Prediction of DNA-binding specificity in zinc finger proteins

SUMEDHA ROY†, SHAYONI DUTTA†, KANIKA KHANNA, SHRUTI SINGLA and DURAI SUNDAR*

Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, New Delhi 110 016, India

**Corresponding author (Fax, +91-11-26582282; Email, sundar@dbeb.iitd.ac.in)*

†*These authors contributed equally to this work.*

Zinc finger proteins interact via their individual fingers to three base pair subsites on the target DNA. The four key residue positions −1, 2, 3 and 6 on the alpha-helix of the zinc fingers have hydrogen bond interactions with the DNA. Mutating these key residues enables generation of a plethora of combinatorial possibilities that can bind to any DNA stretch of interest. Exploiting the binding specificity and affinity of the interaction between the zinc fingers and the respective DNA can help to generate engineered zinc fingers for therapeutic purposes involving genome targeting. Exploring the structure–function relationships of the existing zinc finger–DNA complexes can aid in predicting the probable zinc fingers that could bind to any target DNA. Computational tools ease the prediction of such engineered zinc fingers by effectively utilizing information from the available experimental data. A study of literature reveals many approaches for predicting DNA-binding specificity in zinc finger proteins. However, an alternative approach that looks into the physico-chemical properties of these complexes would do away with the difficulties of designing unbiased zinc fingers with the desired affinity and specificity. We present a physico-chemical approach that exploits the relative strengths of hydrogen bonding between the target DNA and all combinatorially possible zinc fingers to select the most optimum zinc finger protein candidate.

[Roy S, Dutta S, Khanna K, Singla S and Sundar D 2012 Prediction of DNA-binding specificity in zinc finger proteins. *J. Biosci.* **37** 483–491] **DOI** 10.1007/s12038-012-9213-7

1. Introduction

The study of zinc finger domains and their interactions with DNA stems from the need to know how the binding of transcription activators and repressors to the genome regulates the expression repertoire of all genes in the cell. The presence of various protein folds that command sequence-specific binding, such as helix-turn-helix, leucine zipper and zinc finger domain, elicits the desire to use them for therapeutic purposes. The most common DNA-binding motif is the cysteine-histidine (*Cys2-His2*) zinc finger in the human genome and most of the multicellular animals. Zinc finger domain was discovered as a transcription factor in TFIIIA, the very first transcription factor to be isolated, during the transcription of 5S RNA gene by RNA PolIII (Pelham and Brown 1980). This protein, which was first isolated from the oocyte of *Xenopus laevis*, is 40 kDa in size and interacts with a 50 bp region called the internal control region, hence protecting it from enzymatic attack (Miller, *et al.* 1985, Klug 2010).

The most intriguing aspect of these proteins was the presence of repeating motifs (Miller, *et al.* 1985). The structure consists of three structural domains that bind to the internal control region of the 5s RNA gene (Brown 1984). Computational analysis showed the presence of nine tandemly repeated similar units of 30 amino acids each, of which 25 amino acids fold around the Zn ion and the remaining 5 act as linkers for consecutive fingers(Miller,

Keywords. Genome targeting; zinc finger proteins

Abbreviations used: CoDA, Context-Dependent Assembly; HMM, Hidden Markov Model; IAS, Interface Alignment Score; OPEN, Oligomerized Pool Engineering; PWM, Position Weight Matrix; SVM, Support Vector Machine; TF, transcription factor; ZiFiT, Zinc Finger Targeter; ZFN, zinc finger nucleases; ZFP, zinc finger proteins

Published online: 25 June 2012

et al. 1985). The conserved invariant pair of Cys-His residues tetrahedrally coordinates with the zinc ion to fold into an independent structural domain called the finger or the module. Other than these conserved amino acid residues, the hydrophilic residues like Tyr, Phe and Leu add stability to the zinc finger protein by formation of hydrophobic clusters (Klug 2010). The remaining residues of the finger are polar and basic in nature.

1.1 *Zinc finger proteins*

The crystal structure of Zif268 (PDB code 1AAY) corroborates the existence of two anti-parallel beta-sheets and an alpha-helix in the folded individual zinc finger. The presence of anti-parallel beta-sheets, which include the loop formed by Cys-Cys residues, and the alpha-helix, which includes the His-His loop at its –COOH terminal, imparts the uniqueness to the zinc finger held together by the zinc ion. Structural independence of the zinc fingers are considered since they are connected by linkers. In a tandemly repeated array, the zinc finger domains are linked by HC link sequences, which are conserved (Berg 1990). It is called H-C since the first residue of this conserved sequence is His and the last one is Cys. Further, few of the residues form a type II beta-turn. Thereby, for the perfect interaction with the DNA, the array of zinc finger domain connected by the H-C links gives it the characteristic radius and pitch owing to the formation of right-handed superhelix.

