**Background**

In our quest to transdifferentiate the gut cells into insulin-producing cells, we propose to transport our 3 transcription factors PDx1, MAFA and NGN3, as a single polyprotein sequence, from our bacteria through the T3SS injection. This was done since sending the individual proteins via T3SS would have posed problems like one protein hampering the transfer of another. For the same purpose, appropriate linkers were selected and a protease was chosen which would ultimately cleave the polyprotein at particular sequences to free the 3 TFs to perform their individual roles in the transdifferentiation process.

In this regard, it is important to understand and verify that the function of the 3 TFs is not lost/altered during this process. In other words, we need to make sure that the active domains of our TFs are unaffected by the polyprotein formation and cleaving process. For this, we take the help of structural modelling.

**Fundamental Basis:**

### **Protein folding**

Protein folding is the physical process by which a polypeptide chain folds into a highly specific and biologically functional three-dimensional structure from a random coil. This three-dimensional conformation of a protein molecule is determined by its amino acid sequence. Every protein molecule exists as an unfolded entity immediately after translation. It is because of the various non-covalent interactions between the amino acids that the protein folds and adopts a structure called the native state.

### **Poly-proteins**

Poly-proteins are proteins that can be cleaved to produce more than one functionally active polypeptides. In our case, the 3 TFs are linked together to form a polyprotein.

In most cases, the functionality of such proteins is not affected because of the fusion process. It is because of intrinsic protein domains modularity. The part of a polypeptide that corresponds to a given domain can be removed or added to the rest of the molecule without hampering its native functional roles.

However, it is highly advised to predict the three-dimensional structure of fusion protein or the artificially attached proteins after the linkers have been cleaved by the chosen protease. The knowledge of the spatial organization of any given protein and the effect of the remnant linker residues on the active site of the protein is extremely useful for understanding the function (and its changes (if any) and for the rational modifications of the proteins.

**Methods/Servers**

1. [Protein Data Bank](https://www.rcsb.org/): The **Protein Data Bank** (**PDB**) is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids.
2. [Uniprot](https://www.uniprot.org/): **UniProt** is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects.
3. [I-TASSER](https://zhanglab.ccmb.med.umich.edu/I-TASSER/): It is a bioinformatics method for predicting three-dimensional structure models of protein molecules from amino acid sequences.
4. [Swiss Model](https://swissmodel.expasy.org/): It is a structural bioinformatics web-server dedicated to homology modeling of 3D protein structures.
5. [PyMOL](https://pymol.org/): *PyMOL* is a user-sponsored molecular visualization system on an open-source foundation, maintained and distributed by Schrödinger.
6. [PSIPRED](http://bioinf.cs.ucl.ac.uk/psipred/) : PSI-blast based secondary structure PREDiction is a method used to investigate protein structure.
7. [YASARA](http://www.yasara.org/): Yet Another Scientific Artificial Reality Application (YASARA) is a computer program for molecular visualisation, modeling, and dynamics. We have used it for energy minimization of our structures.
8. [MolProbity](http://molprobity.biochem.duke.edu/) : The Ramachandran plots have been generated using MolProbity. It is most complete for crystal structure of proteins and acts as an active validation tool that produces coordinates, graphics and numerical evaluations.

The evaluation/analysis was based on RC plot, RMSD & C-Score.

**C Score**: C-score is a confidence score for estimating the quality of predicted models by I-TASSER. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. (C-score is typically in the range of [-5,2]).

**RMSD**: The **root-mean-square deviation of atomic positions** (or simply **root-mean-square deviation**, **RMSD**) is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins. The RMSD of two aligned structures indicates their divergence from one another.

**Ramachandran Plot**: The Ramachandran plot is a plot of the torsional angles - phi (φ)and psi (ψ) - of the residues (amino acids) contained in a peptide. The plot indicates which torsional angles are allowed and provides an insight into the structure of peptides.

.**Results:**

**Sequences: The native protein sequences were obtained from PDB or Uniprot.**

Sequences of Native Proteins

Pdx1: MNGEEQYYAATQLYKDPCAFQRGPAPEFSASPPACLYMGRQPPPPPPHPFPGALGALEQG

SPPDISPYEVPPLADDPAVAHLHHHLPAQLALPHPPAGPFPEGAEPGVLEEPNRVQLPFP

WMKSTKAHAWKGQWAGGAYAAEPEENKRTRTAYTRAQLLELEKEFLFNKYISRPRRVELA

VMLNLTERHIKIWFQNRRMKWKKEEDKKRGGGTAVGGGGVAEPEQDCAVTSGEELLALPP

PPPPGGAVPPAAPVAAREGRLPPGLSASPQPSSVAPRRPQEPR

MafA:

MAAELAMGAELPSSPLAIEYVNDFDLMKFEVKKEPPEAERFCHRLPPGSLSSTPLSTPCS

SVPSSPSFCAPSPGTGGGGGAGGGGGSSQAGGAPGPPSGGPGAVGGTSGKPALEDLYWMS

GYQHHLNPEALNLTPEDAVEALIGSGHHGAHHGAHHPAAAAAYEAFRGPGFAGGGGADDM

GAGHHHGAHHAAHHHHAAHHHHHHHHHHGGAGHGGGAGHHVRLEERFSDDQLVSMSVREL

NRQLRGFSKEEVIRLKQKRRTLKNRGYAQSCRFKRVQQRHILESEKCQLQSQVEQLKLEV

GRLAKERDLYKEKYEKLAGRGGPGSAGGAGFPREPSPPQAGPGGAKGTADFFL

Ngn3:

MTPQPSGAPTVQVTRETERSFPRASEDEVTCPTSAPPSPTRTRGNCAEAEEGGCRGAPRK

LRARRGGRSRPKSELALSKQRRSRRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLT

KIETLRFAHNYIWALTQTLRIADHSLYALEPPAPHCGELGSPGGSPGDWGSLYSPVSQAG

SLSPAASLEERPGLLGATFSACLSPGSLAFSDFL

Sequences of Proteins with Residues

Pdx1: SGGSGSGSGMNGEEQYYAATQLYKDPCAFQRGPAPEFSASPPACLYMGRQPPPPPPHPFPGALGALEQGSPPDISPYEVPPLADDPAVAHLHHHLPAQLALPHPPAGPFPEGAEPGVLEEPNRVQLPFPWMKSTKAHAWKGQWAGGAYAAEPEENKRTRTAYTRAQLLELEKEFLFNKYISRPRRVELAVMLNLTERHIKIWFQNRRMKWKKEEDKKRGGGTAVGGGGVAEPEQDCAVTSGEELLALPPPPPPGGAVPPAAPVAAREGRLPPGLSASPQPSSVAPRRPQEPRGGSGSGSGPKKKRKVGGSGSGSGENLYFQ

MafA:

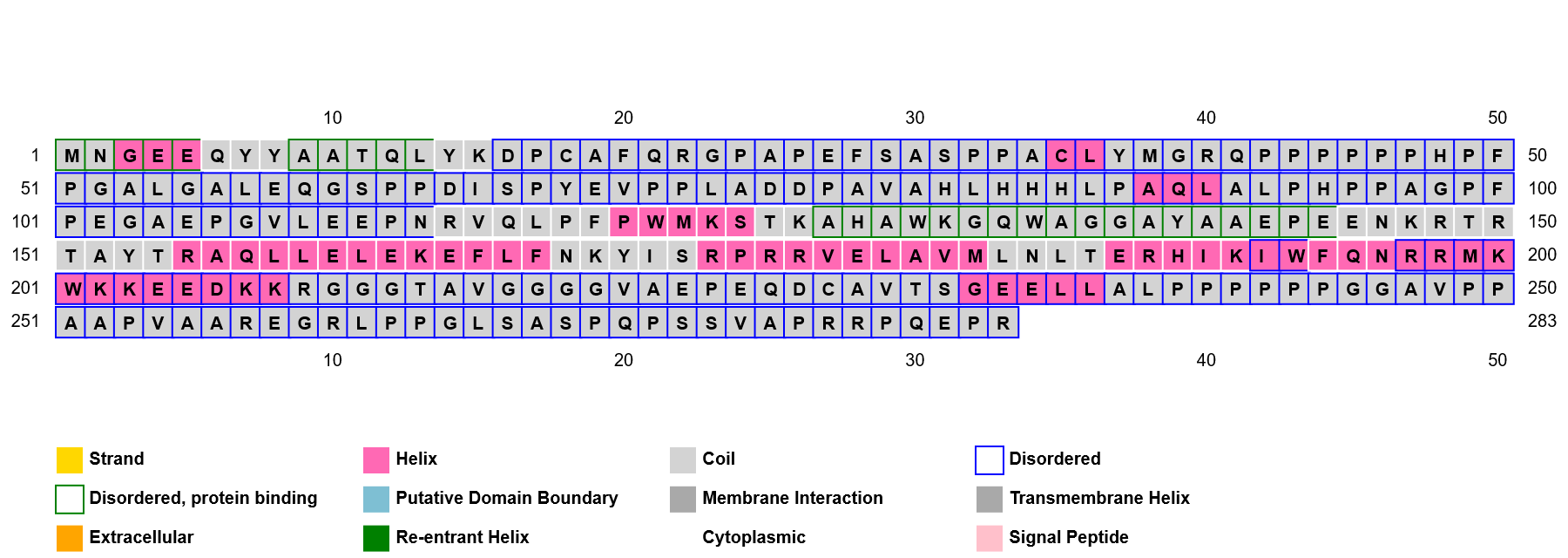
SGGSGSGSGMAAELAMGAELPSSPLAIEYVNDFDLMKFEVKKEPPEAERFCHRLPPGSLSSTPLSTPCSSVPSSPSFCAPSPGTGGGGGAGGGGGSSQAGGAPGPPSGGPGAVGGTSGKPALEDLYWMSGYQHHLNPEALNLTPEDAVEALIGSGHHGAHHGAHHPAAAAAYEAFRGPGFAGGGGADDMGAGHHHGAHHAAHHHHAAHHHHHHHHHHGGAGHGGGAGHHVRLEERFSDDQLVSMSVRELNRQLRGFSKEEVIRLKQKRRTLKNRGYAQSCRFKRVQQRHILESEKCQLQSQVEQLKLEVGRLAKERDLYKEKYEKLAGRGGPGSAGGAGFPREPSPPQAGPGGAKGTADFFLGGSGSGSGPKKKRKVGGSGSGSGENLYFQ

