UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

In silico modeling of protein-protein interactions of human *Krüppel-like zinc finger 10 gene* product with transcriptional inhibitors SnoN and Ski: analysis of gene polymorphisms' effects as the potential reason for cardiac hypertrophic myopathies

Author:
Sophia KRAVCHUK

Supervisor:
Olexandr KORCHYNSKY

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Sophia KRAVCHUK, declare that this thesis entitled, "In silico modeling of protein-protein interactions of human *Krüppel-like zinc finger 10 gene* product with transcriptional inhibitors SnoN and Ski: analysis of gene polymorphisms' effects as the potential reason for cardiac hypertrophic myopathies" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"If we knew what it was we were doing, it would not be called research, would it?"

- Albert Einstein

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

In silico modeling of protein-protein interactions of human *Krüppel-like zinc finger 10 gene* product with transcriptional inhibitors SnoN and Ski: analysis of gene polymorphisms' effects as the potential reason for cardiac hypertrophic myopathies

by Sophia KRAVCHUK

Abstract

In this thesis, it will be explored the structure of the human *Krüppel-like zinc finger 10 gene* and protein, its single nucleotide polymorphisms, and how they affect the protein. It will be shown the interactions between Krüppel-like zinc finger 10 with Ski and with SnoN proteins. The analysis of single nucleotide polymorphisms' potential influence on the repressive function of the protein complex and its connection to hypertrophic cardiomyopathy will be discussed.

Acknowledgements

First of all, I want to thank the Armed Forces of Ukraine for giving me the opportunity to write this thesis, for every single day when I wake up because of the sound of my alarm and not of explosions, for being able to see my family and friends, to walk in the park or even drink coffee. There will never be enough to repay for your deeds and sacrifice. Thank you from the bottom of my heart!

I am also very grateful to my supervisor Olexandr Korchynsky for helping to choose the theme of the research, for guidance throughout the work, for being patient enough to explain to me every required detail, and for providing directions to learn and giving valuable advice.

Last but not least, I would like to thank my friends Julia, Sofia, and Olena for providing unconditional support, for helping to find the motivation to continue my work or even to start over, and for all the discussions and meetups.

Contents

List of Figures

List of Tables

List of Abbreviations

KLF	**K**rüppel like **f**actor
SKI	**S**loan-**K**ettering **I**nstitute
SKIL	**S**loan-**K**ettering **I**nstitute **L**ike
pdb	**P**rotein **D**ata **B**ank
SNP	**S**ingle- **N**ucleotide **P**olymorphism
HCM	**H**ypertrophic **c**ardio**m**yopathy
PTTG1	**P**ituitary **T**umor-**T**ransforming **G**ene 1
TIEG1	**T**ransforming growth factor β-**i**nducible **e**arly **g**ene 1
WT	**w**ild- **t**ype
TGFβ	**T**ransforming **g**rowth **f**actor β
TF	**T**ranscription **f**actor
bps	**b**ase **p**air**s**
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
PRODIGY	**PRO**tein bin**DI**ng ener**GY** prediction
IC	**I**nterfacial **C**ontact
Å	**Å**ngströms ($10-10$ m)
I-TASSER	**I**terative **T**hreading **ASSE**mbly **R**efinement
HADDOCK	**H**igh **A**mbiguity **D**riven protein-protein **DOCK**ing

List of Symbols

R	ideal gas constant (in kcal K-1mol-1)	$\mathrm{kcal\,K^{-1}\,mol^{-1}}$
T	temperature	K
ΔG	predicted free Gibbs energy	$\mathrm{kcal\,mol^{-1}}$
K_d	dissociation constant	mol

Dedicated to my parents...

Chapter 1

Introduction

1.1 Problem

Hypertrophic cardiomyopathy (HCM) is a disease in which the heart muscle becomes thick (hypertrophied). The thickened part of a heart makes the ventricles more minor and unable to pass through the right amount of blood, so the heart has to make more efforts to maintain the proper pressure and oxygen saturation.

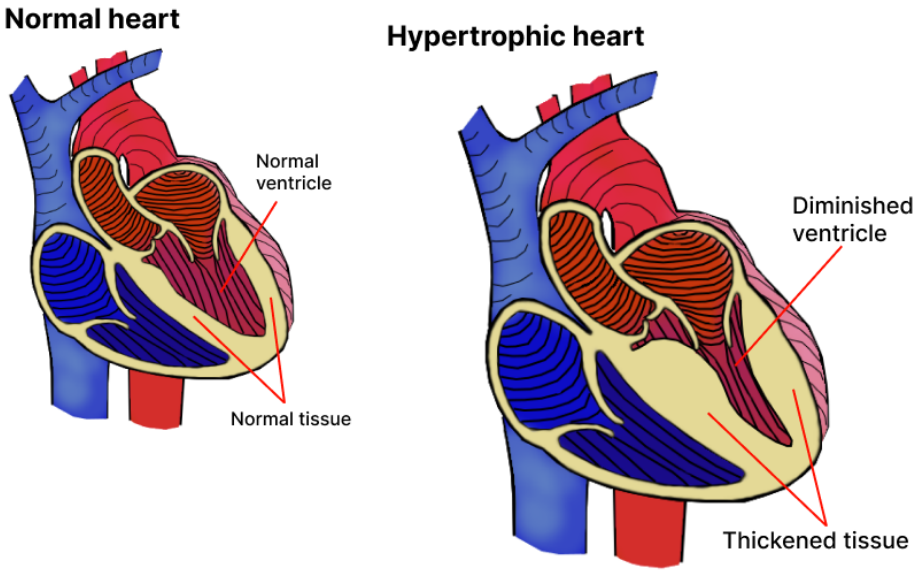


FIGURE 1.1: Hearts comparison (normal and HCM)

People who have HCM suffer chest pains, breathing difficulties, dizziness, or faintings, and the heart is racing and beating abnormally. So this disease does affect the quality of life; HCM patients should maintain a proper diet, restrict their physical activity, and avoid drinking alcohol or smoking. However, even sticking to all the restrictions and rules cannot guarantee normal life conditions. HCM has no cure and can cause sudden cardiac arrest and heart failure.

Hypertrophic cardiomyopathy affects an estimated 600,000 to 1.5 million Americans, or 1 in 500 people. It's more common than multiple sclerosis, affecting 1 in 700 people. [6].

HCM is most often caused by abnormal genes and is inherited as an autosomal dominant trait. It results from mutations in one of several genes. To date, nine genes

have been identified as HCM causing. Up to 60% of patients who suffer HCM have a pathogenic mutation on genetic testing.[15]

Gene	**Protein**	**Frequency (%)**
Cardiac myosin-binding protein C	MYBPC3	30–40
β cardiac myosin heavy chain	MYH7	20-30
Cardiac troponin T	TNNT2	5-10
Cardiac troponin I	TNNI3	4-8
Regulatory myosin light chain	MYL2	2-4
Essential myosin light chain	MYL3	1-2
α tropomyosin	TPM1	<1
α cardiac actin	ACTC1	<1
Muscle LIM protein	CSRP3	<1

TABLE 1.1: Pathogenic mutations of HCM

Mutations in *Krüppel-like factor 10* have also been identified in human HCM patients, but roles for *KLFs* in heart failure and vascular disease remain under investigation.[5] Recent studies have hypothesized that *TIEG1* could be a novel candidate gene in the pathogenesis of genotype-negative HCM in humans. [4]

1.2 Motivation

In the last decade, the revolution in computer technology has made bioinformatics a significant part of microbiology. Now it is merely impossible to imagine a microbiological task without bioinformatical methods and tools, protein models, or data from an extensive scale sequencing of genomes. The combination of bioinformatics and microbiology nowadays potentiates a new leap in medical studies.

I have always been interested in medicine and even dreamed of becoming a doctor. That is why I would be eager to be able to use the knowledge and skills that I have gained in the university to help people understand the cause of HMC and maybe help some scientists find a cure or prevent the progression of the disease.

1.3 Goal

The main goal of this work is to determine whether the SNPs in the KLF10 protein are affecting its complexes with Ski and SnoN proteins and inhibiting their repression function. To achieve it would be required to:

- model structures of KLF10, Ski and SnoN proteins
- predict protein-protein interactions of KLF10 with Ski
- predict protein-protein interactions of KLF10 with SnoN

1.4 Thesis structure

- **Chapter 2. Background**

This chapter contains basic information about the TGFβ signaling pathway, KLF10, Ski, SnoN genes, and proteins. It also describes the function of the KLF10-Ski-SnoN complex.

- **Chapter 3. Related works**

This chapter describes two pieces of research based on which this thesis was conceived.

- **Chapter 4. Approach**

This chapter contains a description of the work done during this research, the main steps, and inferences.

- **Chapter 5. Tools**

This chapter describes all the tools and algorithms required for this thesis.

- **Chapter 6. Conclusions and future work**

In this final chapter, the results of the research are reviewed. In the end, possible options for further studies are considered.

Chapter 2

Background

Further chapters and statements can not be introduced without an understanding of some basic processes and instances. To better comprehend this research, it should be determined what the TGFβ signaling pathway is and, how it affects the cardiac tissue, what connection is between it and KLF10, Ski, SnoN proteins.

2.1 Biological functions of TGFβ family cytokines

"The transforming growth factor-beta (TGFβ) signaling pathway is involved in many cellular processes in the human organism, including cell growth, cell differentiation, cell migration, apoptosis, cellular homeostasis, and other cellular functions."[27] TGFβ signaling plays context-dependent roles; for most normal cells, it inhibits cell growth and works as a tumor suppressor, while in the late stages of cancer, TGFβ signaling promotes invasion and metastasis.

However, TGFβ overexpression in heart tissues is associated with fibrosis and hypertrophy.

2.1.1 TGFβ signal transduction: Molecular mechanism

Activation of the pathways is primarily regulated by the release of active TGFβ from the latent complex stored in significant amounts in most tissues. Activation of a small fraction of TGFβ is capable of generating the maximal cellular response.

The membrane of a cell has type I and type II receptors. To start the signaling, they bind with a TGFβ ligand (TGFβ, Activin, Nodal, Growth, and differentiation factors (GDFs), Anti-mullerian hormone (AMH), Bone morphogenetic proteins (BMPs)). Then when the pathway is initiated, the complex binds with the intracellular signaling regulatory molecule; for this pathway, it is one of the R-SMADs (SMAD1, SMAD2, SMAD3, SMAD5, SMAD8/9). Also, it needs a molecule called Co-SMAD (SMAD4), which helps drive the signaling pathway once R-SMADs are active.

The phosphorylated R-SMAD/Co-SMAD complex enters the nucleus, binds transcription promoters/cofactors, and causes the transcription of DNA.[2, 16]

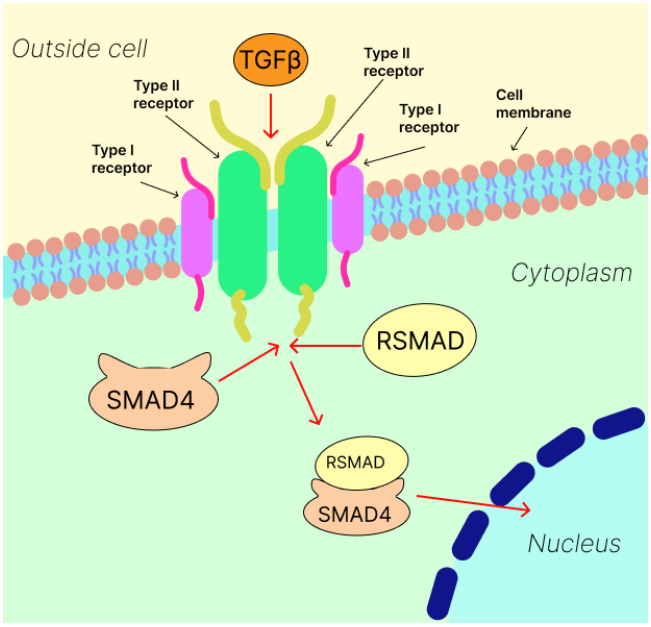


FIGURE 2.1: Schematic representation of transforming growth factor (TGF)-β signaling.