The crystallized structure of mouse transcript factor Zif268 has three finger domains (Pavletich and Pabo 1991). This protein's alpha-helix interacts with the DNA major groove where each finger interacts with three successive DNA bases at −1, 3 and 6 amino acid residue positions on the helix respectively, via hydrogen bond interaction (figure 1). The 2 amino acid residue position on the alpha-helix interacts with the triplet on the adjacent strand (secondary strand) called cross-strand interaction (Fairall*, et al.* 1993), hence adding significant specificity to the interaction, and the refined model emphasizes the protein's ability to bind to a 4 bp overlapping subsite. In the above pattern of recognition, the residue at position 6 on the alpha-helix contact's with the 5′ base of the primary strand, the residue at position 3 with the middle base and the residue at position −1 with the 3′ base. This helix that initiates a novel method of DNA recognition is called the recognition helix and the DNA code is called the recognition code. Hence, these 7 conserved amino acid residues ensure tertiary folding, whereas the variable residues are responsible for the specificity of each domain.

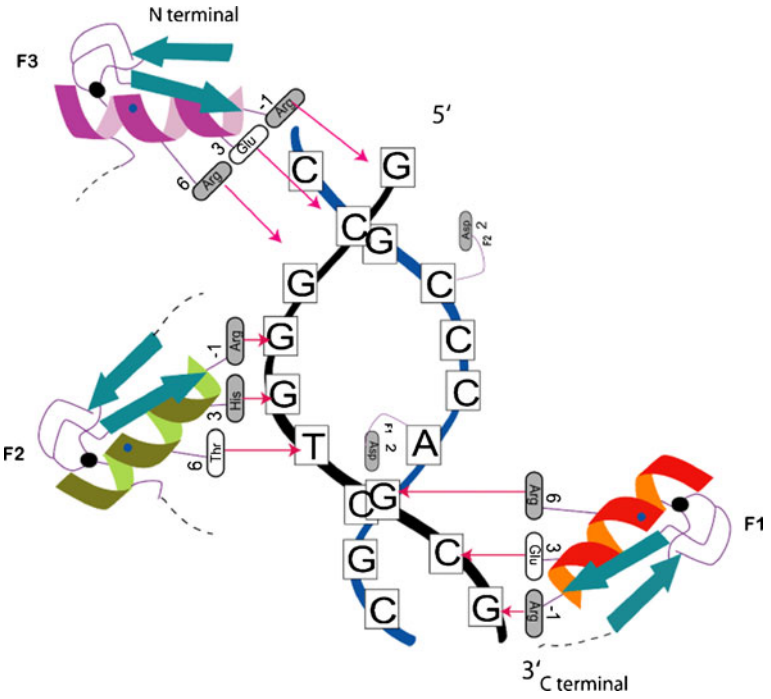

Figure 1. A zinc finger protein binding to its specific target DNA site. A *ββα*-three zinc finger domain protein (Zif268) wrapping around the target DNA; the zinc finger amino acid side chains at the alpha-helical positions −1, 3 and 6 makes sequence-specific hydrogen bond interactions with the nucleotides on the primary DNA strand (shown in black); a cross-strand interaction from the amino acid Asp at position 2 (of finger 1 and finger 2) to the complementary DNA strand (blue) is also depicted. The three fingers are labeled as F1, F2 and F3 with the linkages between consecutive fingers shown by extended broken lines. The tetrahedrally coordinated zinc atoms are shown as small black spheres.

Concluding from a broader domain where the specificity of the residues to the corresponding DNA sequence in quantifying the strength of a DNA–protein interaction is paramount, mutating even a single residue can result in altered affinity or loss of binding (Wright and Lim 2007). This deduction can hence form the major basis of a new approach to develop engineered zinc fingers based on physico-chemical parameters of the ZFP–DNA interaction, where mutating the key residue positions randomly could generate the desired zinc finger protein (ZFP) for our target DNA with maximum affinity and specificity. Inter-finger side chain–side chain interactions contribute significantly to the DNA–protein interface stability. In case of Zif268, inter-finger interactions between Thr-52 and Arg-74 aids the Arg-74–Asp-76 interaction, thereby orienting the Arg-74 residue to interact with the target DNA. The DNA recognition also heavily banks on the intra-finger interactions amidst the side chains of individual fingers (Fairall, *et al.* 1993). The trade-off between affinity and specificity considering the DNA–protein interaction gives insight into the very application of zinc fingers for regulating gene expression.