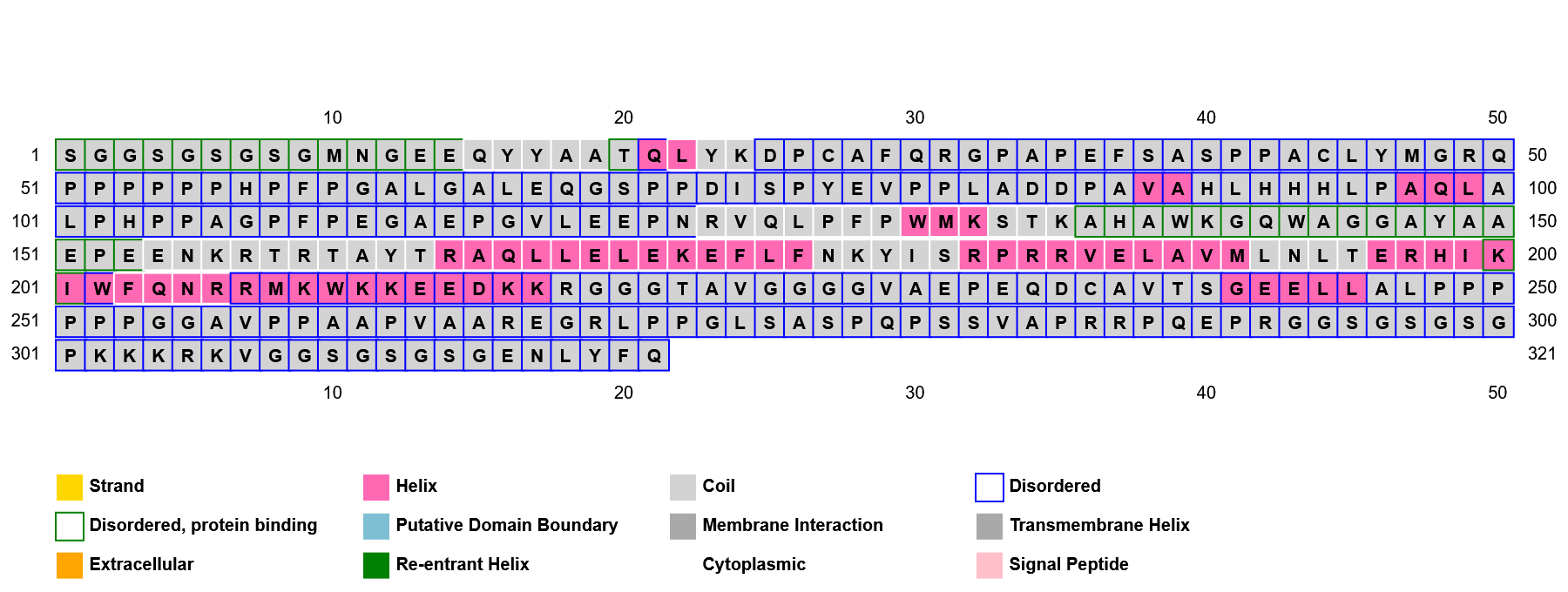
Ngn3:

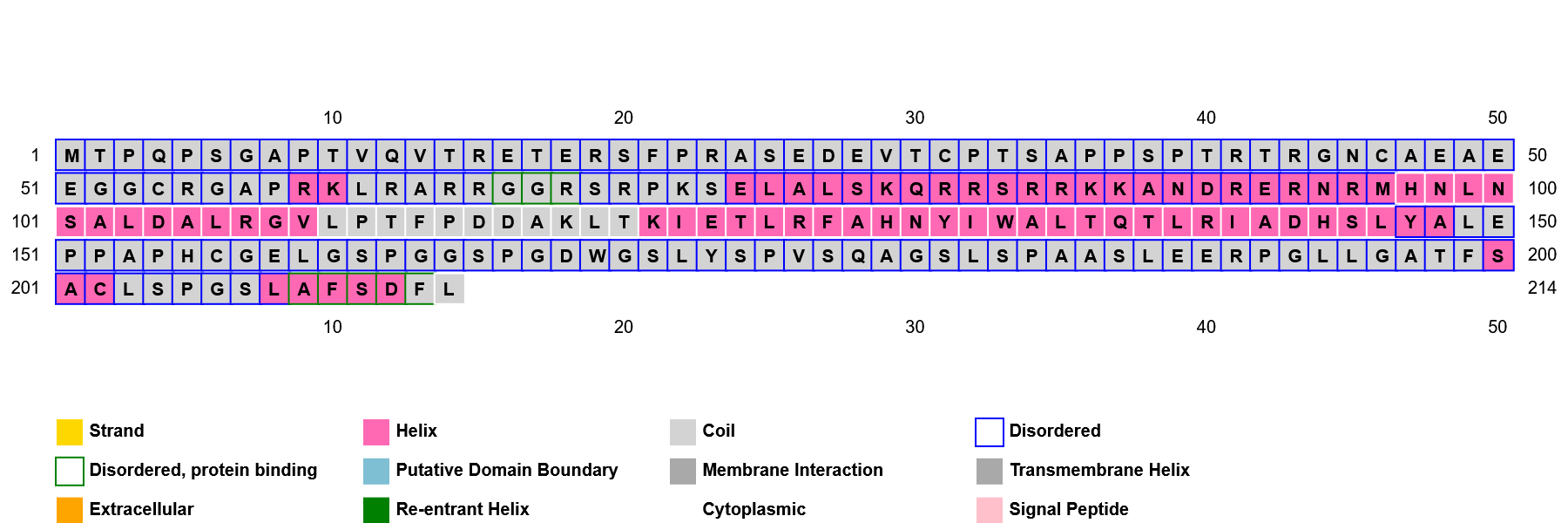
SGGSGSGSGMTPQPSGAPTVQVTRETERSFPRASEDEVTCPTSAPPSPTRTRGNCAEAEEGGCRGAPRKLRARRGGRSRPKSELALSKQRRSRRKKANDRERNRMHNLNSALDALRGVLPTFPDDAKLTKIETLRFAHNYIWALTQTLRIADHSLYALEPPAPHCGELGSPGGSPGDWGSLYSPVSQAGSLSPAASLEERPGLLGATFSACLSPGSLAFSDFLGGSGSGSGPKKKRKVGGSGSGSGENLYFQ **-----------------**

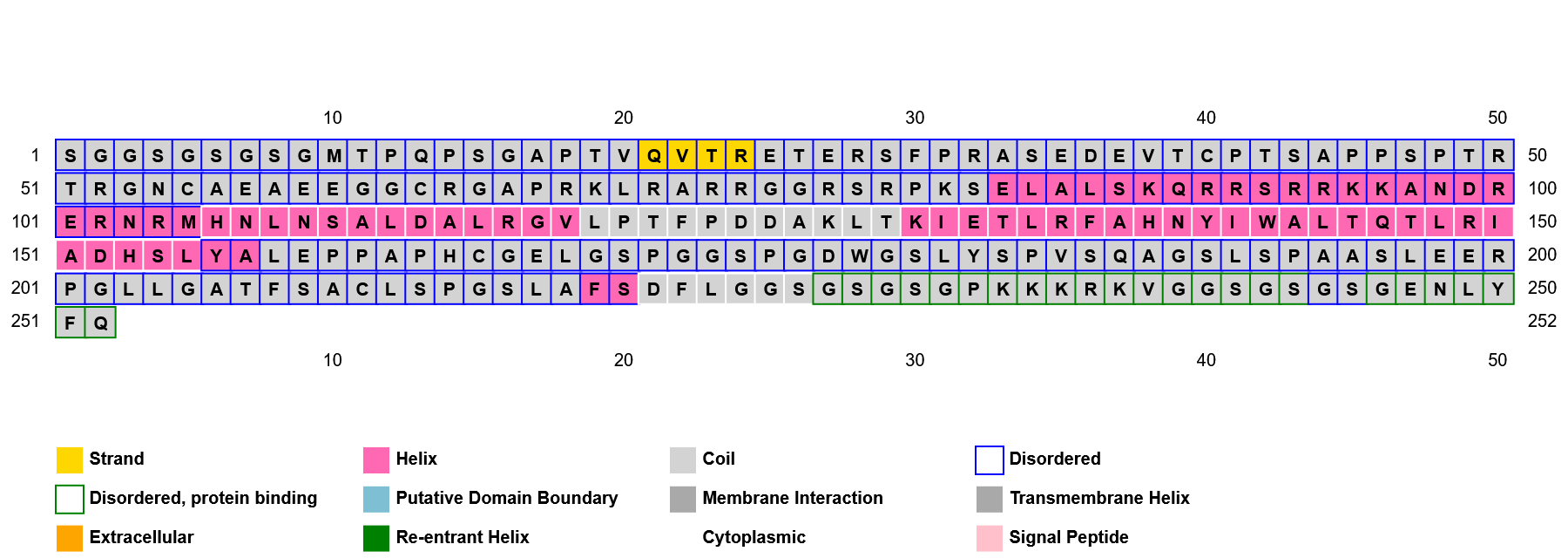
**PSIPRED:** The secondary structures were predicted successfully using this server. A large part of our TFs except for the active domains were predicted to be just random coils which was later validated when we obtained the structures from I-Tasser.

