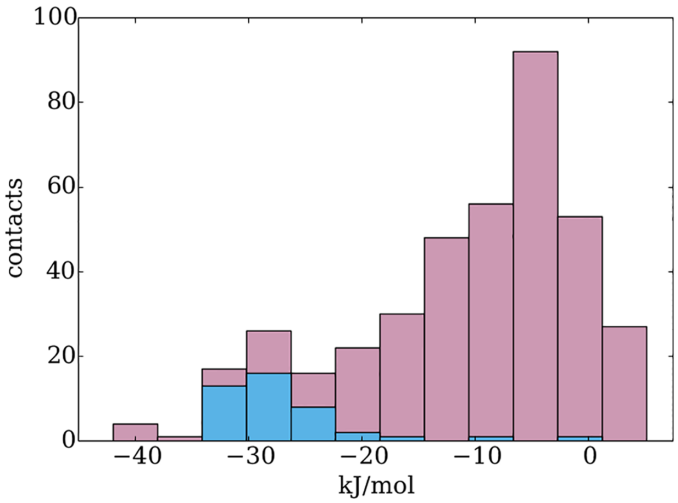

Fig 3. Interaction energy profile example: dAMP—asparagine dimers. The pink bars display the interaction energies of the entire distribution; the blue bars show the interaction energies provided by the members of its most stabilising cluster. The latter are shown in dark blue in Fig 2. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$.

doi:10.1371/journal.pone.0158704.g003

is utilised by many protein—DNA complexes and as such is not bound to be functional only under the unique local environment of a single protein family.

2. The cluster to which the dimer belongs must correspond to an attractive and most energetically favourable arrangement of the two partners.

3. Little to no other contacts other than those belonging to the distinct low-lying cluster are to be present within its interaction energy range. This criterion has two consequences. First, it enables identification of specificity-determining dimer geometries based on the respective interaction energies. Second, it implies that all dimers within that particular interaction energy range are highly sterically specific, as they could have been identified as forming a cluster.

4. The previous criteria specify energetically distinct geometries within the respective distributions. For an amino acid *A* to uniquely distinguish between individual DNA bases, the interaction energies found in dimers from the identified distinct low-lying cluster must also be lower (signed) than those provided by any contacts of *A* with any other base type. In other words, the stabilisation of the complex *A*—*B*, where *B* is the recognised DNA base, adopting a conformation falling to the distinct cluster, must be greater than the interaction energy found for any dimer of *A* with any other base type. This distinction is to be made for each of the nucleotide edges (Hoogsteen, Watson-Crick, sugar-phosphate) separately, as it may be possible for an amino acid to uniquely distinguish between different bases in each these regions.

Only when all these criteria are met can the coupling between the energetic and geometrical aspects of specificity be assumed. The interaction energy profile shown in Fig 3 already illustrates some of these distinctive characteristics; for a demonstration of the selectivity criteria involving all DNA bases, see S1–S4 Figs. The following sections cover the application of these rules to the various sets of amino acid—DNA base (dNMP) dimers and the identification of the distinct low-lying clusters described above.

Table 3. DNA base—amino acid dimer types that can contribute to direct recognition.

Rel. permittivity	1	4	80
Adenine	**N^H,Q^H,K^s,T^s**	**N^H,Q^H,K^s,T^s**	**N^H,Q^H,K^s,T^s**
Guanine	**R^H**,D^W	**R^H,D^W**	**R^H,D^W**
Cytosine			
Thymine		**T^H**	**T^H**

The results for different dielectric environments are shown. Only the complexes in which the amino acid is in direct contact with the base moiety were considered. The dimer types for which an exceptionally good agreement between the interaction energy profile characterics and the criteria of specificity is observed are shown in bold. The superscript shows which edge of the DNA base (nucleotide) is contacted by the amino acid: ^{H}—Hoogsteen edge; ^{s}—sugar edge, ^{P}—phosphate group, ^{dis}—dispersion (stacking) interaction with the DNA base.

doi:10.1371/journal.pone.0158704.t003

Observed binding preferences

Contacts with DNA bases. The simplest set to be analysed consists of only those side chain—DNA base dimers that feature a direct interaction between the two residues. The elimination of the charged sugar-phosphate group, for now, restricts the search space to those contacts in which its contribution is not essential for stabilisation of the interaction. Table 3 presents interactions that meet the above described criteria of specificity. These contacts were identified by visual inspection of the interaction energy profiles in the most numerous set constructed after identical protein—DNA complexes had been removed. The dimers that meet the criteria without exception are shown in bold; the other contacts appear significant, but some ambiguity in complying with the rules remains (for example, few non-clustering dimers provide similar interaction energies).

This set of side chain—DNA base dimers should predominantly be viewed as a control group, as it contains all the structural data necessary to recognise the contacts traditionally thought of as being involved in the direct readout. Indeed, the adenine—asparagine, adenine—glutamine, and guanine—arginine motifs were successfully recognised as forming specific contacts in all environments, supporting our hypothesis that the specificity can be observed by coupling the energetic and geometric features. The interaction energy profiles of these canonical amino acid—DNA base dimers can be seen in S5–S7 Figs. In addition, there were several other dimer types (adenine—lysine, adenine—threonine, guanine—aspartate, and thymine—threonine) with interaction energy profiles that shared the distinctive characteristics described above. Given that the properties of isolated DNA bases can be quite different from those of their nucleotide forms, these contacts will be investigated in greater detail if shown to be significant when the complete DNA residues are considered (see below). It should be noted that some of these dimers were already identified and explored in our previous work on DNA base interactions in the gas phase [41].

Base-directed contacts with dNMPs. We next investigated how addition of the sugar-phosphate group changes the binding preferences of the residues. Only dimers in which the side chain is in contact with the DNA base moiety were still considered. Table 4 shows the dNMP—amino acids dimer types that meet the stated criteria of specificity. As before, these analyses were performed on the set of protein—DNA complex structures obtained after discarding identical entries.