1.2 *Applications of engineered ZFP*

The work on zinc finger–DNA interaction (Miller, *et al.* 1985) showed the existence of a new protein fold for nucleic acid binding as well as a novel principle of DNA recognition, which, upon exploitation, forms the basis for engineering novel zinc fingers. Fingers with different triplet specificity can be engineered by mutating the key amino acid residues, hence enabling specificity in DNA recognition by ensuring a large number of combinatorial possibilities (Isalan, *et al.* 1998). Further, linking these modules or fingers as they function independently can ascertain the recognition of longer DNA stretches (Liu and Stormo 2008). The two main methods currently used to generate engineered zinc fingers arrays are modular assembly (Liu and Stormo 2008) and bacterial selection methods (Joung, *et al.* 2000, Durai, *et al.* 2006). Another method called bipartite selection improves the specificity of the domains to bind to a target DNA of interest by fusing individual zinc fingers from two pools of engineered zinc fingers (Davis and Stokoe 2010).

1.2.1 *Zinc finger nucleases:* Fusion of the engineered zinc finger domain with non-specific nuclease domains can cause double-stranded breaks in the target DNA, and these molecular scissors are called zinc finger nucleases (ZFN) (Kim, *et al.* 1996, Durai, *et al.* 2005, Cathomen and Joung 2008). The ZFNs consist of a Cys2-His2 zinc finger domain fused to the nuclease domain of a type IIS restriction endonuclease *Fok1* that cleaves double-stranded DNA non-specifically. It is the zinc finger domain that imparts the specificity to the ZFN for its target DNA. Since double-stranded cleavage by *Fok1* is only possible upon its dimerization, the need for the ZFNs to dimerize is a pre-condition for its specific nuclease activity (Mani, *et al.* 2005). ZFNs can be used for targeted mutations as well as gene correction. In case of targeted mutation, ZFNs have been employed to target the first coding exon of CCR5, and their introduction in T-cell lines results in its reduced expression in the cells as well as protection from HIV infection (Reynolds, *et al.* 2003). Gene correction is the inclusion of the corrected sequence by a donor construct encoding it, via homologous recombination, into the respective region of the defective gene. ZFNs hence play an important role in gene correction via homologous recombination in case of treating monogenic disorders such as SCID, Gaucher's disease, x-linked disorders, etc. (Urnov, *et al.* 2005, Klug 2010). ZFPs and ZFNs have therapeutic uses as well, especially where there are successful reports of ZFPs used to target VEGF-A promoters for treating diabetic nephropathy (Liu, *et al.* 2001) and amyotrophic lateral scelerosis (Bae, *et al.* 2003). Developing isogenic cell lines that discern on the basis of presence or absence of drug of interest enhance drug discovery by use of ZFPs (Davis and Stokoe 2010).

2. Prediction tools for zinc finger proteins

One of the key goals in zinc finger engineering has been to produce proteins that can specifically recognize a predetermined DNA sequence. The experimental strategies for engineering zinc finger proteins either through selection or by rational design, although useful, are time consuming, expensive and use techniques not accessible to all laboratories. The regularity in structures of zinc fingers which recognize and bind to nucleotide base triplets, where only the few amino acids at fixed positions interacting with the DNA strand vary, offer the possibility of devising and using a prediction algorithm. As a result of this, ZFPs benefit mostly by computational design. ZFP prediction tools would be valuable for researchers interested in designing specific zinc finger transcription factors and ZFNs for several biological and biomedical applications including targeted gene regulation, enzyme engineering, genome editing, gene therapy, etc. The computational design methods can be divided into prediction tools of two types. The first is based on experimental data which can use either the sequence-based or structure-based approach. The second one captures the fundamental science behind the interaction, in this case, chemical binding and specificity. A list of available online Web tools for prediction of DNA-binding specificity in ZFP is presented in table 1. Some such bioinformatics models and their

Table 1. Web tools for predicting DNA-binding specificity in zinc finger proteins

