Disordered Transactivation Domain Prediction Analysis of *Homo sapiens* and *Mus musculus* Transcription Factors

Samantha Renard, Margaret Campbell, Shih-Ting Huang, Dr. Elizabeth Komives

Abstract

p53 and NFkB have been shown to contain a disordered transactivation domain (TAD). To find other transcription factors that are predicted to also include a disordered TAD, 223 total transcription factors from both *Homo sapiens* and *Mus musculus* were scanned for intrinsically disordered regions (IDRs), then found scanned for TAD regions. After the protein data was logged in a Google Sheet, p50 and p65 were expressed by E.coli using a recombinant plasmid, a process which would be repeated on the other proteins that were predicted to carry out further experiments to confirm the predictions. Out of the 126 *Mus Musculus* proteins logged, fifteen proteins had disordered TAD regions with over an 85% match, and there were four perfect matches. Out of the 107 *Homo sapiens* proteins logged, nine proteins had disordered TAD regions with over an 85% match, and three were perfect matches. It is important to note that TAD prediction only covers the 9 amino acid long TAD and not other type of TAD.

Introduction

Even though it is commonly thought that proteins require a predetermined folded structure to function correctly, there are many proteins that are intrinsically disordered or unstructured. These intrinsically disordered proteins (IDPs) can be partially or completely unstructured, or normal, structured proteins that have intrinsically disordered regions (IDRs). Two transcription factors, p53 and NFkB, both have disordered transactivation domains, or TADs. The TAD of a transcription factor contains the binding sites for other proteins, so the possibility of a TAD also being an IDR could open the door for new discoveries concerning the role of the domain and their functioning mechanism. p53, a transcription factor with a key role in tumor suppression, has usually been experimented on without it's transactivation domain, but the Komives' Lab recent hypothesis that the TAD folding and unfolding might change the way the transcription factor is commonly studied.

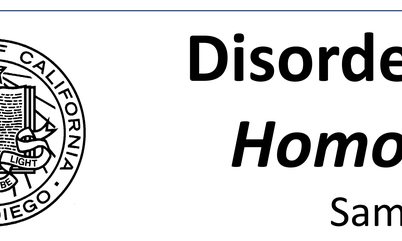

FIGURE 1: Intrinsically disordered protein

The purpose of the database created in the experiment was to find other transcription factors present in *Homo sapiens* and *Mus musculus* that included 9 amino acid long intrinsically disordered transactivation domain. In the future, this experiment could be continued by performing protein expression on the other proteins with predicted disordered TADs using the same procedure for expressing p50 and p65.

Database Compilation

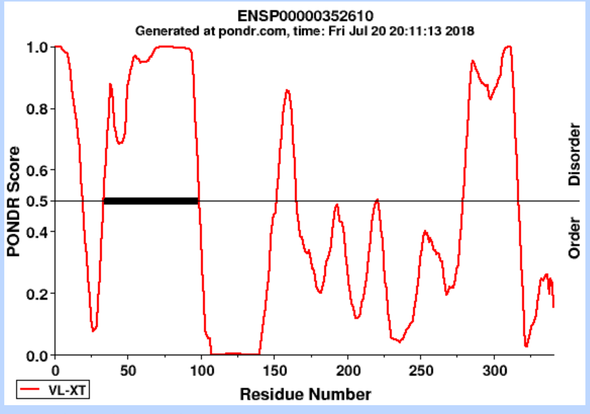

Figure 2: PONDR's prediction of IDRs in *Homo sapiens* protein ENSP00000352610. The bolded region, with a predicted disorder segment ranging from 34-99, has very high probability of being disordered (average strength: 0.8941). The Residue Number refers to the predicted disorder segment's placement in the amino acid sequence of the protein, and the PONDR score refers to the probability of different regions being disordered.

Database Compilation (continued)

There are several databases, like PONDR (Predictor Of Natural Disordered Regions) that look at the amino acid sequence of a protein and predict the IDRs of the protein based on the properties of the individual amino acids and its relationship with neighboring amino acids within the sequence. Some amino acids, like alanine and glutamic acid, are disorder-promoting, hydrophilic, charged amino acids. Others, like cysteine and phenylalanine, are ordered, hydrophobic, uncharged amino acids. When databases like PONDR notice a region of a protein sequence with several disorder-promoting amino acids together with some possibly ambiguously-charged amino acids, these databases make charts similar to the one in Figure 2.