The specific recognition of adenine by asparagine or glutamine features a bidentate hydrogen bond between the side chain amide group of the amino acid and the C6 amino group/N7 atoms of the base. These canonical interactions remain the most energetically favourable even

Table 4. dNMP—amino acid dimer types that can contribute to direct recognition.

Rel. permittivity	1	4	80
dAMP	$\mathbf{N^H}$,Q^H	N^H,$\mathbf{Q^H}$,K^s,T^s	N^H,Q^H,$\mathbf{K^s}$,T^s
dGMP	R^H	R^H,D^W	$\mathbf{R^H}$,$\mathbf{D^W}$,L^{dis}
dCMP	$\mathbf{I^{dis}}$	I^{dis},K^s	K^s
TMP	S^H,T^H,$\mathbf{Y^H}$	T^H,$\mathbf{Y^H}$	H^{dis},$\mathbf{T^H}$,$\mathbf{Y^H}$

The results for different dielectric environments are shown. Contacts featuring interactions only with the sugar-phosphate backbone were excluded. Significant interactions not present in Table 3 are underlined. The dimer types for which an exceptionally good agreement between the interaction energy profile characterics and the criteria of specificity is observed are shown in bold. The superscript shows which edge of the DNA base (nucleotide) is contacted by the amino acid: ^{H}—Hoogsteen edge; ^{s}—sugar edge, ^{P}—phosphate group, ^{dis}—dispersion (stacking) interaction with the DNA base.

doi:10.1371/journal.pone.0158704.t004

in the presence of the sugar-phosphate group, as long as the low dielectric constant (ε = 1 or 4) of the local environment is assumed. Despite meeting all of our specificity requirements, the interaction of adenine with threonine was realised in too few contacts to allow deeper statistical investigation. This interaction is realised in the minor groove *via* a single hydrogen bond between the side chain hydroxyl group of the amino acid and the N3 atom of the base.

When a water-like (ε = 80) dielectric environment is assumed, the interaction between adenine and lysine is the only one to display energetically and geometrically distinctive characteristics. This interaction features a single hydrogen bond between the terminal amino group of the amino acid and the N3 atom of the base. In addition, a set of van der Waals contacts is present between the aliphatic lysine side chain and the 2'-deoxyribose atoms. It should be noted that the members of a single cluster in the distribution of cytosine—lysine dimers adopt an analogous geometry (Tables 4 and 5). However, these latter contacts form a significant cluster only when the most redundant set of protein—DNA complexes is considered, and the respective binding motif is thus not widely utilised.

The previously identified clusters in the distributions of asparagine, glutamine, and threonine lost some of their distinctive characteristics with the increased dielectric constant. Notably, several non-cluster contacts began to offer the same interaction energies as those found in the clusters. Thus, we hypothesise that the specific interaction with the charged lysine plays a role in recognition of adenine over a larger distance, while the nucleic acid remains enveloped

Table 5. dNMP—amino acid dimer types that can contribute to direct recognition.

Rel. permittivity	1	4	80
dAMP	$\mathbf{N^H}$,$\mathbf{Q^H}$,T^s	$\mathbf{N^H}$,$\mathbf{Q^H}$,K^s	N^H,$\mathbf{Q^H}$,$\mathbf{K^s}$,T^s
dGMP		$\mathbf{R^H}$,$\mathbf{D^W}$,L^{dis}	$\mathbf{R^H}$,$\mathbf{D^W}$,L^{dis}
dCMP	$\mathbf{Q^P}$,$\mathbf{I^{dis}}$	$\mathbf{Q^P}$,$\mathbf{I^{dis}}$	$\mathbf{Q^P}$,K^s
TMP	$\mathbf{Q^P}$,S^H,T^H,$\mathbf{Y^H}$	$\mathbf{Q^P}$,S^H,T^H,$\mathbf{Y^H}$	$\mathbf{Q^P}$,S^H,T^H

The results for different dielectric environments are shown. Contacts featuring interactions only with the sugar-phosphate backbone were included. Significant interactions not present in Table 4 are underlined. The dimer types for which an exceptionally good agreement between the interaction energy profile characterics and the criteria of specificity is observed are shown in bold. The superscript shows which edge of the DNA base (nucleotide) is contacted by the amino acid: ^{H}—Hoogsteen edge; ^{s}—sugar edge, ^{P}—phosphate group, ^{dis}—dispersion (stacking) interaction with the DNA base.

doi:10.1371/journal.pone.0158704.t005

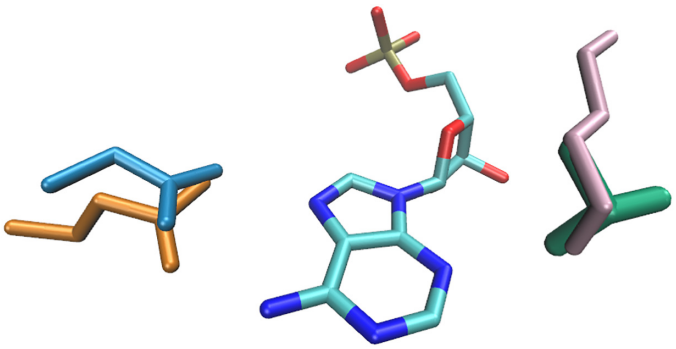

Fig 4. Recognition of adenine by asparagine (blue), glutamine (orange), lysine (pink), and threonine (green).

doi:10.1371/journal.pone.0158704.g004

by a hydration shell. When the electric properties of the interfacial region are described by a high dielectric constant, interactions with amino acids other than lysine do not provide similar levels of base-specific resolution. The increased relative selectivity in a water-like environment compared to the low dielectric was also observed in other motifs involving charged amino acids (see below).

The atomic-level details of the specific interactions involving adenine are shown in Fig 4. The interaction energy profiles of the abovementioned dimers can be seen in S9–S12 Figs.