Tool	Features	URL	Reference
SVM Model	Quantifies degree of DNA-binding preference of particular ZFP–DNA pair to, even in the absence of known binding sites	*http://compbio.cs.princeton.edu/zf*	Persikov *et al.* 2009
ZIFIBI	Predicts C_2H_2 ZFP binding sites in cis-regulatory regions of target DNA by searching against gene name or the SWISS-PROT database	*http://bioinfo.hanyang.ac.kr/ZIFIBI/frameset.php*	Cho *et al.* 2008
ZiFiT	Searches given DNA against chosen standard dataset(s) for modular design of ZFP. Options for Context-Dependent Assembly and Oligomerized Pool Engineering are also available	*http://zifit.partners.org/ZiFiT*	Sander *et al.* 2007
ZiF-BASE	Exhaustive database of natural and engineered proteins; allows search of DNA sequence for binding sites and known ZFPs	*http://web.iitd.ac.in/~sundar/zifbase*	Jayakanthan *et al.* 2009
ZiF-Predict	Allows modular or synergistic prediction of 2- and 3-finger C_2H_2 ZFPs (and ZFNs) that bind to specificied target DNA sites	*http://web.iitd.ac.in/~sundar/zifpredict*	Molparia *et al.* 2010
Zinc Finger Tools	Searches for potential target sites within DNA; predicts ZFP for target; predicts target site for given ZFP and finds similarity of particular DNA with target site	*http://www.zincfingertools.org*	Mandell and Barbas 2006

approaches for prediction of DNA-binding specificity in zinc fingers are described below:

2.1 *Structure-based approach (2001)*

A computational scheme that uses knowledge-based parameters for amino acid–base interactions based on data from crystallographic information of protein–DNA has been described. By applying these parameters to specified binding models, a score that reflects the stereo-chemical complementarity and structural compatibility between a protein sequence and a DNA site can be evaluated. The advantage of this procedure is that it can be used for the prediction of binding sites for newly identified proteins that are clustered to a defined family based on binding data. However, it does not always predict the known site at the first rank because it does not consider other factors that affect binding, such as the sequence context of the binding sites and coupled interactions. Secondly, possible position-dependent effects that are specific to each binding motif are masked (Mandel-Gutfreund, *et al.* 2001).

2.2 *Simple physical model (2004)*

This model does not require any prior knowledge of structure or sequence. Protein design approaches like the combination of simple physical models of macromolecular energetics and rapid algorithms for sampling side-chain conformations provide a powerful quantitative description of protein–DNA interfaces in their entirety. An all-atom description of both the DNA and protein is used. The model uses a simple physically based energy function, fixed DNA and protein backbone conformations, and a rotamer-based description of protein side-chain conformation. This model does not include electrostatic and water-mediated interactions dictating affinity and specificity. Also, multiple protein–DNA binding modes have to be considered. This model requires improvements by utilizing backbone sampling and docking techniques. Prediction in structurally homologous complexes is limited by specificity (Havranek, *et al.* 2004).

2.3 *Zinc Finger Tools (2006)*

Zinc Finger Tools is a multiple-utility Web server that can scan a given DNA sequence for consecutive DNA triplets that can be targeted with zinc finger domains, design a zinc finger protein for a valid DNA sequence and, most importantly, predict the binding sites in ZFP. The user inputs the amino acid sequence of the ZFP and not the DNA sequence. The algorithm is such that it recognizes only helices and ignores all other sequence to minimize the impact of poor sequence quality or extraneous sequences. This tool can also be used to ensure that the intended targeted site of a designed ZFP is correct (Mandell and Barbas 2006).

2.4 *Zinc Finger Targeter (2007)*

Zinc Finger Targeter (ZiFiT) was developed by the Zinc Finger Consortium as an effort to provide a simple and easy tool for ZFP and ZFN design. It is a popular Web tool that provides an integrated modular design approach by

incorporating three different datasets enumerating zinc finger-binding patterns for independent modules developed by Barbas, Sangamo and ToolGen (Segal, *et al.* 1999, Dreier, *et al.* 2001, Liu, *et al.* 2002, Bae, *et al.* 2003, Dreier, *et al.* 2005) . The user may enter the query DNA sequence and choose one or more of these sets. The data from the chosen sets is then used to identify the best DNA target site in the query and the corresponding zinc finger arrays are returned as a text file. Scores are given alongside predictions as an indication of their chances of success, as measured by a bacterial two-hybrid assay. Recently, options of using Context-Dependent Assembly (CoDA) (Sander, *et al.* 2011) and Oligomerized Pool Engineering (OPEN) (Maeder, *et al.* 2008) for design have been added. The Consortium advises the use of CoDA due to its simplicity over OPEN and much higher success likelihood than the modular approaches (Sander, *et al.* 2007).

