A universal deep-learning model for Zinc Finger design enables Transcription Factor reprogramming[1]

Ruben Solozabal

March 2023

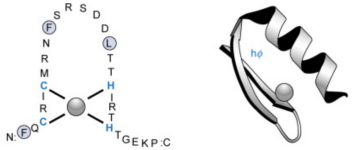
[1] Ichikawa, David M., et al. Nature Biotechnology (2023)

Outline

Outline

Introduction

- **Zinc finger (ZF)** is a small structural motif that is characterized by the coordination of one or more zinc ions (Zn2+) in order to stabilize the fold.



Cys2His2 zinc finger motif, consisting of an α helix and an antiparallel β sheet. The zinc ion is coordinated by two histidine residues and two cysteine residues. Rec. M. Isalan, in Encyclopedia of Biological Chemistry, 2013

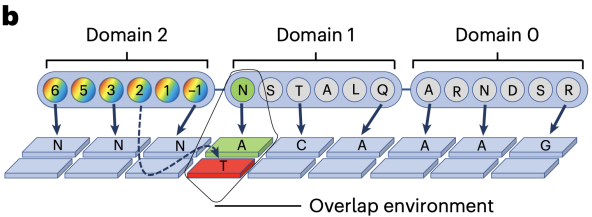
- ZFs are a family of TFs. On humans 50% TFs use ZFs to attach DNA.
- Each finger binds to **3 DNA base-pairs**. Multiple fingers can be combined in **ZFs arrays** to attach to larger sequences.
- A six-finger array provides up to 18 bp of specificity that gives single-target resolution in the human genome.
- ZFs proteins can be designed to **control the expression** of any desired target gene. Opposed to the many genes that TFs would influence.

Paper motivation

- **Zinc fingers (ZFs)** can be engineered to bind specific target sequences in the genome and regulate gene expression.
  - While CRISPR–Cas and transcription activator-like effector (TALEN), the size of these proteins complicates delivery. ZFs require < 170 amino acids to specify a unique sequence in the human genome.
- The paper describes the screening of 49 billion protein–DNA interactions and the development of a machine-learning model, **ZFDesign**, that solves ZF design for any genomic target.
- ZFDesgn is used to **reprogramming natural ZF TFs** to provide tools that can activate or repress target genes that are sufficiently small for multiplexed delivery with minimal risk of immunity.

Selection of ZF specificity and compatibility

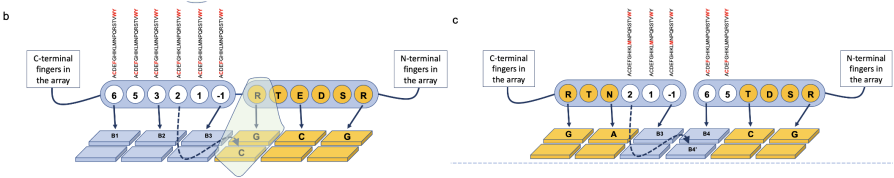
- The difficulty in designing Zinc Fingers resides in designing compatible neighboring fingers.
- Influence on adjacent ZFs can be explained by the multiple side chains of adjacent ZFs that bind DNA in close proximity to one another.
  - 'Overlap' at position 6 of an N-terminal helix can be within hydrogen-bonding distance of positions -1 and 2 side chains of its C-terminal neighbor.
- Libraries exist for a set of interface diversity containing compatible ZF α-helices based on a single neighboring domain.



Interactions between adjacent α-helices and DNA

Selection of ZF specificity and compatibility

- Two general approaches have been used to engineer ZFs, although the engineering of ZFs remains a challenge.
  1. Engineering one finger at a time by the selection of functional variants from ZF libraries.
     - Screen of all amino acid combinations at the six critical positions of the ZF alpha-helix in a single-adjacent-finger context.
  2. Second approach focuses on the interface between adjacent ZFs.
     - Evaluate the complexity of compatibility at the interface between ZFs.
- There is a combinatorial explosion of amino acids to evaluate when screening ZFs α-helices.