After recognizing the predicted disorder segments of specific proteins (which were listed on transcriptionfactor.org), these same proteins were entered into the Nine Amino Acids Transactivation Domain 9aaTAD Prediction Tool. This predictor predicted the presence of TADs (transactivation domains), and the predicted TADs were compared to the predicted IDRs to see if any of the protein's TADs were disordered. Well-known transcription factors, like p53 and NFkB both had known disordered 9 amino acid-long TADs, so the database looked for disordered 9 amino acid-long TADs in the entered protein sequence. In Figure 3, the same *Homo sapiens* p53 protein used earlier, ENSP00000352610, was predicted by the Nine Amino Acids Transactivation Domain 9aaTAD Prediction Tool to have three matches, but two of the three predicted TAD regions had lower than an 85% match, so they were not included in our database. The perfect TAD match predicted, however, was also in the predicted IDR of the protein, meaning that ENSP00000352610 had one predicted disordered TAD. This process was repeated for 107 *Homo sapiens* proteins and for 126 *Mus musculus* (house mouse) proteins. The results of the data are later explained in the conclusion and featured in Tables 1-2.


Matches (3):

Sequence	Start	End	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12		
ETFSDLWKL	17	25	-	+	+	-	+	+	+	+	+	+	+	+	83% match	Graph
QAMDDLMLS	38	46	+	+	+	+	+	+	+	+	+	+	+	+	Perfect match	Graph
DDIEQWFTE	48	56	-	+	+	-	-	+	+	+	+	+	+	+	75% match	Graph

Figure 3: The Nine Amino Acids Transcription Domain 9aaTAD Prediction Tool's prediction of TADs in Homo sapiens protein ENSP00000352610. The perfect match sequence, beginning with the residue numbers of 38-46, can be found in PONDR's predicted IDR located from 34-99.

p50 and p65 Protein Expression

The purpose of this lab was to use an expressive vector designed for gene expression to target p50FL and p65FL. Transcriptionfactor.org, the database used, predicted several possible disordered TADs, but we were unable to test each of the 200 or so proteins analyzed. Instead, p50FL and p65FL were expressed, and the process that was used would be replicated for the other proteins in their production.

First, the E.coli BL21 cells were transformed, then plated onto AMP plates. This included placing cell culture samples into a plasmid solution containing our recombinant plasmid, which was genetically modified to encode the required proteins and enzymes that would protect the cells from antibiotics. The plasmid of an E.coli is their circular DNA that contains the genetic functions. Specific restriction enzymes were used to cut specific places on the plasmid to allow easier genetic-sequence insertion. This recombinant plasmid would also allow, later, the E.coli to produce the protein of interest. To get the cells to transform, they were transitioned between heat and cold to put the cells under stress, increasing their intake of plasmids. The cell cultures were then placed in the shaking incubator to receive necessary oxygen and to be moved around to obtain imperative nutrients.

When the cell cultures were placed onto AMP plates, the antibiotic ampicillin killed E.coli that had not taken in the recombinant plasmid; therefore insuring that they would not be used for the experiment. The cultures were inoculated and placed to grow again in the shaking incubator until the solution density reached the optimal density of 0.6 - 0.8, determined by a spectrophotometer.

When the cell cultures became large enough to express the protein, the cultures were removed from the shaker and placed on ice to stop further growth. With the new plasmid, which had modified lac Operon, the E.coli now produced the desired protein due to the presence of the inducer, IPTG. The lac Operon on the plasmid is modified to code for the desired protein instead of lactose digesting enzymes. The lac Repressor keeps the RNA polymerase from producing unnecessary lactose-digesting enzymes. Inducers like allolactose can temporarily bind to the repressor so this part of the plasmid is transcribed and translated. To make the E.coli constantly produce our protein, they were induced with IPTG, which would act similar to allolactose and bind to the lac repressor. The IPTG binded to the lac repressor, therefore the E.coli continuously produced the desired proteins, p50FL and p65FL.