2.5 *PWM prediction: Sensitivity to docking geometry (2007)*

Transcription factor (TF) binding-specificity is described by a consensus sequence or Position Weight Matrix (PWM). Given a TF–DNA complex structure, a scoring function is used to evaluate relative affinities. For scoring, approaches like knowledge-based structural potentials or all-atom modelling of complexes are used. Since PDB for every complex is not available, homology modelling of complexes is employed. Conserved stereo-specific H-bond interactions are informative for template selection. This requires similarity of docking geometry, which is quantified in terms of Interface Alignment Score (IAS). The IAS score and prediction accuracy are related. For modelling side chains and bases, residue confirmations are iteratively minimized by selecting rotamers, generated by wriggling algorithm that yielded lowest energy of complex (Siggers and Honig 2007).

2.6 *ZIFIBI (2008)*

Publicly available data was used to predict the interaction patterns between the amino acids of the C_2H_2 zinc finger domains and nucleotides. Then a 3 Position Weight Matrix (PWM) was constructed for positions −1, 3 and 6 of the alpha-helix and a Hidden Markov Model (HMM) can be used to calculate the most probable state path of three nucleotides sequences. ZIFIBI provides functions to search DNA binding sites and by the gene name, SWISS-PROT ID or SWISS-PROT access number for specific protein and to search target genes. These computations are used to predict C_2H_2 zinc finger transcription factor binding sites in cis-regulatory regions of their target genes. The ZIFIBI database contains proteins with potential binding sites for zinc finger proteins that have not yet been experimentally identified. The average Euclidean distance of ZIFIBI was 0.613929, which is lower than those found in other studies, and its predictions were similar to other studies, thereby demonstrating its superiority (Cho, *et al.* 2008).

2.7 *SVM-based approach (2009)*

Support Vector Machine (SVM) is a state-of-the-art classification technique. Using canonical binding model, the C_2H_2 zinc finger protein–DNA interaction interface is modelled by the pairwise amino acid–base interactions. Using a classification framework, known examples of non-binding ZF–DNA pairs are incorporated. Using a linear kernel, information about relative binding affinities of ZF–DNA pairs is incorporated. A polynomial SVM also captures dependencies among the canonical contacts. SVMs search for a weight vector w that best separates binding and non-binding proteins. The advantage of the polynomial kernel over the linear SVM may suggest the limitation of the originally used canonical representation. It features vectors into a higher dimensional space, thereby making possible implicit inclusion of higher order interactions not listed in the original canonical model. But, use of the polynomial kernel does not allow the incorporation of relative binding information (Persikov, *et al.* 2009).

2.8 *ZiF-Predict (2010)*

ZiF-Predict is a Web tool that is based on artificial neural network and helps in predicting recognition helices for C_2H_2 zinc fingers binding to specific DNA targets. An exhaustive dataset of 7-residue-long recognition helices of three-finger ZFPs, ZFNs and their corresponding triplets reported in literature were used. In this user-friendly interface, users can input a DNA sequence and select the option to predict two or three zinc fingers for the same. This Web tool also incorporates both the molecular prediction and the synergistic interactions between the fingers, a feature not available elsewhere. For instance, depending on the position of the finger motifs, binding affinities to the target sequence may differ. The network consisted of an input layer followed by two hidden layers and a single output neuron (Molparia, *et al.* 2010).

3. Zinc finger prediction using a physico-chemical approach

There have been constant attempts to generate a recognition code or a 1:1 map relating each of the 64 triplets to the corresponding recognition helix, somewhat similar to the codon : amino acid table. It was expected that as more and more data about zinc fingers are collected, a recognition

code would clearly emerge. The basic fallacy with this approach is the degenerate binding of zinc fingers to multiple DNA target sites, albeit with different affinities. It has been found that the zinc finger motifs bind synergistically instead of a modular fashion, whereas the recognition code neglects any position-specific dependencies on the binding preferences. Moreover, there is not sufficient data available to conclusively draw a direct correlation for such sequence-based approaches to be successful. Experimental approaches like the phage display method overcome the inherent assumptions in the previous approach and rely solely on the experimentally observed affinity (Rebar and Pabo 1994). However, the labour-intensive, time-consuming and expensive method gradually propelled the need for computational prediction tools that directly quantify the affinity and specificity between the target DNA and the zinc finger to make predictions. Such physico-chemical approaches draw 'true inferences' by solely using information from 3-D structures, without relying on any previously available biased or limited data. Therefore, while saving on time, money and effort, it quantifies actual binding forces to determine the relative specificity and may be used to generate a ranked list of ZFPs that will most likely bind to the target DNA.

