Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

ZiF-Predict: A Web Tool for Predicting DNA-Binding Specificity in C2H2 Zinc Finger Proteins

Bhuvan Molparia[1#], Kanav Goyal[2#], Anita Sarkar[1#], Sonu Kumar[1#], and Durai Sundar[1*]

[1]*Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India;*
[2]*Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Delhi, New Delhi 110016, India.*

Genomics Proteomics Bioinformatics 2010 Jun; 8(2): 122-126. DOI: 10.1016/S1672-0229(10)60013-7

Abstract

Engineering zinc finger protein motifs for specific DNA targets in genomes is critical in the field of genome engineering. We have developed a computational method for predicting recognition helices for C2H2 zinc fingers that bind to specific target DNA sites. This prediction is based on artificial neural network using an exhaustive dataset of zinc finger proteins and their target DNA triplets. Users can select the option for two or three zinc fingers to be predicted either in a modular or synergistic fashion for the input DNA sequence. This method would be valuable for researchers interested in designing specific zinc finger transcription factors and zinc finger nucleases for several biological and biomedical applications. The web tool ZiF-Predict is available online at http://web.iitd.ac.in/~sundar/zifpredict/.

Key words: artificial neural network, C2H2 zinc fingers, ZiF-Predict

Introduction

Zinc finger proteins (ZFPs) contain a finger-shaped fold that coordinates zinc ions with a combination of cysteine and histidine residues for their fold stabilization. Their structure comprises of two anti-parallel β-sheets and an α-helix. Each finger binds to a three-base pair DNA template in the major groove. Changing the amino acids at certain key positions generates zinc finger (ZF) motifs with different triplet sequences and varying binding specificities (*1, 2*). Binding to a longer DNA sequence improves the specificity of the ZF attaching to its target site and only a DNA sequence having six triplets (or more) is unique in the human genome. Custom-designed ZFPs provide a powerful platform technology since other functional domains like non-specific *Fok* I cleavage domain (N), transcription activator domain (A), transcription repressor domain (R) and methylase (M) can be fused to the ZFPs to form zinc finger nucleases (ZFNs), zinc finger transcription activators (ZFAs), zinc finger transcription repressors (ZFRs) and zinc finger methylases (ZFMs), respectively. Custom-designed ZFs have tremendous applicability in targeted gene regulation, enzyme engineering, genome editing, gene therapy, *etc* (*3*).

The determination of DNA-binding specificity in ZFs is functionally relevant. Identification of the ZFPs to specific DNA-target site can only been accomplished through laborious and time-consuming

# Equal contribution.
*Corresponding author.
E-mail: sundar@dbeb.iitd.ac.in

experiments. Therefore, computational identification may provide a good alternative. In this study, we have developed a method for the prediction of DNA-binding specificity in ZFPs that focuses on identifying the target (binding) sites of the classical C2H2 fingers in the user's input DNA sequence and predicting the protein recognition helices that bind to them using artificial neural network. The method has been implemented online as ''ZiF-Predict'' at the site http://web.iitd.ac.in/~sundar/zifpredict hosted at the IIT Delhi website.

Method

Salient features

ZiF-Predict allows modular as well as synergistic ZFP predictions. It is based on an artificial neural network model, built and operated on MATLAB (R2008a) to simulate complex relationships between the user's input and the training data, classifying them into different groups, distinguishing random seven amino acid stretches from those present in the recognition helix and making predictions based on specific and non-specific interactions. It is designed using PHP, HTML and Java™ Development Kit 1.2. ZiF-Predict can also predict specific ZFNs for target DNA-binding sites of the form [$(NNN)_3$-(spacer)-$(NNN)_3$], the spacer being 4-6 base pairs in length and N = G, A, T or C.

Training algorithm

For determining DNA-binding specificity in ZFPs, we chose a seven amino acid stretch in the ZFP α-helix. Our algorithm considers the positions -1, +3 and +6 relative to the start of the ZFP α-helix (key positions in determining binding specificity) as these are particularly variable in different ZFPs. For classification purposes, we created a multi-layered neural network model, which was iteratively trained such that it finds the best possible hyperplane that can separate different categories of N-dimensional input vectors and formulates a weight matrix for distinguishing between the various classes.