1. For PDX1   
   a)Native sequence

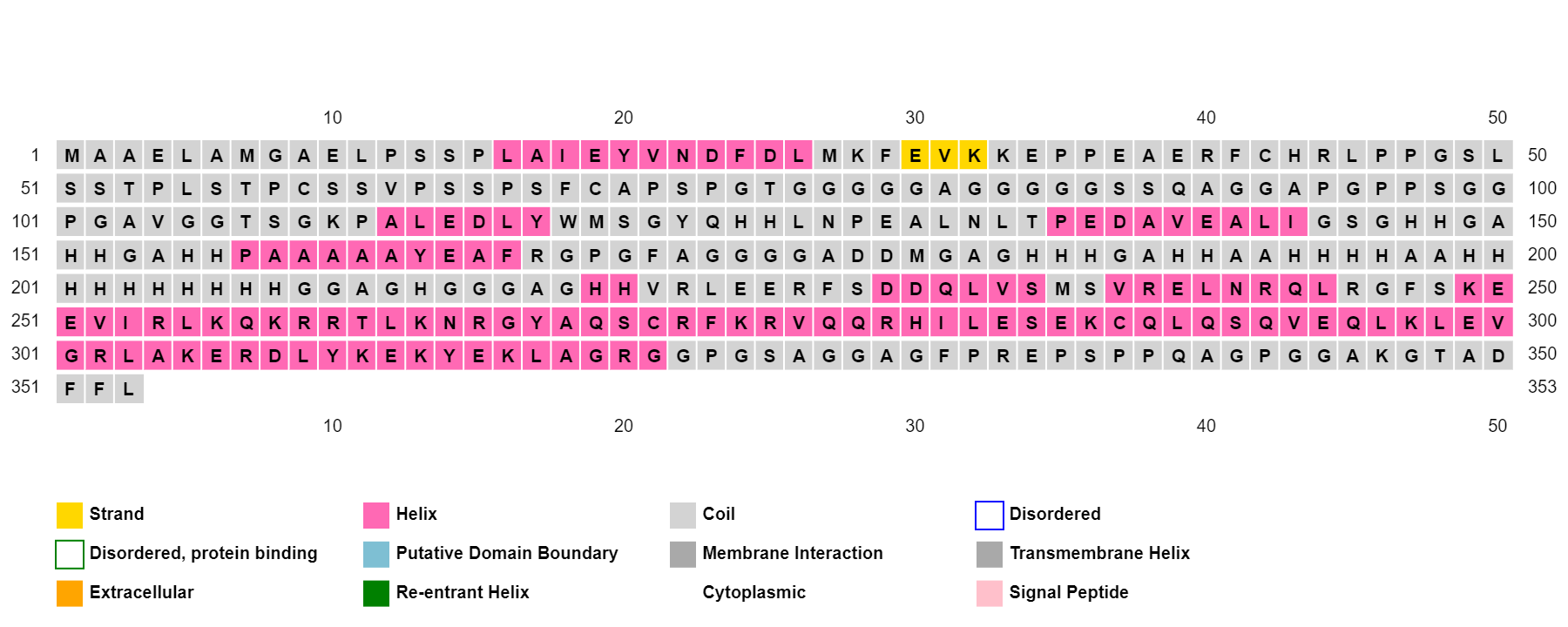


b) with remnant residues from linkers and protease cleavage  


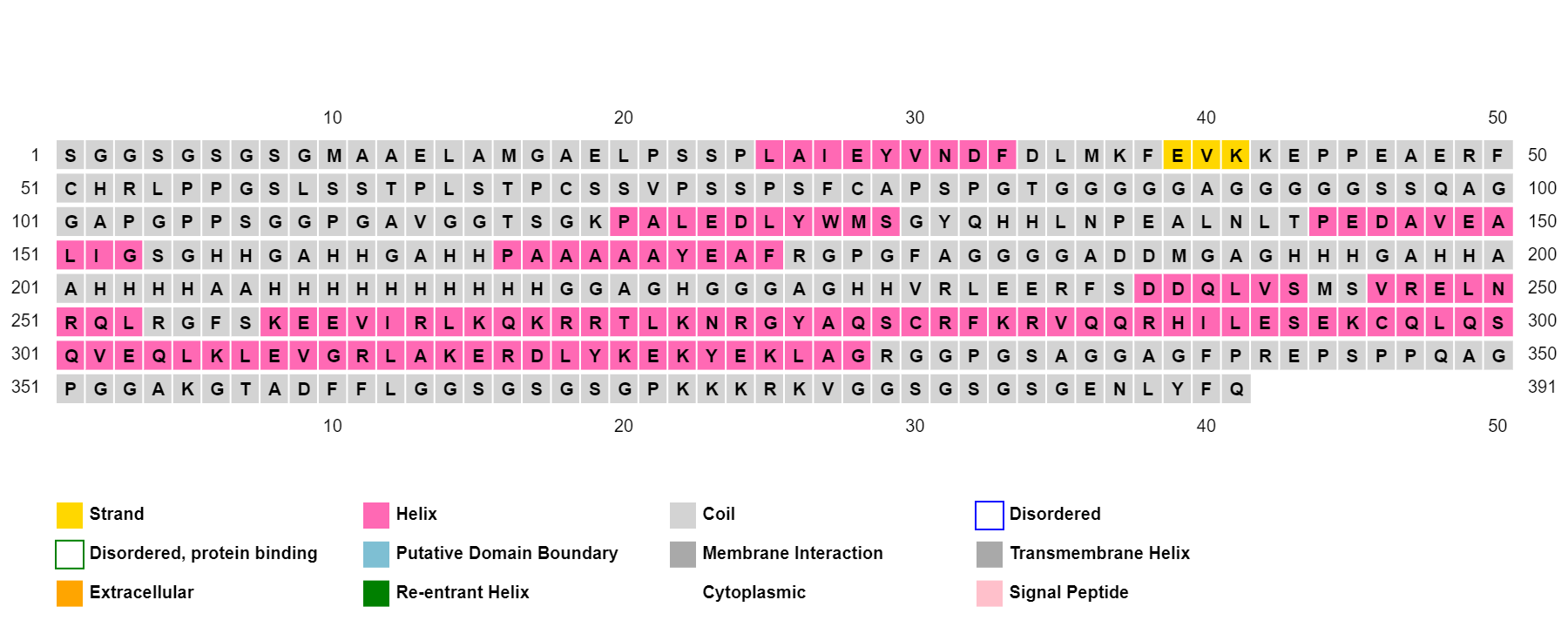
1. NGN3  
     
   a)Native sequence  
     
   

b) With additional residues  
  


1. MAFA  
     
   a)Native sequence



b) With additional residues



-------------------------

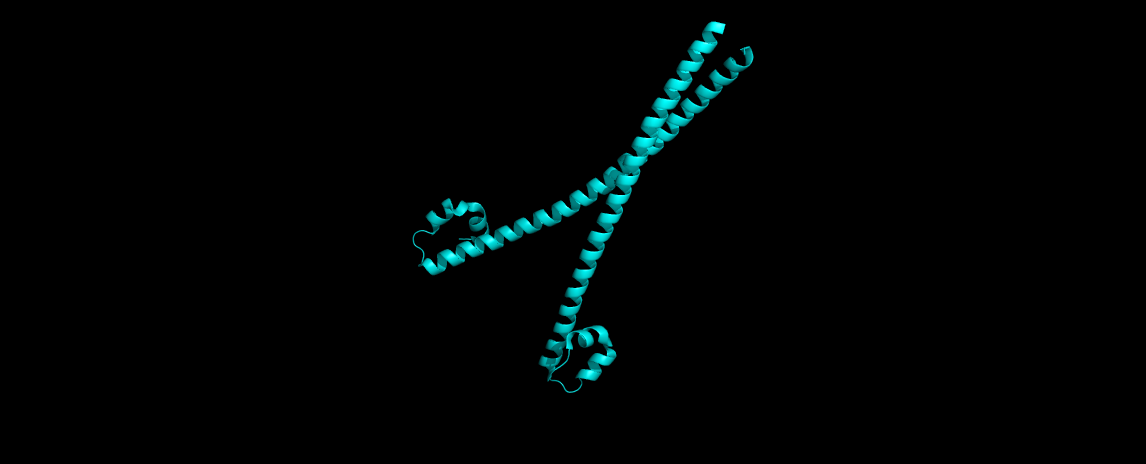
CRYSTAL STRUCTURES FOR HOMEODOMAINS: After going through literature and searching on protein data bank (PDB), we found that the complete crystal structures for our TFs are not readily available. In most cases, just the structure of the active domains are available (PDB IDs: **2H1K** for PDX1, **4EOT** for MafA and a previously published [**Swiss Model** for NgN3](https://swissmodel.expasy.org/repository/uniprot/Q9Y4Z2?csm=2E80623E46F88B83) ). So,