The canonical recognition of guanine by arginine is realised *via* a bidentate hydrogen bond between the terminal guanidino group of the amino acid and the O6 and N7 atoms of the base. This interaction also displays the distinctive characteristics only when the solvent effects are taken into account. Partial screening of the atomic charges is needed for the recognition in this case because the fine details of the interaction with the base are otherwise lost due to the dominant electrostatic attraction of the amino acid with the DNA backbone.

The specific interaction of aspartate with guanine features a bidentate hydrogen bond between the terminal carboxylic group of the amino acid and the N1 and N2 amino group atoms of the base. This binding obviously interferes with the Watson-Crick pairing between DNA bases. A deeper analysis of the complexes from which this contact originates (PDB IDs 1JB7, 1OMH, 1PO6, 1XJV, and 3ZH2) reveals that the motif is utilised in the recognition of aptameric, telomeric, or otherwise strained DNA structures. While likely not involved in routine sequence recognition, this highly stabilising interaction contributes to and can even be crucial for the recognition of non-canonical forms of DNA. This exceptional case illustrates the robustness of our general criteria of specificity. Without any prior information about the structures present in the set, we were able to find a group of non-homologous proteins which, nonetheless, featured the same binding motif involved in the sequence recognition in the respective complexes. Interestingly, the interactions of aspartate and glutamate with guanine *via* the Watson-Crick edge were found to provide to the most favourable binding free energies of all amino acid—DNA base dimer types [39].

The atomic-level details of the specific interactions involving guanine are shown in Fig 5. The interaction energy profiles of the two dimers can be seen in S13 and S14 Figs.

The possible recognition of guanine by asparagine or glutamine through the sugar edge did not display the clustering characteristics. Similarly, there were no distinct interactions involving cytosine as a base, especially when the solvent effects were taken into account and the more restrictive sequence identity criteria were applied. In our previous work, we were able to identify distinct clusters of asparagine and tyrosine side chains forming contacts with cytosine *via* a single hydrogen bond featuring the O2 atom of the DNA base as an acceptor [41]. The low

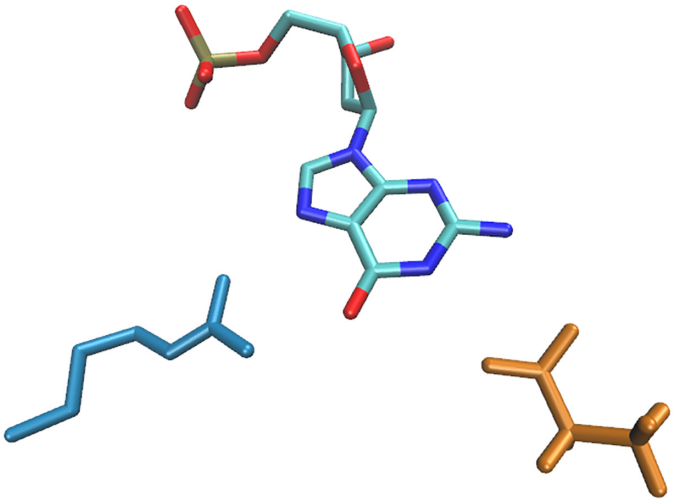

Fig 5. Recognition of guanine by arginine (blue) and aspartate (orange).

doi:10.1371/journal.pone.0158704.g005

population of these clusters (6 and 4 contacts for asparagine and tyrosine dimers, respectively) meant that the removal of only few protein—DNA complexes from the data set due to redundancy put the population these motifs below the threshold needed for the detection of the clusters. As we have now taken into consideration the presence of the clusters across various redundancy levels when identifying the distinct interaction motifs, these contacts do not appear in Table 3. It is, however, possible that, given a larger data set, the significance of these motifs could become apparent. The absence of the motifs involving guanine identified in ref. [41] follows the same reasoning.

A cluster of isoleucine that displays preference towards cytosine *in vacuo* consists of contacts featuring van der Waals interactions involving almost all atoms of the amino acid side chain and almost all atoms of the nucleotide (Fig 6; the corresponding interaction energy profile is shown in S15 Fig). We found that the dGMP—leucine dimers adopt a similar geometry. However, the most energetically favourable cluster lacks some of the distinctive characteristics in this case. Although it is known that specific hydrophobic amino acids are crucial for the stabilisation of some repressor/operator complexes [14], there has not been any sign of a universal one-to-one correspondence between the interacting residues. It is, of course, possible that the

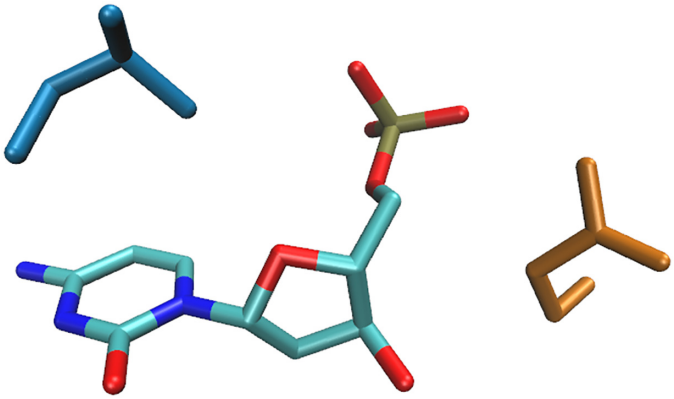

Fig 6. Interactions of cytosine with isoleucine (blue) and glutamine (orange).

doi:10.1371/journal.pone.0158704.g006

current treatment of the solvent effects is inadequate for the complete description of the stabilisation provided by this binding motif.

