8-2022

Modeling Electrostatics in Molecular Biology and Its Relevance With Molecular Mechanisms of Diseases

Mahesh Koirala
mkoiral@g.clemson.edu

MODELING ELECTROSTATICS IN MOLECULAR BIOLOGY AND ITS RELEVANCE WITH MOLECULAR MECHANISMS OF DISEASES

A Dissertation

Presented to

the Graduate School of

Clemson University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Physics

---

by

Mahesh Koirala

August 2022

Accepted by:

Dr. Emil Alexov, Committee Chair

Dr. Feng Ding

Dr. Hugo Sanabria

Dr. Joshua Alper

Dr. Brian Dominy

ABSTRACT

Electrostatics plays an essential role in molecular biology. Modeling electrostatics in molecular biology is a complicated task due to the water phase, mobile ions, and irregularly shaped inhomogeneous biological macromolecules. A particular approach to calculating electrostatics in such a system is to apply the Poisson-Boltzmann equation (PBE). This dissertation presents the popular DelPhi package that solves PBE using a finite-difference method and delivers the electrostatic potential distribution throughout the modeling box. Details descriptions of some of the examples DelPhi can handle, and their accuracy with the analytical solutions are presented. Receptor–ligand interactions are involved in various biological processes; therefore, understanding the binding mechanism and ability to predict the compulsory mode is essential for many biological investigations. So, this dissertation also presents the use of electrostatics for the proper orientation guidance and the pulling force to deliver the ligand close to the receptor. We use the newly developed DelPhiForce steered Molecular Dynamics (DFMD) approach to model the binding of barstar to barnase and demonstrate that the first-principles method can also model the binding. It shows DFMD can successfully dock barnstar to barnase even if both initial positions and orientations are entirely different from the correct one.

Single nucleotide polymorphism (SNP) is a single nucleotide variation in the genome. Change of single amino acid in the corresponding protein due to these variations results in single amino acid variation (SAV). SAVs cause intense

alterations of the related biological processes and thus can be associated with many human diseases. This dissertation reflects the use of existing computational approaches to model the effects of SAVs to reveal molecular mechanisms related to human diseases. We use our supervised in-house combinatory in-silico predictor method to investigate the impact of unclassified missense mutations in the MEN1 gene found in breast cancer tissue issues. We also examine the biophysical and biochemical properties to predict the effects of these missense variants on the menin protein stability and interactions. The results are compared with the impact of known pathogenic mutations in menin causing neoplasia. Together with our in-silico consensus predictor, we classify missense mutations in menin protein found in breast cancer tissue into pathogenic and benign, thus, suggesting an indicator for early detection of elevated breast cancer risk.

In the end, we show the role of intravesicular pH in melanosome maturation and formation. Here, we computationally investigate the pH-dependent stability of several membrane proteins and compare them to the pH dependence of the strength of TYR. We confirm that the pH optimum of TYR is neutral, and we also find that proteins that are negative regulators of melanosomal pH are predicted to function optimally at neutral pH. In contrast, positive pH regulators were expected to have an acidic pH optimum. Our findings are consistent with previous work that demonstrated a correlation between the pH optima of stability and activity. They are compatible with the expected activity of positive and negative regulators of melanosomal pH. Furthermore, our data suggest that disease-causing variants impact the pH

dependence of melanosomal proteins; this is particularly prominent for the OCA2 protein. In conclusion, melanosomal pH appears to affect the activity of multiple melanosomal proteins.

DEDICATION

I dedicate this dissertation to my loving parents (Madhu Prasad Koirala & Saraswoti Koirala), beautiful daughter (Sophie Koirala) and lovely wife (Manisha Poudel).

ACKNOWLEDGMENTS

I express my sincere thanks to my supervisor, Dr. Emil Alexov, for his unconditional support and guidance throughout my Ph.D. It's my pleasure to have him as my advisor, who has helped me gain expertise in biophysics, structural biophysics, and bioinformatics. His encouragement, selfless and unbiased support are incomparable. I am praised for his logistic and scientific discussions and grateful for his moral and emotional support. It is one of the best parts of life to have an advisor like him.

I am thankful to the members of my doctoral committee- Dr. Feng Ding, Dr. Hugo Sanabria, Dr. Joshua Alper, and Dr. Brian Dominy for their overall knowledge of the subject matter and help in discussion on several occasions. I also express gratitude to Dr. Hugo Sanabria and Dr. Joshua Alper for teaching me biophysics courses. Their advice and suggestions are priceless, and the path to completing my Ph.D.

I am deeply grateful to Dr. Lin Li from The University of Texas as El Paso and Dr. Yunhui Peng from the National Institute of Health for their support and help in joining the lab and learning the Delphi and other tolls in the lab. Without their help, it would not be possible to finish my several projects. I want to thank all the present and past lab members- Dr. Bhoua Wu, Dr. Swagata Pahari, Dr. Shailesh Kumar Panday, Dr. Gen Li, Dr. Mihiri Hewa, Dr. Satishkumar Ranganathan Ganakammal, Dr. Nayere Tajielyato, Dr.Arghya Chakravorty, Jacob Jeffries and Shannon Bonomi for their continuous support, discussions, and contributions in my research projects.

I also thank the faculty of the pandA at Clemson for helping me pursue a fundamental understanding of Physics. In that spirit, I thank Dr. Jason Brown, Dr. Pooja Puneet, Dr. Linsin The, and Dr. Daniel Thomson for having me as their TA during this time and for their understanding and support even in different unusual circumstances. I also want to thank the office staff- Amanda, Lori, Rise, Celeste, and Kim for their excellent administrative support. I am also grateful for the resources provided by Clemson University, especially Palmetto Computing. I am joyfully thankful to all my friends, especially Abhishek, Pan, Anirudh, Peshal, Prakash, Shailendra, Nawaraj and Sachet.

Not the least, I am incredibly thankful to my parents for their unconditional love and countless sacrifice. The love and support from my wife, Manisha Poudel, helped me achieve this success. I am grateful to her for being at my side all the time and giving birth to our beautiful daughter, Sophie Koirala. I am thankful to Ajay Kunwar & Machhindra Basnet for being our family away from home and supporting me in every step. I am fortunate and grateful for having you all by my side all the time.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

This chapter covers the introduction to the fundamental ideas and tools that are important for the foundation of the work presented in this dissertation. It discusses the role of electrostatics in molecular biology, including protein stability and protein-protein interactions. The other sections of this chapter introduce Single Nucleotide Polymorphisms (SNPs) and their association Zwith different diseases highlighting their effects on proteins functions. In the end, it describes the role of pH on protein stability and activity.

1.1 Role of Electrostatics in Molecular Biology

Electrostatics is an essential component in the various biological process because each atom of biomolecules carries a charge and thus the electrostatic interactions are present at atomic levels of detail[1]. Electrostatic interactions dominate other forces when atoms or molecules are apart at a longer distance than the typical bond length. They facilitate molecular recognition, and biological activities of the biomolecules. For example, electrostatic forces participate in protein folding and binding, protein -DNA/RNA interactions, ion binding, dimerization, and protein-microtubule binding[2]. Electrostatics is also responsible for the pH dependence on the stability of biomolecules. Electrostatics is the major component of total solvation energy of biomolecules[3]. Furthermore, electrostatics is associated with the disease since disease-causing mutants frequently alter wild-type interactions [4].

1.2 Single Nucleotide Polymorphisms (SNPs)

Mutation is a change in the nucleotide sequence of DNA, which occurs during its replication, recombination, or other processes. The mutation results in substitution, deletion, or insertion. Many mutations are repaired before protein synthesis occurs, so they do not affect the function of the corresponding biomolecules. Some of the mutations positively affect the related organism, called beneficial mutations. However, other mutations may drastically reduce the organism's ability to survive, and such mutations are called harmful mutations[5]. The single nucleotide variation in the genome resulting in single amino acid variation is known as Single Nucleotide Polymorphisms (SNPs)[6, 7]. They can change the corresponding biological process causing various diseases[8, 9]. The human population shares about 99.5 percent of DNA codes, and the remaining 0.5 percent makes it unique among the organisms. So, the study of SNPs to reveal the molecular mechanism associated with human disease is essential. SNPs are of two types, synonymous and non-synonymous. Synonymous mutations do not change the amino acid (AA) sequence of the protein. Missense and nonsense mutations are two types of non-synonymous mutations. Missense mutations change AA in the protein, and nonsense creates a stop codon and results in premature truncation [10]. This dissertation will discuss the effects caused by different missense mutations that are associated with various diseases.