2.1.2 Inhibition of TGFβ-SMAD signals

In addition to R-SMADs and Co-SMAD, there is a third SMAD protein family, namely the inhibitory SMADs (I-SMADS): SMAD6 and SMAD7. They play a crucial role in regulating TGF beta signaling and are involved in negative feedback. [27]

I-SMADs can inhibit the TGFβ signaling pathway through various mechanisms. First, SMAD7 can form a stable complex with Type I receptors, leading to inhibition of R-SMAD phosphorylation and the formation of the R-SMADs-Co-SMAD complex. SMAD7 can also use other ligases, translocate them into the cytoplasm and cause the degradation of the receptor. Besides, SMAD7 can serve as an adaptor protein for some ligates to promote the degradation of type I receptors and R-SMADs and SMAD4. [28]

2.2 KLF10-Ski-(SnoN) complex as a transcriptional repressor

A repressor is a protein or a complex of proteins that inhibits the expression of one or more genes. It binds to the gene's promoter region, which prevents the production of messenger RNA (mRNA). A DNA binding protein may play a role of either an activator or repressor of gene expression, depending on the cellular context. This blocking or reducing of expression is called repression. Repressor proteins are essential for the regulation of gene expression in cells. [17, 25]

In the usual way, without repression, the expression of a gene goes as follows: a transcription factor (TF), a protein or complex of proteins, binds to specific enhancer or promoter regions of DNA sequence. Then TFs use a variety of mechanisms for the regulation of gene expression. (Figure 2.2)

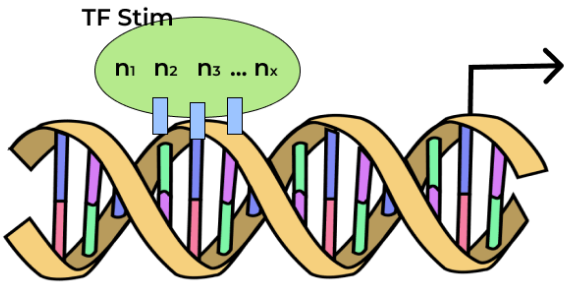


FIGURE 2.2: Molecular mechanism for gene expression stimulation by transcription factors

The expression of genes can be regulated by repressing a TF with a specific inhibitor. Some inhibitors can bind to a TF and stop the transcription. (Figure 2.3 A)

However, some inhibitors cannot bind to transcription factor stimulators by themselves and form a complex with other proteins. Krüppel-like factor 10 protein plays a role of a 'bridge' between TF stimulator and inhibitors. (Figure 2.3 B)

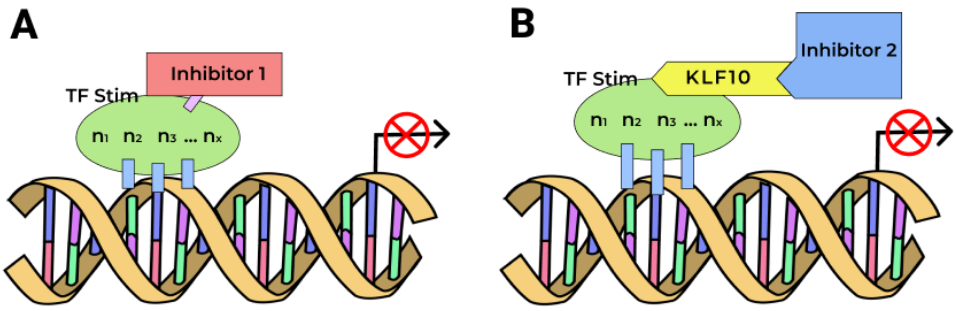


FIGURE 2.3: Molecular mechanisms of gene expression blocking
(A) A general mechanism for blocking of gene expression by transcriptional repressors
(B) Blocking of gene expression by KLF10-inhibitor complex

A complex of KLF10 and Ski, SnoN proteins play a repressor role for the TGFβ signaling pathway. They affect some SMAD family proteins and withdraw the cells out of the tumor-suppressing control. (Figure 2.4)

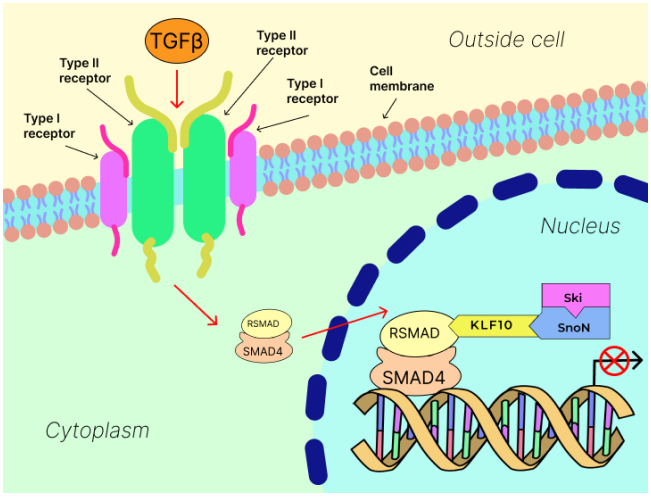


FIGURE 2.4: TGFβ/SMAD signal transduction repressing by a KLF10-Ski-SnoN complex

2.3 Krüppel-like factor 10

Gene

Krüppel-like factor 10 (KLF10/TIEG1) is a protein-coding gene that is located on Chromosome 8 (102,648,784-102,655,725 reverse strand). It has four exons, is annotated with 22 domains, its transcript length is 2,915 bps, translation length is 480 residues. [9]

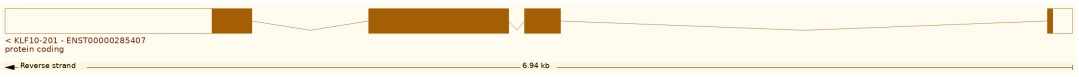


FIGURE 2.5: Krüppel-like factor 10 [9]

KLF10 gene has several essential domains:

- Repression domain
- SIAH-binding domain
- SH3 binding domain (proline-rich region)
- Zinc finger domain (DNA binding domain)

[4]

The SIAH-binding domain is essential for proteasomal degradation of *TIEG1*, three unique repression domains, R1, R2, and R3 are crucial for its repressor activities, and the SH3 binding domain is for protein-protein interactions.

Protein

The human KLF10 protein has a 480 amino acid sequence. It also has all the domains that are in the gene. In this research, the most crucial is the repression domains. KLF10 works as an effector protein in TGFβ mediated cell growth control and differentiation. KLF10 plays a vital role in regulating many genes essential for

various cellular functions, including cell proliferation, differentiation, and apoptosis. KLF10 is involved in several different types of gene expression in multiple cell types and serves as a target gene for many signaling pathways.

FIGURE 2.6: Krüppel-like factor 10 Protein *(purple region - repression domains)*

2.4 Ski & SnoN

2.4.1 Sloan-Kettering Institute protein Structure (Ski)

Gene

The *Ski gene* in the human organism is located on Chromosome 1 (2, 228, 319 - 2, 310, 213 forward strand). Its length is 81 894 nucleotides, has 7 exons, is annotated with 29 domains, transcript length is 6,083 bps, and translation length is 728 residues. [10] This gene encodes the nuclear protooncogene protein. It functions as a repressor of TGF-beta signaling and may play a role in neural tube development and muscle differentiation.[3]

FIGURE 2.7: Sloan-Kettering Institute protooncogene [10]

Protein

The human Ski protein has a 728 amino acid sequence and three main domains. It is expressed both inside and outside of the nucleus.[26] The N-terminal Ski-dachshund homology domain (Ski-DHD) is located in residues 91–192. This domain has lost the ability to bind directly to DNA but interacts with R-SMADs and other transcriptional regulators. The Ski protein also has a SAND-like domain (residues 219–312) responsible for SMAD4 binding. Also, the transformation domain is localized in the first 304 residues and is responsible for cell transformation and transcriptional gene repression.[19]

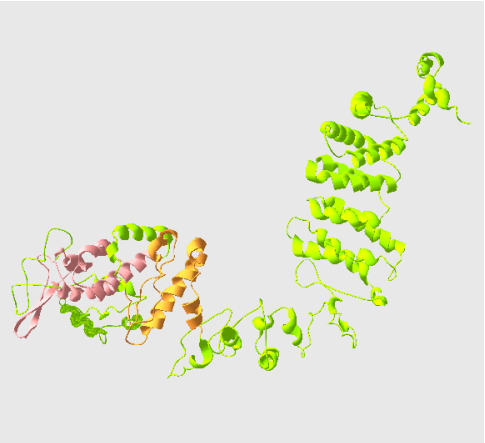


FIGURE 2.8: Sloan-Kettering Institute Protein *(pink region - Ski-DHD, orange region - SAND-like domain)*

2.4.2 Sloan-Kettering Institute like protein Structure (Skil or SnoN)

Gene

SKI like protooncogene (SNO, SnoA, SnoN) is located on the Chromosome 3 (170,357,715 - 170,396,835 forward strand). its length is 39120 nucleotides, it has seven exons, is annotated with 34 domains, transcript length is 7,155 bps, and translation length is 684 residues. [8]

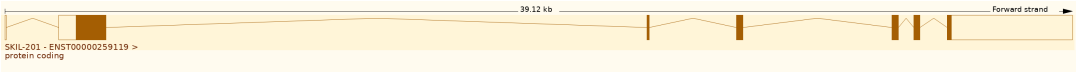


FIGURE 2.9: Sloan-Kettering Institute Like protooncogene [8]

Protein

SnoN is a 684 amino acid sequence protein with four conserved domains: minimal transformation domain, SAND-like domain, DHD domain (residues 137 to 243), and SMAD-binding domain (residues 85 to 88 and 260 to 357). [13, 19]

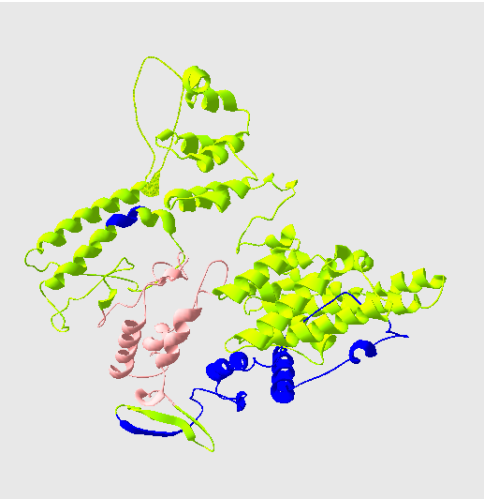


FIGURE 2.10: Sloan-Kettering Institute Protein *(pink region - Ski-DHD, blue region - SMAD-binding domain)*

Chapter 3

Related works

This research is mainly based on two other works. The problem of this work was conceived after reading and analyzing the article by Mayo Clinic, "TGFβ-Inducible Early Gene-1 (TIEG1) Mutations in Hypertrophic Cardiomyopathy" written by J. Martijn Bos, Malayannan Subramaniam, John R. Hawse, and others. Furthermore, the article "Transcriptional cofactors Ski and SnoN are major regulators of the TGFβ/SMAD signaling pathway in health and disease", written by Angeles C. Tecalco-Cruz, Diana G. Ríos-López, and others. So this research is based on their work and hypothesis.

3.1 Research of Mayo Clinic

"TGFβ-Inducible Early Gene-1 (TIEG1) Mutations in Hypertrophic Cardiomyopathy" [4]

The scientists of the different departments of the Mayo Clinic observed the connection between missense variants of KLF10 and HCM in mice organisms. Rajamannan reported that male *TIEG1* mice exhibited late-onset HCM compared to wild-type mice. So they have hypothesized that *TIEG1* could be a novel candidate gene in the pathogenesis of genotype negative HCM in humans, possibly through a loss of its repression on PTTG1 expression.

In that study, they analyzed 923 unrelated patients with HCM and all their translated regions of *TIEG1*. In these observations were discovered six novel missense variants of KLF10 (A12T, M27T, E137K, A204T, T216A, and S225N) that were supposed to be the reason for the disease.

Also, in that research, the scientists claimed that only one SNP (M27T) is in a repression domain on the protein and the other five (A12T, E137K, A204T, T216A, and S225N) are not, which means they cannot play a role in disabling of a repression function of KLF10. (Figure 3.1)

They also explained that KLF10 potentiates the TGFbeta signaling pathway by repressing the inhibitory SMAD7 gene.

WT TIEG1 repressed SMAD7-promoter activity by 70%. In contrast, two putative TIEG1-HCM variants—A12T and S225N— altered SMAD7-promoter expression with S225N-TIEG1 showing significantly increased activation of SMAD7 promoter activity as compared to wild-type.