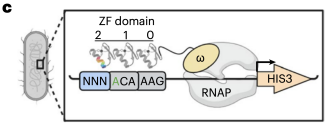


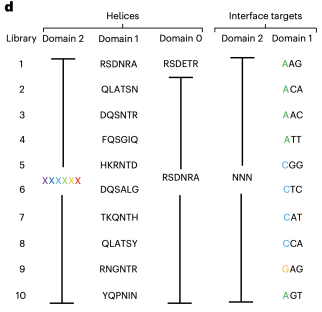
Libraries for screening ZFs interfaces

Outline

ZFs libraries screening

- The paper screens **10 ZF libraries**, each presenting the randomized C-terminal helix in a unique adjacent finger environment defined by the adjacent ZF helix
- Screen ZFs libraries across each of the 64 possible **3bp target**.
- In total, **49 billion protein–DNA** interactions are screened from ten libraries, across 12 sets of 64 selections per library, for 768 independent selections.





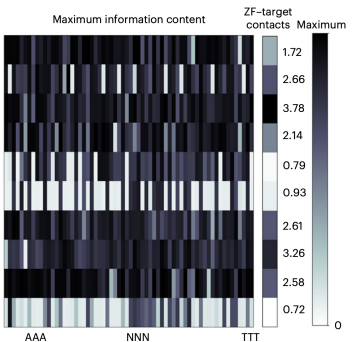
Dataset selection

- Screening shows global and target-specific differences induced by library environments.
- First selection of helices ranged from 128,000 to 1 million per library.
- To distinguish selections that have low enrichment because of overlapping but unique strategies, maximum information content is used at a single helical position of a cluster.
  1. Each sequence was assigned to the cluster associated with the PWM [1].
  2. Quantify the information content across MOTIFs generated from the different sequence clusters: *"a successful selection should produce clusters where at least* ***one helical position*** *with high information content"*.
  3. Shannon entropy for each helix was calculated based on the number of reads associated with each possible encoding nucleotide sequence. Helices with fewer than ten reads or entropy < 0.07 were removed.

Analyzing the ability of ZFs to bind specific targets

Screening reflects the ability of ZFs to bind any 3-bp target in a wide range of adjacent finger.

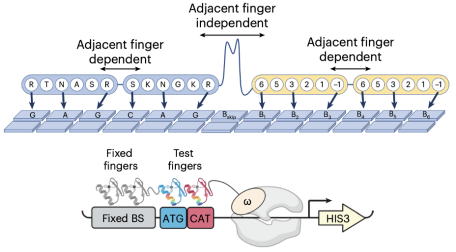
- For the 10 interfaces screened 39–100% of the 3-bp target selections led to successful enrichment of ZFs
- For each of the 64 three-base-pair targets, at least eight different library contexts resulted in successful enrichment of ZFs.



Enrichment across selections.

Double-helix ZFs datasets

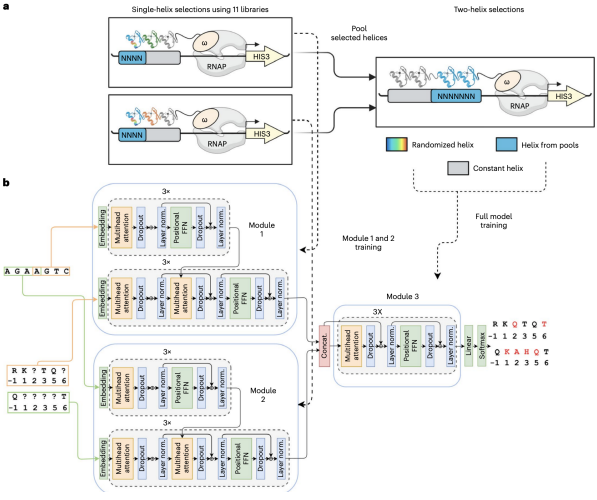
- Accounting for single-helix ZFs design represents only a small number of potential adjacent finger influences.
- To test greater variability at the interface, 200 two-finger libraries were created by assembling pools of helices successfully selected to bind each 3-bp half-site of a 6-bp target.
- This linker prefers a base to be skipped between the binding sites of the fixed and screened ZF pairs, reducing the potential influence of fixed ZFs and encouraging these pairs to work as independent modules.



