Potential application of FoldX force field based protein modeling in zinc finger nucleases design

HE ZuYong[1], MEI Gui[2], ZHAO ChunPeng[3] & CHEN YaoSheng[1*]

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510006, China;
[2]Institute of Animal Sciences, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China;
[3]College of Animal Science and Technology, South China Agricultural University, Guangzhou 510640, China

Engineered sequence-specific zinc finger nucleases (ZFNs) make the highly efficient modification of eukaryotic genomes possible. However, most current strategies for developing zinc finger nucleases with customized sequence specificities require the construction of numerous tandem arrays of zinc finger proteins (ZFPs), and subsequent largescale *in vitro* validation of their DNA binding affinities and specificities via bacterial selection. The labor and expertise required in this complex process limits the broad adoption of ZFN technology. An effective computational assisted design strategy will lower the complexity of the production of a pair of functional ZFNs. Here we used the FoldX force field to build 3D models of 420 ZFP-DNA complexes based on zinc finger arrays developed by the Zinc Finger Consortium using OPEN (oligomerized pool engineering). Using nonlinear and linear regression analysis, we found that the calculated protein-DNA binding energy in a modeled ZFP-DNA complex strongly correlates to the failure rate of the zinc finger array to show significant ZFN activity in human cells. In our models, less than 5% of the three-finger arrays with calculated protein-DNA binding energies lower than -13.132 kcal mol^{-1} fail to form active ZFNs in human cells. By contrast, for arrays with calculated protein-DNA binding energies higher than -5 kcal mol^{-1}, as many as 40% lacked ZFN activity in human cells. Therefore, we suggest that the FoldX force field can be useful in reducing the failure rate and increasing efficiency in the design of ZFNs.

zinc finger nuclease, FoldX, force field, protein mutation, modeling

A zinc finger nuclease (ZFN) consists of an engineered Cys$_2$-His$_2$ zinc finger protein (ZFP) fused to a non-specific cleavage domain of the type IIS endonuclease *Fok* I. Each finger recognizes a specific 3–4 bp region of DNA via a single α-helix. Therefore, a pair of ZFNs with four fingers linked in tandem can bind to a sequence as long as 30 bp with high specificity. Through the dimerization of the cleavage domains of *Fok* I, a pair of ZFNs can actively introduce a double-strand break (DSB) at the target site, and subsequently a variety of custom-modifications can be in-

troduced into the eukaryotic genomes through the repair of the DSB (Figure 1). When supplied with elaborately designed donor constructs, a pair of ZFNs can improve the efficiency of gene knock-out or knock-in by several orders of magnitude via the homologous recombination (HR) repair pathway [1], and without donor constructs, can induce gene disruption by non-homologous end joining (NHEJ) repair which is based on an error-prone gap repair pathway that can lead small deletions or insertions to the target locus [2] (Figure 2). ZFNs have enabled highly efficient genomic modifications in mammalian cells, and have been successfully employed to produce gene knock-out rats [3]. Thus,
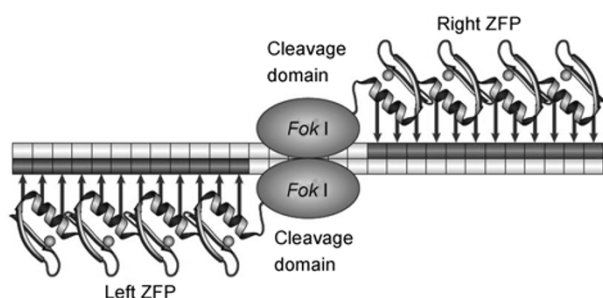

Figure 1   A schematic diagram of a pair of zinc finger nucleases (ZFNs) bound to the target DNA sequence. Each ZFN consists of the non-specific cleavage domain of *Fok* I coupled to a four-finger array designed to specifically bind to a "half-site" of the full target site. After dimerization, the nuclease is activated and cuts the 5–6 bp "spacer " sequence which separates the two "half-site". Adapted from [4].

ZFNs provide a novel tool that overcomes some of the difficulties in making gene knock-out farm animals, which, currently, is mainly based on the selection of relatively rare HR events that occur in primary cells at frequencies as low as 10^{-6}–10^{-9} [5].

ZFNs have heralded the coming of a new era in mammalian genetic modification. However, making functional ZFN pairs is a complex process that involves the construc-