They found that only one of the variants (S225N-TIEG1) significantly altered TIEG1's regulation of the SMAD7 promoter compared to WT, suggesting a minor role for the TIEG1-SMAD7 inhibition in this disease.

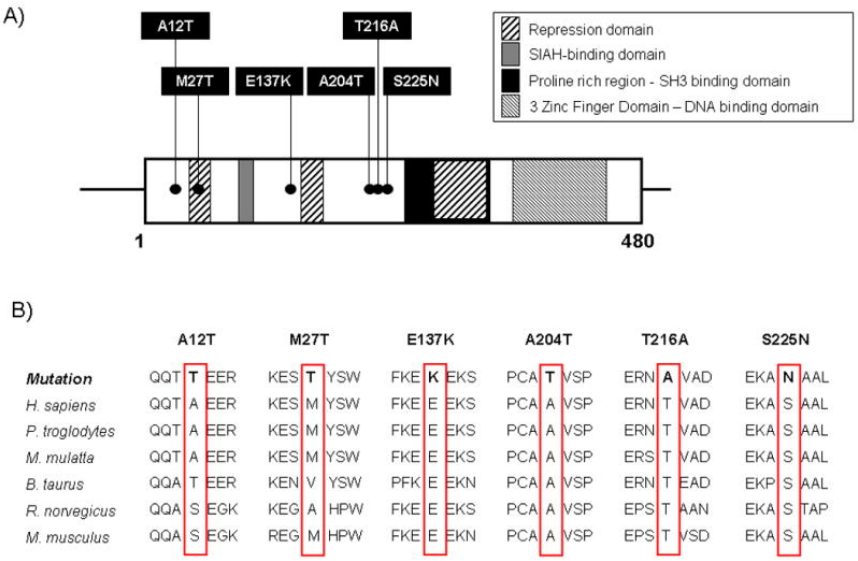


FIGURE 3.1: Topology of TIEG1-protein and sequence conservation
(A) *Schematic representation of the 480-amino acid containing TIEG1 protein indicating the identified missense variants implicated with HCM*
(B) *Conservation across species of novel, HCM-associated TIEG1-variants.*(reproduced from: Bos et al, 2012 [4]).

Also, that study showed that 5 of our 6 TIEG1 variants resulted in a significant increase in PTTG1 promoter activity relative to WT TIEG1. Moreover, they hypothesized that PTTG1 protein expression might be a pathway biomarker in HCM.

3.2 Research of Instituto de Investigaciones Biomédicas at Universidad Nacional Autónoma de México

"Transcriptional cofactors Ski and SnoN are major regulators of the TGFβ/SMAD signaling pathway in health and disease" [19]

This research explains the structure of Ski and SnoN genes and proteins that is already shown in Chapter 2 (2.4). Also, the researchers have written about the main functions of Ski gene:

- activity throughout embryonic development (formation of the central nervous system (CNS), skeletal muscle, and limbs)
- role in normal myelination of axons
- promotes muscular development by enhancing the expression of muscle regulatory factors
- decrease body fat mass
- and others

Furthermore, SnonN functions are:

- modulate neuronal branching and positioning during axonal growth and regeneration

- it is involved in muscle cell differentiation by regulating the formation of muscle fibers
- and others

Moreover, forming a Ski-SnoN complex, these proteins can have many other functions. Notably, the Ski and SnoN proteins are downregulated in fibrosis and cancer metastasis but upregulated in tumor growth.

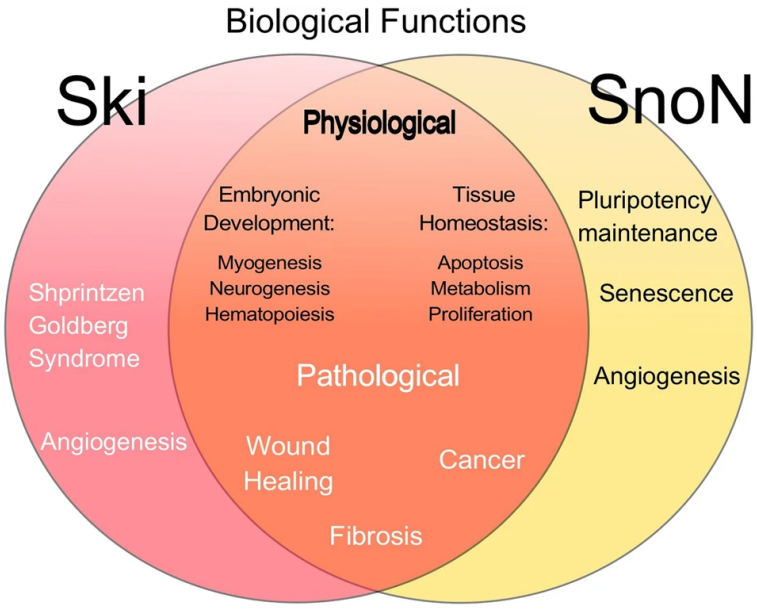


FIGURE 3.2: Ski and SnoN functions in health and disease. The specific or common biological functions of Ski and SnoN are linked to some physiological (black text) or pathological processes (white text) (reproduced from: Tecalco-Cruz et al, 2018[19])

That work also described the effect of Ski and SnoN proteins on the TGFβ signaling pathway. Ski and SnoN expression is enhanced during liver regeneration, probably to neutralize TGFβ/SMAD antiproliferative actions; Ski and SnoN cofactors may also mediate antifibrotic responses by blocking TGFβ signals.

In conclusion, it is claimed that the Ski and SnoN proteins play critical roles in health and disease by controlling the outcome of TGFβ and other signaling pathways. Therefore, balanced Ski and SnoN expression levels must be controlled to normalize the TGFβ signaling outcome in diverse pathologies.[19]

Chapter 4

Approach

For this research, the work was divided into main parts: analysis of the SNPs and repression domains of the KLF10 protein, protein modeling of KLF10 WT and all of the six polymorphisms, modeling of Ski and SnoN proteins, running the molecular docking with KLF10 WT and four polymorphisms with of Ski and SnoN proteins, calculating the free Gibbs energy.

Below is described in detail the approach of all the work parts.

4.1 Detection of the repression domain of *KLF10* on its protein

The work on this research started with the analysis of the article "TGFβ-Inducible Early Gene-1 (TIEG1) Mutations in Hypertrophic Cardiomyopathy" (3.1) and the structure of the KLF10 protein. As was explained previously, 6 SNPs of KLF10 are supposed to affect repressor functions of the KLF10: A12T, M27T, E137K, A204T, T216A, and S225N. In the article (3.1), it was shown that five of these SNPs are not in the repression domain, so first of all, it was necessary to find those repression regions on the protein.

4.1.1 Implementation

In the *KLF10 gene* annotation that is shown that there are three repression regions on it (Figure 4.1):

- R1 - Chromosome 8: 102,650,001 - 102,650,400
- R2 - Chromosome 8: 102,651,201 - 102,652,200
- R3 - Chromosome 8: 102,653,801 - 102,654,400

The DNA sequence should be translated into the amino acid sequence to find those repressing regions on the protein. On the site of The National Library of Medicine of The National Center for Biotechnology Information it is possible to get the DNA sequence by the coordinates and then run the basic local alignment search tool x (BLASTx) with it.

BLAST is an algorithm that compares the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. It takes as an input two sequences (in FASTA or Genbank format), the first sequence is a gene or a protein, and the second is a database to search in.

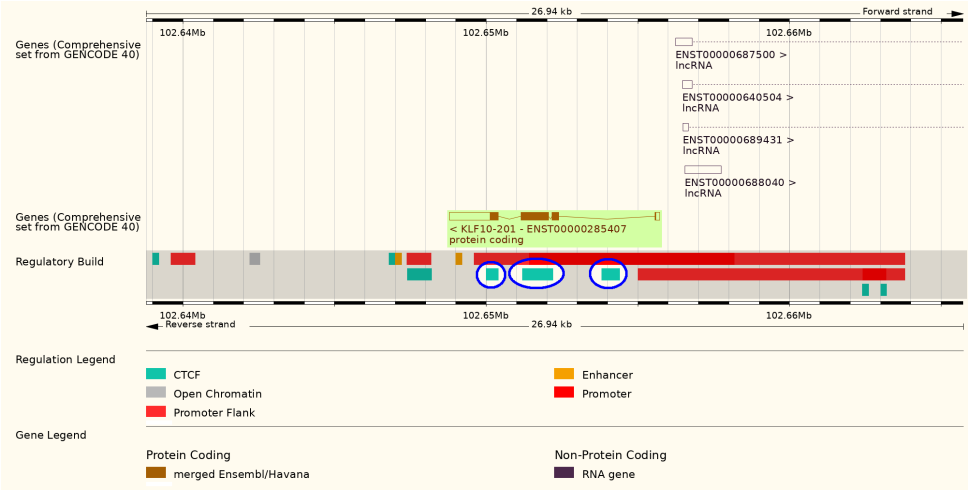


FIGURE 4.1: Human Krüppel-like factor 10 annotation[7]

For example, if one specifies a gene ID, or sequence in a BLAST search window and defines a human database, BLAST will identify sequences within the human genome with a significant homology to the given gene based on the similarity of the sequence. There are different types of BLAST: BLASTn, BLASTp, BLASTx, BLASTpgp, and others.

To find repression domains, the BLASTx was used. It takes DNA sequence as an input, translates it into protein, finds the resemblance with a protein-coding gene in the human genome, and returns the amino acid sequence of a specific protein.

After analysis of BLASTx results, it became clear that two of three gene repression regions are protein-coding and appear in the KLF10 protein. The first region, R1, covers amino acids 79-377, and the second, R2, - 395-472.

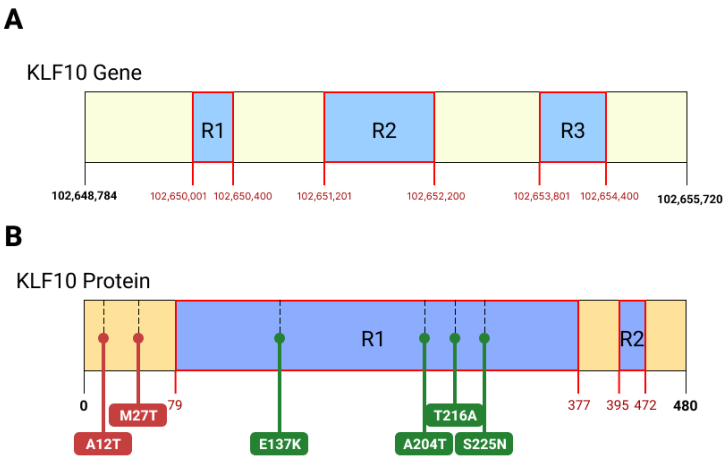


FIGURE 4.2: Repression domain on KLF10
(A) Location of repression regions on *KLF10 Gene* (A Chromosome sequence with introns)
(B) Location of repression regions on KLF10 Protein

The third repression region is not found in KLF10 protein because it is located in the *KLF10 gene's* intron area.

4.1.2 Inference

So, after this finding, it is clear that SNPs A12T and M27T of KLF10 are not in the repression domain (Figure 4.2 B) and cannot affect its inhibiting functions. From now on, the subject of the research will be the other four SNPs (E137K, A204T, T216A, S225N).

In WT KLF10, the 137th residue is E (GLU, Glutamic Acid), 204th - A (ALA, Alanine), 216th - T (THR, Threonine), and 225th - S (SER, Serine).

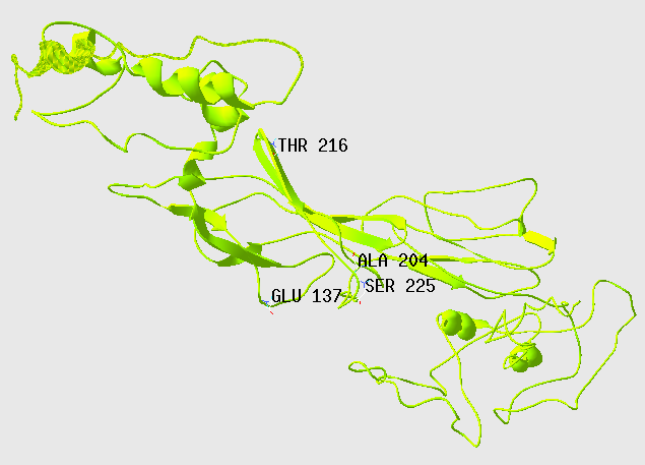


FIGURE 4.3: KLF10 WT structure with labeled studied residues

In the polymorphisms that are studied in this research, these amino acids are changed respectfully: 137th - K (LYS, Lysine), 204th - T (THR, Threonine), 216th - A (ALA, Alanine), and 225th - N (ASN, Asparagine). These SNPs are changing the protein, its roles, and structure and can be the reason for KLF10 losing its repressing functions.