Paired format for two-finger selections using the base-skipping linker.

Architecture

- **ZFDesign** is a modern machine learning for designing Zinc Fingers takes into account the compatibility of neighboring fingers.
- Uses a novel **hierarchical transformer** architecture



Architecture details

- The model comprises **two hierarchical modules**.
- The model first is trained on single-helix selections to predict residues in partially masked helices that bind 4-mer nucleotide sequences that includes the target 3-mer and the overlap base.
- The embeddings generated from the first modules are fed into a second module that learns inter-helix compatibility.
- The full model is trained on two-helix selection data to predict residues in partially masked helix pairs that bind 7-mer nucleotide sequences.
- During training the nucleotide target is provided as well as a partially masked ZF sequence (50% of core residues were masked).
- **Cross-entropy loss** is evaluated based on output probabilities.

Architecture details

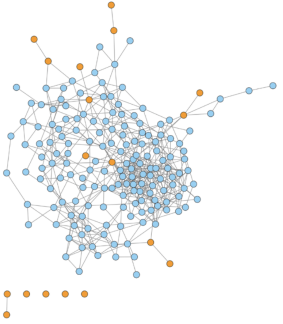
- Architecture of the three modules is based on the Transformer.
- The model uses attention layers that relate nucleotide bases to helical residues
- An encoder generates a high-dimensional representation for each base in a nucleotide 4-mer.
- A decoder then generates predictions for each core residue in a ZF helix.
- Only modification: the decoder operates bidirectionally.
- Model hyperparameters:
  - All attention layers repeated 3 times and each attention layer had 4 heads.
  - The model-embedding dimension was 128.
  - Value- and key-embedding dimensions for the dot-product attention is 256.
  - The hidden dimension in the FFN was set to 128.
  - Dropout percentage of 0.3

Training dataset

- Selections performed with 10 libraries against 192 different nucleotide 4-mers. In total, the dataset included 2,071,764 data points. The number of selected helices ranged from 128,000 to 1 million per library screened.
- Data points were split into training, test and validation datasets at proportions of 80, 10 and 10%.
- The full model was trained using data from helix-pair B1H selections performed against nucleotide 7-mers. An initial dataset of selections against 189 7-mers was split into training and validation datasets at proportions of 90 and 10%, respectively. This dataset contained a total of 327,792 data points.

Validation dataset

- To ensure that the validation set was sufficiently different from the training dataset, a graph was generated where nucleotide 7-mers were represented as nodes, and edges connected 7-mers within two base substitutions from each other.
- Nodes with the lowest degree in the graph, and their neighbors, were then added to the validation dataset

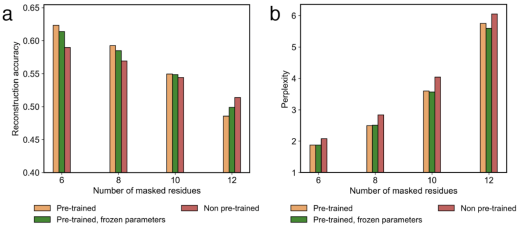


Graph representation of the seven-mer sequences in the training and validation datasets.

Effect of pre-training on model performance

Comparison of reconstruction accuracies when the model is pre-trained on single-helix selection:

- The parameters for modules 1 and 2 are randomly initialized,
- transferred from the pre-training step,
- or transferred and from the pretraining step and frozen



Quantification of the effect of pre-training on model performance

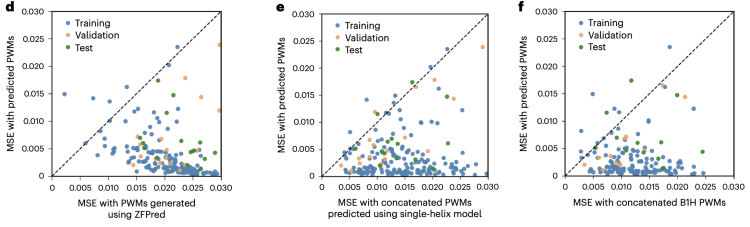
- Pretraining modules 1 and 2 require 1.3 million iterations; training the full model required 3.4 million iterations

Outline

Resutls

ZFDesign generates compatible ZF pairs

- ZFDesign is used to generate ZF sequences to target 6-mers on the testing dataset.
- Starting from an empty sequence, the model is run once for each amino acid in the ZF helix pair. At each iteration an amino acid is predicted, and this prediction is provided as context in subsequent iterations.
- As baseline comparison: ZFPred [2], the single-finger models to generate ZF sequences for each DNA 3-mer, and concatenated PWMs.