tion of libraries containing numerous tandem arrays of zinc fingers, and the subsequent largescale *in vitro* validation of the DNA binding affinities and specificities of these arrays via bacterial two-hybrid (B2H) selection [6]. The labor and expertise required have limited the broad application of ZFNs. Using molecular modeling we can build 3D models of engineered ZFP-DNA complexes and calculate the binding energy between a ZFP and its target DNA sequence. This approach may help preclude zinc finger arrays with poor affinities to their target loci, and reduce the complexity involved in the production of functional ZFN pairs. In this study, we used the crystal structure of the Zif268 protein-DNA complex (1AAY.pdb) [7] as the basis for molecular modeling, and employed the FoldX algorithm [8] to build 3D models of 364 complexes of zinc finger arrays with their target DNA sequences. We used the arrays designed by Maeder *et al.* [9] using OPEN (oligomerized pool engineering) platform to target 19 different sites located within three endogenous human genes (*VEGF-A*, *HoxB13*, and *CFTR*), an endogenous plant gene (tobacco *SuRA*), and the *EGFP* reporter gene. The protein-DNA binding energies in the modeled complexes were calculated. Based on a nonlinear and linear regression analysis, we found that the calculated protein-DNA binding energies in the models
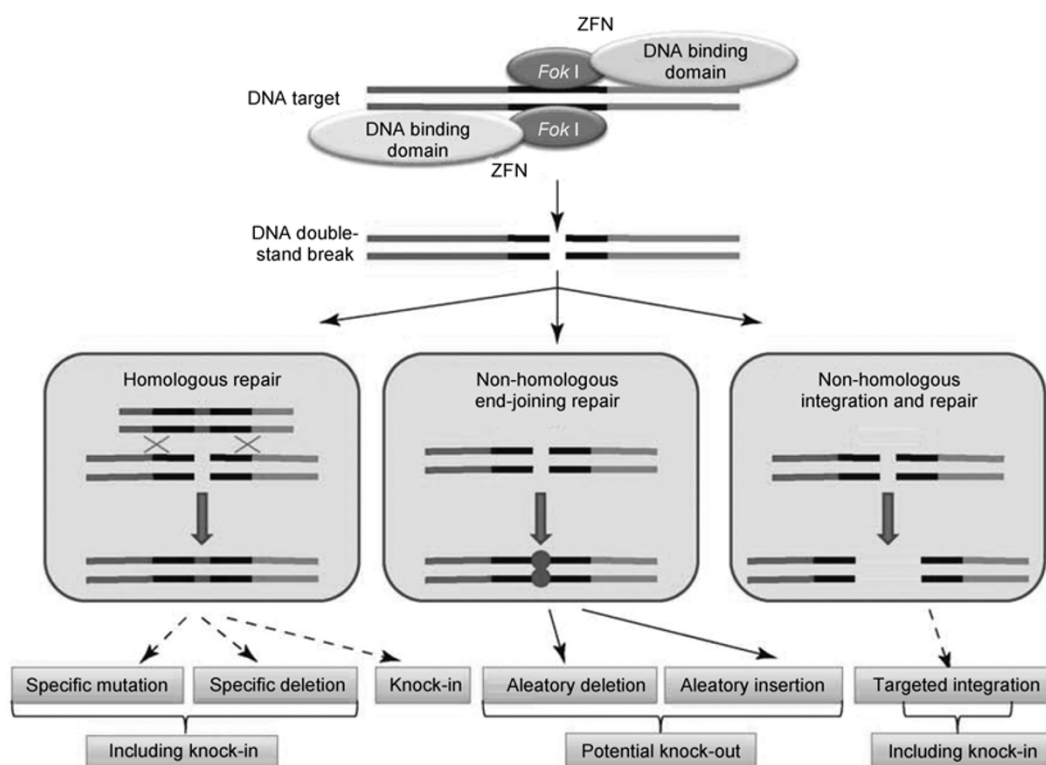

Figure 2   Genomic modifications using zinc finger nucleases (ZFNs). Gene knock-out and knock-in can be created during cellular repair of a double-strand break (DSB) induced by a pair of ZFNs, either through the homologous recombination (HR) repair pathway or the non-homologous end-joining (NHEJ) pathway. When an investigator-designed donor construct is supplied, gene knock-out or knock-in will occur via HR. If no donor construct is presented, gene disruption occurs through NHEJ, an error-prone repair process for re-ligating two break ends, which leads to small deletions and insertions at the target locus. Even non-homologous sequences can be inserted into the break site during the repair process. Adapted from [5].

strongly correlate to the failure rates of the zinc finger arrays as active ZFNs in human cells. In addition, we built 3D models of the 56 zinc finger arrays designed by Foley *et al.* [10] using the OPEN platform to recognize six different sites located within six endogenous zebra fish genes (*kdr*, *DT*, *HIF*, *Telo*, *TfR2*, and *Grck*). When the data for the two zinc finger arrays (a total of 420 models) were combined and analyzed, the regression results were largely unchanged. Based on the results of the statistical analysis, we conclude that FoldX can be successfully employed to preclude zinc finger arrays with lower affinities to their target DNA sequences, saving substantially on the laborious screening of functional ZFN pairs.