The distinctive characteristics of the interactions of serine, threonine, and tyrosine with thymine become prominent only after the addition of the sugar-phosphate group to the DNA base. All of these binding motifs involve a hydrogen bond between the donor hydroxyl group of the amino acid side chain and one of the phosphate group acceptor oxygen atoms. The C5 methyl group of thymine sterically stabilises these motifs by interacting with the hydroxyl group oxygen atom from the opposite side. This interaction is not possible in contacts with the other DNA bases. In addition, the hydrophobic effect may stabilise the interaction of the two methyl groups in contacts involving threonine. This aditional stabilisation may be the cause of the higher population and stereospecificity of the motif involving threonine compared to that containing serine.

The atomic-level details of the specific interactions involving thymine are shown in Fig 7. The interaction energy profiles of the abovementioned dimer types can be seen in S16–S18 Figs.

DNA backbone-directed contacts. Finally, we considered distributions involving all amino acid—dNMP dimers, including those featuring solely contacts with the DNA backbone. Table 5 presents those dimer types in which the amino acid side chains form clusters displaying the distinct properties described above. These preferences were found on the set from which identical protein entries had been discarded.

In addition to the previously identified dimers, two notable new interactions appeared. The contacts of cytosine and thymine with glutamine feature a single hydrogen bond between the side chain amide group of the amino acid and one of the phosphate group oxygen atoms. No interaction with the nitrogenous base moieties are present. These contacts display the distinct characteristics *in vacuo* as well as in the tested dielectrics. We are unsure why the amino acids prefer the pyrimidine-containing nucleotides. Similar interactions involving the purine nucleotides are not present in this set. This interaction motif can be observed for the pyrimidine bases even after applying the most strict redundancy-culling criteria. It is possible that the apparent preference displayed by these dimers is the result of an inadequate sampling of the configurational space realised in the currently available structures of protein—DNA complexes (see below).

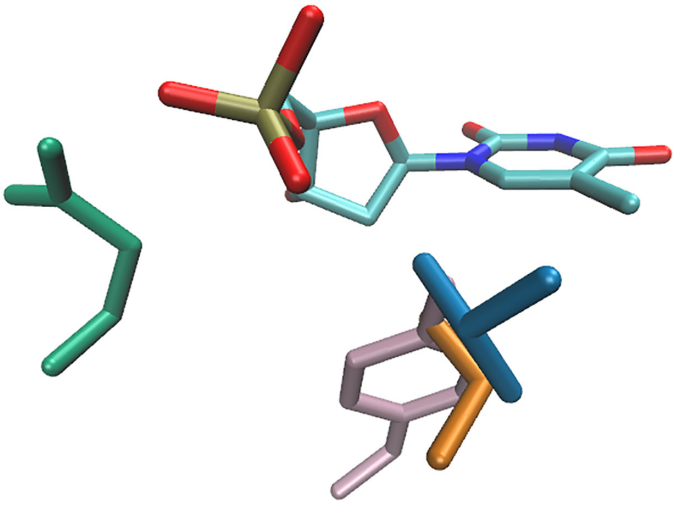

Fig 7. Recognition of thymine by threonine (blue), serine (orange), tyrosine (pink), and glutamine (green).

doi:10.1371/journal.pone.0158704.g007

The interaction energy profiles of these dimers can be seen in S19 and S20 Figs. Atomic-level details of the interactions are shown in Figs 6 and 7 for cytosine and thymine, respectively.

Possible extensions of the model

Structural data. A clear drawback of our interaction energy profile-based specificity definition is that only sufficiently represented motifs will be detected. If this analysis is to be considered complete, we must assume that all amino acid preferences for DNA bases can already be detected in the binding modes realised in the currently available structures of protein—DNA complexes. It is unfortunate that the number of amino acid—DNA residue dimers is still an order of magnitude lower compared to the number of amino acid—amino acid contacts available from high-quality protein structures [63]. This insufficiency is most apparent when using more strict homology-reducing criteria. While most of the motifs presented here can still be found in these less redundant sets, reduction of their population by a half (or more) often makes the presented interaction energy profile-based technique inappropriate due to the insufficient resolution of the histograms.

Water-mediated interactions. While the treatment of the solvent as a continuous dielectric offers a significant improvement over the gas phase calculations, it is inappropriate for the treatment of interactions that naturally involve a bridging water molecule. Previous work found that almost one fifth of all contacts involve a solvent-mediated interaction, most of them directed at the DNA backbone [27]. The crucial role of well-ordered water molecules in sequence recognition has been described in the *trp* repressor/operator complex [14]. Unfortunately, large variations in the number of water molecules present in the crystallographic structures exist in the range of resolutions used in this study. For example, an increase in resolution from 2.6 Å to 1.9 Å for the nucleosome core particle has revealed over 2,500 more water molecules [4]. Therefore, even if the solvent molecules that can be found in the protein—DNA structures used in this work were included in the calculations, it is be likely that some natural water-mediated interactions would still be missing.

Calculation of additional binding free energy components. The presented approach to calculating the potential energy of the interaction between two residues provides, of course, a limited approximation of the binding free energy, which is the biophysically relevant potential. In particular, no explicit treatment of the entropic effects was attempted, although a part of the solvent entropy could have effectively been captured by the cavitation component of the GB/SA approach. While the entropy of the solutes could be estimated from normal mode vibrational analysis or molecular dynamics trajectory [64], the determination of this term is beyond the scope of this paper.

Free energy difference calculations with explicit representation of the solvent molecules, such as those recently performed [39], are, of course, the most appropriate computational approach to fully accounting for all components of the binding free energy. The omission of the explicit treatment of the solvent entropy in our paper could have resulted in some motifs, especially those involving non-polar amino acids, being significantly mistreated. In fact, very few of the herein identified distinct clusters involve interactions which do not feature a dominant electrostatic component. On the other hand, the interaction energies, and especially the interaction energy differences, of the identified dimers (which almost unanimously feature hydrogen bonds) can be expected to be reasonably close to the biophysically relevant energy values.