# 2H1K

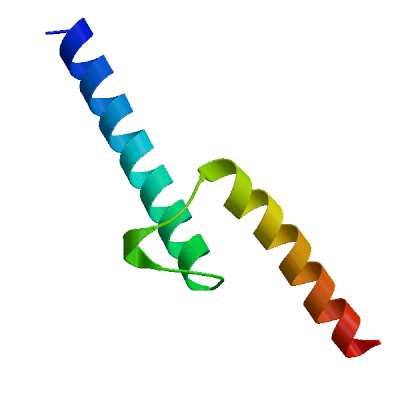


#### Crystal structure of the Pdx1 homeodomain

4EOT



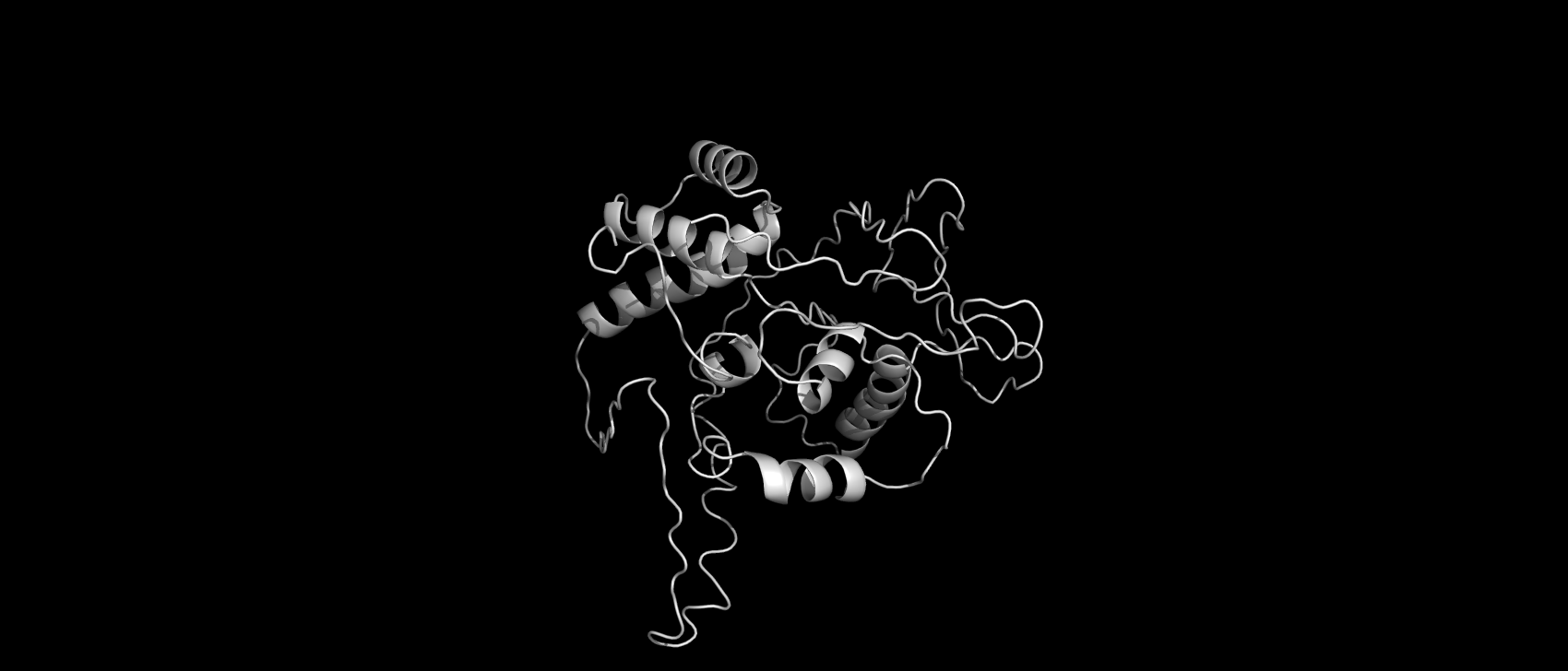
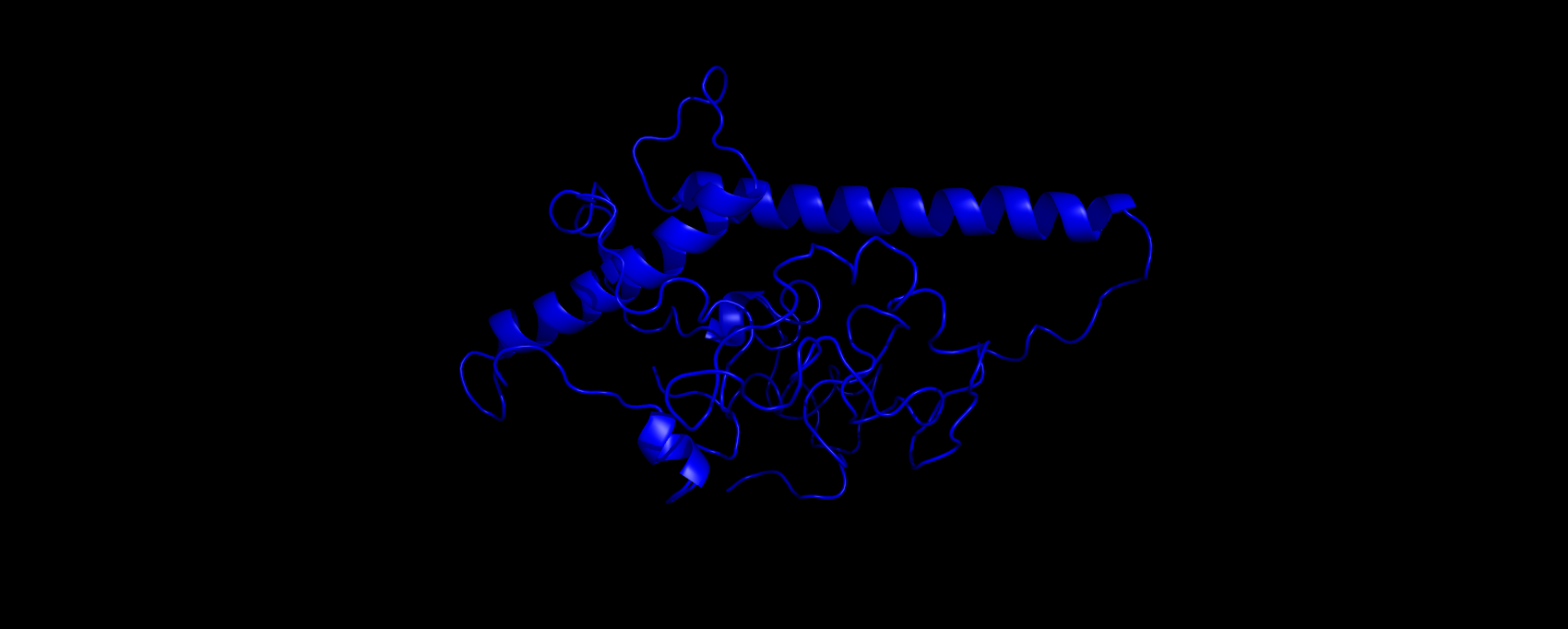
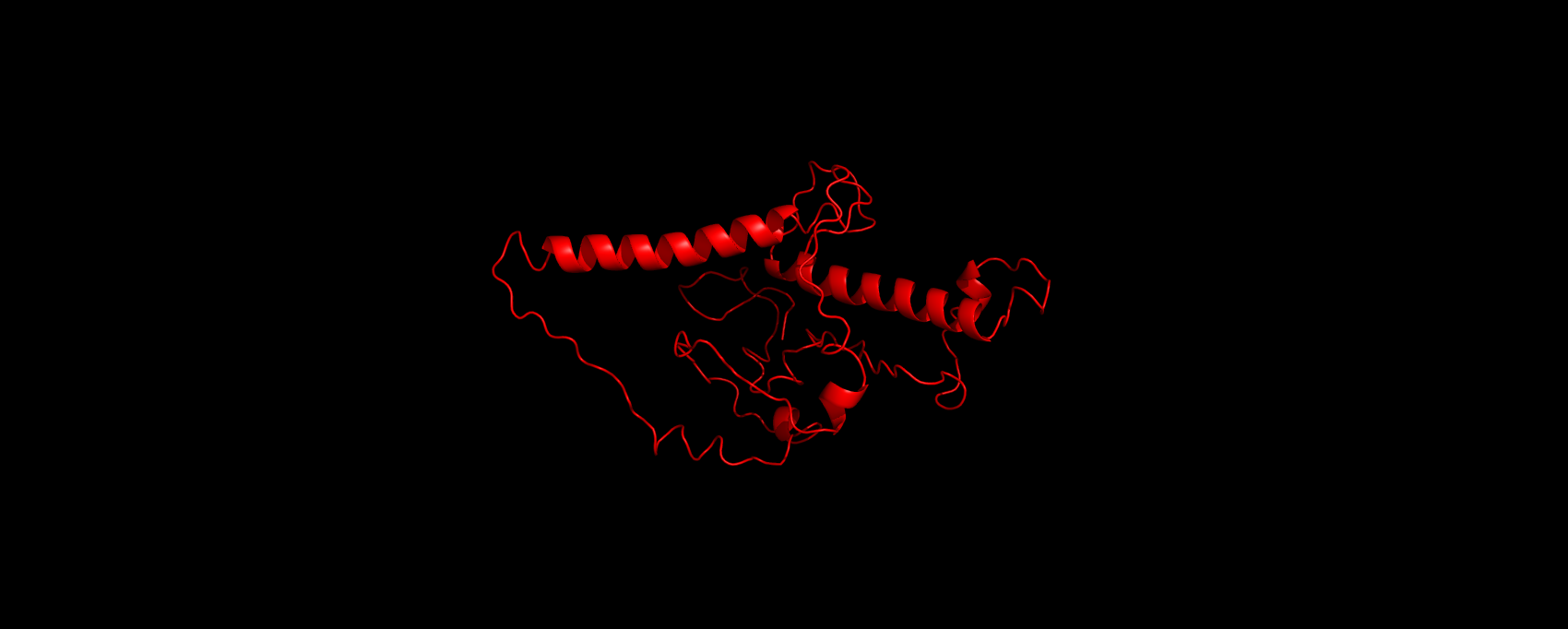
Crystal structure of Mafa homeodomain