Predicted and real error on the logos, reported as MSE on predicted position weight matricies (PWMs).

Details on de novo design of ZF–helix pairs

- For improving the sequence generation adapts an A*-based sampling [3] method and a temperature-dependent sampling procedure.
- The A*-based sampling involves iteratively maintaining a priority queue of partially masked sequences. At every iteration, the top partially masked sequence is taken from the priority queue and passed through the network.
- The following equation is used to assign heuristically a priority to each partially masked sequence:

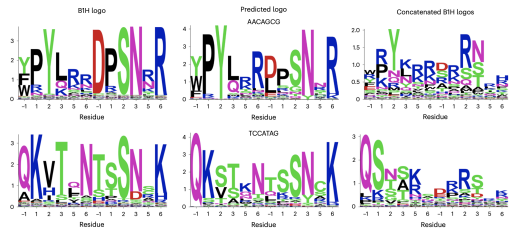
$$p_j = \sum_{i=1}^{j} \log(p_i) + \sum_{j}^{12} \log(p^*)$$

where p_i denotes the probability assigned to the prediction made at iteration i, and j denotes the number of predicted residues. p^* denotes the expected maximum probability that would be assigned by the network to later predictions. This parameter can be tuned to move the search closer to a *greedy* or *breadth-first search*.

Resutls

ZFDesign generates compatible ZF pairs

- ZFDesign is used to generate ZF sequences to target 6-mers from our test dataset.
- Starting from an empty sequence, the model is run once for each amino acid in the ZF helix pair. At each iteration an amino acid is predicted, and this prediction is provided as context in subsequent iterations.
- As alternative baseline comparisons, we first used the single-finger models to generate ZF sequences for each DNA 3-mer and concatenated them.

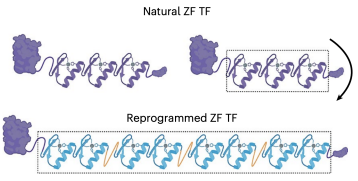


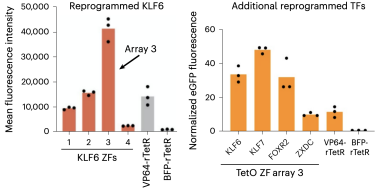
Predicted logos, real B1H logos and concatenated single-helix B1H logos for test set sequences.

Resutls

Reprogramming of human transcription factors

- This approach presents the designed ZFs in the exact context in which ZFs would occur naturally in the human protein, minimizing the risk of immunogenicity.
- This experiment replaces KLF6 ZFs with a series of ZF arrays designed to bind the operator sequence.
- Also, TFs other than KLF6 can also be reprogrammed to bind the same sequence and regulate the target.





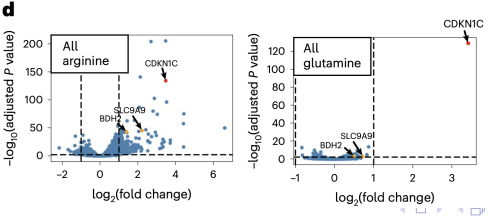
TF protein KLF6 is activated with four ZF arrays designed to bind the target sequence.

Resutls

Specificity and genome-wide regulatory activity of RTF

- ZFDesign enables the reprogramming of TFs for either activation or repression.
- To test the **precision** of regulation, RNA-seq is used to quantify RTF off-target regulation.
- KLF6 RTF activators for gene CDKN1C: 268 misregulated genes
- Substitution of 2, 4, or 8 amino-acids at mutant versions of CDKN1C: off-target genes was reduced to 1.