The lack of a large repository of naturally occurring/known ZFP–DNA complexes acts as a major hindrance for exploiting natural binding patterns to understand and model engineered ZFPs to a given DNA target. The first step that logically followed was the creation of an organized, exhaustive database of natural and engineered proteins from various protein databases as well as literature. Zif-BASE is such a database, which also facilitates the search of a zinc finger corresponding to a binding site present in a stretch of a raw DNA sequence input. The tool provides supplementary information such as the source organism (wherever applicable), the 7-alpha-helix recognition residues as well as the 3-D structure (Jayakanthan*, et al.* 2009). The next step that followed was using the available knowledge to predict the C_2H_2 zinc finger that is most likely to bind to the given DNA. This tool, ZiF-Predict, as discussed in the previous section, is based on a multilayered artificial neural network model that is trained on the known zinc finger binding data from Zif-BASE (Molparia*, et al.* 2010). With adaptive learning, and choice of making both modular and synergistic zinc finger predictions, it was a good starting point. However, the scope of such a sequence-based design approach, based purely on propensity and favoured interactions as derived from known data, is restricted by the limited data available for ZFPs.

Due to the inherent limitation of homology or sequence-based prediction tools, the focus here was shifted to a more realistic, un-biased, physico-chemical approach to the problem of ZFP prediction. After studying the various interactions among known zinc finger complexes, it was found that the hydrogen bonding forces are primarily responsible for specificity and affinity among the protein and DNA. As a result, an approach that would make predictions based on strength of hydrogen bonds as a measure of affinity has been developed (unpublished data). With the standard Zif-268 as template, mutations were performed at four recognition helix positions (−1, +2, +3 and +6) for each of the three fingers to generate all possible ZFPs and the scores generated determine the rank of each protein for the given DNA. The method also considers the dependency of the position of zinc finger on the interactions unlike other modular approaches. The major benefit of such an approach is that it is objective in quantifying the binding interactions without being prejudiced by the limited and biased current knowledge repository of zinc fingers, giving only predictions with quantified affinity. The approach is, however, dependent upon the inherent assumptions and simplifications of the binding model and energy function. The prediction from such an approach returns the top recognition residues at each finger corresponding to the codon at a particular position in a 9 bp target DNA site. This typically leads to a large fraction of results being returned with similar scores, instead of a unique few, also including false-positives. For a researcher, this poses the problem of selecting from the pool of results and non-specific selections with lower affinities not leading to the desired binding. Experimental randomization of the Zif-268 and pull-down assays using a given DNA sequence suggested only 66% accuracy of the predictions, which needed to be further verified (unpublished data).

In order to utilize the inherent strengths of the structure-based algorithm developed while narrowing the results set with minimum false-positives, we resorted to protein–DNA complex docking. This allowed us to evaluate relative affinities of the predictions for a given query DNA and gave us the final confirmation of the ZFP–DNA complex in PDB format. To get some basic insights, 5 random ZFP predictions for the input DNA sequence of AAAAACAAG from the set of over 1000 results given by our algorithm were docked using the HADDOCK online Web server (*http://haddock.science.uu.nl/services/HADDOCK/haddock.php*) (de Vries*, et al.* 2010). As controls, the affinity of Zif-268 to its consensus sequence AGCGTGGGCGT, as well as to our test sequence, were also compared. The HADDOCK score, which is a weighted sum of various energy components, was used by the software to sort and select the most stable complex conformation. Since the score is derived from the energy of the complex, it may be used as broad indicator of DNA–ZFP affinity. So the HADDOCK score of the most stable complexes of each protein under consideration with its corresponding DNA were compared to judge the relative affinities of the ZFPs for the given DNA. This is based on the assumption that the HADDOCK scores can be compared when the DNA sequence in each complex is same and the proteins themselves vary only at the key residues. As expected, it was found that Zif-268 had lower affinity for

the test sequence as compared to its consensus sequence. The relative order of the predictions by the algorithm did not reflect any order of increasing or decreasing affinity. Out of the 5 predictions, 1 bound weaker, 1 equally and 3 stronger than Zif-268 bound with the query sequence AAAAACAAG. Moreover, 2 out of 5 predictions bound more strongly to the test sequence as compared to Zif-268 with its consensus sequence. This reflects the strength of our structure-based design strategy in its ability to predict ZFPs with greater affinity than found to occur naturally.