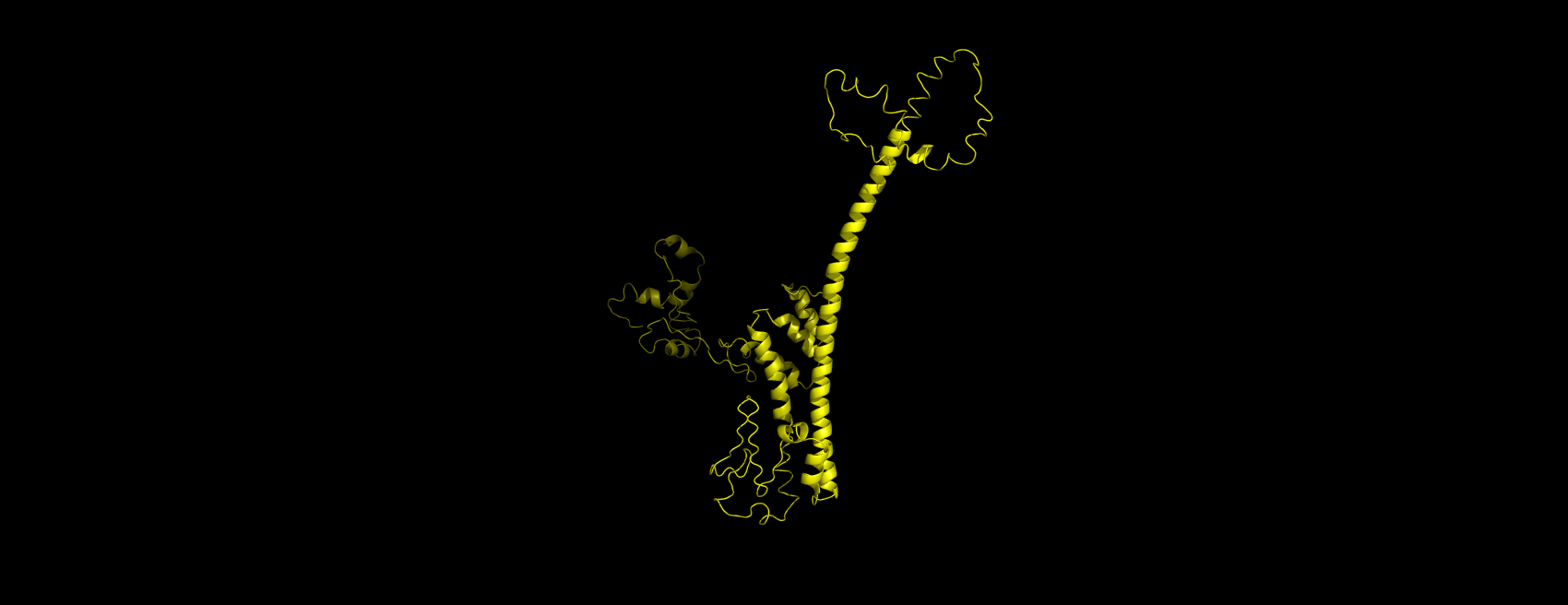
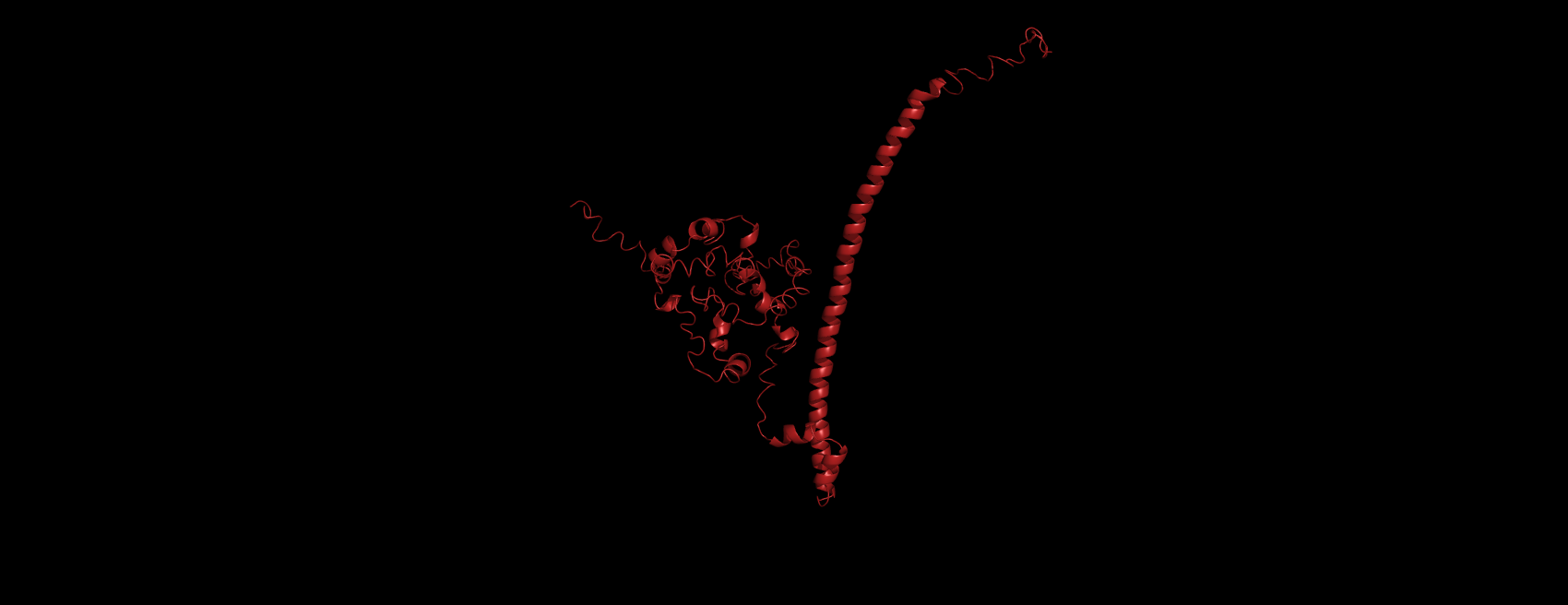
Swiss model for Ngn3

**Models obtained from I-TASSER:**

Since the complete crystal structures were not available, we obtained Structural models for all three TFs (with and without the additional residues from linker sequences) using I-TASSER and their energies were minimized using YASARA server. The structures obtained showed that a large part of our structures are just random coils as predicted by PSIPRED.