Disease-causing mutations affect the function of the corresponding proteins. They can affect proteins stability and their binding. They can also alter the conformation dynamics, hydrogen bonding, salt bridges, electrostatic interactions, and the pH dependence of the proteins[4, 5, 11-14]. There is a relation between the properties mentioned above and the propensity given mutation to be pathogenic [15-19]. Therefore, the ability to correctly predict the change of wild-

type properties of the corresponding macromolecules due to mutations is critical for disease diagnostics.

1.3 Protein-Protein Interactions

Protein-protein interactions (PPIs) are abundant in the cell, it is estimated that each protein on average interacts with four partners. It results in homodimers, heterodimers, enzyme inhibitors, and antibody-protein complexes[2, 20-22]. They are involved in various biological processes, so it is essential to understand the binding mechanism and predict the binding mode of the biomolecules. The deviation of the wild-type patterns and related physical properties are related to different diseases. PPIs alter the kinetic properties of the protein, and they are also the standard mechanism to allow for substrate channeling, formation of binding sites, and drug discovery[2, 22-24].

Modeling PPIs is a two-fold problem: predicting the binding pose and revealing the binding trajectory. There are two major approaches for predicting binding pose as homology modeling and docking, while much less attention has been paid to predicting the binding pathway. This is very important for cases that require significant conformational changes affecting backbone motion and opening of the binding pocket[25-28]. This dissertation discusses the molecular dynamics (MD) simulation technique taking the electrostatic interactions into account and addresses PPIs' binding trajectory and pose.

1.4 pH dependence of Protein Stability

The pH of a solution is an essential characteristic of many biological processes. pH controls the macromolecular activities, and for every macromolecule, there is a particular pH at which the macromolecule is the most stable and its activity is maximum. Macromolecular interactions are also pH dependents, and there is typically a pH optimum at which the binding affinity is maximum [14, 29, 30]. Subcellular compartments within a cell have different pHs, reflecting their function, from low pH in lysosomes to high pH in peroxisomes. Increasing the scale of this idea, pH plays a crucial role in body organ function and varies from very acidic in the stomach to neutral in the blood. All the above examples indicate that the regulation and maintenance of pH are essential for many biological phenomena. The pKa values of titratable groups in macromolecules determine the pH dependence of their stability, interactions, and enzymatic activity. Any changes in the pKa's may affect the wild-type properties of the corresponding macromolecule[30-32]. This dissertation explains the pH dependence on the stability of the melanosome proteins for the melanin formation and functions.

CHAPTER TWO

MODELING ELECTROSTATICS IN MOLECULAR BIOLOGY

This chapter discusses popular software DelPhi and its associated resources to model biomolecules' electrostatic potentials and energies. It explains some features, examples with tutorials, and benchmarking results against analytical solutions.

2.1 DelPhi

As discussed in the previous chapter, electrostatics is an essential component in various biological processes. The modeling of electrostatics is complex process in molecular biology which is complicated due to the presence of water phase, mobile ions, and irregularly shaped inhomogeneous biological macromolecules. To overcome these difficulties, the Poisson-Boltzmann method is developed, which is a particular approach to calculating electrostatics potential by using the Poisson-Boltzmann equation (PBE)[33, 34]. DelPhi is the popular package that solves PBE using a finite difference method and delivers the electrostatics potential throughout the modeling box[35-37]. The new Delhi C++ is an object-oriented PBE package supporting various multiprocessing and memory distribution levels. The multiprocessing results in a significant improvement in computational time. The memory distribution approach reduces RAM requirement for large macromolecular assemblages and permits large-scale modeling in Linux clusters with moderate architecture. The traditional implementation of PBE in molecular biology considers that the biological macromolecules are low dielectric cavities immersed in the

water phase, described as a continuum medium with a large dielectric constant. In this protocol, ions are point charges that obey the Boltzmann distribution and can be present outside the so-called Stern layer. This implementation of PBE is two-dielectric PBE. In contrast, the Gaussian-based smooth dielectric function treats the solute and solvent on the same footage, and the continuous dielectric function describes the entire computational space. This models the macromolecules as inhomogeneous objects, and there is no sharp boundary between solute-solvent. Ions in such a scenario are point charges obeying Boltzmann distribution. Still, the argument of the Boltzmann function has a desolvation penalty which does not allow ions to propagate into the macromolecular interior unless there is a cavity. This approach is Gaussian PBE[37-39]. Both techniques are used in various computational investigations and show that they can be used for computing biologically measurable quantities.

DelPhi is available to all users worldwide. It is available in a standalone version (*http://compbio.clemson.edu/lab/delphisw/*) and also can be used as the webserver (*http://compbio.clemson.edu/sapp/delphi_webserver/*). DelPhi's other resources include downloadable packages and webservers[40]. DelPhiPka [41]uses DelPhi to calculate the protonation states of polar residues of protein and DNA/RNA at a given pH. DelPhiForce[42, 43] uses DelPhi to calculate electrostatics forces between protein-ligand, protein-protein, and protein-DNA/RNA. There are packages available within the DelPhi to study the effect of missense mutations on the folding free energy of Protein (SAAFEC)[44] and the binding affinity of the protein-protein complexes (SAAMBE)[45] and protein-DNA complexes[46] (SAMPDI).

2.2 Benchmarking Against Analytical Solutions

This section shows the benchmarking results of DelPhi delivered electrostatic energies and their comparison with analytical solutions. This benchmarking is an ultimate test of the accuracy of the DelPhi protocol and computer code. Several cases for which analytic solutions can be delivered are shown in Figure 1-Figure 3. These include the electrostatic component of solvation energy of a sphere immersed in solvent (Figure 1), the total electrostatic energy of two charged sphere in a dielectric cavity (Figure 2), electrostatic energy of a spherical charge moving across semi-infinite dielectric solvent (Figure 3: A and B) and the same for a spherical charge approaching a cylinder (Figure 3: C and D). The semi-infinite dielectric solvent is modeled as a box with dimensions 25x25x25A, and thus represents an approximation. The result clearly indicates that DelPhi C++ delivered energies are matching analytical solutions.

It should be, clarified that in case of Figure 1, electrostatic solvation energy of a sphere, the datapoint at scale=0.5 [grid/A] in case of sphere radius equal to 1A is not provided in the graph. The reason is that finite-difference algorithm requires the size of the object larger than the grid resolution (scale = 0.5 [grid/A] corresponds to grid resolution of 2A).

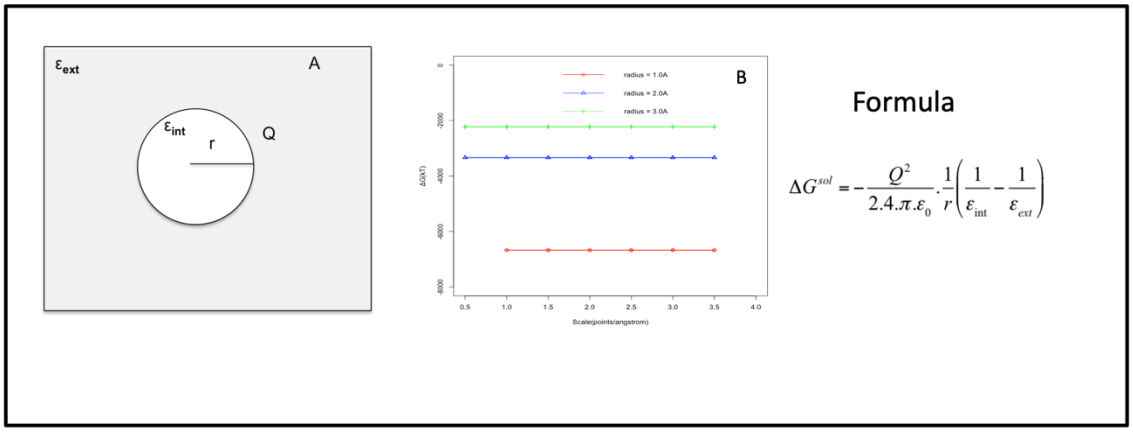

Figure 1. (A) Schematic illustration of example 1: A charged sphere with low dielectric constant is inside a Media with a high dielectric constant. (B) Electrostatic component of solvation energies calculated by Delphi against the analytical solution