1   Materials and methods

1.1   Molecular modeling of ZFP-DNA complexes

FoldX (http://foldx.crg.es) is an empirical force field developed for a fast and quantitative evaluation of the effect of mutations on the stability of proteins, protein complexes, and protein-DNA complexes [8]. The core engine of FoldX is the calculation of the free energy of a protein or complex based on its high-resolution crystal structure. The following empirical equation is used to calculate the free energy (in kcal mol^{-1}) of unfolding a protein:

$$\Delta G = a\Delta G_{vdw} + b\Delta G_{solvH} + c\Delta G_{solvP} + d\Delta G_{wb} + e\Delta G_{hbond}$$
$$+ f\Delta G_{el} + g\Delta G_{kon} + hT\Delta S_{mc} + l\Delta G_{clash}.$$

The different energy terms used in this equation are weighted using empirical data obtained from protein engineering experiments. The interaction between two molecules is evaluated by calculating the free energy of binding ($\Delta G_{binding}$). To measure the binding energy of a complex AB, FoldX calculates the free energies of the AB complex and of the interacting molecules A and B separately. The interaction energy is then calculated using the following equation:

$$\Delta G_{binding} = \Delta G_{AB} - (\Delta G_A + \Delta G_B).$$

The 1.6 Å refined structure of the Zif268 protein-DNA complex containing three zinc fingers (1AAY.pdb) [7] was selected as the basis for modeling. Employing the Build-Model mutation engine of FoldX, the backbones of the three zinc fingers of Zif268 were each mutated to match the sequences of the engineered zinc fingers used on the OPEN platform [9]. Briefly, the amino acids FRPYACPVESC-DRRFS at the N-terminal end of the α-helix in the first zinc finger (F1) of Zif268 were sequentially mutated to GERPFQCRICMRNFS, and the amino acids HIRIHT at the C-terminal end of the α-helix were sequentially mutated to HTRTHT. Similarly, the amino acids GQKPFQCRIC-MRNF at the N-terminal end of the α-helix in the second

zinc finger (F2) of Zif268 were sequentially mutated to GEKPFQCRICMRNFS, and amino acids HIRIHT at the C-terminal end of the α-helix were sequentially mutated to HLRTHT. Finally, the amino acids GEKPFACDICGRKFA at the N-terminal end of the α-helix in the third zinc finger (F3) of Zif268 were sequentially mutated to GEKPFQC-RICMRNFS, and amino acids HTKIHLR at the C-terminal end of the α-helix are sequentially mutated to HLRTHLR. Thus, a mutant complex with the zinc finger backbones used on the OPEN platform was created. The nine base pairs of DNA and seven amino acids within the zinc finger α-helix on the interface of the mutant complex were then mutated into the corresponding bases and amino acids of the 420 engineered zinc finger arrays with their target DNA sequences in the two OPEN studies. These arrays were designed to target 25 different sites located within three endogenous human genes (*VEGF-A*, *HoxB13*, and *CFTR*), six endogenous zebra fish genes (*kdr*, *DT*, *HIF*, *Telo*, *TfR2*, and *Grck*), one endogenous plant gene (tobacco *SuRA*), and the *EGFP* reporter gene (Appendix Table S1 in the electronic version). The binding energy between a zinc finger protein and its target DNA in each mutant complex was computed using the Complex Analysis function of FoldX.

The 3D figures of protein-DNA complexes were generated using VMD (http://www.ks.uiuc.edu/Research/vmd) and rendered using POV-Ray (http://www.povray.org).

1.2   Statistical analysis

To obtain a more accurate result, from the total zinc finger arrays in the study by Maeder *et al.* [9], those with a high basal rate or those that activate transcription with more than 40% coefficient of variance in the bacterial two-hybrid (B2H) assay were removed for further statistical analysis (Appendix Table S1 in the electronic version). The scatter plots of protein-DNA binding energy in the modeled complexes against the activated transcription rates measured by B2H were generated using SPSS 15.0. For the correlation analysis, the chi-square test and the one-way ANOVA test were implemented in SPSS 15.0. It has been shown that zinc finger arrays that are activated by 1.57-fold or less in the B2H assay fail to form functional ZFNs in human cells [11]. Therefore, here we considered those arrays with activated transcription rates of less than 1.57-fold as inactive arrays. For regression analysis, we counted the percentage of inactive arrays within each two-unit free energy of binding (kcal mol^{-1}), and took that percentage as the probability of one zinc finger array failing to form an active ZFN in human cells. The calculated binding energies from modeled complexes were plotted against the probability of one zinc finger array failing to form an active ZFN in human cells. Finally, a curve was generated by nonlinear regression of the plotted points using the following equation:

$$y = \frac{1}{\theta} e^{\frac{x+\alpha}{\theta}}$$

where *y* is probability of zinc finger array that activates 1.57-fold or less in B2H. θ and α are regression coefficients. *x* is binding energy between zinc finger array and its target DNA in the modeled complex.

Taking the logarithm value on each side of the equation, it can be converted into a linear regression equation as follows:

$$\ln y = ax + b$$

where *y* is probability of zinc finger array that activates 1.57-fold or less in B2H. *a* and *b* are regression coefficients. *x* is binding energy between zinc finger array and its target DNA in the modeled complex.

Nonlinear and linear regression analysis was implemented in MatLab 7.1.