1. PDX1 : Grey: without residues (-4.21) and magenta: with additional residues(-4.34).
2. NGN3: Red: Without residues (-4.79)and Blue: With residues(-4.79)  
     
     
   

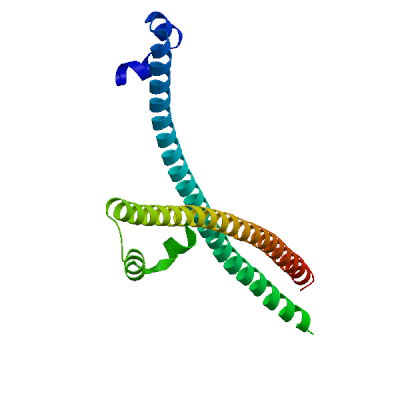
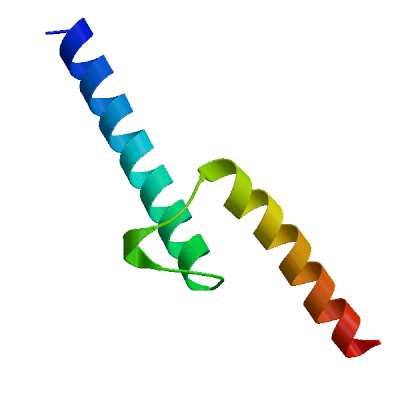
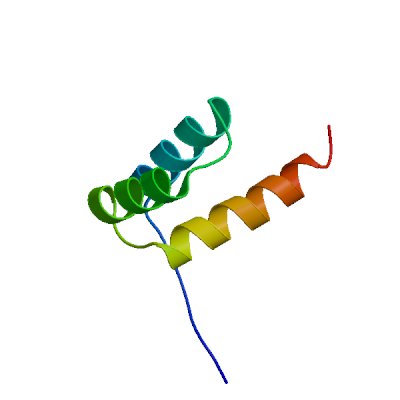
1. MafA : FireBrick (without resides C=-3.96) and Yellow (with residues, C=-2.92)



----------------------

------

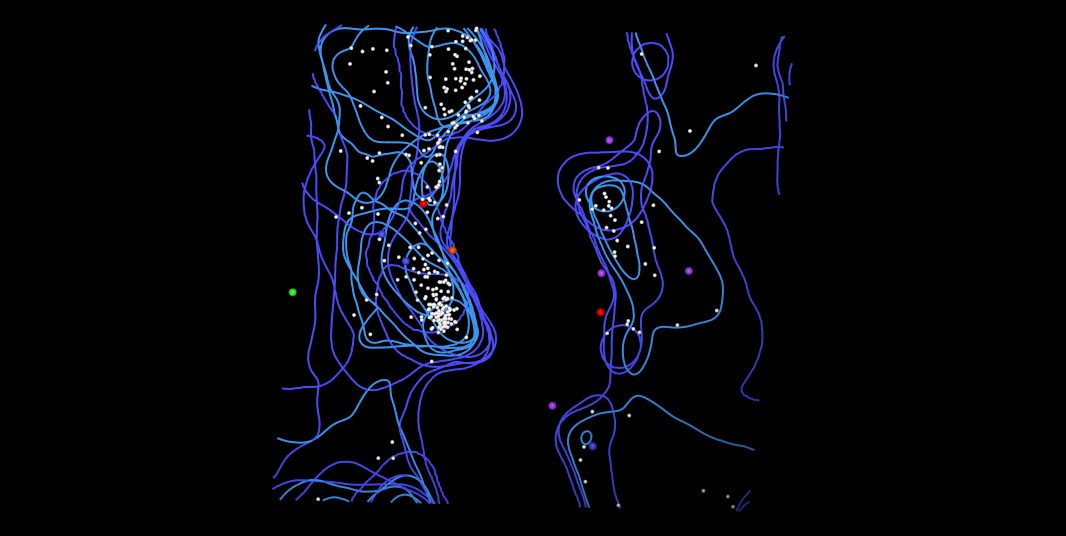
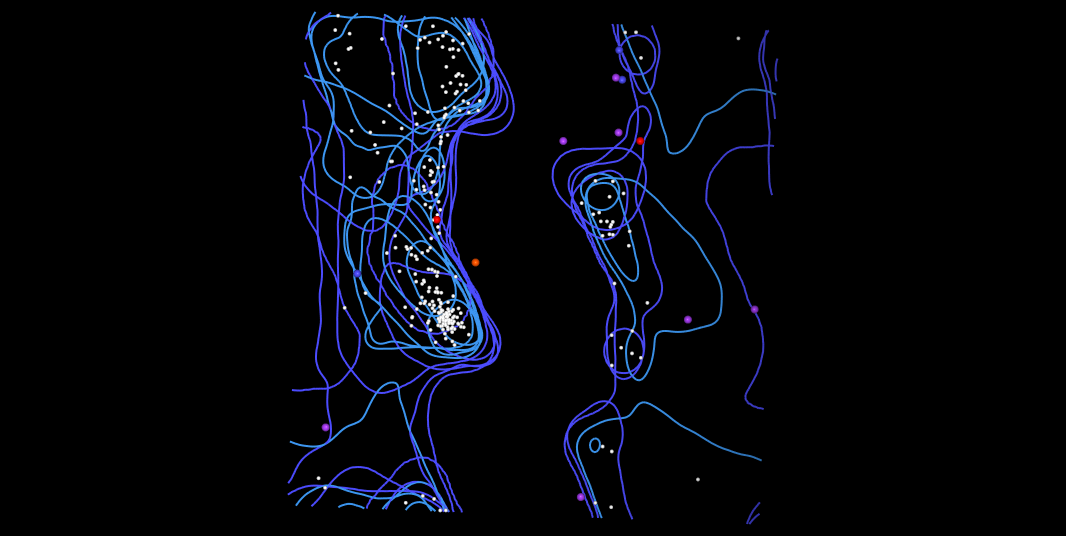
**Swiss Model:** We also usedHomology modeling via the Swiss Model server to obtain the structure for our TFs. This was done as an alternative approach to the one via I-TASSER. However, the homology modeling was not very useful for our case since the complete structures were not obtained via this method irrespective of the templates used. Hence we went ahead with the i-Tasser structures for our analysis. ( Swiss Models for PDX1, NGN3, MAFA (LtoR))



**Ramachandran plots**:

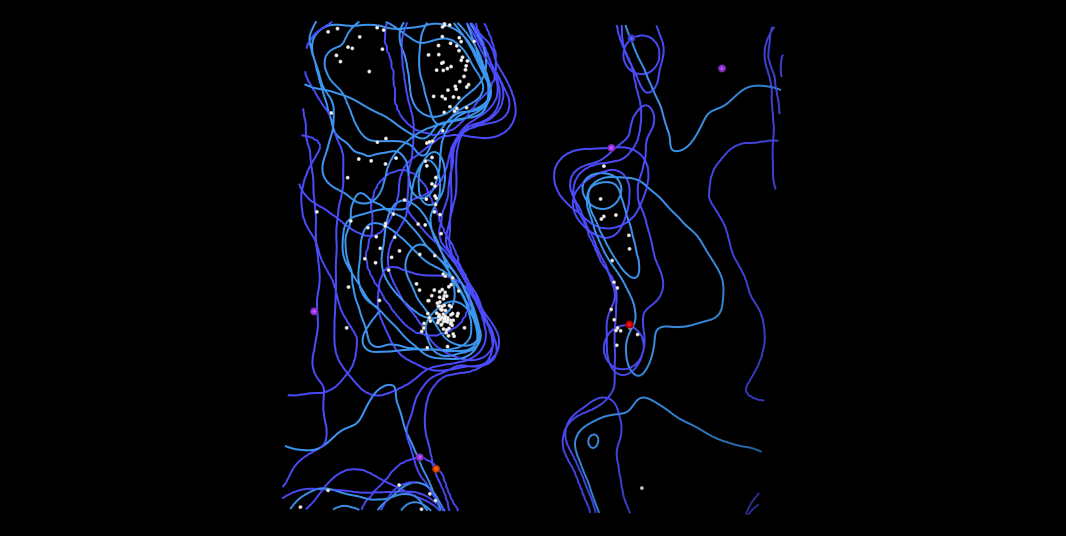
The RC plots for the 2 structures (with and without residues) were compared for each of the TF to see if there are any significant differences in the secondary structures. The results obtained are as follows:

PDX1



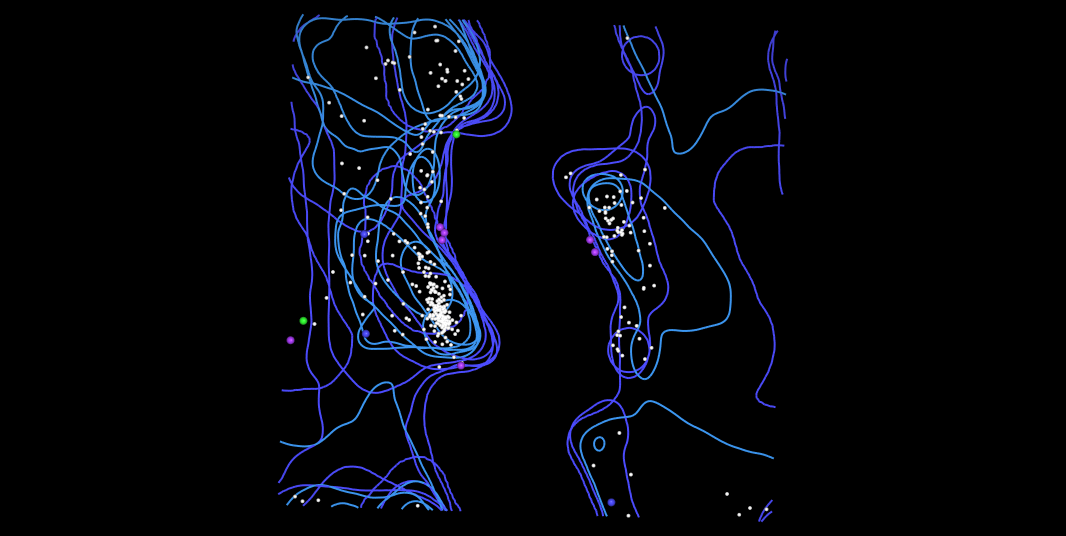
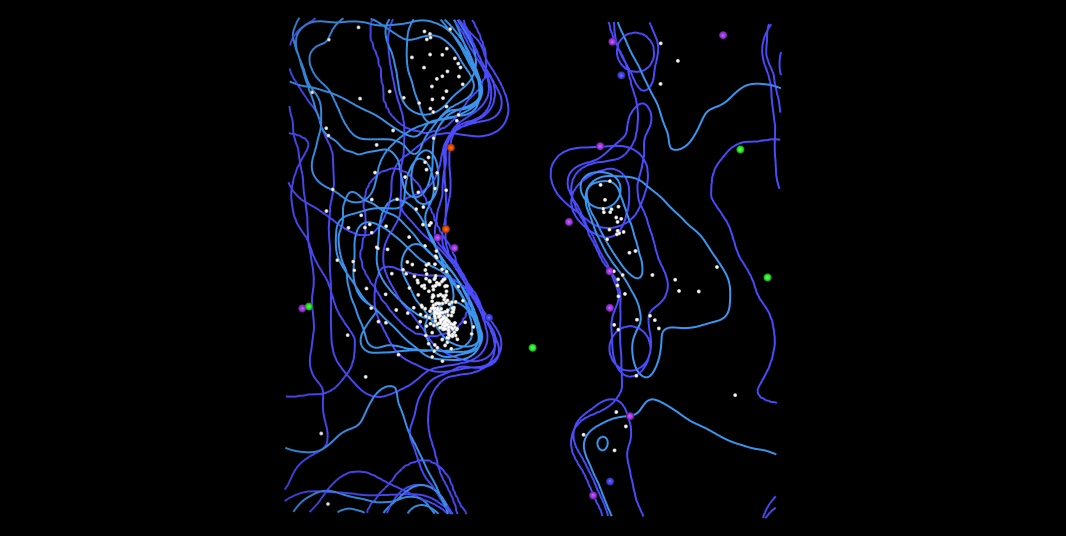
(RC plots for PDX1 Native (left) and with residues( right) , no major differences in the secondary structures noted)

NGN3



(RC plots for NGN3 Native (left) and with residues( right) , no major differences in the secondary structures noted)

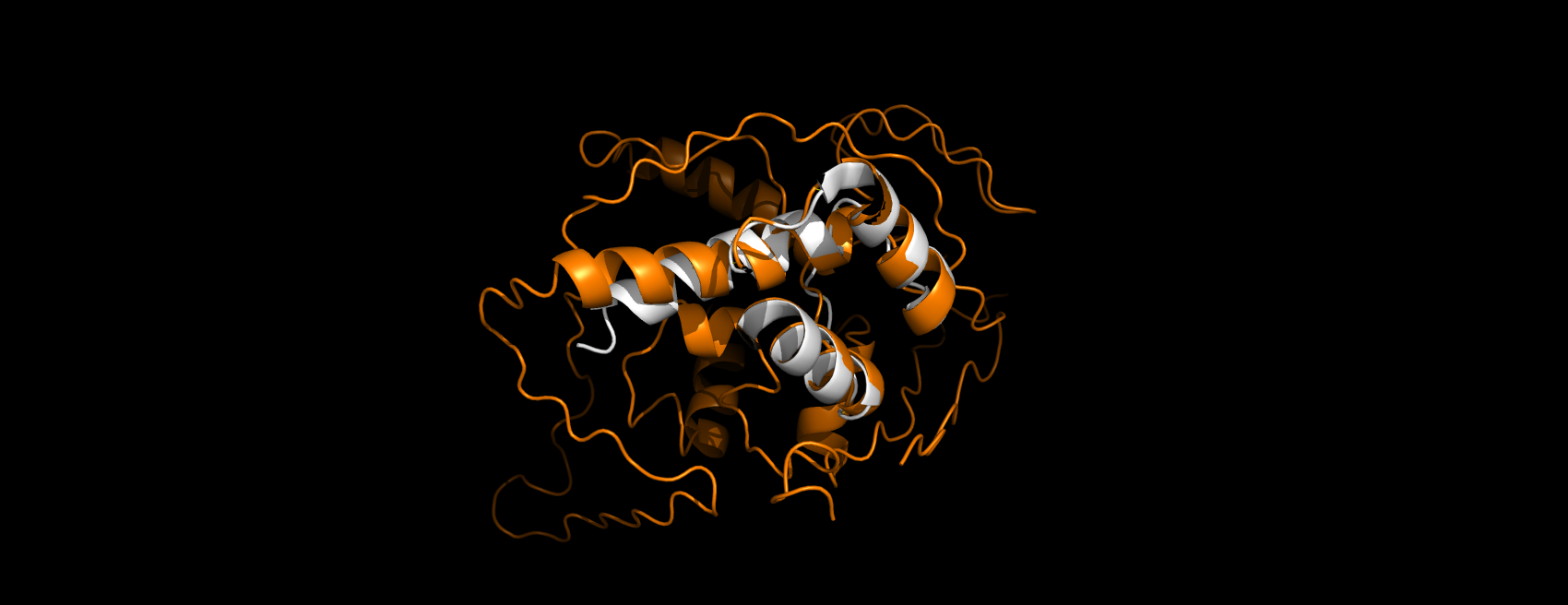
MAFA



(RC plots for MAFA Native (left) and with residues( right) , no major differences in the secondary structures noted)

**Alignment**: The structural alignments were done using PyMOL to analyze if the structure was significantly affected by the addition of linker residues. The active domains remain unchanged for the proteins with linker residues. Since most of the remaining parts of the proteins are random coils as seen from PSIPRED results and I-TASSER models, they didn’t align as expected. The RMSD Values obtained were under acceptable limits for the active domains

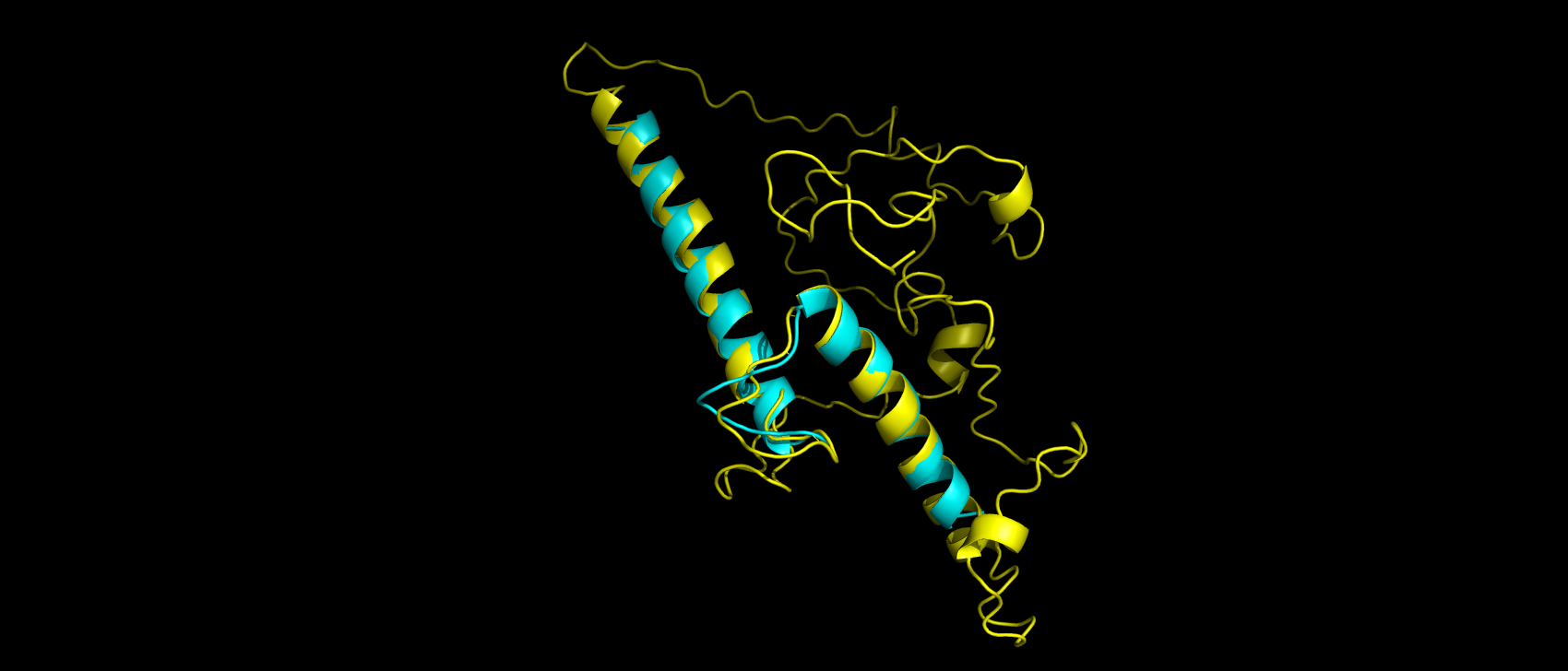
**Alignment of I-tasser models with crystal structures of active domain**