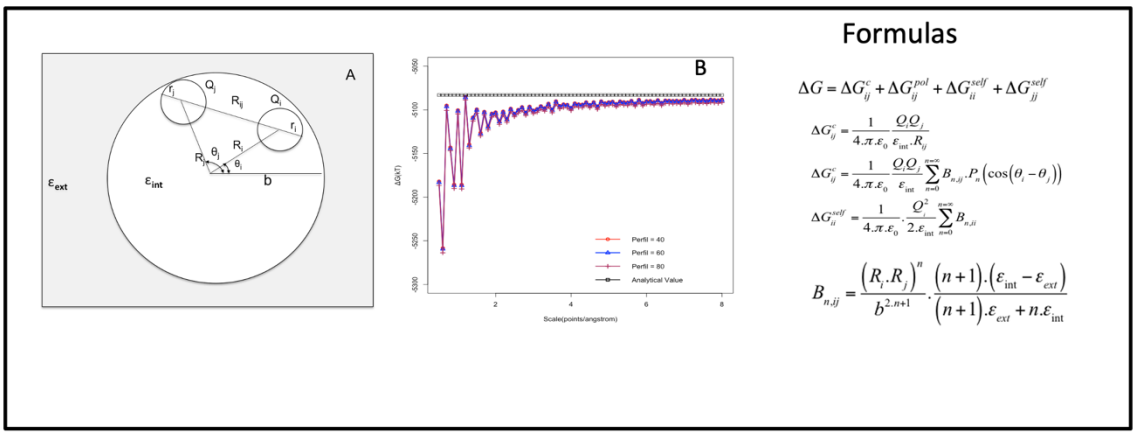

Figure 2. (A) Schematic illustration of example 2: a cavity with low dielectric constant is inside a media with high dielectric constant. Two charged atoms are located inside the cavity. (B) Electrostatic component of solvation energy obtained from Delphi compared with analytical solution.

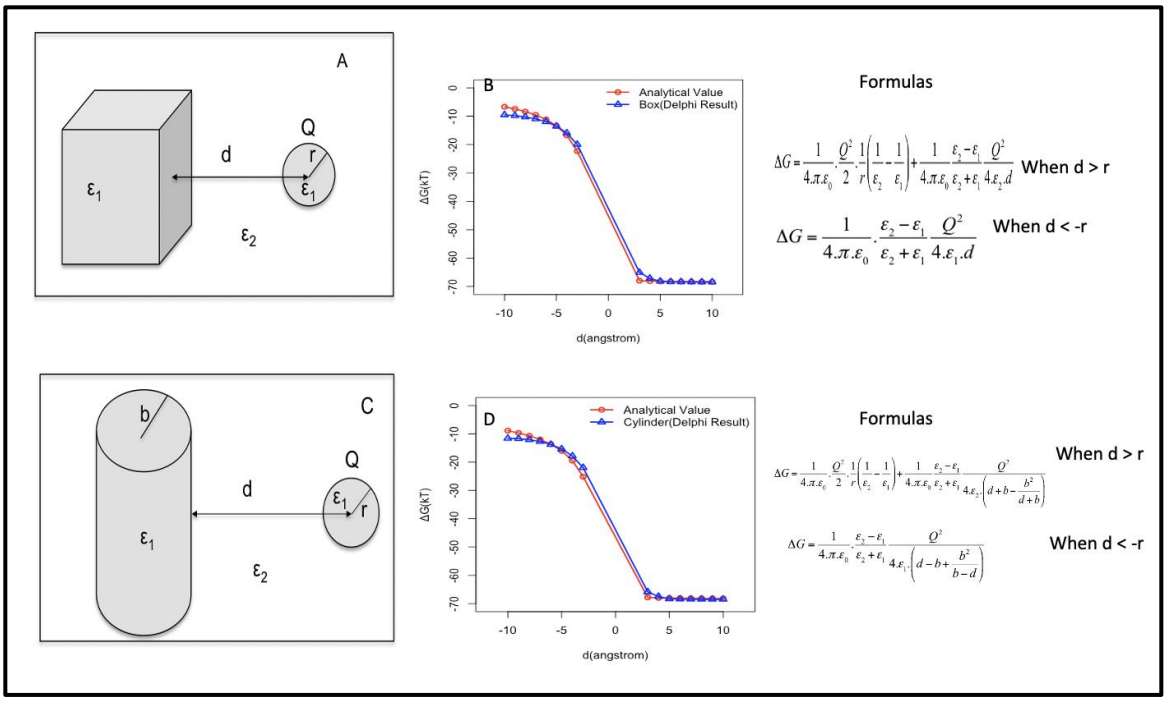

Figure 3. (A) Schematic illustration of example3: A sphere in semi-infinite dielectric plane (Box). A sphere with low dielectric constant ε_1 is initially positioned in region with high dielectric constant ε_2 and moves into the region with low dielectric constant ε_1 inside the box.(B) Electrostatic component of solvation energy derived from Delphi compared with analytical solution.(C) Schematic illustration of cylinder in two dielectric medium: A charged sphere with low dielectric constant ε_1 initially positioned with high dielectric constant ε_2 and moves into region of low dielectric constant ε_1 inside the cylinder.(D) Electrostatic component of solvation energy obtained from Delphi compared with analytical solution.

2.3 Tutorial Examples of DelPhi

DelPhi reads parameters, instructions, and other necessary details from a parameter file. The parameter file contains names of files providing coordinates of biomolecules and if necessary, the files with atomic charges and radii. Details instructions of parameter file and other details of the syntax and methods are provided in online tutorial

(*https://github.com/delphi001/delphi_tutorial_livecoms/tree/master/LiveCoMS_Examples).*
Here I describe two sets of examples showing how to set up parameter and other files to compute various electrostatic quantities.

2.3.1 Computing solvation energy

In this section, I describe how to calculate electrostatic component of solvation energy, i.e, ΔG_p for a single atom and a real protein, as shown in Figure 4.

The case of single atom (a charged sphere) is considered because there is an analytical solution via Born formula of charged ion, and the numerical solution provided by DelPhi can be compared with the analytical. This also provide example of computing electrostatic solvation energy using traditional two-dielectric model as well as the Gaussian smooth dielectric function PBE. The corresponding files can be found in directory *Example_3.1.1.* and sub directories there in.

I will begin with the simplest case of charged sphere (*see Example-3.1.1/EX1*) using traditional two-dielectric model. Consider a charged sphere of radius 3A carrying a charge +1 e.u. (electron units). Let's consider that internal and external dielectric constants are 1 and 80, respectively. Thus, the analytical solution is -92.33 kT. DelPhi can be executed using the files provide in the above-mentioned directories as:

$DELPHI_EXE_param_charged_sphere.prm\ >Charged_sphere.log

After the run is completed, the user can find the corrected reaction field energy from the output information in the log file, which is:

Energy > Corrected reaction field energy: -92.50 kT

Thus, the numerical solution delivered DelPhi matches the analytical solution. Users can check the sensitivity of results by changing their scale parameter.

Applying Gaussian-based smooth dielectric function approach on the same problem, the charged sphere, results in different polar solvation energy. First, there is no analytical solution since there is no longer hard sphere, but rather a spherical object with smooth dielectric function having minimum value at the center of the sphere and smoothly reaching 80 in bulk solvent. DelPhi can be executed with this example as:

$DELPHI_EXE_param_charged_sphere_gauss.prm\ > charged_sphere_gauss.log

The calculated polar solvation energy is in the log file:

Energy > Corrected reaction field energy: -211.20 kT

The other example of how to calculate polar solvation energy of a protein, all the related files are kept in the directory *Example_3.1.1/Ex2/*. The DelPhi run for computing it via traditional two-dielectric approach can be initiated using command below:

$DELPHI_EXE param_protein.prm > protein.log

After the run, the user can find the output information in the log file:

Energy > Corrected reaction field energy: -1005.07 kT

Similarly, modeling the same protein with Gaussian-based smooth dielectric function can be done by invoking DelPhi and appropriate parameter file as:

$DELPHI_EXE param_protein_gauss.prm \ > protein_gauss.log

After the run is completed, one finds the output information in the log file:

Energy > Corrected reaction field energy: -3583.05 kT

The slight difference in “corrected field energy” values for traditional two-dielectric and Gaussian-based methods is because traditional two-dielectric approach is intending to deliver solvation energy of a rigid molecule, while Gaussian-based methods intends to mimic the effect of conformational changes. Use of traditional two-dielectric model with higher than 1.0 dielectric constant represents a case of treating conformational changes uniformly, i.e, uniform flexibility. However, macromolecules are not homogeneous objects. Instead, Gaussian-based model assigns local dielectric values depending on atomic packing and thus reflects inhomogeneous nature of proteins.