One possible validation of our approach would be the comparison of the calculated interaction energies with the free energy differences derived from statistical potentials obtained by

evaluating the relative probabilities of the various interaction motifs. Protein—nucleic acid complexes were previously explored using this approach by Mandel-Gutfreund and Margalit [25] and others [65–67]; for a thorough list of references related to the study of protein structures see ref. [68]. Such an approach could, in principle, be used to correctly describe not only the motifs in which the enthalpic contribution to the free energy of binding is dominant (as can be assumed for the motifs described in this paper), but also the motifs in which binding is driven by hydrophobic and other entropic effects. This would require that the conformational space of all amino acid—nucleotide dimers be adequately sampled by the dimer geometries extracted from the currently available structures of protein—DNA complexes. Based on the populations of some of the motifs presented in this article, we think that this requirement could hardly be met as of now.

Interactions of larger residue blocks. As this study was focused on finding the binding preferences at the one-to-one correspondence level, contacts spanning multiple base steps or featuring interactions with both DNA residues in a base pair were not explicitly treated as such. It has been shown that the assumption of additivity of individual amino acid—mononucleotide interactions is a reasonable approximation in the search of DNA binding sites [69]. If the preferences towards oligonucleotide blocks were to be probed in as exhaustive a manner as done for individual DNA bases, it would become apparent that the number of contacts provided by the currently available protein—DNA structures would not be sufficient for a reasonable analysis. If one were to also consider the preferences of larger peptide blocks, the sheer number of possible sequence variations would quickly become greater than the number of available contacts altogether. It is, however, very well possible that the extension of the presented methodology to cover the interactions of these larger biomolecular fragments could reveal additional interaction motifs significant for the process of direct sequence readout. Alternatively, the energetics of interactions with neighbouring base steps or amino acids could, in some cases, disrupt the observed binding preferences. Explicit solvent MD simulations of selected oligopeptide/oligonucleotide complexes currently seem to be the best theoretical approach to investigate the binding preferences involving these larger fragments.

Role of non-specific contacts. Depending on the applied redundancy-culling criteria, only between one tenth and one quarter of all amino acid—DNA residue dimers were found in the clusters. It appears rational to ask what is the role of the remaining contacts. The only thing that can be said based on our study about these non-clustering dimers is that they do not massively participate in binding motifs involved in the direct readout of single DNA bases. On the other hand, given that the protein side of the interaction interface is occasionally limited to only several amino acids, there may be little space left for random noise. These remaining contacts can thus well serve as modulators of the recognized motifs involved in the direct readout, or, alternatively, be involved in facilitating the shape recognition or other, more complex phenomena. This can only be decided in the context of each individual DNA-binding protein.

Evolution of the protein—DNA interface. In this work, we have shown that several amino acid—DNA nucleotide combinations considerably extend the library of motifs that can be utilised in direct sequence recognition. The significance and conservation of these motifs across various protein families may have had unknown consequences on the evolution of transcription factors and their cognate DNA sequences. We have so far made very general predictions about the interactions between individual residues without probing the original biomacromolecules. Indeed, our next logical step will be to investigate the relationship between the utilisation of the specific motifs and the source protein structures.

Supporting Information

S1 Fig. Demonstration of the criteria of specificity: the interaction energy profile of the dAMP—glutamine dimer. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$. Only those dimers in which the amino acid interacts with the base moiety of the nucleotide were considered in the construction of the profile. No two 100% identical proteins were present in the set from which the dimers were extracted. The pink histograms show the interaction energy profile of the entire distributions; the blue histograms display the interaction energy profile of its most energetically stabilising cluster. Note how the cluster in this distribution meets the specificity criteria:

1. it represents the most favourable arrangement of the partners within the distribution,
2. very few other (*i.e.*, non-cluster) contacts within the profile provide similar interaction energies as the cluster's members,
3. the interactions with the other DNA bases (S2–S4 Figs), do not contain a significant number of contacts with similar interaction energies.
(TIF)

S2 Fig. Demonstration of the criteria of specificity: the interaction energy profile of the dCMP—glutamine dimer. The pink histograms show the interaction energy profile of the entire distributions; the blue histograms display the interaction energy profile of its most energetically stabilising cluster. Note how the character of the cluster (the shape and position of the cluster profile relative to the profile of the distribution) differs from that of the cluster in dAMP—glutamine distribution (S1 Fig). The selection of the data set for the construction of the profile and other computational details are the same as in S1 Fig.
(TIF)

S3 Fig. Demonstration of the criteria of specificity: the interaction energy profile of the dGMP—glutamine dimer. The pink histograms show the interaction energy profile of the entire distributions; the blue histograms display the interaction energy profile of its most energetically stabilising cluster. Note how the character of the cluster (the shape and position of the cluster profile relative to the profile of the distribution) differs from that of the cluster in dAMP—glutamine distribution (S1 Fig). The selection of the data set for the construction of the profile and other computational details are the same as in S1 Fig.
(TIF)

S4 Fig. Demonstration of the criteria of specificity: the interaction energy profile of the TMP—glutamine dimer. The pink histograms show the interaction energy profile of the entire distributions; the blue histograms display the interaction energy profile of its most energetically stabilising cluster. Note how the character of the cluster (the shape and position of the cluster profile relative to the profile of the distribution) differs from that of the cluster in dAMP—glutamine distribution (S1 Fig). The selection of the data set for the construction of the profile and other computational details are the same as in S1 Fig.
(TIF)

S5 Fig. Interaction energy profile of a canonical amino acid—DNA base dimer involved in the direct readout: adenine—asparagine. The energetically lowest lying cluster (blue) shows distinctive characteristics, as defined in text and in S1 Fig legend. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$. Only those dimers in which the

amino acid interacts with the base moiety were considered in the construction of the interaction energy profiles. No two 100% identical proteins were present in the set from which the dimers were extracted.
(TIF)

S6 Fig. Interaction energy profile of a canonical amino acid—DNA base dimer involved in the direct readout: adenine—glutamine. The energetically lowest lying cluster (blue) shows distinctive characteristics, as defined in text and in S1 Fig legend. The selection of the data set for the construction of the profile and other computational details are the same as in S5 Fig.
(TIF)