Network architecture and datasets

The ZiF-Predict network consists of an input layer followed by two hidden layers and a single output neuron (**Figure 1A**). The inputs (a_j) are multiplied by the weights (w_{ji}) assigned to the connections between neurons of two adjacent layers. These weights get modified after every input to best fit the problem. The sum of the weighted inputs and bias forms the input to each neuron in the first layer, Σ_j ($w_{ji}a_j$) (**Figure 1B**). ZiF-Predict neurons use the Tan Sigmoid function as their differentiable transfer/activation function (T) to generate their output (a_i). The network is a feed-forward back-propagation network. The weights and biases are calculated as

$$w_{j\ new} = w_j + \alpha \times I_j \times Err,$$

where α is the learning rate, I_j is an input-dependent term and Err is the error. This network follows the above-mentioned gradient descent training system with momentum and adaptive learning rule.

Recognition helices of ZFP that specifically recognize each of the ANN, GNN, CNN or TNN triplets in DNA with high affinity and sequence specificity are not completely available. ZiF-Predict has been created to predict DNA-recognition specificity in ZFs for all the possible DNA targets. Having the ZF motif modules for all the 64 DNA triplets in hand will enable rapid design of the desired ZFNs needed for targeting and modifying a specific locus within the human genome.

In ZiF-Predict, the artificial neural network was trained and tested on an exhaustive dataset of seven-residue-long recognition helices of ZFPs (both natural and engineered ones) and the synergy was taken into account between the adjacent fingers and their positions. This is a highlight of ZiF-Predict because it can predict the recognition helices of ZFP for any DNA triplet.

Momentum

Momentum is a functional feature of the neural network that helps in providing faster convergence. In ZiF-Predict, it allows the network to respond not only to the local gradient, but also to recent trends in the error surface. Acting like a low-pass filter, momentum allows the network to ignore small features in the

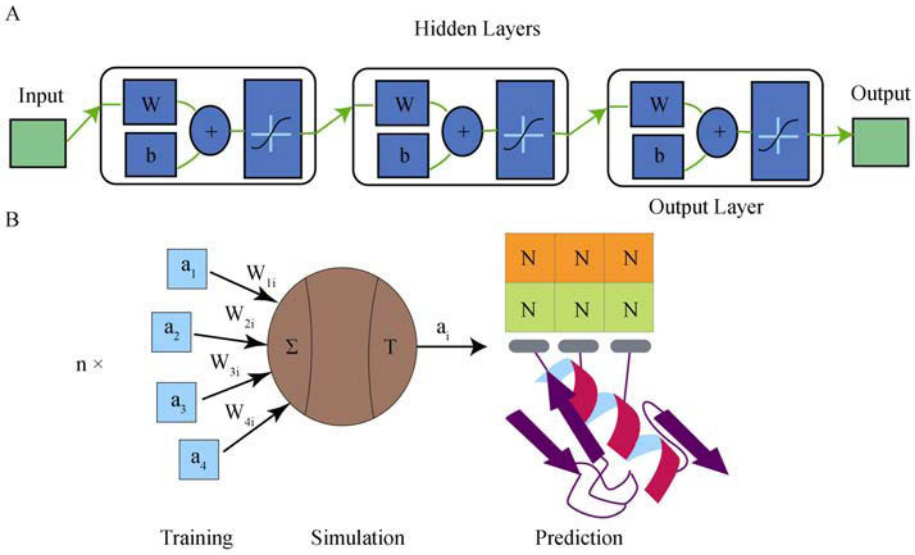

Figure 1 Graphical representation of ZiF-Predict. **A.** The schematic diagram of the artificial neural network model for prediction. **B.** Working concept of the artificial neural network (a_j=inputs, w_{ji} =weights, b=bias, ΣT=Tan sigmoid function).

error surface. Without momentum a network may get stuck in a shallow local minimum. Momentum is added to back-propagation learning by making weight changes equal to the sum of a fraction of the last weight change and the new change suggested by the back-propagation rule. The magnitude of the effect that the last weight change is allowed to have is mediated by a momentum constant, *mc*, which can be any number between 0 and 1. When the momentum constant is 0, a weight change is based solely on the gradient. When the momentum constant is 1, the new weight change is set to equal the last weight change and the gradient is simply ignored. If the new performance function on a given iteration exceeds the performance function on a previous iteration by more than a predefined ratio, the new weights and biases are discarded, and the momentum coefficient is set to zero.