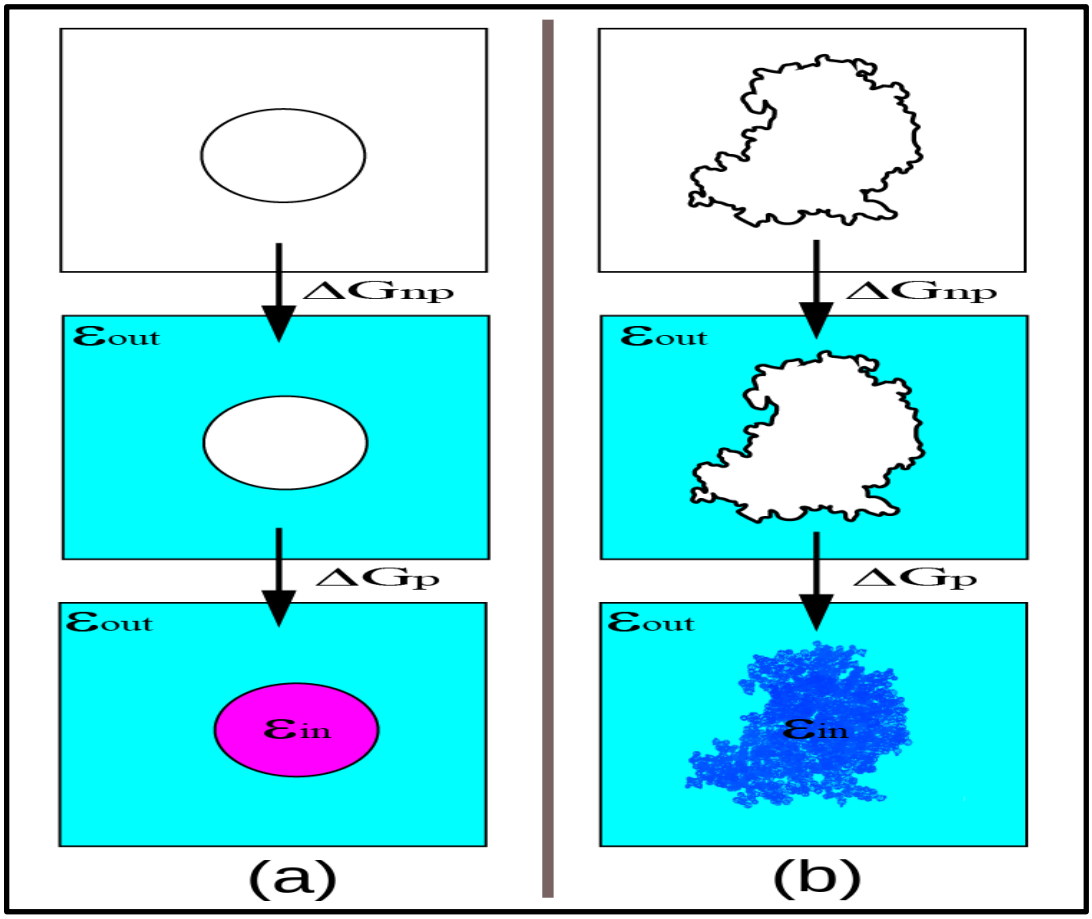

Figure 4.The schematic representation of computation of solvation for (a) a charged sphere and (b) a protein. The solvation of a molecular system consists of two components, non-polar component $\Delta G_{np}$and polar component ΔG_{p}. The non-polar components account for the energy required to create a cavity to accommodate the molecular system (solute) in the solution or moving solute from gas phase into solvent keeping solute atoms partial charges turned off, and polar component accounts energy required to turn on the partial charges of solute atoms in the solvent.

2.3.2 Computing electrostatic potential, energy of interaction and forces between two sets of atoms

In this section, I will describe how to calculate the electrostatic potential, energy of interaction and force between two sets of atoms using '*frc'* option. Scheme is shown in Figure 5.

In the parameter file we need to add command "*site (argument)",* specially in this case *site (a,p, f).* This makes Delphi to report the potentials and electrostatic field component at the positions of the subsets of atoms in the *frc* file. This calculation requires two types of *pdb* files, one is "*atoms.pdb*" which contains co-ordinates of all atoms which contribute to the electrostatic potential, and another is "*atom-1.pdb*" which contains a dummy atom to specify the coordinate at which electrostatic potential and electric fields are required to be computed. The charge and size information for all atoms present in *atoms.pdb* have to be provided in charge and size files *atoms.crg* and *atoms.siz* respectively. In present case we attempt to calculate electrostatic potential and electrostatic field at the origin due to a system of two atoms A1 and A2 whose charges are $q_1 = 10\ e_c$ and $q_2 = 20e_c$ respectively, the size of both atoms is $1.0\dot{A}$, and coordinates of A1 and A2 are (5.0, 5.0,0.0) and (5.0,0.0,0.0) respectively. Here charges are in unit of charge

of proton i.e., e_c and distance and coordinates are in $\dot{A}$. A schematic representation of the system is shown in Figure 5.

The analytical expression for such system is known and shown in equation 1. Therefore, we can benchmark DelPhi results against analytical value. Since DelPhi unit of energy is $\frac{K_{BT}}{e}$, we shall convert the analytical energy also in $\frac{K_{BT}}{e}$ for comparison. Using Boltzmann constant $K_B = 1.38 \times 10^{-23} JK^{-1}$, absolute temperature T = 297.33K (default temperature in DelPhi), elementary charge $e_c = 1.6 \times 10^{-19} C$, and external dielectric constant $\epsilon_{ext} = 80$, the potential computed from equation 1 comes to be 38.035 $\frac{K_{BT}}{e}$, while potential computed from DelPhi is 38.1317 $\frac{K_{BT}}{e}$, which shows a relative error ~10^{-3} showing an excellent agreement to analytical value.

$$\emptyset = \frac{1}{4\pi\varepsilon_0\varepsilon_{ext}}\left(\frac{q_1}{d_1} + \frac{q_2}{d_2}\right) \quad (1)$$

Similarly, analytical values of X- and Y- components of electric field E_x and E_y at position of A3 due to systems of charges A1 and A2 are -6.613 $\frac{K_B T}{e_c \dot{A}}$ and -0.993 $\frac{K_B T}{e_c \dot{A}}$, while value calculated from DelPhi are -6.7208 $\frac{K_B T}{e_c \dot{A}}$ and -1.0031 $\frac{K_B T}{e_c \dot{A}}$ respectively, which are also in good agreement to analytical values. All the required fields are provided in the directory *Example_3.1.4/.*

Users can run DelPhi for this example using following command:

$DELPHI_EXE param_frc.prm > delphi_frc.log

The output file i.e., *atoms.frc* contains necessary lines with all information regarding potential and components of electric field. They can be further used to calculate energy of interaction and force between two atoms. We can change the grid size and scale according to the need.