I-tasser model for Pdx1 aligned with it’s homeodomain (PDB ID: 2H1K) **RMSD: 0.955**



I-tasser model for Mafa aligned with it’s homeodomain(PDB IB:4EOT) **RMSD:1.799**

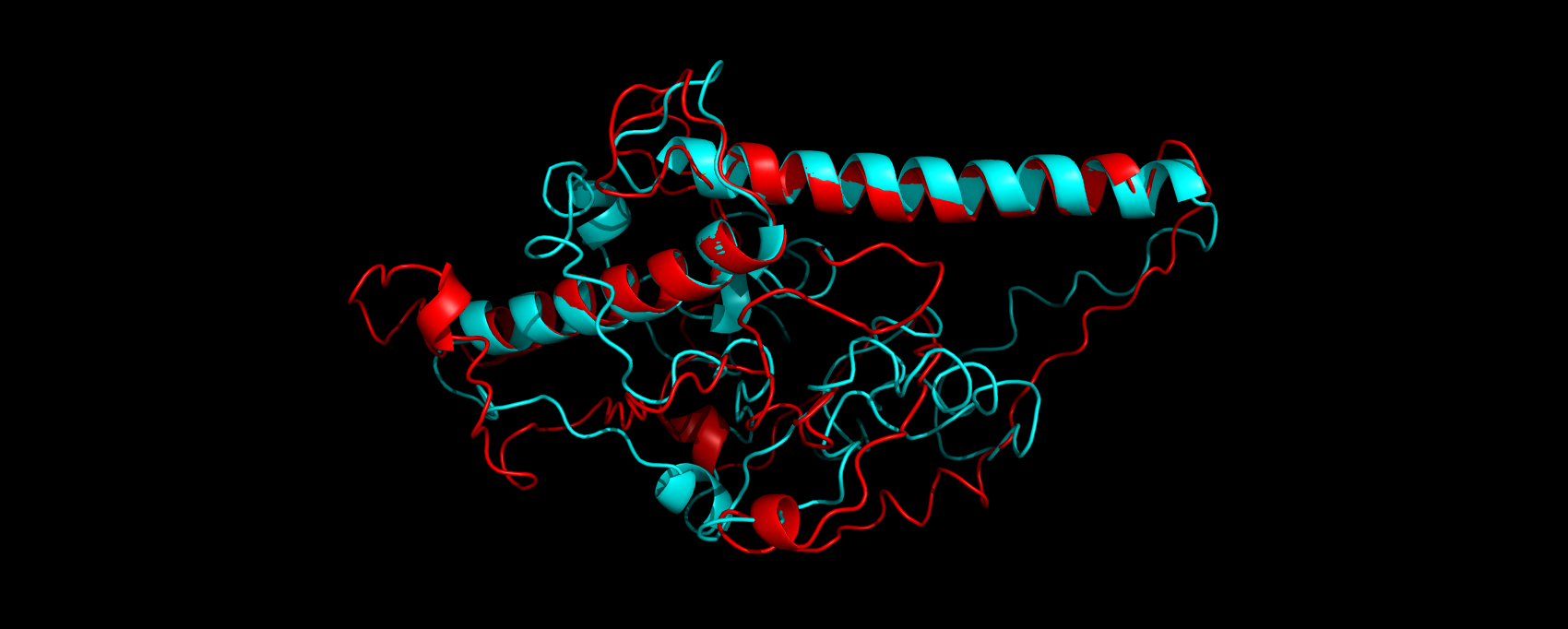


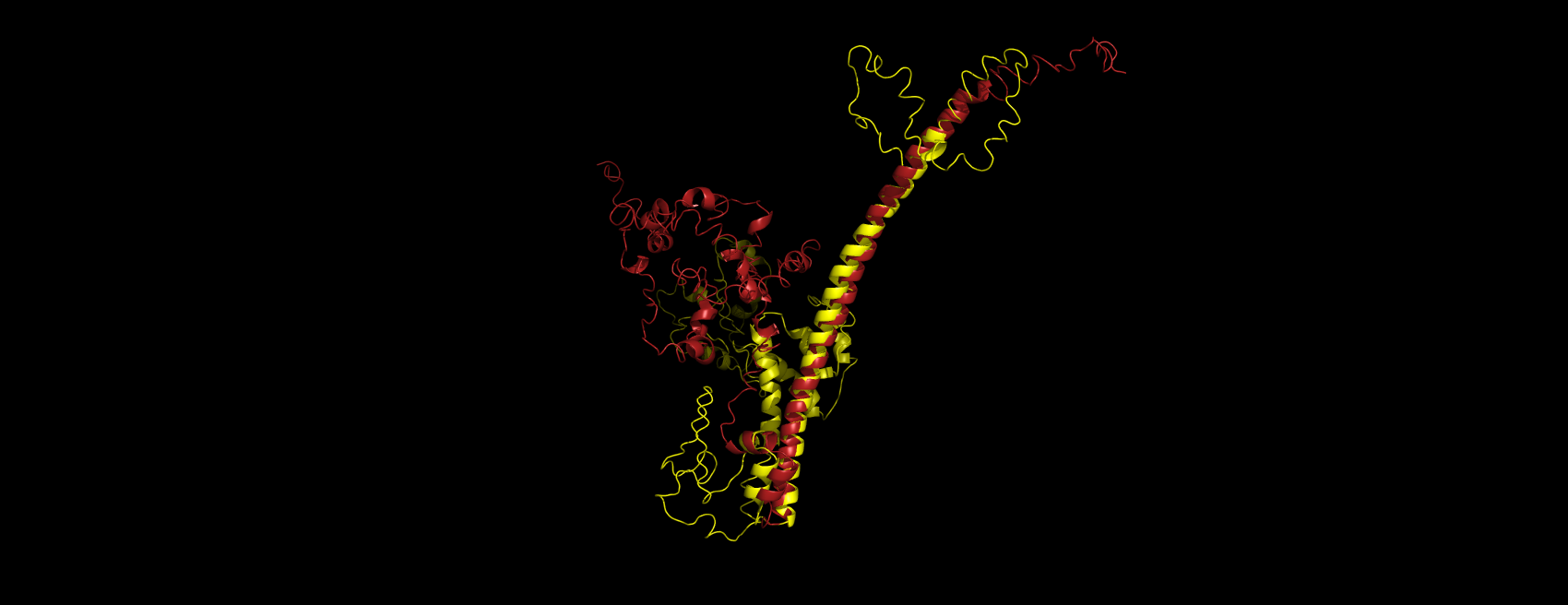
I-tasser model for Ngn3 aligned with Swiss model for Ngn3 **RMSD: 0.713**

**Alignment of I-TASSER Structures.**



Alignment of PDX1 (with residues (magenta) and without residues (grey))  
**RMSD: 0.879 (for active domains only)**

  
Alignment of Ngn3 (without residues (red) and with residues (cyan))  
**RMSD: 0.774 (for active domains only)**

****

Alignment of MafA(without residues (firebrick) and with residues (yellow))  
**RMSD: 1.501 (for active domains only)**

**Conclusion:**

It is quite evident from the results obtained from the structural models and the alignments that the active domains of our TFs remained unchanged with the addition of linker residues to form the polyprotein and subsequent cleavage by the protease. This goes along with our expectation that the small 8 linker residues added on either end would not affect the active domain. The same is corroborated by the low RMSD values for active domain alignment and the Ramachandran plots. The model, therefore, shows that our TFs don’t lose their functionality in the process of transferring them in the form of a polyprotein. Based on the above predictions we have no reason to suspect that functional properties of the fusion proteins will be inhibited by steric hindrance or structural inflexibility.

(Although this is a valid computational prediction, we understand that it is always a better practice to express the proteins and check for their activity experimentally which we could not do this year.)

**References:**

* http://2009.igem.org/Team:Warsaw/Modelling/Structural
* Williams et al. (2018) MolProbity: More and better reference data for improved all-atom structure validation. Protein Science 27: 293-315
* Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K (2009) Proteins 77 Suppl 9, 114-122 PMID19768677.
* A Roy, A Kucukural, Y Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5: 725-738 (2010)
* Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1), W296-W303 (2018)
* Psipred Server :Buchan DWA, Jones DT (2019). The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Research <https://doi.org/10.1093/nar/gkz297>
* The PSIPRED secondary structure prediction method : Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.