2   Results and discussion

2.1   The 3D structures of modeled ZFP-DNA complexes

The refined Zif268 protein-DNA complex (1AAY.pdb) was successfully mutated to form all the ZFP-DNA complexes in this study using the BuildModel function of FoldX. As shown in Figure 3, the amino acids Arg(−1), Glu(+3), and Arg(+6) in the F1 α-helix of 1AAY which directly interact with their target site GCG were mutated to Thr(−1), Agr(+3), and Ile(+6), respectively, to form the F1 α-helix of EG292L_1. These amino acids recognize a new target site GGT (292–294) in the coding region of the *EGFP* gene. The detailed 3D structure, including the critical α-helix and β-sheet motifs, of the mutant complex is highly similar to the original Zif268 protein-DNA complex. Therefore, the 1.6 Å refined structure of the Zif268 protein-DNA complex containing three zinc fingers (1AAY.pdb) may provide an excellent framework for the molecular modeling of zinc finger-DNA complexes.
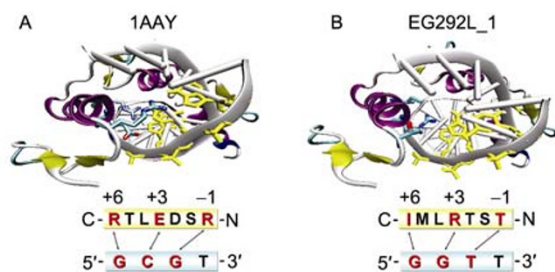

Figure 3 The stereo structures of protein-DNA complexes. A, The Zif268 protein-DNA complex (1AAY). B, The mutant complex which targets GGT (292–294) in the coding region of the *EGFP* gene. Sky-blue indicates the three amino acids at positions −1, +3 and +6 of the F1 α-helix, which directly interact with the 3 bp target subsites (yellow).

2.2   Distributions of the protein-DNA binding energies calculated from the modeled ZFP-DNA complexes, and their transcriptional activation ability tested in the B2H assay

In the B2H assay, the majority of zinc finger arrays activate transcription by 10-fold or less, and partial arrays activate it by 1.57-fold or less, indicating that these arrays will form inactive ZFNs in human cells [11] (Figure 4A and D). Zinc finger arrays with target sites composed of three GNN subsites (where N is any base) are, as a whole, more effective than arrays with target sites composed of at least one TNN subsite. Most zinc finger arrays with target sites composed of three GNN subsites activate transcription between 1.57-fold and 10-fold, and a considerable percentage of arrays activate transcription between 10-fold and 15-fold (Figure 4A, B and E). By contrast, most zinc finger arrays with target sites composed of at least one TNN subsites activate transcription by less than 5-fold, and few arrays activate it by over 10-fold (Figure 4A, C and F). These results indicate that engineered zinc finger arrays targeting sequences that contain one or more TNN subsites will have lower affinities for their target sites, and are, therefore, likely to increase the failure rate for making functional ZFNs in eukaryotic cells. This is consistent with the experimental results reported by other investigators [12]. A computational design method will help reduce the risk of failure in making functional ZFNs targeting DNA sequence containing TNN subsites.