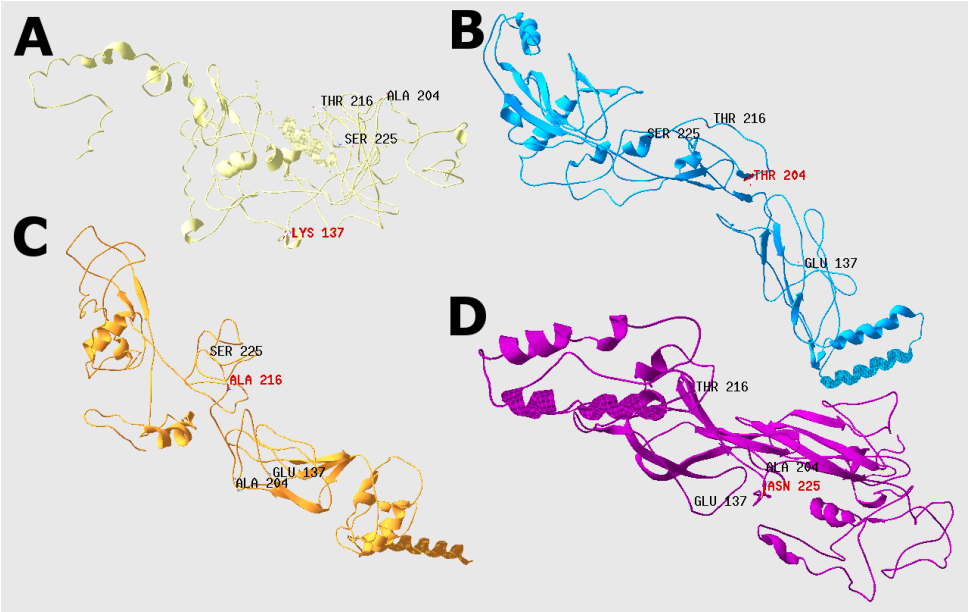


FIGURE 4.4: KLF10 SNPs' structures with labeled studied residues
(A) *E137K* **(B)** *A204T* **(C)** *T216A* **(D)** *S225K*

4.2 Running molecular docking on KLF10s polymorphisms and Ski protein

Molecular modeling is a method that predicts the preferred orientation of one molecule to a second in a stable complex. Molecular docking generates different possible three-dimensional structures of a complex that are ranked using a scoring function in the software.

The structures are ranked depending on Gibbs free energy, and the best model is detected and picked.

More about this algorithm is described in Chapter 5.4.

For this research it was done five molecular dockings:

- KLF10 WT & Ski (Figure 4.5)
- KLF10 E137K & Ski (Figure 4.8a)
- KLF10 A204T & Ski (Figure 4.8b)
- KLF10 T216A & Ski (Figure 4.8c)
- KLF10 S225N & Ski (Figure 4.8d)

As a result, it was received five models of possible complexes.

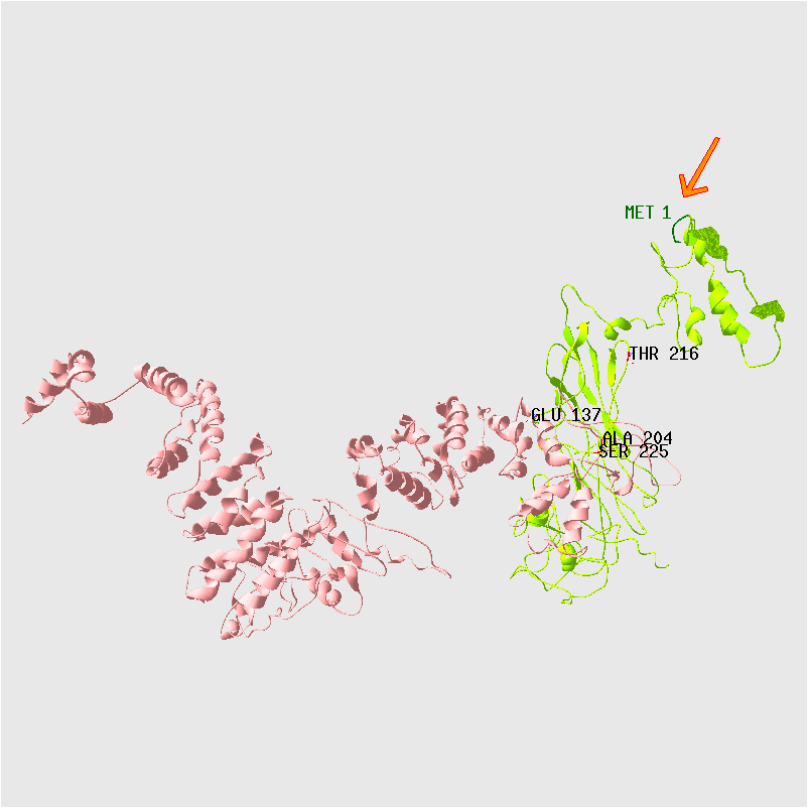
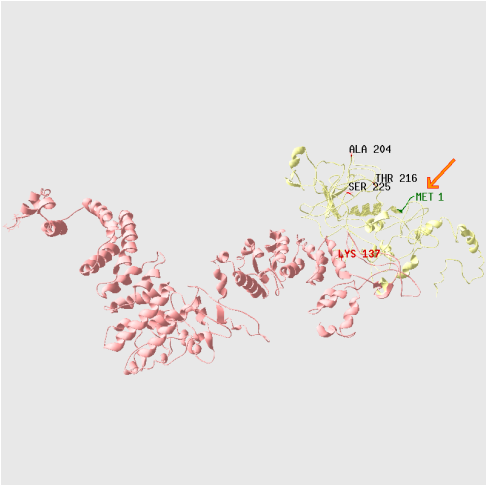
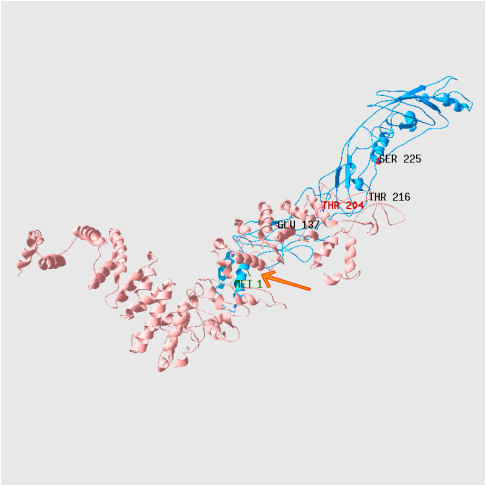


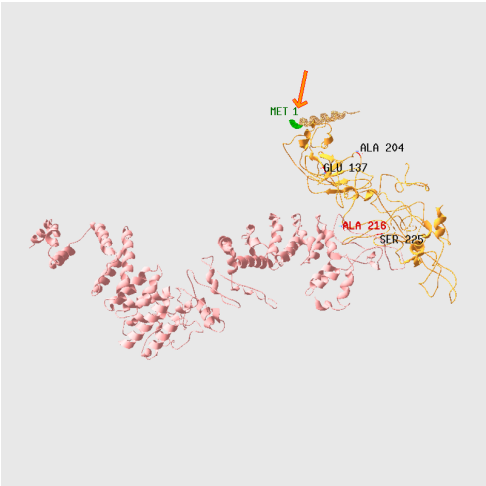
FIGURE 4.5: KLF10 WT-Ski complex structure with labeled studied residues
(Orange arrow indicates the position of the first amino acid residue of KLF10 polypeptide,
KLF10 is colored in green, Ski - red)



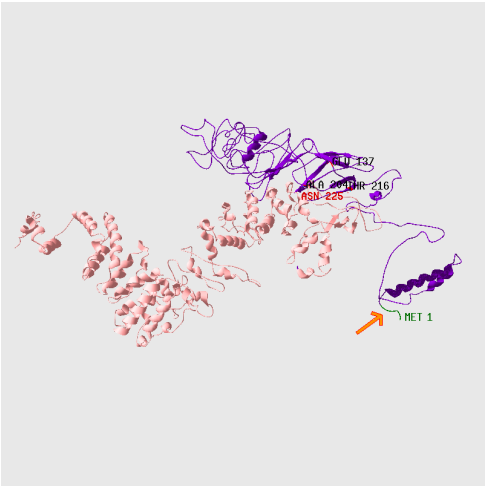
(A) *KLF10 E137K Ski complex structure*



(B) *KLF10 A204T Ski complex structure*



(C) *KLF10 T216A Ski complex structure*



(D) *KLF10 S225N Ski complex structure*

FIGURE 4.6: KLF10 SNPs and Ski complex structures with labeled studied residues
(Orange arrow indicates the position of the first amino acid residue of KLF10 polypeptide,
Ski is colored in red, KLF10 E137K - yellow, KLF10 A204T - blue, KLF10 T216A - orange, KLF10 S225N - purple)

It is now clear that SNPs studied in this research affect the interaction of KLF10 and Ski.

KLF10 T216A Ski complex and KLF10 S225N Ski complex differ. The Ski is fixed on the images, but KLF10 is pivoted at 180 degrees.

4.3 Running molecular docking on KLF10s polymorphisms and SnoN protein

After the analysis of Ski protein, asimilar molecular dockings of KLF10's snip variants with SnoN protein were run:

- KLF10 WT & SnoN (Figure 4.7)
- KLF10 E137K & SnoN (Figure 4.8a)

- KLF10 A204T & SnoN (Figure 4.8b)
- KLF10 T216A & SnoN (Figure 4.8c)
- KLF10 S225N & SnoN (Figure 4.8d)

As a result, it was received five models of possible complexes.

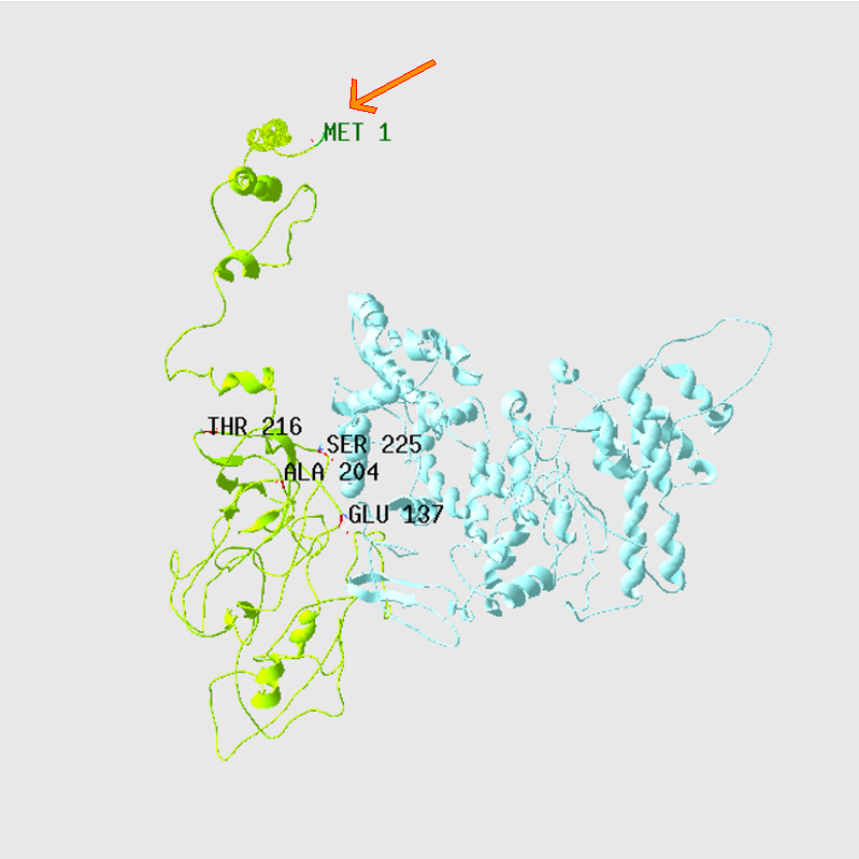
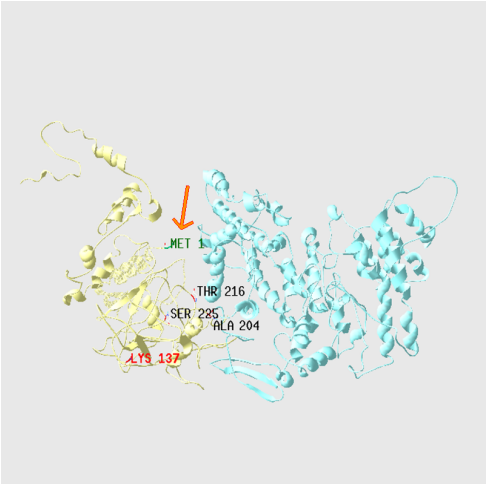