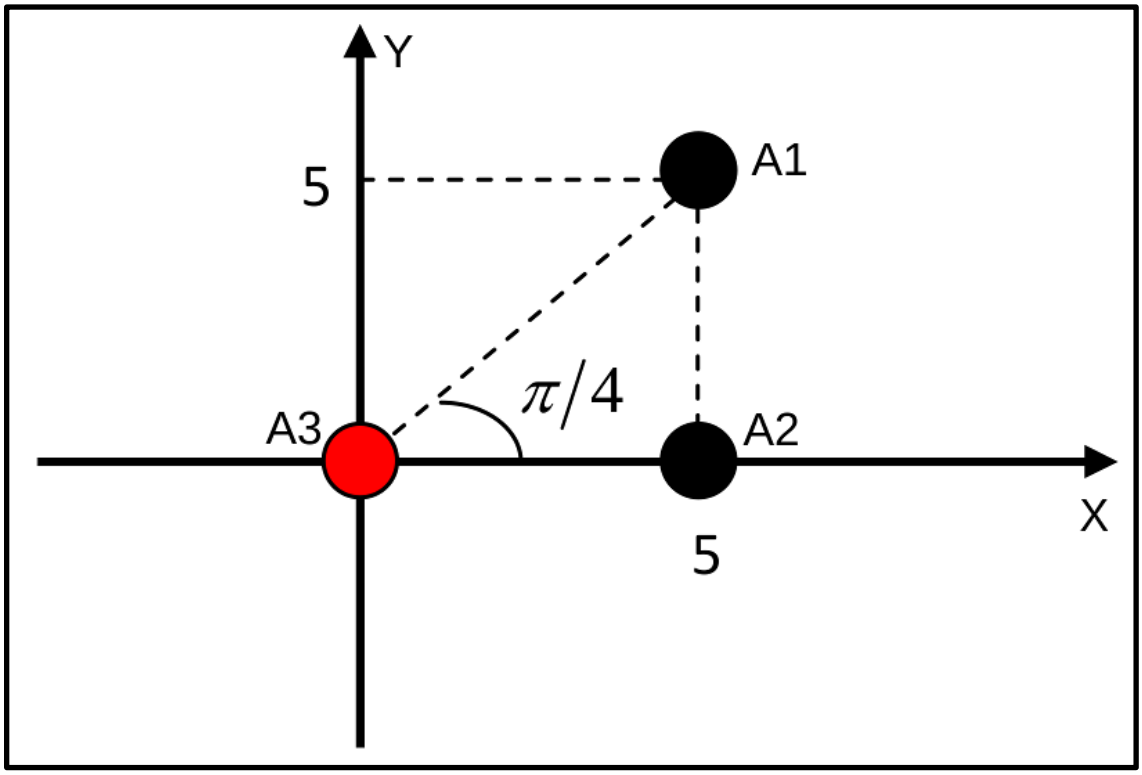

Figure 5.The schematic representation of setup of example system for electrostatic potential, energy of interaction and force between two sets of atoms (A3 due to A1 and A2).

CHAPTER THREE

STUDY OF PROTEIN-PROTEIN INTERACTIONS

Ab-initio binding of barnase–barstar with DelPhiForce steered Molecular Dynamics (DFMD) approach

3.1 Introduction

Protein-protein interactions (PPIs) are essential for numerous biological processes such as cell to cell interactions, metabolic control and development [47]. This importance is demonstrated by the fact that deviations of the wild-type patterns of PPIs and the corresponding biophysical characteristics of protein complexes were shown to be associated with diseases [23, 48-54]. Upon the binding a protein-protein complex is formed and depending on the participating monomers the complexes are classified as homo-dimers , hetero-dimers, enzyme-inhibitor and antibody-protein complexes [55], just to mention some. Furthermore, these complexes are characterized according to shape and size [56], complementarity between interfaces [57, 58], residue interface propensities [59, 60] including hydrogen bonding [61] and secondary structure [62]. Among the driving forces causing complex formation, the electrostatic interactions are the only long-range interactions and thus play primary role in guiding the monomers toward the binding. The electrostatics energies and forces were shown to play major role in different

molecular phenomena such as PPIs, ion binding [53, 63-65], protein folding and stability [66, 67], protein -DNA/RNA interactions [68-70], and protein-microtubule binding [48, 71-73].

Modeling PPIs is a two-fold problem: predicting the binding pose and revealing the binding trajectory. There are many approaches for predicting binding pose, including homology modeling [25] and docking [26], while much less attention has been paid to predicting the binding pathway. The last is very important for cases involving binding that requires large conformational changes involving backbone motion [27] and opening of the binding pocket [28].

The technique capable in principle of addressing both, the binding trajectory and the binding pose, is molecular dynamics (MD) simulations of PPIs [74]. Essential component in these MD simulations is the treatment of water phase: either with explicit solvent waters [75] or Generalized Born (GB) implicit model [76, 77]. Explicit solvent model is computationally too expensive when large complexes are involved, and monomers are separated at large distances. Alternative is the GB model, which is much less computationally demanding. However, the default setup of GB model involves cut-offs for both of Born radii and pair-wise energies calculations. So, if the distance between receptor and ligand is large than the cut-off distance, the default setup of GB model will result in omitting contributions of distant atoms and thus affecting the modeling [78]. Increasing the cut-offs in MD setup presumably should solve this issue, however it results in dramatical slowdown of the calculations and in addition, still does not model electrostatic interactions correctly, as demonstrated in ref. [78].

To address such a deficiency, in our previous work, we introduced a hybrid method, the DelPhiForce steered MD (DFMD), that takes advantage of the accurate calculations of electrostatic forces at each atom of the ligand and ports these forces into steered MD [78]. It was demonstrated that DFMD correctly predicts the binding trajectory (the entrance of the binding pocket) and binding pose of a small ligand (spermidine) bound to a receptor (spermine synthase) [78]. However, DFMD has never been tested on PPIs. Here we extend the applicability of DFMD by applying it to model binding of barstar to barnase. We show that DFMD is capable of guiding barstar-barnase complex formation irrelevant of the starting positions and orientations of separated monomers.

3.2 Materials and Methods

3.2.1 Delphi

Delphi is a popular software that solves the Poisson-Boltzmann equation (PBE), which uses finite difference method to compute electrostatic potential distribution throughout the modeling box [36, 79, 80].

3.2.2 DelPhiForce

DelPhiForce (http://compbio.clemson.edu/delphi-force) is available within Delphi package to calculate the electric field, forces, and energy of a two-molecular system. In our study it is used to compute the electric force between receptor and ligand. The receptor's atoms are charged using amber force field [81] and ligand's atoms are kept neutral. Then Delphi uses finite

difference approach to solve the PBE and computes the electrostatic potential at each atom of ligands due to charge in the receptor atoms, and from there to deliver the atomic electrostatic forces [42, 43].

3.2.3 DelPhiForce steered Molecular Dynamics (DFMD) simulations

We used DFMD [78] method to study the binding of ligand to receptor, the barstar binding to the barnase. The DFMD method uses an approach that combines long range electrostatic force via DelPhiForce and NAMD MD simulation package [82]. The ligand (barstar) was offset 30 Angstroms away from of its bound position to assure that the only none zero force is the long-range electrostatic force. Then the electrostatic force on each atom of barstar was calculated with DelPhiForce [42, 43]. The forces are then given to steered MD module of NAMD [82]. Simulations were done with Generalized Born implicit solvent model (GBIS) [83]. The ion concentration was 0.15M, solvent dielectric was 80 and temperature was maintained at 300 K using a Langevin thermostat. In addition, cut off was taken at 18 Å and periodic boundary conditions were applied with a cubical box of 150 Å. Constant pulling forces were applied in the simulations where the direction and magnitude of the steered forces were calculated by DelPhiForce [42, 43, 83]. The steered force was recalculated and updated every 500 steps of the simulations. The simulations were done using default value of steered electrostatic force range: low limit 0.2 kcal $mol^{-1}Å^{-1}$ and upper limit 1.18 kcal $mol^{-1}Å^{-1}$.

3.2.4 Selection and preparation of barnase and barstar monomers

The barnase-barstar complex structure was taken from Protein Data Bank (PDB ID: 1BRS)[84] and we selected chain 'C' as barnase and chain 'F' as barstar. The PDB file had two mutations, C80A and C82A with respect to wild-type and these mutations has been mutated back to wild-type residues. The reverse mutations were performed using UCSF Chimera [85] where rotameric state with highest occupancy in Dunbrach library of rotamers is utilized.