S7 Fig. Interaction energy profile of a canonical amino acid—DNA base dimer involved in the direct readout: guanine—arginine. The energetically lowest lying cluster (blue) shows distinctive characteristics, as defined above. The selection of the data set for the construction of the profile and other computational details are the same as in S5 Fig. The "envelope" of non-cluster contacts in the profile is caused by the symmetry of the arginine guanidino group: four energetically equivalent orientations of the side chain involving the guaninidino group as hydrogen bond donor exist; however, the cluster consists of only one of those. One of these alternative orientations is shown in S8 Fig.
(TIF)

S8 Fig. dGMP—arginine dimer: one of four energetically equivalent geometries. These geometries contribute to the "envelope" of non-cluster contacts covering the cluster profile (blue) in S7 Fig. Compare with Fig 5 (blue) in the main text.
(TIF)

S9 Fig. Interaction energy profile of a dimer involving dAMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dAMP—asparagine. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$. Only those dimers in which the amino acid interacts with the base moiety of the nucleotide were considered in the construction of the interaction energy profile. No two 100% identical proteins were present in the set from which the dimers were extracted.
(TIF)

S10 Fig. Interaction energy profile of a dimer involving dAMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dAMP—glutamine. The selection of the data set for the construction of the profile and other computational details are the same as in S9 Fig.
(TIF)

S11 Fig. Interaction energy profile of a dimer involving dAMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dAMP—threonine. The selection of the data set for the construction of the profile and other computational details are the same as in S9 Fig.
(TIF)

S12 Fig. Interaction energy profile of a dimer involving dAMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dAMP—lysine. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 80$. The selection of the data set for the construction of the profile is the same as in S9 Fig.
(TIF)

S13 Fig. Interaction energy profile of a dimer involving dGMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dGMP—arginine. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 80$. Only those dimers in which the amino acid interacts with the base moiety of the nucleotide were considered in the construction of the interaction energy profile. No two 100% identical proteins were present in the set from which the dimers were extracted. The "envelope" of isoenergetic non-cluster contacts covering the cluster profile is present for the symmetry reasons discussed in the legend of S7 Fig and illustrated in S8 Fig.
(TIF)

S14 Fig. Interaction energy profile of a dimer involving dGMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dGMP—aspartate. The selection of the data set for the construction of the profile and other computational details are the same as in S13 Fig.
(TIF)

S15 Fig. Interaction energy profile of the dimer involving dCMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: dCMP—isoleucine. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 1$. Only those dimers in which the amino acid interacts with the base moiety of the nucleotide were considered in the construction of the interaction energy profile. No two 100% identical proteins were present in the set from which the dimers were extracted.
(TIF)

S16 Fig. Interaction energy profile of a dimer involving TMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: TMP—serine. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$. All amino acid—nucleotide dimers were considered in the construction of the interaction energy profile. No two 100% identical proteins were present in the set from which the dimers were extracted.
(TIF)

S17 Fig. Interaction energy profile of a dimer involving TMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: TMP—threonine. The selection of the data set for the construction of the profile is the same as in S16 Fig.
(TIF)

S18 Fig. Interaction energy profile of a dimer involving TMP in which the energetically lowest lying cluster displays some of the distinctive characteristics: TMP—tyrosine. The selection of the data set for the construction of the profile is the same as in S16 Fig.
(TIF)

S19 Fig. Interaction energy profile of a dimer in which the energetically lowest lying cluster displays some of the distinctive characteristics: dCMP—glutamine. The binding motif realised in the distinctive cluster features only an interaction with the DNA backbone. The interaction energies were calculated in an environment with dielectric constant $\varepsilon = 4$. All amino acid—nucleotide dimers were considered in the construction of the interaction energy profile. No two 100% identical proteins were present in the set from which the dimers were extracted.
(TIF)

S20 Fig. Interaction energy profile of a dimer in which the energetically lowest lying cluster displays some of the distinctive characteristics: TMP—glutamine. The binding motif realised in the distinctive cluster features only an interaction with the DNA backbone. The selection of

the data set for the construction of the profile is the same as in S19 Fig.
(TIF)

Acknowledgments

The authors would like to express gratitude to Mgr. Jiří Hostaš, Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, for his continuous supply of *ab initio* results.

Author Contributions

Conceived and designed the experiments: DJ RAL JV. Analyzed the data: DJ JV. Wrote the paper: DJ RAL JV.

References

1. Smith ML, Chen IT, Zhan Q, O'Connor PM, Fornace AJ Jr. Involvement of the p53 tumor suppressor in repair of UV-type DNA damage. Oncogene. 1995; 10: 1053–1059. PMID: 7700629
2. Drabløs F, Feyzi E, Aas PA, Vaagbø CB, Kavli B, Bratlie MS, et al. Alkylation damage in DNA and RNA-repair mechanisms and medical significance. DNA repair. 2004; 3: 1389–1407. doi: 10.1016/j.dnarep.2004.05.004 PMID: 15380096
3. Stojic L, Brun R, Jiricny J. Mismatch repair and DNA damage signalling. DNA repair. 2004; 3: 1091–1101. doi: 10.1016/j.dnarep.2004.06.006 PMID: 15279797
4. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. J. Mol. Biol. 2002; 319: 1097–1113. doi: 10.1016/S0022-2836(02)00386-8 PMID: 12079350
5. Balasubramanian S, Xu F, Olson WK. DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. Biophys. J. 2009; 96: 2245–2260. doi: 10.1016/j.bpj.2008.11.040 PMID: 19289051
6. Battistini F, Hunter CA, Moore IK, Widom J. Structure-based identification of new high-affinity nucleosome binding sequences. J. Mol. Biol. 2012; 420: 8–16. doi: 10.1016/j.jmb.2012.03.026 PMID: 22472421
7. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur. J. Biochem. 1977; 80: 319–324. doi: 10.1111/j.1432-1033.1977.tb11885.x PMID: 923582
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28: 235–242. doi: 10.1093/nar/28.1.235 PMID: 10592235
9. Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, et al. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. Mol. Cell. 2008; 32: 878–887. doi: 10.1016/j.molcel.2008.11.020 PMID: 19111667
10. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 2009; 19: 556–566. doi: 10.1101/gr.090233.108 PMID: 19158363
11. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein—DNA recognition. Annu. Rev. Biochem. 2010; 79: 233–269. doi: 10.1146/annurev-biochem-060408-091030 PMID: 20334529
12. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. U. S. A. 1976; 73: 804–808. doi: 10.1073/pnas.73.3.804 PMID: 1062791
13. Kim Y, Geiger JH, Hahn S, Sigler PB. Crystal structure of a yeast TBP/TATA-box complex. Nature. 1993; 365: 512–520. doi: 10.1038/365512a0 PMID: 8413604
14. Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, et al. Crystal structure of *trp* repressor/operator complex at atomic resolution. Nature. 1988; 335: 321–239. doi: 10.1038/335321a0 PMID: 3419502
15. Hegde RS, Grossman SR, Laimins LA, Sigler PB. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. Nature. 1992; 359: 505–512. doi: 10.1038/359505a0 PMID: 1328886