Position –5 arginine substitutions	0	2	4	8
Genes with altered expression	275	211	121	1

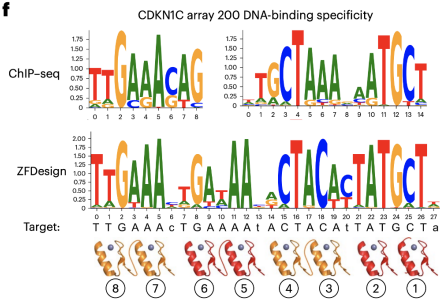


RNA-seq results for CDKN1C.

Resutls

Specificity and genome-wide regulatory activity of RTF

- Characterize the DNA-binding specificity of the ZF arrays.
- ChIP–seq peaks contained two independent motifs, suggesting that the base-skipping linker allows modular, independent binding.
- Despite the specificity, subsets of ZFs appear to drive binding genome-wide. This is attributed to the modular design of the arrays.

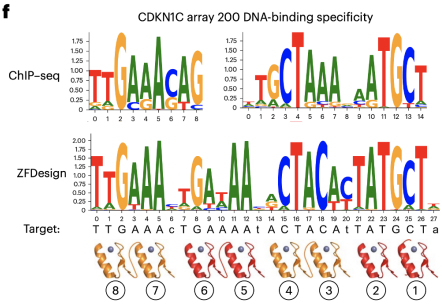


Specificity of CDKN1C by ChIP–seq.

Resutls

Specificity and genome-wide regulatory activity of RTF

- In future work, ZF arrays with conventional linkers that **do not skip** bases will be used to reduce off-target binding.
- Reducing the independence of the binding of ZF pairs will result in higher specificity of designed arrays.

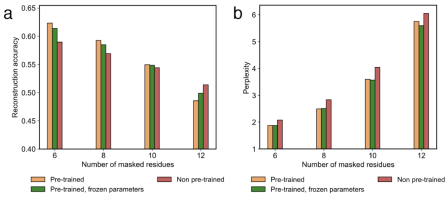


Specificity of CDKN1C by ChIP–seq.

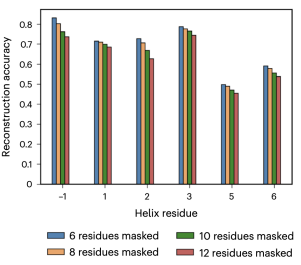
Outline

Ablation studies

- (Left). Comparison of reconstruction accuracies when the model is pre-trained on single-helix selections and re-trained, re-trained with parameters of the single-helix modules frozen, and not pre-trained.
- (Right). Helix sequence reconstruction accuracy with different numbers of masked residues.



Quantification of the effect of pre-training on model performance.



Outline

Conclusions

- The study presents ZFDesign, a hierarchical attention-based model trained on comprehensive screens of ZF–DNA interactions.
- ZFDesign captures the influence of **two adjacent finger** environments and provides a general design model for ZF arrays.
- **Design limitations** remain because the model was trained on two-finger selections that sampled $< 5\%$ of the possible 6-bp targets, and single-finger selections did not sample all interfaces. Therefore, the model is more confident in some domains at the overlap position.
- Finally, presents a design method that allows the seamless **reprogramming of natural TF** to target a sequence of interest in the DNA.
- RTFs can produce activation and repression of genes similar to CRISPR-based tools.

References I

[1] TaeHyung Kim, Marc S Tyndel, Haiming Huang, Sachdev S Sidhu, Gary D Bader, David Gfeller, and Philip M Kim.
Musi: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets.
Nucleic acids research, 40(6):e47–e47, 2012.

[2] April L Mueller, Carles Corbi-Verge, David O Giganti, David M Ichikawa, Jeffrey M Spencer, Mark MacRae, Michael Garton, Philip M Kim, and Marcus B Noyes.
The geometric influence on the cys2his2 zinc finger domain and functional plasticity.
Nucleic Acids Research, 48(11):6382–6402, 2020.

[3] Andrew R Leach and Andrew P Lemon.
Exploring the conformational space of protein side chains using dead-end elimination and the a* algorithm.
Proteins: Structure, Function, and Bioinformatics, 33(2):227–239, 1998.