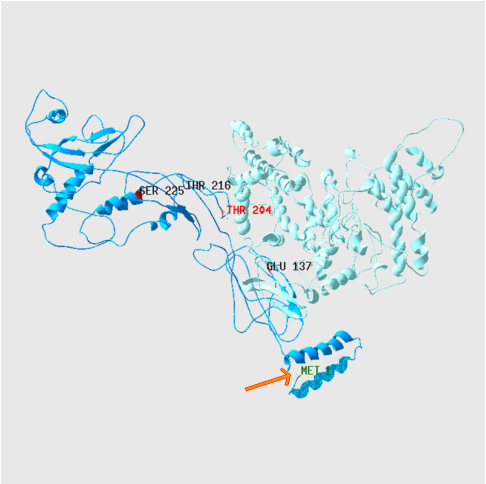
FIGURE 4.7: KLF10 WT SnoN complex structure with labeled studied residues
(Orange arrow indicates the position of the first amino acid residue of KLF10 polypeptide,
KLF10 WT is colored in green, SnoN - cyan)

With SnoN, snips of KLF10 also have a particular influence on the interaction interface.

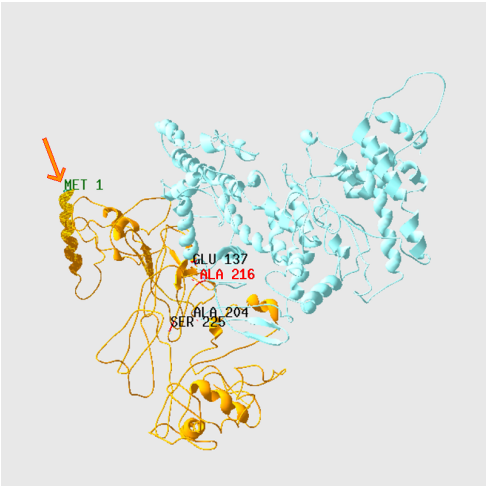
KLF10 A204T with SnoN and KLF10 S225N with SnoN complexes have the slightest difference.



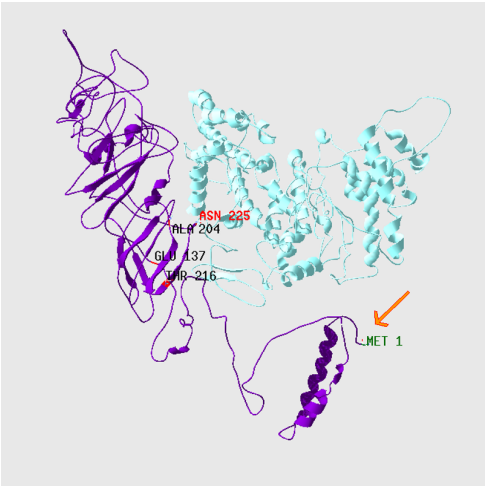
(A) *KLF10 E137K SnoN complex structure*



(B) *KLF10 A204T SnoN complex structure*



(C) *KLF10 T216A SnoN complex structure*



(D) *KLF10 S225N SnoN complex structure*

FIGURE 4.8: KLF10 SNPs and SnoN complex structures with labeled studied residues
(Orange arrow indicates the position of the first amino acid residue of KLF10 polypeptide,
SnoN is colored in cyan, KLF10 E137K - yellow, KLF10 A204T - blue, KLF10 T216A - orange, KLF10 S225N - purple)

4.4 Determination of Gibbs free energy

4.4.1 Implementation

After having all five complexes, the analysis of differences was done. It became clear that snips within KLF10 have a particular influence on the complex.

The possible way to determine the stability of the connection is to calculate the Gibbs free energy of the complex.

So, the Gibbs free energy of modeled complexes of KLF10 WT and KLF10 SNPs with Ski and SnoN was predicted and calculated using the PRODIGY system.

The results were as follows:

Gibbs Free Energy (kcal mol-1)		
Complex	**With Ski**	**With SnoN**
KLF10 WT	-12.8	-9.5
KLF10 SNP **E137K**	-11.5	-8.3
KLF10 SNP **A204T**	-15.1	-14.3
KLF10 SNP **T216A**	-12.8	-12.2
KLF10 SNP **S225N**	-13.9	-14.5

TABLE 4.1: Calculated Gibbs Free Energy for received complexes after molecular docking

4.4.2 Inference

The calculated interaction energies show that the connection between Ski and T216A SNP of KLF10 has the same amount of Gibbs free energy as a Ski and KLF10 WT complex. That result means that SNP T216A does not affect the KLF10 and Ski interaction which means it does not affect this complex's repression functions.

Moreover, KLF10 SNP E137K and Ski complex, as well as KLF10 SNP E137K with SnoN, have a smaller absolute value of interaction energy which means that the repression function of the complex is inhibited. Therefore KLF10 SNP A204T and KLF10 SNP S225N with Ski and also SnoN, as well as KLF10 SNP T216A with SnoN, have a more considerable absolute value of energy which means that the repression function of the complex is increased.

Chapter 5

Tools

For this research, it was required to perform several bioinformatics procedures (protein structure prediction, model animation, gene translation, protein-protein interaction prediction). There are many tools for that, and this chapter describes which ones out of those were actually used.

5.1 Prediction of a protein structure

Iterative Threading ASSEmbly Refinement (I-TASSER) is a bioinformatics method for predicting the three-dimensional structure model of protein molecules from amino acid sequences. [22]

It detects structure templates from the Protein Data Bank using fold recognition (or threading).

5.1.1 I-TASSER server

I-TASSER server is an online platform that implements the I-TASSER algorithms for protein structure and function predictions. It models predictions of the three-dimensional structure of protein molecules from their amino acid sequences. [14]

I-TASSER server's method

Firstly, the server has an amino acid sequence and tries to find template proteins of similar parts from the PDB library.

In the second step, all the parts selected from the PDB library are placed into the full-length model using replica-exchange Monte Carlo simulation, with the Gibbs free energy as a score.

Then the best model is selected, and the modeling performance is done once again to remove the steric clash and refine the global topology of the cluster centroids.

After that, the model with the most considerable absolute amount of Gibbs free energy is selected, and atomic-level structure refinement is done.[30, 31, 29]

I-TASSER server's output

As a result I-TASSER server in the output returns up to five full-length atomic models.

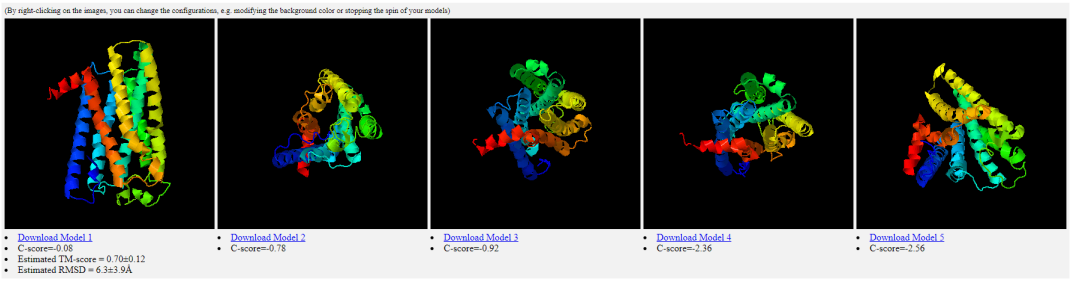


FIGURE 5.1: Example of five predicted by I-TASSER models

Usually it takes up to 1-2 days from submitting a sequence to receiving the prediction results. But it depends on the protein size and the amount of sequences are accumulated in the queue.

Some of the models that were required for this research was being predicted in 6 days.

5.2 Protein Data Bank file format

The Protein Data Bank (pdb) file format is a format that represents three-dimensional structures. The pdb format was invented in 1976 as a format of Protein Data Bank, and all the data on biological macromolecules there was saved using it. The pdb format includes description and annotation of protein, its coordinates, occupancy, and temperature factor.

It has a fixed number of columns (80 columns) based on the width of the computer punch cards that were previously used to exchange the coordinates. [23]

5.2.1 Example of reading pdb file

As a example it is the first six rows of KLF10 pdb-structure.

```
ATOM  1  N    MET  A  1  60.257  124.922  164.326  1.00  6.35
ATOM  2  CA   MET  A  1  59.878  125.222  162.949  1.00  6.35
ATOM  3  HA   MET  A  1  60.284  124.426  162.327  1.00  6.35
ATOM  4  CB   MET  A  1  58.345  125.191  162.810  1.00  6.35
ATOM  5  HB1  MET  A  1  57.981  124.213  163.127  1.00  6.35
ATOM  6  HB2  MET  A  1  58.080  125.325  161.760  1.00  6.35
```

The first ATOM line describes the alpha-N atom of the first residue of peptide chain A, which is a Methionine residue. The first three floating point numbers are its x, y and z coordinates in units of Å. The next two columns are the occupancy, and temperature factor, respectively. [23]

5.3 PDB viewers

5.3.1 Tested PDB viewers

As the three-dimensional protein models are saved in pdb format, there are several tools that can read and show them. For this research it was analyzed some of them:

- IQmol
- Swiss-PDB Viewer
- PyMol

IQmol is a free open-source molecular editor written using the Qt libraries. It is available for OS X, Windows and Linux platforms. Its main features are a molecular editor, surface generation and animations (vibrational modes and reaction pathways). [1]

Swiss-PdbViewer is an application created by Swiss Institute of Bioinformatics (SIB) at the Structural Bioinformatics Group at the Biozentrum in Basel. It provides an interface allowing to analyze several proteins at the same time. The proteins can be colored and modified in order to compare structural alignments and their active sites. (Figure 5.2) [11]

PyMol is an open-source but proprietary molecular viewer. It is written in the programming language Python. Its main function are rendering and animating 3D structures. [18, 24]

After the analysis of the available features of these tools, for this research the Swiss-PDB Viewer was chosen.

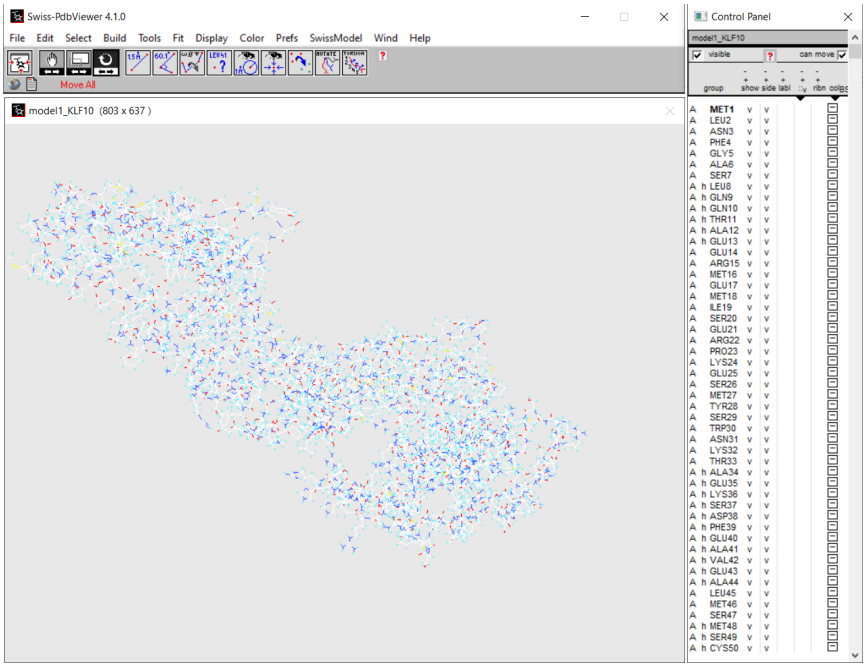


FIGURE 5.2: Swiss-PDB Viewer screen shot

5.4 HADDOCK as a molecular docking tool

HADDOCK (High Ambiguity Driven protein-protein DOCKing) is a docking approach for modeling biomolecular complexes. HADDOCK uses a data-driven method of docking (using experimental restraints such as information derived from biochemical, biophysical, or bioinformatics methods), which performs remarkably better than ab initio docking tools.

