Designing an ID-like Protein *In Silico*

Joshua Sodicoff, Xiaoqiang Huang[1], Yang Zhang[1]

[1]Department of Computational Medicine and Bioinformatics - University of Michigan Medical School, Ann Arbor, MI

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

UROP

Introduction

Protein design has been rapidly developing as a field with the advent of new computational techniques. Programs used to rely entirely on physical calculations to predict structure and develop a residue sequence for a given scaffold. My laboratory's design program, EvoDesign, uses data on homologous structures and a neural network to guide design more effectively.

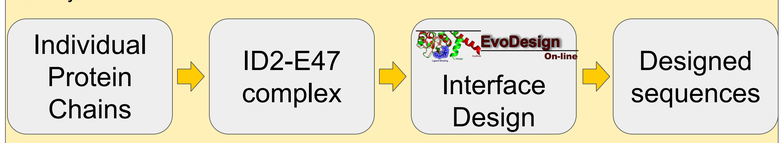
The ID family of proteins all share a set of characteristics that make them strong candidates for remodelling. All 4 proteins are 161 residues long or less, meaning that protein design programs are more able to accurately use them as scaffolds as compared to larger proteins. They also all share a similar structure and incorporate the helix-loop-helix motif, allowing for the application of homologous sequences and folds for realistic modelling of structure.

Langlands, Yin et al. studied their method of action and relative binding affinities in 1997. Their role as Class I and II factor antagonists comes as a result of dimerization. Class I factors must homodimerize or heterodimerize with Class II factors to bind with DNA. As these factors rely on the presence of a basic region of residues on both proteins in the dimer to bind to DNA, dimerization with an ID protein, which lacks a basic region, prevents this. Perk, Iavarone, et al. explained their relation to cancer in 2005.

Methods

In order to complete this study, several computational tools were used. In my original analysis of protein structures and interactions, I extensively utilized the Protein Database as well as pyMol. The Protein Database offers structural information for both individual proteins as well as naturally occuring complexes, as well as sequence information and data on similarity and function. Once I was sure of my target, I used the Database as the source of sequence and structural information for both my scaffold, in the form of ID2 from the 4AYA structure as well as E47 from 2YPB. PyMol allows for the manipulation of PDB (Protein Database) files in a more open environment. I used it to separate the necessary chains from their complexes and study their docking. Once I had my scaffold and receptor in PDB form, I used EvoDesign and I-TASSER to design and model my designed sequences. Both of these programs were designed by my laboratory. EvoDesign takes in a protein scaffold and creates a set of residue sequences expected to have a similar structure. I-TASSER takes in a residue sequence and creates several structures. The resulting models are predicted to have greater stability and binding affinity with E47 than the wild-type protein and I am looking into experimental verification.

At the most basic computational level, this study required a large number of variables and units. However, for the purposes of comparison and output, only a few of these should be discussed. In processing a structure, EvoDesign relies primarily on the use of homologous structures as a way of more effectively predicting a structure. Two statistics, the TM-score and sequence identity relay information on similarity between the scaffold and homologous structures. TM-score is a metric between 0 and 1 that conveys similarity between structures, whereas sequence identity refers the the percent of residues shared between the scaffold and homologous protein. Sequence identity is also used on the output of the program, in addition to several metrics of normalized relative error for structure, solvent accessibility, and torsion angle. Normalized relative error is a calculation based on the error between the neural-network predictions and the scaffold as well as the error of the predictions from the sequence of the scaffold. These metrics provide some information on the expected reliability of the prediction. For individual designs, EvoDesign also offers information on energy, with a more negative number referring to a higher level of predicted structural stability.



Results

Figure 1: ID2-E47 complex

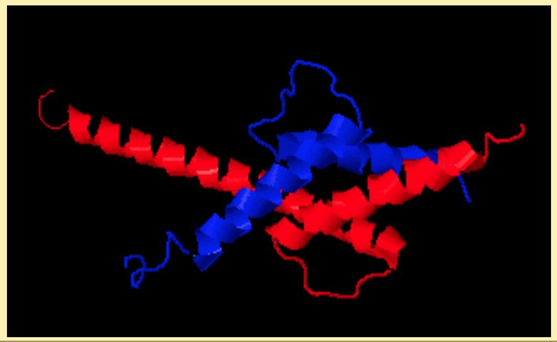


Table 1: Data on Homologous Sequences

Rank	PDB Hit	Entry	TM-score	Iden.
1	1nkpA	Myc/Max	0.8	0.18
2	2ypaA	SCL/E47/LMO2/LDB1	0.77	0.26
3	2ql2C	E47/NeuroD1	0.75	0.77
4	2h7oA	Rho-GTPase binding domain of YpkA	0.74	0.09
5	1nlwA	Mad/Max	0.74	0.19
6	1mdyA	MyoD	0.73	0.28
7	3ojaA	LRIM1/APL1C	0.72	0.05