To correlate the affinity with structural differences, the complex of the test DNA with Zif-268 was superimposed with the complexes of the strongest and weakest binding proteins. The RMSD values (1.93 for the strongest and 1.36 for the weakest) reflected that both the predicted proteins differ significantly from the template Zif-268 structure. This difference in structure is what translates into the difference in affinity. Of the two, the weaker one had a structure more similar to Zif-268 than the stronger one. This is an interesting observation, raising the possibility that this deviation in structure from Zif-268 renders it greater affinity. While visualization of these superimposed complexes in PyMOL, showing significant twisting of the DNA molecule in the process, raises doubts on the reliability of the RMSD values, it will be interesting to study a direct correlation of the structure with affinity.

4. Discussion

Since the ZFP–DNA interaction seems to appear quite redundant and flexible, rendering the understanding of this fundamental mechanism at the molecular aspect is Herculean and the prediction algorithms based on structural information would be a boon to developing engineered ZFPs for use in human therapeutics. The advancements in computational approaches have significantly helped in the engineering of zinc finger proteins by putting the experimental databases to effective use. They are faster, universally accessible and cheaper. Development of prediction tools based on the physico-chemical approach offers the advantage that it is free of bias introduced by limited or uneven research in the field. However, such an approach, with the present understanding of protein–DNA interactions, can only hope to eliminate most of random possibilities and provide a practical starting point to experimental studies, which it cannot completely substitute. Hence, deriving patterns out of structure analysis of the existing PDB structures gives us a better understanding of the structure–function relationship that can benefit in predicting the target DNA stretch for a ZFP or vice versa. Hence, sequence information, cross-strand interaction, intra- and inter-finger interactions are imperative factors governing the specificity of the ZFP–DNA interaction, the strength of which banks majorly on hydrogen bond energy, Van der Waal energy, water-mediated interactions, experimental binding data and computer simulations data. A combined approach to the study of all the above factors and their application in developing an algorithm would competently predict plethora of engineered ZFPs with decreasing specificity or affinity or both for our target DNA stretch, thus emphasizing its use in genome targeting and in therapeutic applications.

The development of our physico-chemical-based prediction tool was an effort to bridge the gap between insufficient knowledge of existing ZFPs and the challenge of unbiased design of ZFPs with high affinity and selectivity. The concern, however, is that the presentation of a large set of 'possible' ZFPs for each input DNA makes the choice of a few candidate proteins for experimental purposes difficult and ambiguous. To overcome this issue, the results from the tool need to be further screened and ranked, limiting the output to at most the top 10 sequences. Incorporating a final docking step in the algorithm will allow the otherwise equally ranked proteins to be distinguished based on affinity and hence report only the top 10 scorers. For screening and improving the results, we are also considering the inclusion of neural-network-based scores from the ZiF-Predict software for taking the middle path between structural and sequence-based prediction. Cumulative scores will probably give better predictions. For the same, the generic properties of DNA targets bound by DNA-binding proteins in general and zinc fingers in particular will be studied. Additionally, since ZFPs are known to naturally prefer G-rich DNA targets (Isalan 2012), a DNA target-search algorithm to identify and give preference to binding sites with high G percentage (only if present) will help predict more specific and strongly binding ZFPs without necessarily biasing the choice of the DNA for the user. A broader aim of performing the docking is also to gain insight into the factors and properties that play a crucial role in determining the affinity of a ZFP towards the particular DNA sequence. A large-scale study, with the aid of machine learning, will help elucidate the common factors among the highest scorers. It is reasonable to expect that performing a docking for a large number of predictions for a given DNA from both physico-chemical as well as other sequence-based prediction methods would provide further insights. While there are strong proponents of both schools, this would serve as a mean to quantify the accuracy and strength of predictions of each method. As mentioned above, it would also be interesting to consider a combined, holistic approach.

Acknowledgements

KK and SS were recipients of the Summer Undergraduate Research Award (SURA) from IIT Delhi. This study was

made possible in part through the support of a grant from the Lady Tata Memorial Trust, Mumbai, and the Department of Biotechnology (DBT) under the IYBA scheme to DS. Computations were performed at the Bioinformatics Centre, supported by the DBT.