HADDOCK can deal with protein-protein, protein-nucleic acids, and protein-ligand complexes.

It has several available interfaces; for this research, the Easy interface was used.

The Easy interface provides a basic level of control. To start the docking user needs to provide two pdb-structures (or enter the ID of them in the PDB Library) and define the sets of active and/or passive residues of the complex.

The HADDOCK server validates the input and puts the request in a queue. It also gives the option to download the file of given parameters and the link to the status page, where the progress of the docking run will be shown.

The result includes several clusters of models and their ranking, so the most conceivable option could be chosen.[21, 12] It usually takes about 12 hours to retrieve the results.

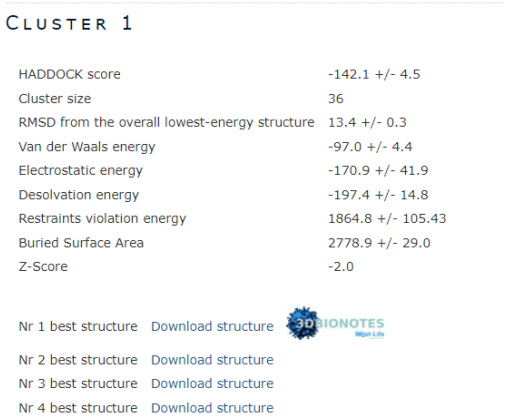


FIGURE 5.3: HADDOCK example of output page

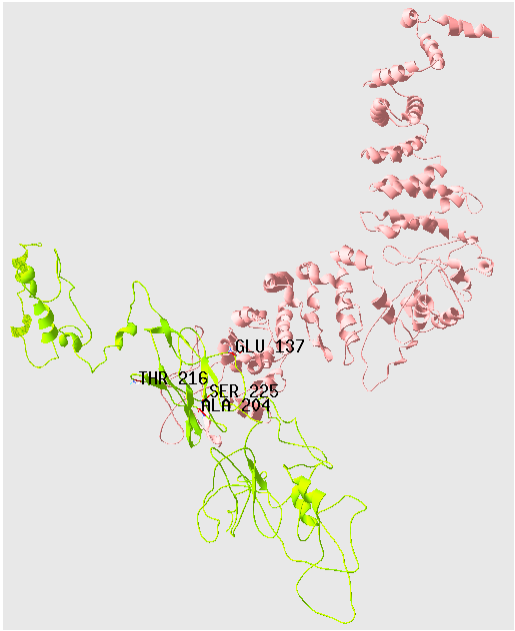


FIGURE 5.4: HADDOCK example of predicted complex

5.5 PRODIGY System

PRODIGY (PROtein binDIng enerGY prediction) system is a collection of web services that predicts the Gibbs free energy and the dissociation constant for given models.

The system has three modes:

- PRODIGY (for Protein-Protein models)
- PRODIGY-LIG (for Protein-small Ligand models)
- PRODIGY-CRYSTAL (for Proteins in crystallographic)

To use the PRODIGY tools, the user needs to upload a PDB structure of a complex or enter the ID of its PDB entry. The request will be processed in a few minutes, and the results will be calculated.[20]

5.5.1 Methods of calculations

PRODIGY system has determined a direct correlation between the number of interfacial contacts in a protein-protein complex and the binding strength. So to calculate the connection power, the ICs have to be counted. Two residues are considered to be in contact if a pair of (any) atoms belonging to the two residues are within a defined cut-off distance (5.5 Å). Then, all the ICs are being classified into three groups based on their physicochemical properties:

- polar: C, H, N, Q, S, T, W
- apolar: A, F, G, I, L, V, M, P, Y
- charged: E, D, K, R

The Gibbs free energy is calculated using the following formula:

$\Delta G_{(predicted)}$ **= -0.09459 ICs**$_{charged/charged}$ **-0.10007 ICs** $_{charged/apolar}$ **+0.19577 ICs** $_{polar/polar}$ **-0.22671 ICs** $_{polar/apolar}$ **+0.18681 % NIS**$_{apolar}$ **+0.3810 % NIS**$_{charged}$ **-15.9433**

Where ICsxxx/yyy is the number of interacted ICs (i.e., ICscharged/apolar is the number of ICs between charged and apolar residues), and NIS are Non-Interacting Surface properties.

After that, having determined ΔGpredicted PRODIGY can calculate the dissociation constant for the predicted complexes via the following formula:

$\Delta G = RT \ln Kd$

where R is the ideal gas constant (in kcal K-1mol-1), T is the temperature (in K), and ΔG is the predicted free energy. By default, the temperature is set at 298,15 K (25.0 °C) but can be modified by the user.[20]

Chapter 6

Conclusion and future works

6.0.1 Discussion

With all the work and inferences considered, it is conceivable to claim that the interaction of KLF10 SNPs with both Ski and SnoN proteins differs from KLF10 WT interactions. Unfortunately, nowadays, it is impossible to perform a prediction of three-componential interactions, so there is no way to analyze the complete KLF10-Ski/SnoN complex with myogenic differentiation transcription factors. From the calculated interaction energies, it is possible to conclude that KLF10 T216A is not a most crutial residue at the interactional interface in the KLF10-Ski complex but does affect the KLF10-SnoN interaction. The only SNP with a smaller absolute value in both complexes is KLF10 SNP E137K, which means its complexes KLF10 SNP E137K with Ski and KLF10 SNP E137K with SnoN are less stable and will possibly have a lesser repression effect on the myogenic signaling pathway.

6.0.2 Conclusion

The main conclusions from this research are:

- Cardiac hypertrophy-related SNPs cause changes of the KLF10 protein structure that modify a spatial orientation of its complexes with Ski transcriptional repressor protein;
- Cardiac hypertrophy-related SNPs cause changes of the KLF10 protein structure that modify a spatial orientation of its complexes with SnoN transcriptional repressor protein;
- modifications in the spatial orientation within complexes between SNPs bearing KLF10 and Ski or SnoN proteins are reflected with changes in Gibbs free energy

6.0.3 Future works

Future work is required for more profound research on KLF10's effect on HMC. The next step in the studies could be the research on KLF10 interactions with MyoD and Myogenin.

Appendix A

Additional tables and figures

Repression region	**DNA sequence**	**Amino acid sequence blasted on KLF10**
R1	CGTGCCAGGCTGTGGGGCTTCTGCTTTAAGCCCACGTTGTGGGGC CACAGACTTGCAGTGGAAGCATTTTTACATCACCACTGGCTCCCG CTGAGACCAAAGTTAGTTCTGACTCTTCACTTTCCGGTCTGTCAC TGTGTGGGAGCAGGGGTTGGAGGTAGAGCAATGTCATTTAGCTTG CTCACTTCCATCTGCCAGTTTGGTAGCTTCTTGGCTGATAGATGG CGCCGGGCATGCTTGGTCAAATGGTCACTCCTCATGAACCGCCGG TCACACATGGGGCACGCAAATTTCTTCTCACCCGTGTGGGTTCGC CTGTGTCTGGACAGTTCATCAGAACGGGCAAACCTCCTTTCACAA CCTTTCCAGCTACAGCTGAAAGGCTTTTCTCCTAAAACAA	GEKPFSCSWKGCERRFARSD ELSRHRRTHTGEKKFACPMC DRRFMRSDHLTKHARRHLSA KKLPNWQMEVSKLNDIAL
R2	GCCACATCCTGGGTGGCTACAGATGTGACTCCTTATCCTTGATGA ATCAATCTGAGGAGTGACTTTTGCTGCTGAAGGGGAAAACCCAGG AGCAGGGGCAATGGGAGAGAGTCTGGTGCCATTCGGGCTCACCAC CGGAGGCTTTGAACTCTGCACAACGGGCTGGGGTACCACAAACAT GACAGCGCCTTTGGGGACTTGTGTGCCCATGAACACAACAGGGGG GCAAACGGCTGGTGGCTGGCTGGGAGGAGTGCTGGGAACGACTGT TGTCACAACAGGGTTGTTGGCAGGAAGGGGAACCATCTGGCAGAT GACCGGCATAGGTGGCACTCCCCCTGCAGATACTGCAGGTGGAGA GACCAACACTGACTTCTGTTGTGGGGACACAGGGGCTGGCTGAGA CCTGCAGATGACCGTCTCTGAGGAAGGCACAGAAAAGTCATAAAG TGCAGCACTTGCTTTCTCATCAACATCTGCCACTGTGTTTCTCTC ACATTTGGATCTGTTTGGTGACACAGCGGCACATGGTATGTTCTT TCTTGCAGCCTCAACATTTAGGTGGGTTCTTCTTCTAAAAGAATT GTTCTGATAGTTGAGGATGCTGGCTGCTTTCATTGGGCAGGTCTG GTGGTTACATAGCTGGGCATCAGCTGTATGACGAATCACACTTGT TGCCTGAGCTTTGGGGAGTTTGGGGGCAGATACTGGGCTCTTTTC TTCCTCTTTGAAAGGTGCGGCAATGTGAGGTTTGGCAGTATCTGA GAGTGACTTGAAGTGTACAGTAGATGGCGCTGGTGCCATCAGATT TGACACTTGAGAGGGTTCAAAGTCAGAAGGACTGTAAGGTGGAGT CAAACACTAAAGAAAAGGGAAATACATAGCATGAGAAATCTACAG TTTATTATATAAATTTTAAGAAATTATATAATTTACGTCTATCTT AAAAACAATAATACTTACAAATGCTGGGATTGTATGAAAATCAGG TGTTCCCGGA	PGTPDFHTIPAFCLTPPYSP SDFEPSQVSNLMAPAPSTVH FKSLSDTAKPHIAAPFKEEE KSPVSAPKLPKAQATSVIRH TADAQLCNHQTCPMKAASIL NYQNNSFRRRTHLNVEAARK NIPCAAVSPNRSKCERNTVA DVDEKASAALYDFSVPSSET VICRSQPAPVSPQQKSVLVS PPAVSAGGVPPMPVICQMVP LPANNPVVTTVVPSTPPSQP PAVCPPVVFMGTQVPKGAVM FVVPQPVVQSSKPPVVSPNG TRLSPIAPAPGFSPSAAKVT PQIDSSRIRSHICSHPGCG
R3	ACAGGAAGGGAAACAGGGAGGGGAGGAAAGGGAGTGGGGCGCGAG GCAGGAAAGGGAAGCGCGGGCGGAGGCGCGGGCGGGGGAGGCAGA CGAGGGGCGGGGGCGGACGGCGGGGGAGATCCTAGCTGGCGGAGA CCGACGGATGGGGCCGCCCTAACCTCAATGAGGCTCACGGGTGAG TCACTGGGAACATTCCTCGCCGCTCGCACGTCCCCGCGCCCCTGC CCCCGCCATTGGCCGGCCGGCTCGGCGGCCCGAGCCCGGATTGGC TGTGCGGCCCGGGGCGGGGGCGTGGGCGAGGAGGGGGCGAGGCAT GTGAACAAAGCGTGATCAGCGGCTGCTCCGGGAGGCGCAGGAAAC GTGAACTGGGAATTGCCGCGCCGTCACCTTTCCCCTACTTCCTGA GCTTGCGGGCCCCCGCCCGCGCCCGACCCGGGGGAGGGGCCGCGG GCGCCGCCGCCAACCGCCCGGCCGTGCGCGCCGGGCCCGCGGGGA GGGGCGGGGCTGCGGGCGCCAGGCTCCGGGAGGCGCCCGCCCCTT CCAGCTCCCCGCGAGCCGCCGTTCCCTTAGGAACCGACAGCGGGG CCGGGCCTCGCTCCC	No significant similarity found