The distance between the center of mass (CoM) of barnase and barstar complex is calculated as 23.39 Angstrom using VMD [86]. The CoM of barnase is positioned at the origin of the reference frame and barstar is rotated around the axes passing through the CoM of barnase so that CoM of barstar resides on X- axis (Figure 6). The starting positions for each simulation are obtained by offsetting barstar 30 Å away from its bound position (DD = 30 Å) along X-axis. To generate alternative unbound orientations for barnase and barstar, 64 initial structures with different relative orientations of barstar and barnase were created. The first set was generated keeping barnase orientation fixed, while barstar was rotated around X- and Z- axes passing through its CoM in increments of 90 degrees. Then the same procedure is repeated for 4 orientations of barnase rotating it around Y-axis passing through its CoM (Figure 6a-d).

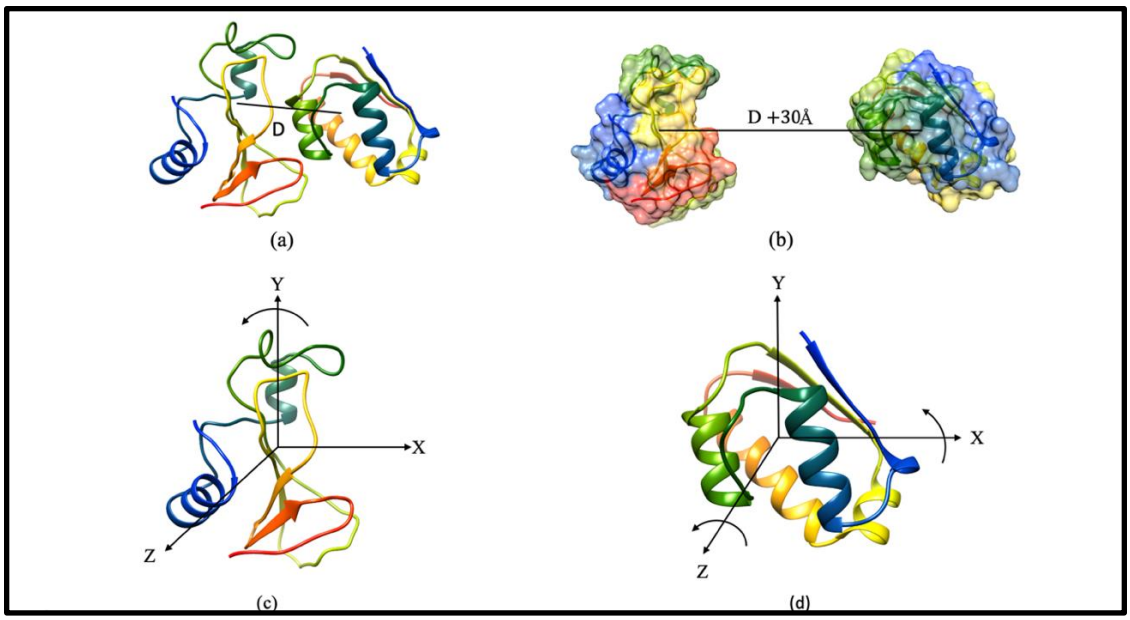

Figure 6. Selection and manipulations of barnase and barstar monomers: (a) experimental structure of barnase-barstar complex with center of mass (CoM) distance denoted by D (left barnase and right barstar); (b) barnase-barstar complex separated by DD=30 Å between centers of mass; (c) different initial orientations of barnase created by rotating barnase about Y axis through its center of mass with increments of 90 degrees steps; (d) different initial orientations of barstar created by rotating barstar about X- and Z- axes though its center of mass with increments of 90 degrees steps.

3.2.5 Ligand RMSD and interface RMSD

Ligand RMSD (L-RMSD) is the RMSD of ligand obtained through the superimposition of the receptor docking model to its crystallographic structure and then computing the RMSD of the ligand [87]. The interface RMSD (I-RMSD) is the backbone RMSD of interface residues of the ligand [88].

3.3 Results

Sixty-four independent DFMD simulations were carried out staring from different positions and orientations of both barnase and barstar. The success of the modeling was monitored via several quantities, mentioned in Methods section: L-RMSD, I-RMSD and the offset from crystallographic bound position (ΔD). Here we provide the results of the simulations by showing several showcase examples and then statistics of the entire set of simulations.

3.3.1 Showcase examples

It is not anticipated that each simulation will result in successful binding, because the simulation time is relatively short (5ns) and the binding process may involve transient binding at wrong places before the correct binding mode is found. Therefore, here we present two successful cases (Figure 7), simply to illustrate that DFMD can bind correctly relatively large proteins even the initial positions and orientations are completely away from the crystallographic ones (the same simulation without DelPhiForce assistance resulted in no binding). The first case is DFMD simulation with initial positions and orientations of monomers as: barnase fixed in its crystallographic position while barstar is moved away by ΔD = 30 Å along X-axis and rotated 90 degree around X-axis and 270 degree around Z- axis (see Figure 6 for more detail). In the second case the initial positions and orientations of monomers are: barnase rotated 90 degrees along Y-axis and barstar moved away along X-axis by ΔD = 30 A and rotated 270 degree along Z- axis. Both cases clearly represent difficult binding initial conditions, especially the second case where both binding interfaces are in completely wrong orientations. Nevertheless, 5ns DFMD simulations are quite successful as shown in Figure 7

The results show the L-RMSD approaches 10 Å limit, which is the cutoff used in Critical Assessment of Predicted Interactions (CAPRI) to indicate acceptable predictions [87]. It is important to observe that once the near-binding-mode is achieved, the L-RMSD does not change much and the system is quite stable (Figure 7a, b). The assessment of the success of simulation is also done via monitoring ΔD. Indeed, one sees that ΔD goes quickly to near zero, indicating the binding partners are situated at almost the same distance as seen in crystallographic structure (Figure 7 a,b). Furthermore, we monitored I-RMSD for both cases (Figure 7a, b). Similar observations as above can be made – the system quickly reaches a binding pose that is very similar to the crystallographic one and stays stable during the rest of simulations. The reason why the system does not exactly reach the crystallographic conformation is the internal flexibility of both the receptor and the ligand. They sample conformations that are several Å RMSD away from their crystallographic structures, an observation that have been reported by other researchers as well [89]. The final positions of barnase-barstar at the end of simulation compared to their native structures are also shown for guidance of the eye (Figure 7c-d).

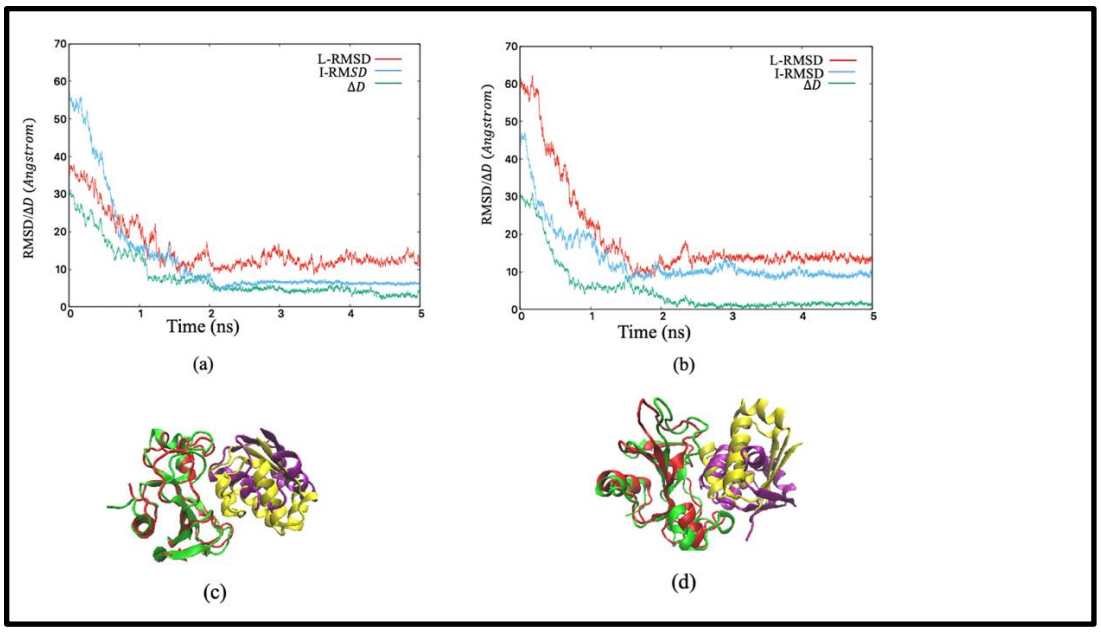

Figure 7. L-RMSD, ΔD and I-RMSD as a function of simulation time (a) L-RMSD , I-RMSD & ΔD of barnase-barstar at different time steps during the simulation with barnase fixed and barstar rotated 90 degree around X-axis and 270 degree around Z- axis passing through CoM; (b) L-RMSD, I-RMSD and ΔD at different time step during the simulation with barnase rotated 90 degree around Y-axis and barstar rotated 270 degree around Z- axis passing through CoM; (c,d): final position of barnase -barstar at the last step of simulation compared to experimental structure when the simulation is performed with initial orientations explained for (a) and (b) respectively: barnase colored green and barstar colored purple, while experimental structures are colored red (barnase) and yellow (barstar).