The binding energies between protein and DNA in all modeled ZFP-DNA complexes exhibit an approximate normal distribution, which is significantly different from the distribution of fold-activation of the zinc finger arrays tested in the B2H assay (Figure 4D). In addition, the calculated correlation coefficient of binding energy and fold-activation is merely 0.225. Therefore, the actual binding activity of one zinc finger array cannot be exactly predicted from the calculated binding energy within the modeled complex using FoldX. A larger experimental test of the binding activities of engineered zinc finger arrays and more detailed structural information may help to validate the free energy equation of FoldX, and improve the accuracy of prediction. For zinc finger arrays with target sites composed of three GNN subsites, the correlation coefficient of binding energy and fold-activation drops to 0.198. By contrast, for zinc finger arrays with target sites composed of at least one TNN subsite, the correlation coefficient of binding energy and fold-activations improves to 0.32 and the distribution patterns of binding energy and fold-activation are similar (Figure 4F). The improved correlation coefficient implies that FoldX may be more useful in designing functional zinc finger arrays for target sites containing at least one TNN subsite.
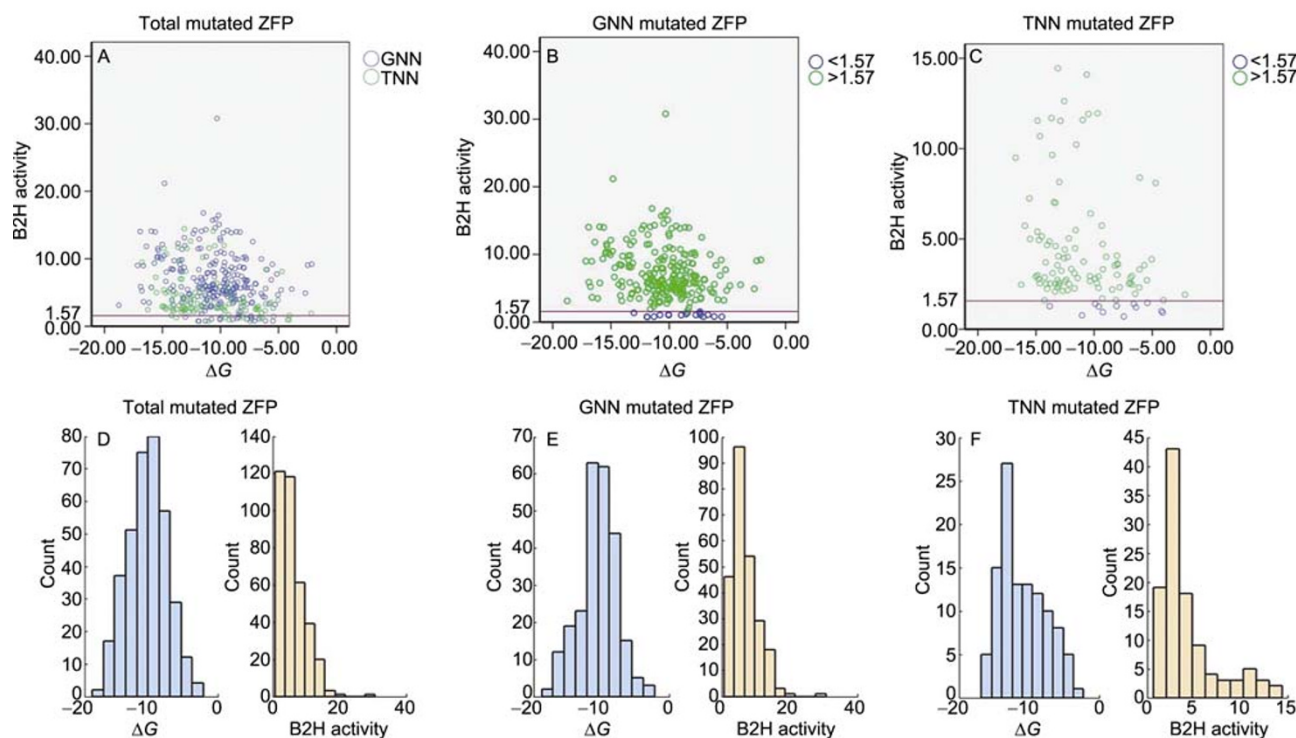
Figure 4   The distribution graph of protein-DNA binding energies calculated from the modeled ZFP-DNA complexes, and transcriptional activation measured in the B2H assay. A, B, and C show the scatter plots of protein-DNA binding energy (ΔG (kcal mol^{-1})) calculated from the modeled complexes against the fold-activation measured by the B2H assay. The purple reference line indicates the threshold for fold-activation (1.57-fold) of a zinc finger array that lacks ZFN activity in human cells. GNN and TNN (where N is any base) indicate the subsites of the zinc finger array target sites. D, E, and F show the number of zinc finger arrays with the DNA-protein binding energy (ΔG (kcal mol^{-1})) and fold-activation indicated. The width of each bar spans two units.

2.3 The calculated protein-DNA binding energy in the modeled ZFP-DNA complexes strongly correlates with the failure rate of the zinc finger arrays that form active ZFNs in human cells

A nonlinear regression analysis on the data derived from the 364 zinc finger arrays in the study by Maeder *et al.* [9] reveals that the calculated protein-DNA binding energy in the modeled ZFP-DNA complex strongly correlates with the failure rate of the zinc finger arrays that show significant activity as ZFNs in human cells (R^2=0.988) (Figure 5A). An engineered zinc finger array with a high binding energy in the modeled complex is likely to fail to form an active ZFN in human cells. If the binding energy is more than −5 kcal mol^{-1} the anticipated failure rate is as high as 40%. However, the anticipated failure rate drops to 5% or less when the binding energy is less than −13.132 kcal mol^{-1} (Figure 5A). A linear regression analysis produces similar results. The binding energy strongly correlates to the failure rate of a zinc finger array forming an active ZFN in human cells (R^2=0.977) (Figure 5D). If the binding energy is less than −12.491 kcal mol^{-1}, the anticipated failure rate would drop to 5% or less (Figure 5D). The nonlinear and linear regression analysis on zinc finger arrays with target sites containing at least one TNN subsite generates similar results to the