16. Rohs R, West SM, Liu P, Honig B. Nuance in the double-helix and its role in protein—DNA recognition. Curr. Opin. Struct. Biol. 2009; 19: 171–177. doi: 10.1016/j.sbi.2009.03.002 PMID: 19362815

17. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein—DNA recognition. Nature. 2009; 461: 1248–1253. doi: 10.1038/nature08473 PMID: 19865164

18. Parker SCJ, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA topography correlates with functional noncoding regions of the human genome. Science. 2009; 324: 389–392. doi: 10.1126/science.1169050 PMID: 19286520

19. Shakked Z, Guerstein-Guzikevich G, Eisenstein M, Frolow F, Rabinovich D. The conformation of the DNA double helix in the crystal is dependent on its environment. Nature. 1989; 342: 456–460. doi: 10.1038/342456a0 PMID: 2586615

20. Jones S, van Heyningen P, Berman HM, Thornton JM. Protein-DNA interactions: A structural analysis. J. Mol. Biol. 1999; 287: 877–896. doi: 10.1006/jmbi.1999.2659 PMID: 10222198

21. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell. 2008; 133: 1266–1276. doi: 10.1016/j.cell.2008.05.024 PMID: 18585359

22. Gaj T, Gersbach CA, Barbas CF. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 2013; 31: 397–405. doi: 10.1016/j.tibtech.2013.04.004 PMID: 23664777

23. Akinrimisi E, Ts'o POP. Interactions of Purine with Proteins and Amino Acids. Biochemistry. 1964; 3: 619–626. doi: 10.1021/bi00893a004 PMID: 14193629

24. Thomas PD, Podder SK. Specificity in protein–nucleic acid interaction. FEBS Lett. 1978; 96: 90–94. doi: 10.1016/0014-5793(78)81069-2 PMID: 729797

25. Berg OG, von Hippel PH. Selection of DNA Binding Sites by Regulatory Proteins: Statistical-mechanical Theory and Application to Operators and Promoters. J Mol Biol. 1987; 193: 723–743. doi: 10.1016/0022-2836(87)90354-8 PMID: 3612791

26. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. Nucleic Acids Res. 1998; 26: 2306–2312. doi: 10.1093/nar/26.10.2306 PMID: 9580679

27. 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. doi: 10.1093/nar/29.13.2860 PMID: 11433033

28. Dror I, Zhou T, Mandel-Gutfreund Y, Rohs R. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. Nucleic Acids Res. 2014; 42: 430–441. doi: 10.1093/nar/gkt862 PMID: 24078250

29. Yang L, Zhou T, Dror I, Mathelier A, Wasserman WW, Gordân R, Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 2014; 42: D148–D155. doi: 10.1093/nar/gkt1087 PMID: 24214955

30. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein–DNA complexes. Nucleic Acids Res. 2010; 38: D91–D97. doi: 10.1093/nar/gkp781 PMID: 19767616

31. Prabakaran P, An J, Gromiha MM, Selvaraj S, Uedaira H, Kono H, et al. Thermodynamic database for protein-nucleic acid interactions (ProNIT). Bioinformatics. 2001; 17: 1027–1034. doi: 10.1093/bioinformatics/17.11.1027 PMID: 11724731

32. Kiliç S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: A database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res. 2014; 42: 156–160. doi: 10.1093/nar/gkt1123

33. Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Res. 1996; 24: 238–241. doi: 10.1093/nar/24.1.238 PMID: 8594589

34. Bonaccorsi R, Pullman A, Scrocco E, Tomasi J. The molecular electrostatic potentials for the nucleic acid bases: Adenine, Thymine, and Cytosine. Theor. Chim. Acta. 1972; 24: 51–60. doi: 10.1007/BF00528310

35. Perahia D, Pullman A. The molecular electrostatic potentials of the complementary base pairs of DNA. Theor. Chim. Acta. 1978; 48: 263–266. doi: 10.1007/BF00549025

36. Šponer J, Hobza P. Nonplanar geometries of DNA bases. *Ab initio* second-order Møller-Plesset study. J. Phys. Chem. 1994; 98: 3161–3164. doi: 10.1021/j100063a019

37. Hobza P, Šponer J. Toward true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations. J. Am. Chem. Soc. 2002; 124: 11802–11808. doi: 10.1021/ja026759n PMID: 12296748

38. Jurečka P, Šponer J, Černý J, Hobza P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid

pairs. Phys. Chem. Chem. Phys. PCCP. 2006; 8: 1985–1993. doi: 10.1039/B600027D PMID: 16633685