TABLE A.1: Repression regions translated and blasted on KLF10 isophorm a

Appendix B

Alternative angles of view on structures of KLF10 SNPs and Ski complex

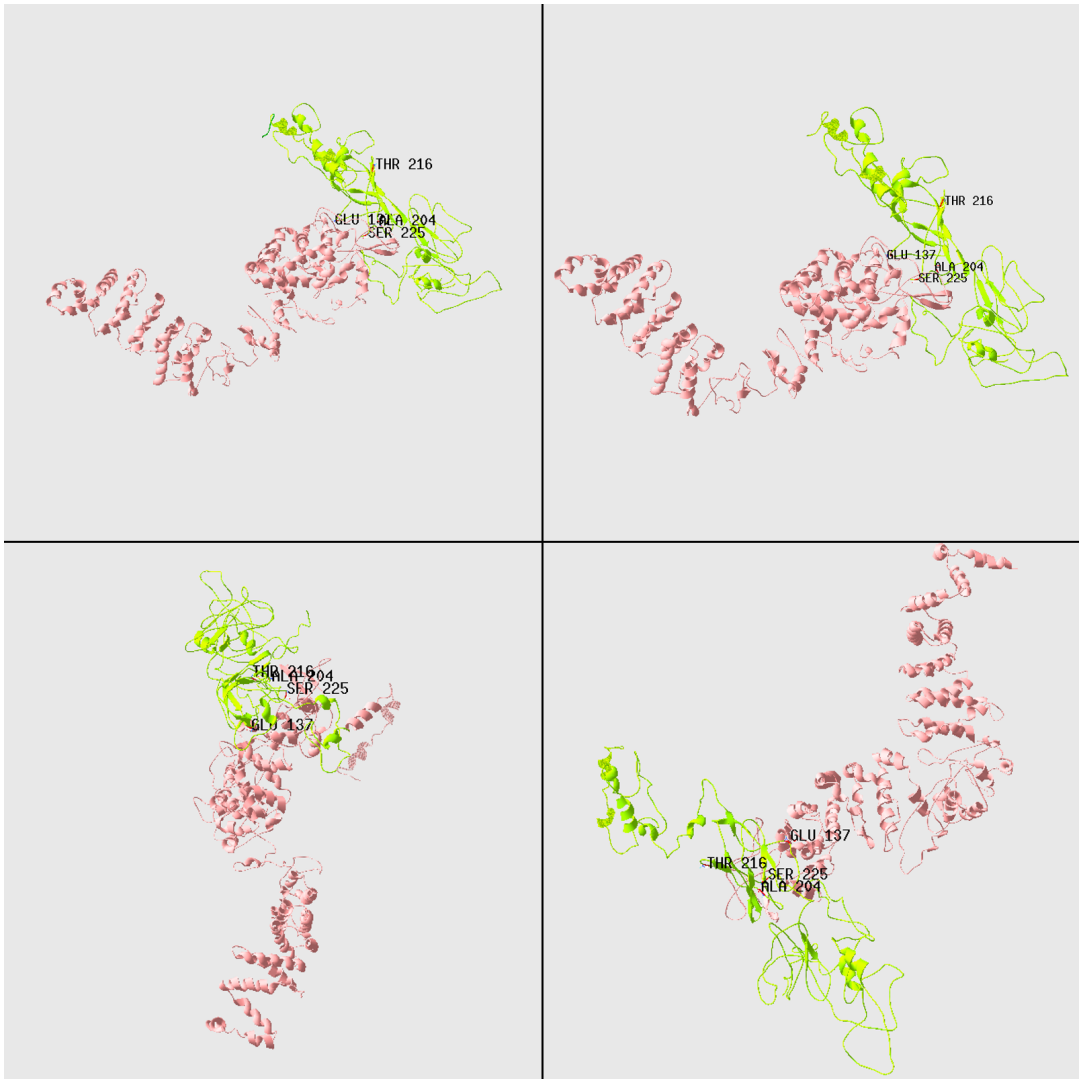


FIGURE B.1: Alternative angles of view on structures of KLF10 WT and Ski complex

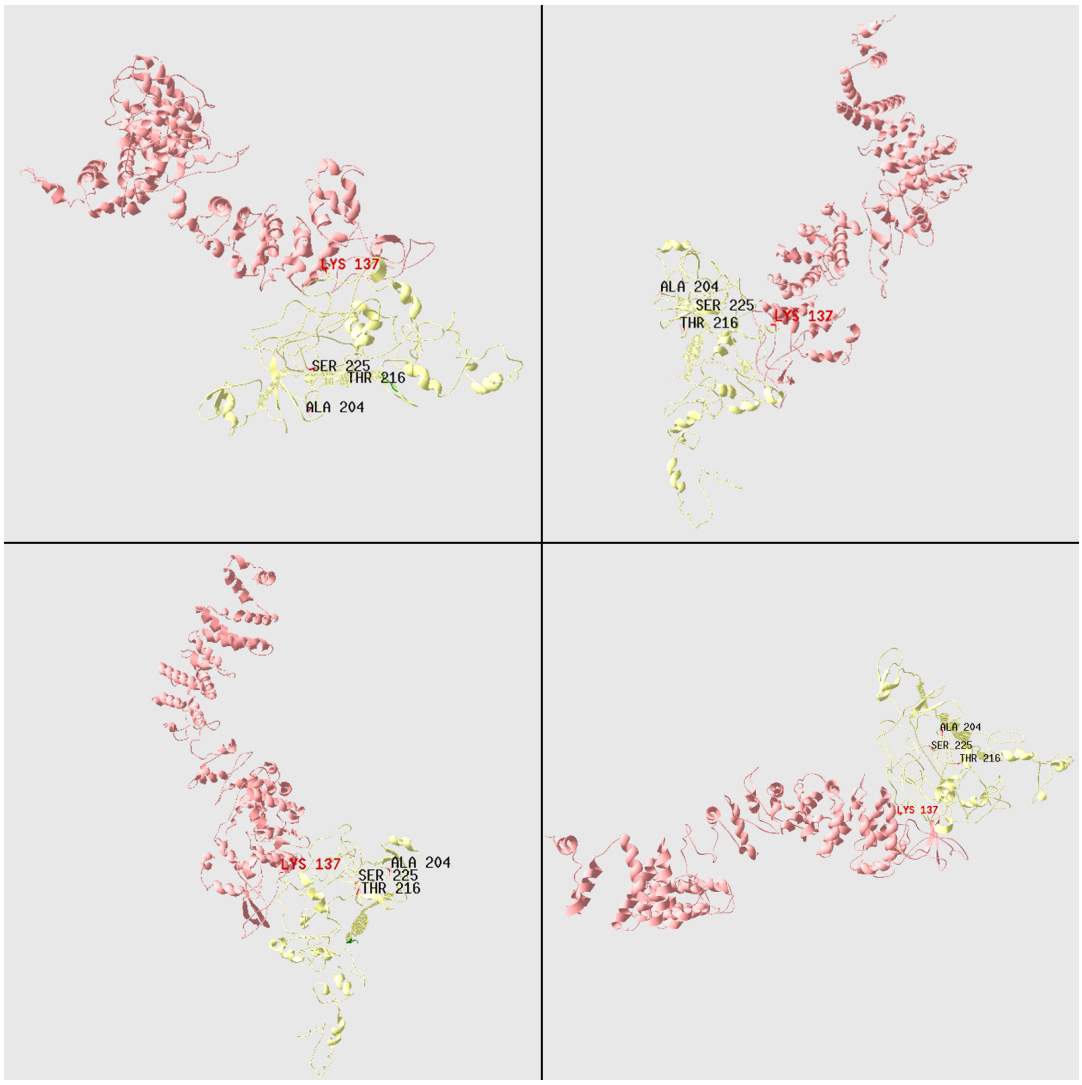


FIGURE B.2: Alternative angles of view on structures of KLF10 E137K and Ski complex

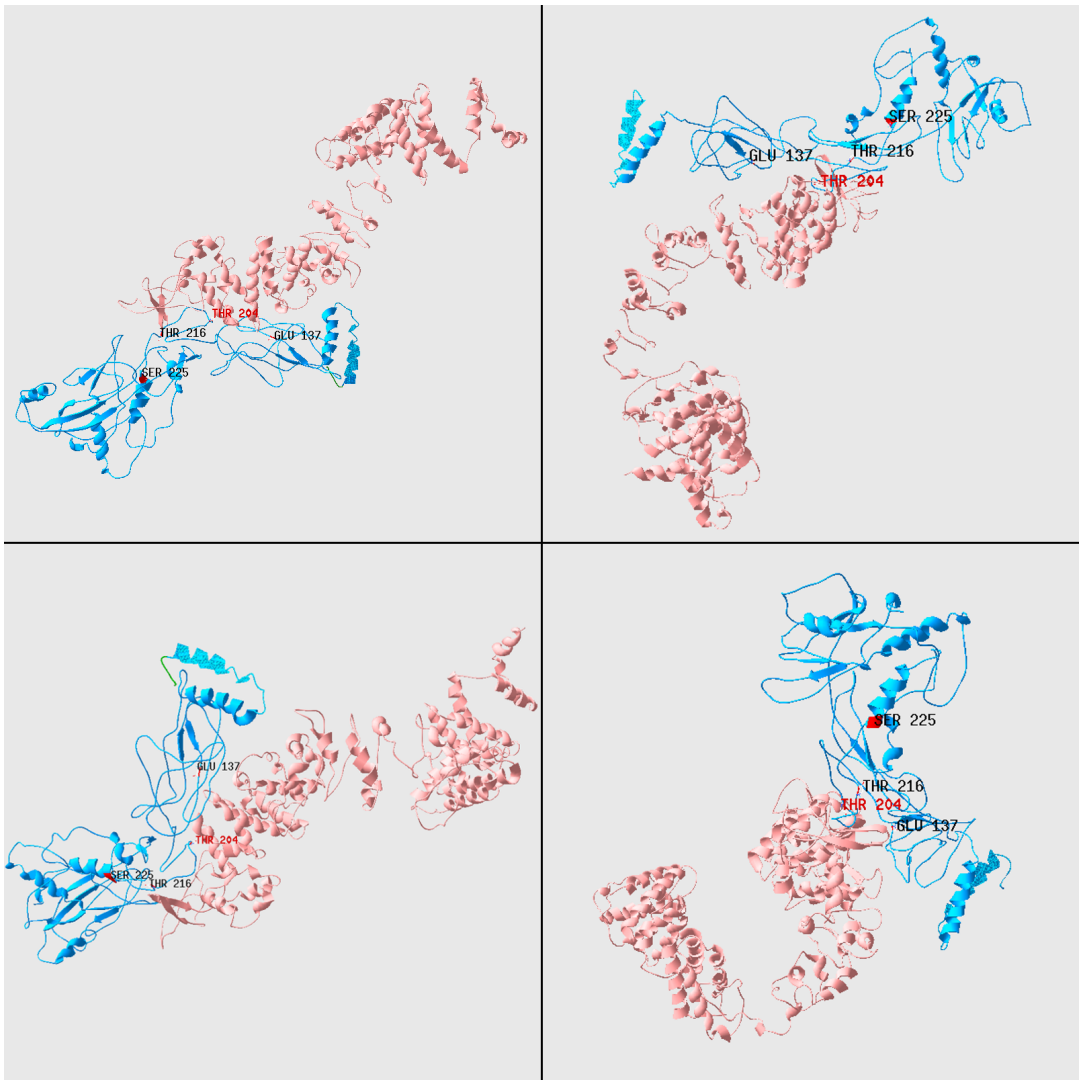


FIGURE B.3: Alternative angles of view on structures of KLF10 A204T and Ski complex