analysis on all the arrays (Figure 5C and F). In this case, the anticipated failure rate is as high as 50% when the binding energy is more than −5 kcal mol^{-1} (Figure 5C). For zinc finger arrays with target sites composed of three GNN subsites, the nonlinear and linear regression analysis both reveal a decrease of the correlation efficient (Figure 5B and F). This is probably because the binding energies for a large number of these arrays are around −10 kcal mol^{-1} and most arrays activate transcription by more than 1.57-fold (Figure 4B and E), making the plot deviate from the fitting curve. Notably, zinc finger arrays with target sites composed of three GNN subsites generally have higher tested affinities, and the failure rate of forming inactive ZFNs is less than 20% when the calculated binding energies are close to −5 kcal mol^{-1}. Therefore, there is less risk involved in designing ZFNs with target sites composed of three GNN subsites. When the data derived from the 56 zinc finger arrays designed to target the zebra fish genome in the study by Foley *et al.* [10] are incorporated for regression analysis, similar results are produced (Appendix Figure S1). For all zinc finger arrays, the protein-DNA binding energy significantly correlates with the failure rate (R^2=0.975). If the binding energy is more than −5 kcal mol^{-1}, the anticipated failure rate is as high as 35%. The anticipated failure rate dropped to 5% or less, when the binding energy is less than −13.3
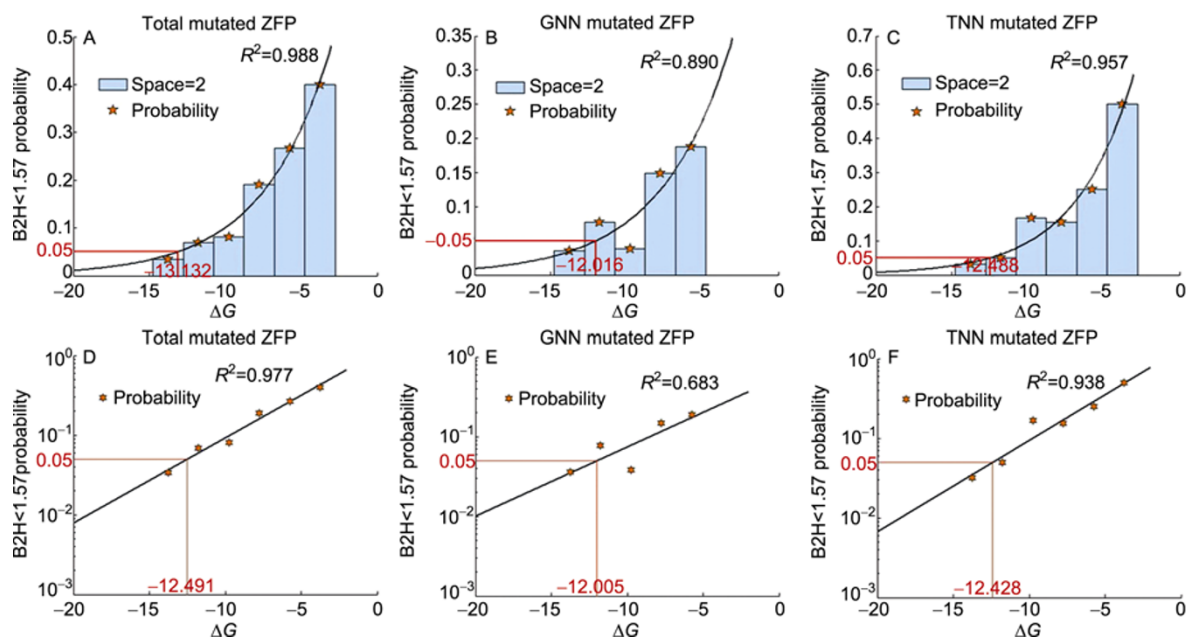
Figure 5 The regression of the probability of a zinc finger array that fails to form an active ZFN in human cells against the binding energy between protein and DNA in the modeled ZFP-DNA complex. A, B, and C show the results of nonlinear regression analysis. Each sky-blue bar spans two units of free energy (ΔG (kcal mol^{-1})). The yellow stars, within two-unit bars, indicate the probability of one zinc finger array failing to form an active ZFN in human cells. D, E, and F show the results of linear regression.

kcal mol^{-1} (Appendix Figure S1A). For zinc finger arrays with target sites containing at least one TNN subsites, the anticipated failure rate is as high as 40% when the binding energy is more than −5 kcal mol^{-1} (Appendix Figure S1C). This supports the previous finding that it is more useful to use FoldX to lower the risk at designing functional zinc finger arrays for target sites containing at least one TNN subsite.

The nonlinear regression analysis reveals that the failure rate of designing a functional ZFN is relatively low when the calculated binding energy of one zinc finger array is less than −12 kcal mol^{-1} and significantly increased when the calculated binding energy is more than −8 kcal mol^{-1} (Figure 5A). A chi-square test confirmed this result. The proportion of zinc finger arrays activating transcription by less than 1.57-fold in the arrays with calculated binding energies less than −12 kcal mol^{-1} is significantly lower than that in the arrays with calculated binding energies more than −5 kcal mol^{-1} ($P<0.001$). In addition, the one-way ANOVA test shows that the fold-activations of zinc finger arrays with calculated binding energies less than −12 kcal mol^{-1} are significantly higher than for zinc finger arrays with calculated binding energies more than −5 kcal mol^{-1} (Figure 6A). The significant difference in calculated binding energies between zinc finger arrays that activate transcription by more than 10-fold and those that active it by less than 1.57-fold in the B2H assay is examined using the one-way

ANOVA test. Interestingly, the average binding energy of the arrays that activate transcription by more than 10-fold activations is close to −12 kcal mol^{-1} and the average binding energy of those arrays that activate transcription by less than 1.57-fold is around −8 kcal mol^{-1} (Figure 6B). This is in agreement with the nonlinear regression results discussed above. Therefore, using FoldX, we successfully improved the efficacy of screening of functional ZFN pairs by precluding zinc finger arrays with binding energies greater than −8 kcal mol^{-1} and focusing on arrays with binding energies of less than −12 kcal mol^{-1}.