References

Bae KH, Kwon YD, Shin HC, Hwang MS, Ryu EH, Park KS, Yang HY, Lee DK, *et al.* 2003 Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat. Biotechnol.* **21** 275–280

Berg JM 1990 Zinc finger domains: hypotheses and current knowledge. *Annu. Rev. Biophys. Biophys. Chem.* **19** 405–421

Brown DD 1984 The role of stable complexes that repress and activate eucaryotic genes. *Cell* **37** 359–365

Cathomen T and Joung JK 2008 Zinc-finger nucleases: the next generation emerges. *Mol. Ther.* **16** 1200–1207

Cho SY, Chung M, Park M, Park S and Lee YS 2008 ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem. Biophys. Res. Commun.* **369** 845–848

Davis D and Stokoe D 2010 Zinc finger nucleases as tools to understand and treat human diseases. *BMC Med.* **8** 42

de Vries SJ, van Dijk M and Bonvin AM 2010 The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5** 883–897

Dreier B, Beerli RR, Segal DJ, Flippin JD and Barbas CF 3rd 2001 Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem*. **276** 29466–29478

Dreier B, Fuller RP, Segal DJ, Lund CV, Blancafort P, Huber A, Koksch B and Barbas CF 3rd 2005 Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.* **280** 35588–35597

Durai S, Bosley A, Abulencia AB, Chandrasegaran S and Ostermeier M 2006 A bacterial one-hybrid selection system for interrogating zinc finger-DNA interactions. *Comb. Chem. High Throughput Screen* **9** 301–311

Durai S, Mani M, Kandavelou K, Wu J, Porteus MH and Chandrasegaran S 2005 Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res.* **33** 5978–5990

Fairall L, Schwabe JW, Chapman L, Finch JT and Rhodes D 1993 The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366** 483–487

Havranek JJ, Duarte CM and Baker D 2004 A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.* **344** 59–70

Isalan M, Klug A and Choo Y 1998 Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry* **37** 12026–12033

Isalan M 2012 Zinc-finger nucleases: how to play two good hands. *Nat. Method.* **9** 32–34

Jayakanthan M, Muthukumaran J, Chandrasekar S, Chawla K, Punetha A and Sundar D 2009 ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics* **10** 421

Joung JK, Ramm EI and Pabo CO 2000 A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **97** 7382–7387

Kim YG, Cha J and Chandrasegaran S 1996 Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. USA* **93** 1156–1160

Klug A 2010 The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.* **79** 213–231

Liu J and Stormo GD 2008 Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics* **24** 1850–1857

Liu PQ, Rebar EJ, Zhang L, Liu Q, Jamieson AC, Liang Y, Qi H, Li PX, *et al.* 2001 Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J. Biol. Chem.* **276** 11323–11334

Liu Q, Xia Z, Zhong X and Case CC 2002 Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *J. Biol. Chem.* **277** 3850–3856

Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, *et al.* 2008 Rapid 'open-source' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell* **31** 294–301

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

Mandell JG and Barbas CF 3rd 2006 Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res.* **34** W516–523

Mani M, Smith J, Kandavelou K, Berg JM and Chandrasegaran S 2005 Binding of two zinc finger nuclease monomers to two specific sites is required for effective double-strand DNA cleavage. *Biochem. Biophys. Res. Commun.* **334** 1191–1197

Miller J, McLachlan AD and Klug A 1985 Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.* **4** 1609–1614

Molparia B, Goyal K, Sarkar A, Kumar S and Sundar D 2010 ZiFPredict: a web tool for predicting DNA-binding specificity in C2H2 zinc finger proteins. *Genom. Proteom. Bioinformatics* **8** 122–126

Pavletich NP and Pabo CO 1991 Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252** 809–817

Pelham HR and Brown DD 1980 A specific transcription factor that can bind either the 5S RNA gene or 5S RNA. *Proc. Natl. Acad. Sci. USA* **77** 4170–4174

Persikov AV, Osada R and Singh M 2009 Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25** 22–29

Rebar EJ and Pabo CO 1994 Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263** 671–673

Reynolds L, Ullman C, Moore M, Isalan M, West MJ, Clapham P, Klug A and Choo Y 2003 Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors. *Proc. Natl. Acad. Sci. USA* **100** 1615–1620

Sander JD, Dahlborg EJ, Goodwin MJ, Cade L, Zhang F, Cifuentes D, Curtin SJ, Blackburn JS, *et al.* 2011 Selection-free zinc-

finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Method.* **8** 67–69

Sander JD, Zaback P, Joung JK, Voytas DF and Dobbs D 2007 Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. *Nucleic Acids Res.* **35** W599–605

Segal DJ, Dreier B, Beerli RR and Barbas CF 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

Siggers TW and Honig B 2007 Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.* **35** 1085–1097

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD and Holmes MC 2005 Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435** 646–651

Wright JD and Lim C 2007 Mechanism of DNA-binding loss upon single-point mutation in p53. *J. Biosci.* **32** 827–839