Adaptive learning

The larger the learning rate, the bigger the step towards the converging point. If the learning rate is made too large, the algorithm becomes unstable. If the learning rate is set too small, the algorithm takes a long time to converge. In this method, first, the initial network output and error are calculated. At each epoch, new weights and biases are calculated using the current learning rate. New outputs and errors are then calculated. If the new error exceeds the old error by more than a predefined ratio, the new weights and biases are discarded. In addition, the learning rate is decreased. Otherwise, the new weights and biases are kept. If the new error is less than the old error, the learning rate is increased.

Datasets

Our network was trained and tested on an exhaustive dataset of seven-residue-long recognition helices of three-finger ZFPs, ZFNs and their corresponding triplets reported in literature (*4-8*) .

Implementation

ZiF-Predict initially searches the input DNA sequence (≥30 bp) for the most likely DNA-binding target (9 and 6 bp for 3 and 2 ZFPs, respectively). The prioritization for target selection and recognition helix prediction is dependent upon triplets favored as binding sites in published data (*4-8*). The result page displays the best ranked recognition helices, based upon hydrogen bonding and van der Waals forces between amino acids at positions -1, +3 and +6 and their corresponding target DNA (specific interactions) as well as the backbone (non-specific interactions) (**Figure 2**).

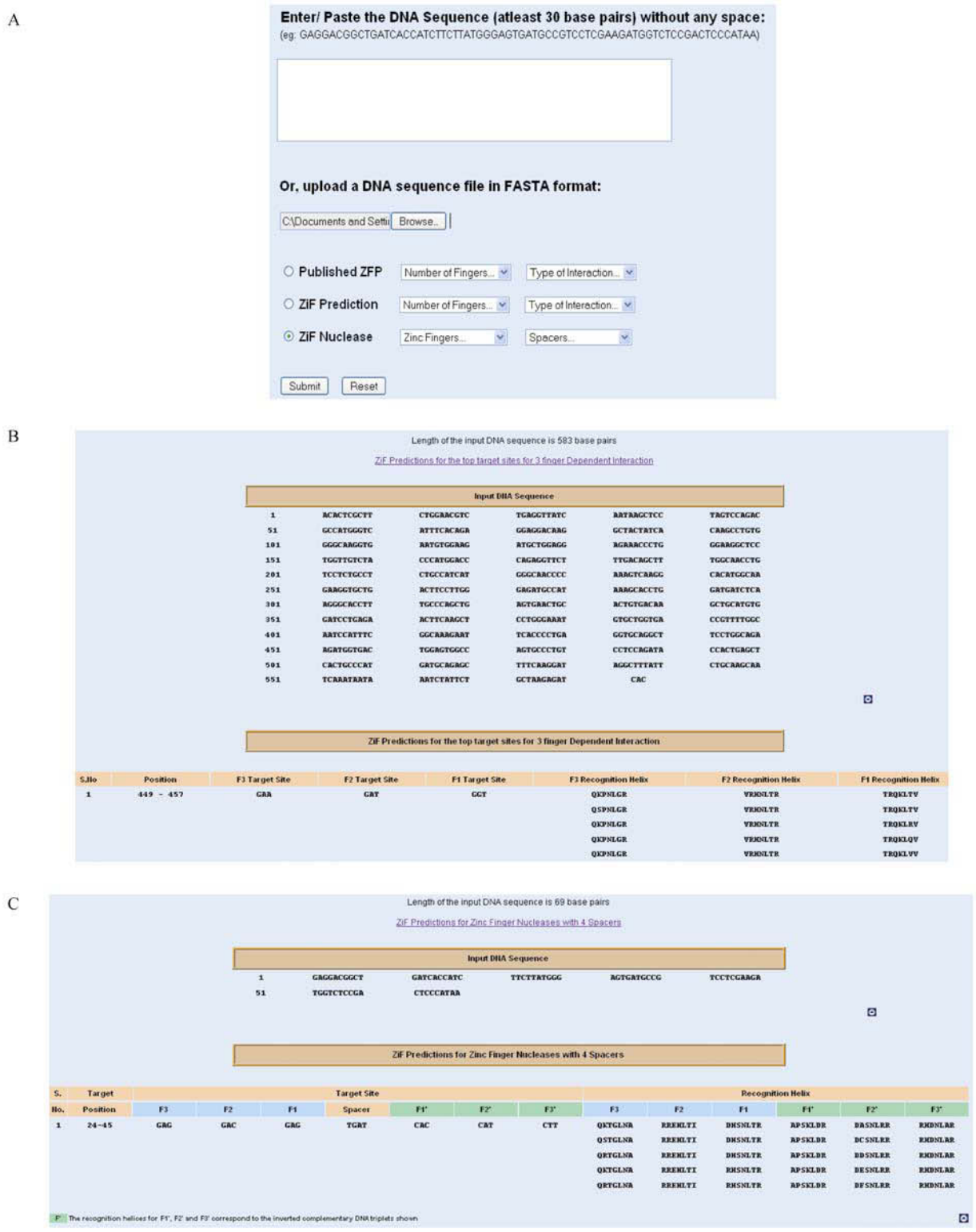

Figure 2 Results of ZiF-Predict using the input DNA sequence. **A.** Sequence input page. **B.** Result page showing the best ranked ZFP recognition helices, based upon hydrogen bonding and van der Waals forces between amino acids at positions -1, +3 and +6 and their corresponding target DNA (specific interactions). **C.** Result page showing the recognition helices for ZFNs comprising of two 3-zinc finger proteins separated of user-input spacer length.

Discussion

We have developed an accurate method for prediction of ZFPs using artificial neural network, trained on experimentally tested ZFPs. A unique feature of this method is that in addition to modular prediction, the synergistic interactions between the fingers have also been considered for prediction, which is not available currently elsewhere. Individual finger motifs at a certain position (say Finger 1, Finger 2 or Finger 3) in a constituent three-finger protein that binds to a target DNA triplet sequence, may not