3 Discussion

The 1.6 Å refined structure of the Zif268 protein-DNA complex containing three zinc fingers (1AAY.pdb) serves as an excellent basis for the molecular modeling of ZFP-DNA complexes using FoldX. The calculated protein-DNA binding energies in the modeled ZFP-DNA complexes are strongly correlated to the failure of the zinc finger array to show significant ZFN activity in human cells. This correlation can be used to preclude engineered zinc finger arrays that have binding energies with their target sites above the calculated threshold, and will help to improve the efficacy of designing a pair of functional ZFNs.
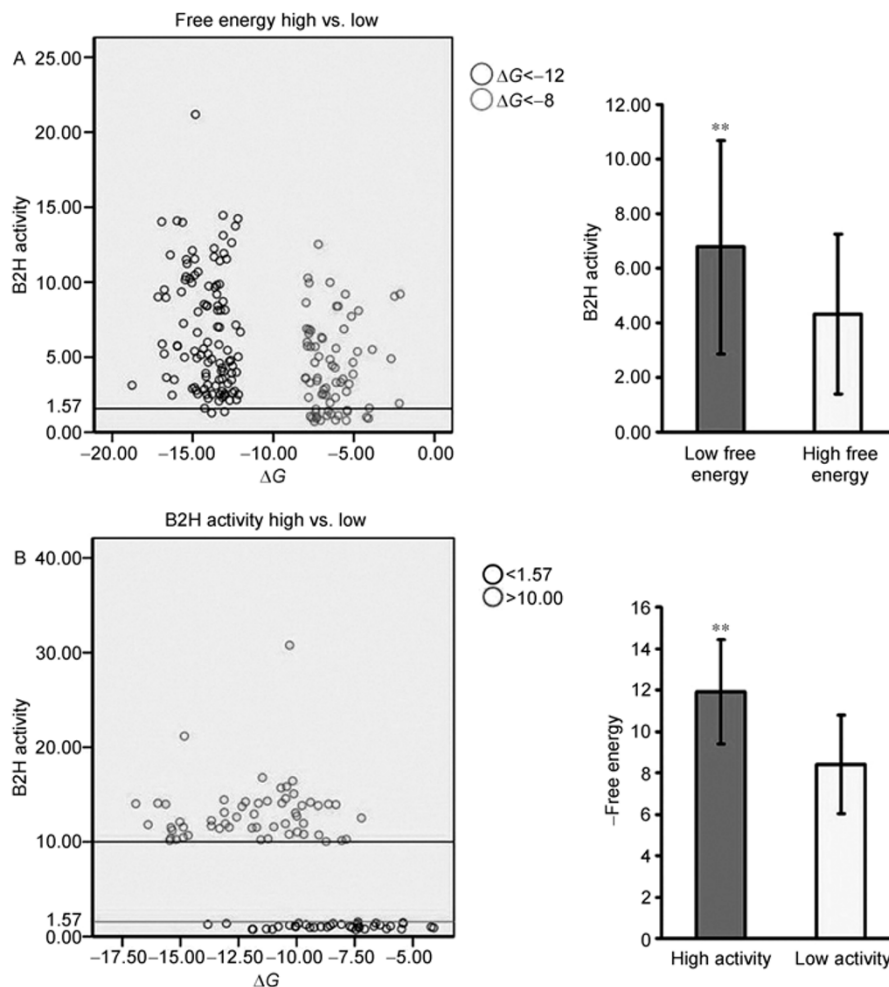
Figure 6 Comparisons of binding activity and binding energy in the modeled complexes. A, Differences in binding activity in the B2H assay between zinc finger arrays with higher and those with lower calculated binding energies in the modeled complexes. B, A comparison of the differences in calculated binding energies in the modeled complexes between zinc finger arrays with higher and those with lower binding activities in the B2H assay. ** indicates a significant difference.