P50 and p65 Protein Expression (continued)

The E.coli culture was placed in the 18° C shaking incubator so it could continue to produce the protein without culture growth or cell death. The broth with the E.coli was then spun down to separate the liquid (supernatant), which included E.coli environment-broth, from the pellet, which only featured the E.coli. The liquid was removed and the formed pellet was resuspended in buffer, then placed into a sonicator. The sonicator used sound waves to burst the E.coli and rip them into pieces so they released the protein they had been building. After another round of centrifuging, the dead E.coli became a pellet, while the supernatant included the required protein.

To purify it, the supernatant was ran through a Ni-NTA column which used three different buffers of different concentrations of imidazole. Imidazole, which resembles the amino acid histidine, binds to the column. Because the protein had a poly-histidine tail, the protein binds to the column. First, the buffer was ran through the column to remove contaminants. Then, the supernatant was ran through the column, and any non-protein would run through the column due to it's low affinity. When the concentration of the buffer was increased, the purified protein itself was pushed through the column. The protein and other things in the supernatant selectively bound to the column, while the supernatant content bound to the column. After, the protein was dialyzed and stored in SEC (size exclusion column) buffer. To ensure that the proteins were present, a gel was ran with standard ladder so the distance each band traveled could identify the protein size. The standard ladder gel was compared to that of the flow-through (things that didn't bind to the column), the wash (the things that bound to the column but not as tightly as the protein), and the elute (hopefully the desired protein).

Results and Conclusion

According to Table 1 (for *Mus musculus*), 126 proteins were scanned, and of 86 of the 126 proteins were more than 85% likely to include 9aa TAD. Out of those 86 proteins, 71 of them had ordered 9aa TADs, while only 15 of them had 9aa TADs in their IDRs. According to Table 2 (for *Homo sapiens*), out of the 107 proteins scanned, 78 were more than 85% likely to include 9aa TAD. 9 of these 78 had disordered 9aa TADs, and the other 69 had ordered 9aa TADs.

The gels showed that the proteins were successfully expressed, and in the future, this procedure could be repeated to confirm that the other proteins that were predicted to have a disordered TAD actually had one. Based on the predicted data, disordered 9aa TADs are relatively uncommon, so p53 and NFkB are really important to further understand.

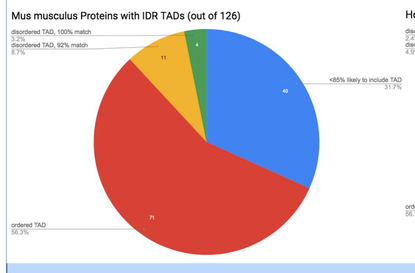

Table 1: *Mus musculus* Proteins with IDR TADs, 126 proteins were scanned.

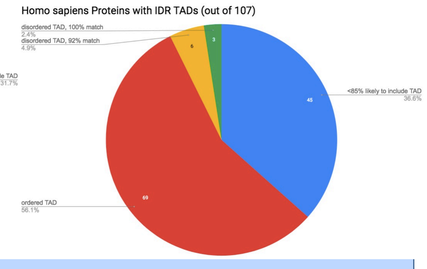

Table 2: *Homo sapiens* Proteins with IDR TADs, 107 proteins were scanned.

Acknowledgements/ References

I would first like to thank Shih-Ting Huang for being an incredible mentor and person, and the overall Komives lab for being so welcoming. I am incredibly grateful for Dr. Elizabeth Komives for starting the Research Scholars program, and for Academic Connections for providing housing and a wonderful community.

Derek Wilson, Varodom Charoensawan, Sarah K. Kummerfeld and Sarah A. Teichmann, DBD - taxonomically broad transcription factor predictions: new content and functionality Nucleic Acids Research 2008 doi: 10.1093/nar/gkm964.
Intrinsically Disordered Proteins. (2013, July 03). Retrieved July 25, 2018, from https://weisgroup.ku.edu/intrinsically-disordered-proteins
Intrinsically disordered proteins. (2018, July 25). Retrieved July 25, 2018, from https://en.wikipedia.org/wiki/Intrinsically_disordered_proteins
Molecular Kinetics, Inc., Washington State University and the WSU Research Foundation
Piskacek, Martin. 9aaTAD Prediction result (2006). Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2009.3984.1> (2009)