3.3.2 Statistical analysis

Statistical analysis shows that most of the simulations are successful despite of the short simulation time. To illustrate the time evolution of the quantities (ΔD, L-RMSD and I-RMSD) that are used in this work to access the success of simulations, we select several time-windows from the DFMD simulations (time-windows 2-3, 3-4, and 4-5ns). Each time window consists of

4000 frames. One anticipates that longer simulations will bring the ligand closer to the crystallographic binding position. Results are shown in Figure 8. Analyzing the offset distance, the ΔD (Figure 8a), one sees that the number of cases with very small ΔD is greater for the time-windows 3-4ns and 4-5ns compared with the time-window 2-3ns. This indicates that indeed as the time of simulations progresses the DFMD brings the partners together close to the experimental positions. Furthermore, Figure 8b shows the time evolution of L-RMSD in these three time-windows separately. Here the trend is not as clear as it is for ΔD. Practically there is no difference of the number of cases with low L-RMSD between the three time-windows. However, there is a tendency in the range of L-RMSD 60-90 Å, that longer simulations provide better outcome, i.e., less cases being with such large L-RMSD. It should be made clear that L-RMSD has three components: L-RMSD due to (1) offset distance, (2) wrong orientation of the ligand and (3) conformational changes of the ligand. In addition, due to the fact that the receptor also undergoes conformational changes during the simulations, the structural superimposition of crystallographic 3D structure of the receptor and the corresponding snapshot is contributing to L-RMSD as well. Finally, the results about I-RMSD are shown in Figure 8c. Here we see both effects mentioned above: an enrichment of case with low I-RMSD for time-windows 3-4ns and 4-5ns compared with case in the time window 2-3ns, and at the same time fewer cases in the regions 60-90 Å. The arguments about the factors contributing to I-RMSD are the same as outlined for L-RMSD.

Overall, the statistical analysis indicates that DFMD facilitates the binding and attempts to bring the apartments to their crystallographic positions in the complex. However, none of 64

simulations resulted in perfect match of the binding mode of barnase-barstar complex as seen in the crystallographic structure, i.e. ΔD = 0 Å, L-RMSD = 0 Å and L=RMSD = 0 Å. One plausible reason for that is in our protocol the structures are not rigid and they sample various conformations (away from crystallographic 3D structures) and thus even they are sampling the crystallographic binding mode, still the corresponding ΔD, L-RMSD and I-RMSD will be different from zero.

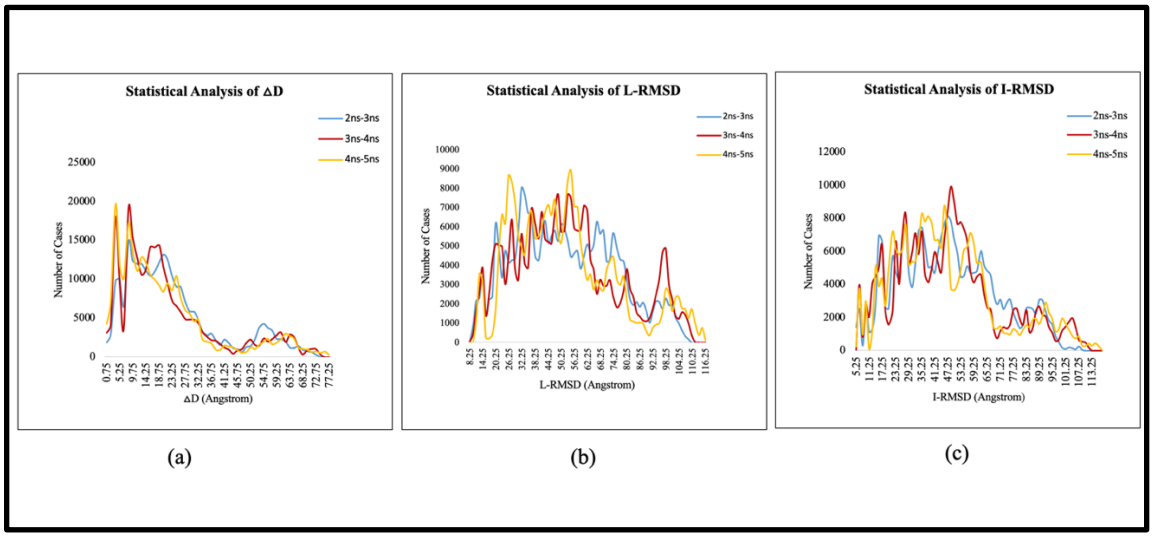

Figure 8. Statistical analysis of the results based on ΔD, L-RMSD and I-RMSD during different time- windows of the simulations (a) analysis of ΔD; (b) analysis of L-RMSD; (c) analysis of I-RMSD

3.3.3 Role of electrostatics

The work was done on diverse set of initial orientations and positions of both barnase and its ligand barstar. We did not observe preference of successful binding with respect to the

corresponding initial conditions. In some cases, as case 2 in the showcase section, the partners were positioned and oriented in completely wrong way (by wrong we mean positions and orientations completely different from crystallographic ones), and still DFMD was able to correctly bring them together very close to the crystallographic binding mode. What is the reason for that? **Error! Reference source not found.** illustrates two cases of initial conditions, an easy case where the barstar is simply moved away from its crystallographic position (**Error! Reference source not found.**a) and a difficult case, where barstar is moved away and rotated as well (**Error! Reference source not found.**b). In the first case (a) the binding interfaces are still facing each other and presumably the recognition should be easier. Indeed, as it can be seen in **Error! Reference source not found.**a, the resulting electrostatic force coming from barnase and acting on barsar is pointing toward the receptor, thus contributing to the association process. In the second case (**Error! Reference source not found.**b), the interfaces are not facing each other and if the barsar is simply pulled toward the barnase, the binding will be totally wrong. However, **Error! Reference source not found.**b indicates that the electrostatic play crucial role by providing a torque that makes barstar to rotate and thus to adopt correct orientation of the binding interface. Such an electrostatic torque effect was shown to be common for various binding partners [42].