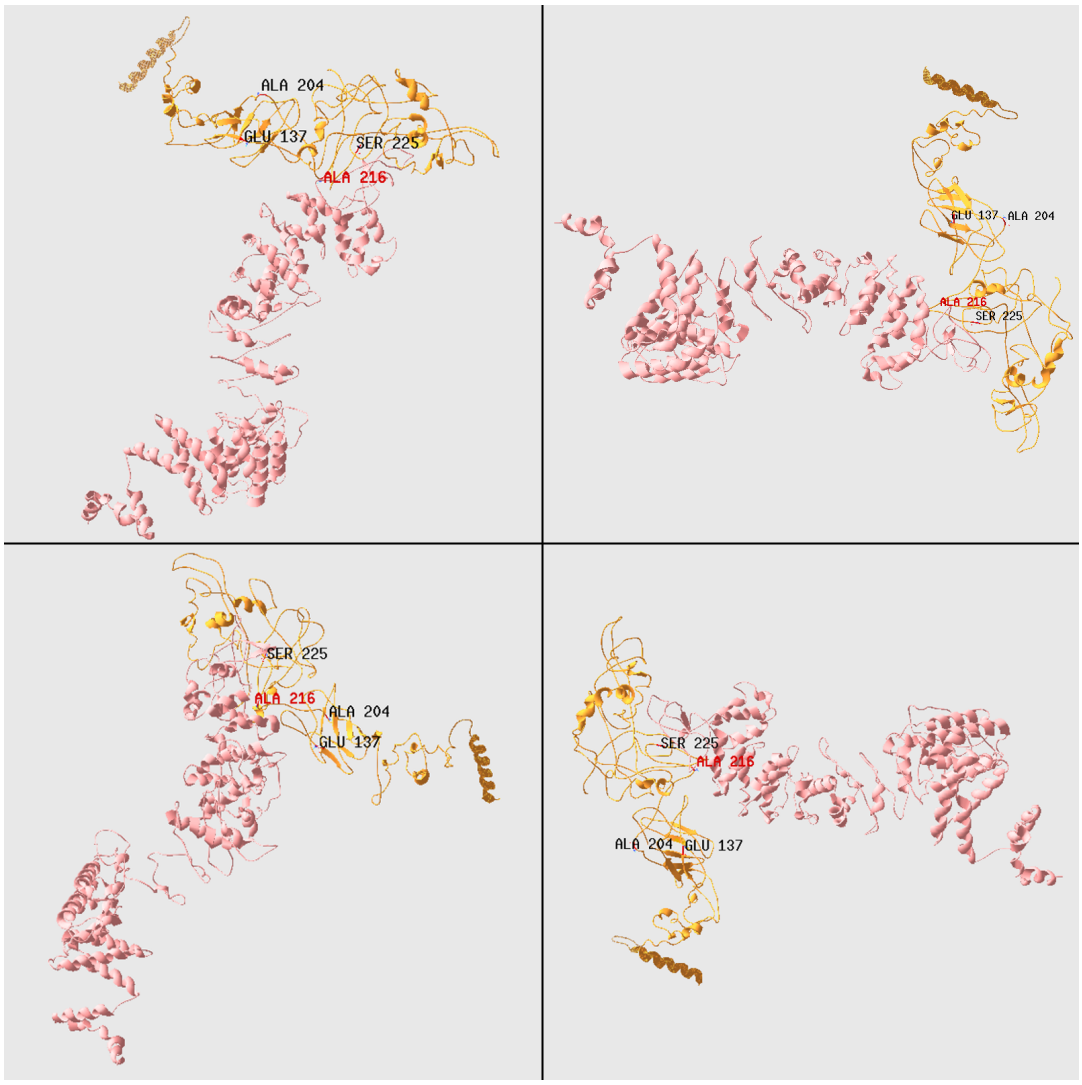


FIGURE B.4: Alternative angles of view on structures of KLF10 T216A and Ski complex

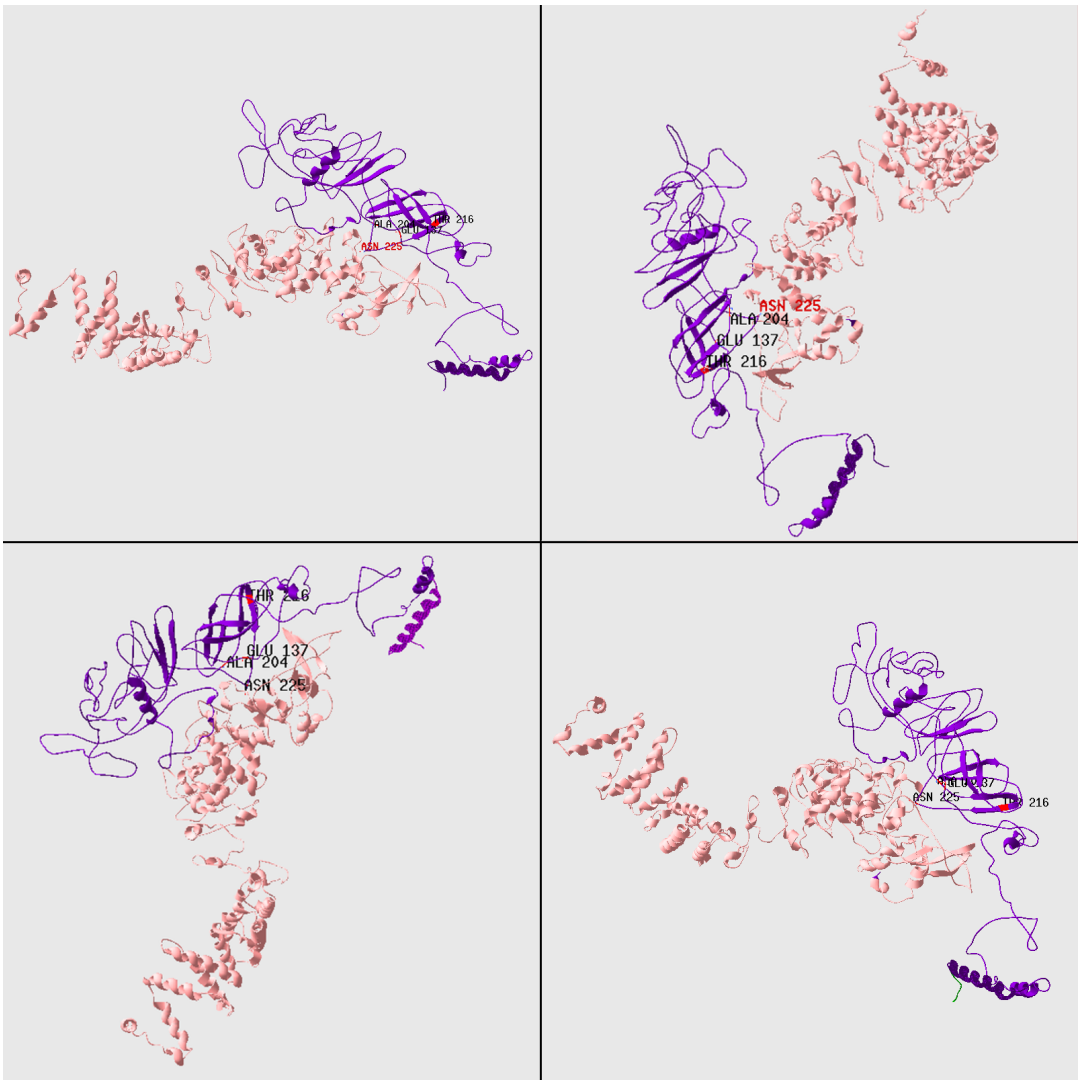


FIGURE B.5: Alternative angles of view on structures of KLF10 S225N and Ski complex

Bibliography

[1] Quantum Chemistry at ANU. *IQmol*. URL: http://iqmol.org/.

[2] YouTube Hussain Biology. *TGF Beta Signalling Pathway*. Last accessed 9 July 2020. 2020. URL: https://www.youtube.com/watch?v=TfPHZqT7b9A.

[3] The National Center for Biotechnology Information. *Transcript: ENST00000378536.5 SKI-201*. Last accessed May 13 2022. 2022. URL: https://www.ncbi.nlm.nih.gov/gene/6498.

[4] J Martijn Bos et al. "TGFβ-inducible early gene-1 (TIEG1) mutations in hypertrophic cardiomyopathy". In: *Journal of cellular biochemistry* 113.6 (2012), pp. 1896–1903.

[5] Michael T Chin. "Ascending KLFs in cardiovascular biology." In: *Trends in Cardiovascular Medicine* 25.4 (2014), pp. 288–290.

[6] ClevelandClinic. *Hypertrophic Cardiomyopathy*. Last accessed 18 March 2022. 2022. URL: https://my.clevelandclinic.org/health/diseases/17116-hypertrophic-cardiomyopathy.

[7] European Molecular Biology Laboratory's European Bioinformatics Institute (Ensembl). *Gene: KLF10 ENSG00000155090*. Last accessed Apr 2022. 2022. URL: http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000155090;r=8:102648784-102655725.

[8] European Molecular Biology Laboratory's European Bioinformatics Institute (Ensembl). *Transcript: ENST00000259119.9 SKIL-201*. Last accessed Apr 2022. 2022. URL: http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000136603;r=3:170357678-170396835;t=ENST00000259119.

[9] European Molecular Biology Laboratory's European Bioinformatics Institute (Ensembl). *Transcript: ENST00000285407.11 KLF10-201*. Last accessed Apr 2022. 2022. URL: http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000155090;r=8:102648784-102655725;t=ENST00000285407.

[10] European Molecular Biology Laboratory's European Bioinformatics Institute (Ensembl). *Transcript: ENST00000378536.5 SKI-201*. Last accessed Apr 2022. 2022. URL: http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000157933;r=1:2228319-2310213;t=ENST00000378536.

[11] Nicolas Guex. *Swiss-PdbViewer aka DeepView*. 2019. URL: https://spdbv.unil.ch/.

[12] Rodrigo V Honorato et al. "Structural biology in the clouds: the WeNMR-EOSC ecosystem". In: *Frontiers in Molecular Biosciences* (2021), p. 708.

[13] Madhav Karthik Kodigepalli. "Scholar Commons Role and Regulation of SnoN/SkiL and PLSCR1 Located at 3q26.2 and 3q23, Respectively, in Ovarian Cancer Pathophysiology Role and Regulation of SnoN/SkiL and PLSCR1 Located at 3q26.2 and 3q23, Respectively, in Ovarian Cancer Pathophysiology". PhD thesis. Sept. 2014.

[14] Zhang Lab. *I-TASSER Protein Structure and Function Predictions*. Last accessed 28 May 2022. 2015. URL: https://zhanggroup.org/I-TASSER/about.html.

[15] Alphonsus C Liew et al. "Hypertrophic cardiomyopathy—past, present and future". In: *Journal of clinical medicine* 6.12 (2017), p. 118.

[16] Warisara Parichatikanond et al. "Therapeutic targets for the treatment of cardiac fibrosis and cancer: focusing on TGF-β signaling". In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 34.

[17] Ph.D. Paul P. Liu M.D. *Repressor*. Last accessed May 15 2022. 2022. URL: https://www.genome.gov/genetics-glossary/Repressor.

[18] DeLano Scientific. *PyMol*. 2006. URL: https://pymol.org/2/.

[19] Angeles C Tecalco-Cruz et al. "Transcriptional cofactors Ski and SnoN are major regulators of the TGF-β/Smad signaling pathway in health and disease". In: *Signal Transduction and Targeted Therapy* 3.1 (2018), pp. 1–15.

[20] The Utrecht University. *PRODIGY method*. 2019. URL: https://wenmr.science.uu.nl/prodigy/method.

[21] GCP Van Zundert et al. "The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes". In: *Journal of molecular biology* 428.4 (2016), pp. 720–725.

[22] Wikipedia. *I-TASSER*. Last accessed 21 May 2022. 2019. URL: https://en.wikipedia.org/wiki/I-TASSER.

[23] Wikipedia. *Protein Data Bank (file format)*. Last accessed 10 February 2022. 2006. URL: https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format).

[24] Wikipedia. *PyMOL*. Last accessed 20 February 2022. 2007. URL: https://en.wikipedia.org/wiki/PyMOL.

[25] Wikipedia. *Repressor*. Last accessed 14 March 2022. 2014. URL: https://en.wikipedia.org/wiki/Repressor.

[26] Wikipedia. *SKI protein*. Last accessed 19 October 2021. 2007. URL: https://en.wikipedia.org/wiki/SKI_protein.

[27] Wikipedia. *TGFbeta signaling pathway*. Last accessed 5 May 2022. 2022. URL: https://en.wikipedia.org/wiki/TGF_beta_signaling_pathway.

[28] Xiaohua Yan, Ziying Liu, and Yeguang Chen. "Regulation of TGF-β signaling by Smad7". In: *Acta biochimica et biophysica Sinica* 41.4 (2009), pp. 263–272.

[29] Jianyi Yang and Yang Zhang. "I-TASSER server: new development for protein structure and function predictions". In: *Nucleic acids research* 43.W1 (2015), W174–W181.

[30] Chengxin Zhang, Peter L Freddolino, and Yang Zhang. "COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information". In: *Nucleic acids research* 45.W1 (2017), W291–W299.

[31] Wei Zheng et al. "Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations". In: *Cell reports methods* 1.3 (2021), p. 100014.