Table 2: Data on Designed Sequences

Design Number	Sequence Identity (%)	Normalized Relative Error			
		Secondary Structure	Solvent Accessibility	Torsion Angle φ	Torsion Angle ψ
1	27.0	1.40	0.13	0.05	0.29
2	20.3	2.40	0.11	0.30	0.32
3	23.0	0.40	0.03	0.05	0.32
4	27.0	1.80	0.08	0.36	0.82
5	23.0	1.60	0.20	0.10	0.69
6	18.9	1.60	0.08	0.15	0.28
7	18.9	2.80	0.55	0.28	0.94
8	18.9	4.20	0.23	0.78	1.54
9	9.5	5.00	0.63	1.09	2.51
10	5.4	2.20	0.54	0.40	0.90

Results

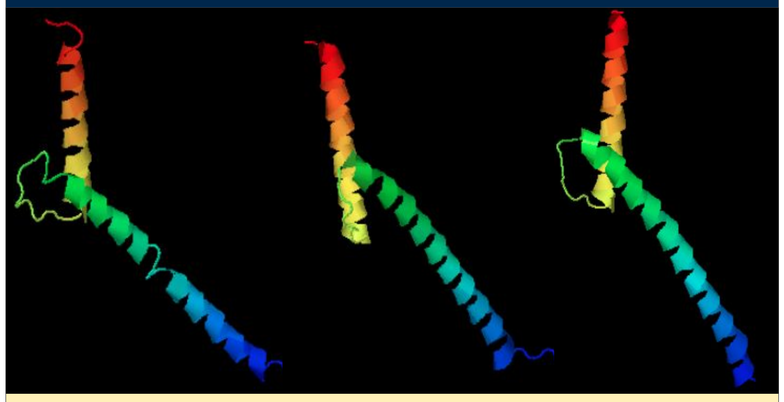


Figure 2: Designs 1, 2, and 3, respectively

Discussion

The *in silico* protein design results suggest that the wild-type structure of ID2, the helix-loop-helix motif, offers the highest likelihood of binding with E47 and other Class I transcription factors. This is highly significant as a structure with significant variation from the wild-type would suggest a more effective method of binding to E47 exists.

Additionally, the results offers increased hope for the correct folding of the protein during *in vivo* expression. Because almost all homologs utilized utilize the basic-helix-loop-helix motif, there is a higher certainty of the designed sequences similarly reflecting this when expressed. Homologs with less closely related structures would provide less evidence of the efficacy of the design.

Conclusion

The outcome of both literature review and *in silico* protein design with EvoDesign suggests that an ID-like inhibitor of binding may be an effective therapeutic and would likely take a similar form to other inhibitors of DNA binding. The data on both homologs and on the normalized relative error also suggests that the design will effectively translate into a functional protein *in vivo*.

The next step in the design process is experimental verification. This process involves the insertion of the designed sequences onto a plasmid and transformation into e. Coli. Once protein is expressed and purified, tests will be conducted to determine the differences in affinity between the wild-type and designed sequences.

Future directions for a project focused on the inhibition of Class I transcription factors will likely include the design of a peptide with similar binding capabilities but fewer residues. Alternatively, a protein that reduces transcription factor activity via a conformational change may also be a valid design approach.

Bibliography

Brender, Jeffrey R., David Shultis, Naureen Aslam Khattak, and Yang Zhang. "An evolution-based approach to De Novo protein design." In *Computational Protein Design*, pp. 243-264. Humana Press, New York, NY, 2017.

Langlands, Kenneth, Xiaoying Yin, Geetha Anand, and Edward V. Prochownik. "Differential interactions of Id proteins with basic-helix-loop-helix transcription factors." *Journal of Biological Chemistry* 272, no. 32 (1997): 19785-19793.

Stighall, Maria, Christina Manetopoulos, Håkan Axelson, and Göran Landberg. "High ID2 protein expression correlates with a favourable prognosis in patients with primary breast cancer and reduces cellular invasiveness of breast cancer cells." *International journal of cancer* 115, no. 3 (2005): 403-411.