bind tightly to the same target if it has a different flanking finger motif. In addition, there will be a difference in binding affinity to the target sequence depending upon the suitable placing of individual finger motifs. Any attempt to build multi-finger peptides by joining individual fingers can now take the advantage of getting information on individual motifs, its position in the constituent ZFP along with its target triplet and neighboring motifs as designed in the synergistic mode of ZiF-Predict.

ZiF-Predict would be handy for rapid screening of gene-specific ZFP and ZFN target sites in plant or mammalian genomes. This will greatly aid genome researchers to design and/or evolve ZFPs for creating custom zinc finger-chimeric proteins for several genomic applications, including human therapeutics.

Acknowledgements

This work was supported partly by the Department of Biotechnology (DBT) under the IYBA scheme and Department of Information Technology (DIT), Govt. of India to DS and partly by a Summer Undergraduate Research Award (SURA) from IIT Delhi to BM and KG. The authors acknowledge the Bioinformatics facility at the DBT-funded Distributed Information Sub Centre at IIT Delhi.

Authors' contributions

BM and SK built the neural network model. KG developed the web interface and the scripts for the prediction tool. AS was responsible for data acquisition, arrangement and designing of the test and training sets. SK, AS and BM devised the schema for the algorithm underlying the prediction system. DS conceived and coordinated the project, guided its conception and design, helped in the interpretation of data, refined the drafted manuscript and gave overall supervision to the project. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Desjarlais, J.R. and Berg, J.M. 1993. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl. Acad. Sci. USA* 90: 2256-2260.

2 Wolfe, S.A., *et al*. 1999. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.* 285: 1917-1934.

3 Durai, S., *et al.* 2005. Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res.* 33: 5978-5990.

4 Jayakanthan, M., *et al*. 2009. ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics* 10: 421.

5 Dreier, B., *et al*. 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.

6 Maeder, M.L., *et al*. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell* 31: 294-301.

7 Liu, Q., *et al.* 2002. Validated zinc finger protein designs for all 16 GNN DNA triplet targets. *J. Biol. Chem.* 277: 3850-3856.

8 Segal, D.J., *et al*. 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.