39. de Ruiter A, Zagrovic B. Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. Nucleic Acids Res. 2014; 43: 708–718. doi: 10.1093/nar/gku1344 PMID: 25550435

40. Pichierri F, Aida M, Gromiha MM, Sarai A. Free-Energy Maps of Base–Amino Acid Interactions for DNA–Protein Recognition. J. Am. Chem. Soc. 1999; 121: 6152–6157. doi: 10.1021/ja984124b

41. Jakubec D, Hostaš J, Laskowski RA, Hobza P, Vondrášek J. Large-Scale Quantitative Assessment of Binding Preferences in Protein–Nucleic Acid Complexes. J. Chem. Theory Comput. 2015; 11: 1939–1948. doi: 10.1021/ct501168n PMID: 26894243

42. Hostaš J, Jakubec D, Laskowski RA, Gnanasekaran R, Řezáč J, Vondrášek J, et al. Representative Amino Acid Side-Chain Interactions in Protein–DNA Complexes: A Comparison of Highly Accurate Correlated Ab Initio Quantum Mechanical Calculations and Efficient Approaches for Applications to Large Systems. J. Chem. Theory Comput. 2015; 11: 4086–4092. doi: 10.1021/acs.jctc.5b00398 PMID: 26575904

43. Wang G, Dunbrack RL. PISCES: A protein sequence culling server. Bioinformatics. 2003; 19: 1589–1591. doi: 10.1093/bioinformatics/btg224 PMID: 12912846

44. Singh J, Thornton JM. Atlas of Protein Side-Chain Interactions, Vols. I & II. Oxford: IRL press; 1992.

45. Singh J, Thornton JM. SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. J. Mol. Biol. 1990; 211: 595–615. doi: 10.1016/0022-2836(90)90268-Q PMID: 2308168

46. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J. Mol. Graph. 1996; 14: 33–38. doi: 10.1016/0263-7855(96)00018-5 PMID: 8744570

47. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970; 48: 443–453. doi: 10.1016/0022-2836(70)90057-4 PMID: 5420325

48. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 1992; 89: 10915–10919. doi: 10.1073/pnas.89.22.10915 PMID: 1438297

49. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16: 276–277. doi: 10.1016/S0168-9525(00)02024-2 PMID: 10827456

50. Berka K, Laskowski RA, Riley KE, Hobza P, Vondrášek J. Representative amino acid side chain interactions in proteins. A comparison of highly accurate correlated *ab initio* quantum chemical and empirical potential procedures. J. Chem. Theory Comput. 2009; 5: 982–992. doi: 10.1021/ct800508v PMID: 26609607

51. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 1995; 117: 5179–5197. doi: 10.1021/ja00124a002

52. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J. Comput. Chem. 2004; 25: 1605–1612. doi: 10.1002/jcc.20084 PMID: 15264254

53. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010; 78: 1950–1958. doi: 10.1002/prot.22711 PMID: 20408171

54. Hess B, Kutzner C, Van Der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. 2008; 4: 435–447. doi: 10.1021/ct700301q PMID: 26620784

55. Chocholoušová J, Feig M. Implicit solvent simulations of DNA and DNA-protein complexes: Agreement with explicit solvent vs experiment. J. Phys. Chem. B. 2006; 110: 17240–17251. doi: 10.1021/jp0627675 PMID: 16928023

56. Gaillard T, Case DA. Evaluation of DNA force fields in implicit solvation. J. Chem. Theory Comput. 2011; 7: 3181–3198. doi: 10.1021/ct200384r PMID: 22043178

57. Kleinjung J, Fraternali F. Design and application of implicit solvent models in biomolecular simulations. Curr. Opin. Struct. Biol. 2014; 25: 126–134. doi: 10.1016/j.sbi.2014.04.003 PMID: 24841242

58. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 1990; 112: 6127–6129. doi: 10.1021/ja00172a038

59. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J. Phys. Chem. A. 1997; 101: 3005–3014. doi: 10.1021/jp961992r

60. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chem. Phys. Lett. 1995; 246: 122–129. doi: 10.1016/0009-2614(95)01082-K

61. Hawkins GD, Cramer CJ, Truhlar DG. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. J. Phys. Chem. 1996; 100: 19824–19839. doi: 10.1021/jp961710n

62. Freedman D, Diaconis P. On the histogram as a density estimator:L2 theory. Z. Wahrscheinlichkeit. 1981; 57: 453–476. doi: 10.1007/BF01025868

63. Berka K, Laskowski RA, Hobza P, Vondrášek J. Energy matrix of structurally important side-chain/side-chain interactions in proteins. J. Chem. Theory Comput. 2010; 6: 2191–2203. doi: 10.1021/ct100007y PMID: 26615945

64. Schlitter J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem. Phys. Lett. 1993; 215: 617–621. doi: 10.1016/0009-2614(93)89366-P

65. Lustig B, Jernigan RL. Consistencies of individual DNA base–amino acid interactions in structures and sequences. Nucleic Acids Res. 1995; 23: 4707–4711. doi: 10.1093/nar/23.22.4707 PMID: 8524664

66. Mirny LA, Gelfand MS. Structural analysis of conserved base pairs in protein–DNA complexes. Nucleic Acids Res. 2002; 30: 1704–1711. doi: 10.1093/nar/30.7.1704 PMID: 11917033

67. Lustig B, Arora S, Jernigan RL. RNA base-amino acid interaction strengths derived from structures and sequences. Nucleic Acids Res. 1997; 25: 2562–2565. doi: 10.1093/nar/25.13.2562 PMID: 9185564

68. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006; 15: 2507–2524. doi: 10.1110/ps.062416606 PMID: 17075131

69. Benos PV, Bulyk ML, Stormo GD. Additivity in protein—DNA interactions: how good an approximation is it?. Nucleic Acids Res. 2002; 30: 4442–4451. doi: 10.1093/nar/gkf578 PMID: 12384591