1 Cathomen T, Joung J K. Zinc-finger nucleases: The next generation emerges. Molecular Therapy, 2008, 16: 1200–1207

2 Lieber M R, Ma Y M, Pannicke U, *et al*. Mechanism and regulation of human non-homologous DNA end-joining. Nat Rev Mol Cell Biol, 2003, 4: 712–720

3 Geurts A M, Cost G J, Freyvert Y, *et al*. Knockout rats via embryo microinjection of zinc-finger nucleases. Science, 2009, 325: 433

4 Urnov F D, Rebar E J, Holmes M C, *et al*. Genome editing with engineered zinc finger nucleases. Nat Rev Genet, 2010, 11: 636–646

5 Le Provost F S, Lillico B, Passet, *et al*. Zinc finger nuclease technology heralds a new era in mammalian transgenesis. Trends Biotechnol, 2009, 28: 134–141

6 Maeder M L, Thibodeau-Begany S, Sander J D, *et al*. Oligomerized pool engineering (OPEN): An 'open-source' protocol for making customized zinc-finger arrays. Nat Protoc, 2009, 4: 1471–1501

7 Elrod-Erickson M, Rould M A, Nekludova L, *et al*. Zif268 protein-DNA complex refined at 1.6 Å: A model system for understanding zinc finger-DNA interactions. Structure, 1996, 4: 1171–1180

8 Schymkowitz J, Borg J, Stricher F, *et al*. The FoldX web server: An online force field. Nucleic Acids Res, 2005, 33: W382–W388

9 Maeder M L, Thibodeau-Begany S, Osiak A, *et al*. Rapid "Open-Source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol Cell, 2008, 31: 294–301

10 Foley J E, Yeh J R J, Maeder M L, *et al*. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by oligomerized pool engineering (OPEN). PLoS One, 2009, 4: e4348

11 Ramirez C L, Foley J E, Wright D A, *et al*. Unexpected failure rates for modular assembly of engineered zinc fingers. Nature Methods, 2008, 5: 575

12 Sander J D, Zaback P, Joung J K, *et al*. An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. Nucleic Acids Res, 2009, 37: 506–515
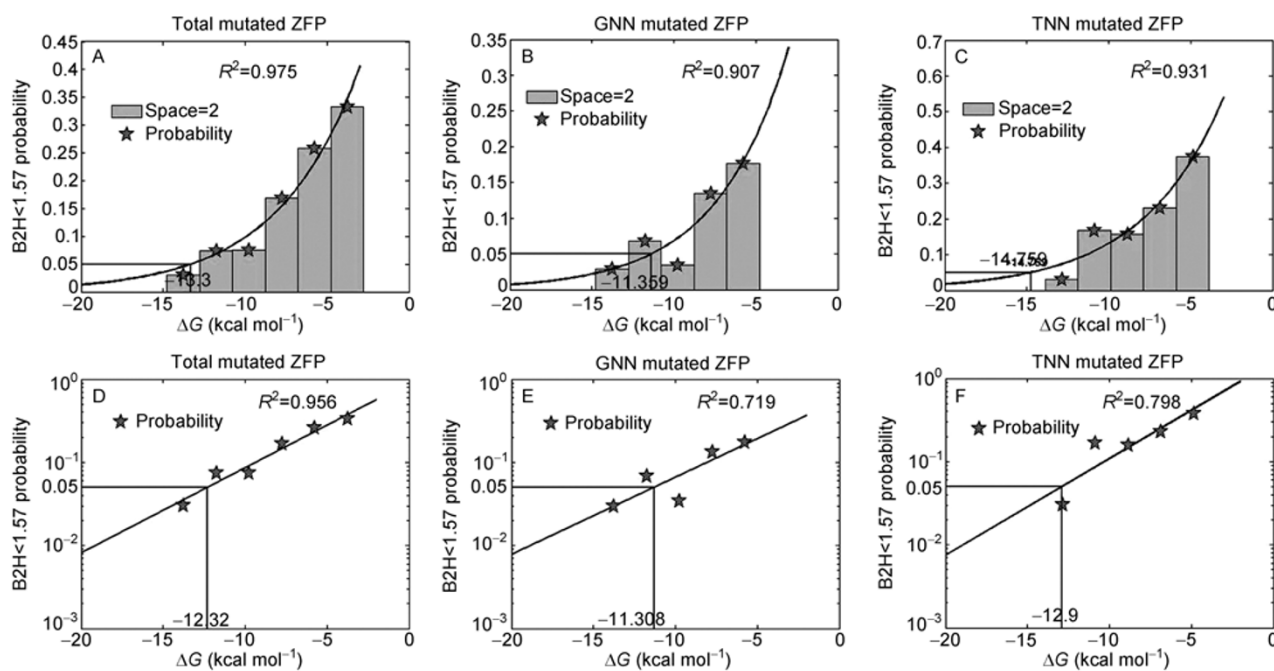
Appendix Figure S1   The regression of the probability of a zinc finger array that fails to form an active ZFN in human cells against the binding energy between protein and DNA in the modeled ZFP-DNA complex (The 56 zinc finger arrays derived from the study by Foley *et al.* are included here [10]). A, B, and C show the results of the nonlinear regression analysis. Each sky-blue bar spans two units of free energy (ΔG (kcal mol^{-1})). The yellow stars, within the two-unit bars, indicate the probability of one zinc finger array failing to form an active ZFN in human cells. D, E, and F show the results of the linear regression analysis.