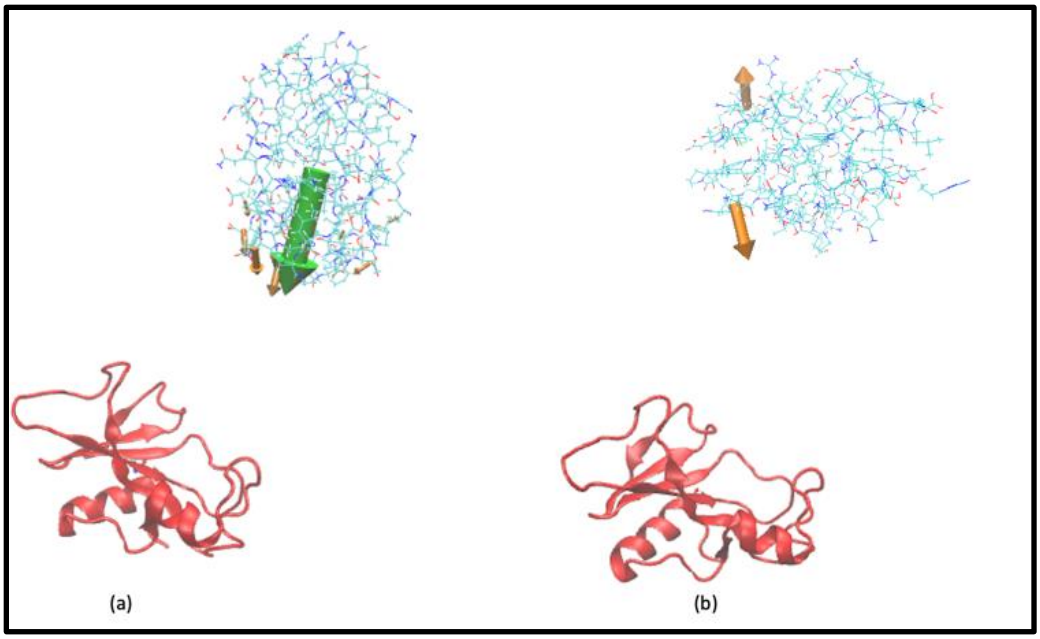

Figure 9. Electrostatic forces generated by barnase (cartoon representation) to barstar placed 30 Å away from its bound position. Orange arrows represent forces acting on each residue, while the green arrow is the total resultant force. The length of the arrow reflects calculated magnitude of the force. The side chain of residues of barstar are represented as line, of which blue and red positively charged residue and negatively charged residue respectively: (a) initial crystallographic structure of barnase-barstar separated by 30 Å from its bound position; (b) electrostatic torque formed when keeping barnase fixed, barstar is rotated around X-axis by 90 degree and around Z-axis by 270 degree and separated by 30 Å to its bound position.

3.3.4 Role of conformational flexibility and longer simulation time

It was mentioned above that none of the simulations resulted in a perfect binding mode. The corresponding quantities, ΔD, L-RMSD and I-RMSD were found to be different from zero in all 64 runs. To probe what is the plausible reason for that, we carried out traditional MD simulations

(without DelPhiForce) staring from the crystallographic structure of barnase-barstar complex. The results of 5ns simulations are shown in Figure 10a. It can be seen that ΔD, L-RMSD and I-RMSD are different from zero, confirming our initial thought that conformational dynamics associated with MD results in deviation from crystallographic structure and thus one cannot expect that DFMD modeling will reproduce the binding mode seen in X-ray experiment.

Furthermore, we tested if the DelPhiForce could be causing problems when monomers are in bound state. For this purpose, we carried out DFMD simulations starting from the X-ray structure of the complex (Figure 10b). One can compare Figure 10a and Figure 10b to observe that there is not significant difference. This indicates that DFMD in bound state delivers results similar to traditional MD and does not cause artifacts.

Finally, one may expect that longer simulations will provide better results. To quickly probe this, we took a particular case of binding mode resulting in small ΔD, L-RMSD and I-RMSD (taken from the last frame of the simulations shown in Figure 10b) and subjected it to additional 5ns NAMD. Results are shown in Figure 10c. Unfortunately, no improvement was observed. This suggests that DFMD should not be targeted to deliver the "perfect" binding mode, but rather to position the binding partners in a near-binding mode.

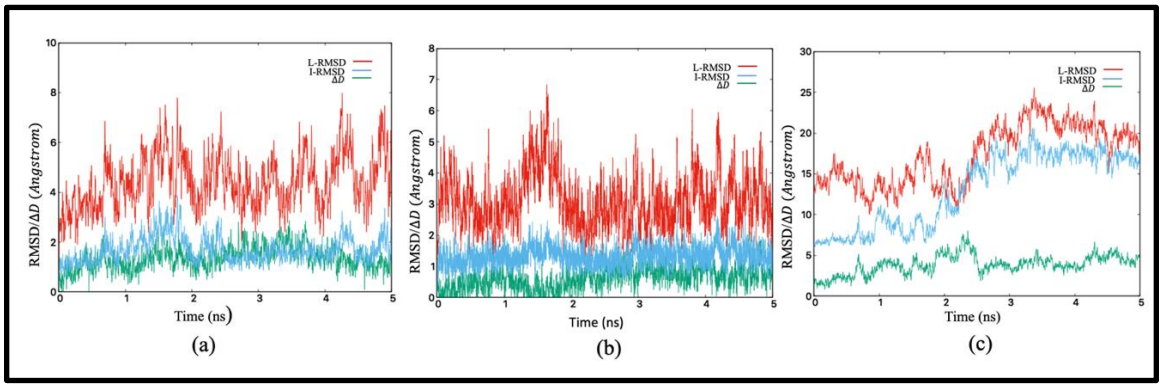

Figure 10. L-RMSD, I-RMSD and ΔD as a function of simulation time: (a) NAMD simulation of bound crystal structure; (b) DFMD simulation of bound crystal structure; (c) NAMD simulation of last frame of the successful DFMD simulation.

3.4 Conclusions

This work showed that DFMD is capable of assisting modeling of protein-protein binding and produces trajectories that describe various scenarios, including case at which the starting positions and orientations are completely different from the binding ones. At the same time, the approach is fast and computationally less expensive than traditional MD-based approaches.

CHAPTER FOUR

EFFECT OF DISEASE-CAUSING DNA VARIANTS IN PROTEIN STABILITY, DYNAMICS, AND INTERACTION

In-silico analysis to identify the role of MEN1 missense mutations in breast cancer

4.1 Introduction

Variants in *MEN1* gene has been found to have a strong association with multiple endocrine neoplasia type 1 that follows an autosomal dominant (AD) inheritance pattern in various tumor syndromes[90]. *MEN1* triggered neoplasia type1 is involved in the development of various tumors such as parathyroid adenomas, duodenopancreatic neuroendocrine tumors, and pituitary (anterior) adenomas with a 94% penetrance by the age of 40[91]. *MEN1* encodes for menin, a 610 amino acid oncosuppressor protein, found in the nucleus. Menin protein interacts with other proteins such as JunD/AP1, Smad3, NFκB, estrogen receptor (ER) and others to participate in various transcription and cell signaling processes[92].

The inheritance of germline MEN1 mutation such as a point mutation or deletion (leading to loss of heterozygosity (LOH)), disrupts the signaling pathway of menin protein, and thus predisposes an individual to develop tumor. The germline mutations occurring in MEN1 that truncate the menin protein result in dysfunctional product and thus affecting its role in tumor suppressor activity[92]. Studies based on mutagenesis performed previously indicate that the LOH of menin protein disrupts its interaction with histone methyltransferase (HMT) mixed lineage leukemia protein (MLL). This mechanism in turn impacts histone associated trimethylation (H3K4me3) transcription activation process thus fails to suppress the development of endocrine tumors[93].

The oncosuppressor menin protein plays a vital role in suppression of ER-positive breast cancer as well[94]. Previous studies outline that the upregulation of menin protein which binds to the ER and thus enhances the ER activity which implicates that it acts as a growth stimulator in ER-positive breast cancer[95]. Menin also regulates ERα-mediated transcription by enhancing H3K4 (histone 3 lysine 4) methylation process leading to breast carcinogenesis and various other tumor progression mechanisms[96]. Loss of heterozygosity (LOH) and disruption in menin protein expression in a Dutch cohort strongly shows involvement of MEN1 in breast cancer carcinogenesis[97].

The goal of this study is to classify missense MEN1 variants reported to be identified in breast cancer tissues from COSMIC (Catalogue of Somatic Mutations in Cancer) database[98] with unknown pathogenicity. This is done by applying our combinatory in-silico predictor[99] approach to annotate the pathogenicity of these variants. We also analyzed the impact of these variants on biochemical and biophysical properties of the menin protein along with the in-house built menin-ER and menin-MLL complexes. In parallel the same is done for curated pathogenic variants causing neoplasia. Comparing the observations made for mutations associated with neoplasia and variants in COSMIC database along with predictions made by our in-silico predictor, we provide classification of pathogenic menin mutations associated with breast tumorigenesis.

4.2 Materials & Methods

4.2.1 Selection of the MEN1 variants

We collected 19 unique missense variants in MEN1 gene (GRch37 version) from COSMIC database that are reported to be identified in breast cancer samples. Among these 19 variants, only 2 are listed in ClinVar[100]and they are annotated as variants of unknown significance (VUS).

In parallel, we obtained 8 MEN1 missense variants associated with multiple endocrine neoplasia type1 from ClinVar database categorized as pathogenic. We will use these pathogenic mutations to compare their effect on menin stability and interactions with the effect caused by variants taken from COSMIC database and 5 benign variants from VariSNP database[101], since there were no benign variants reported in ClinVar.

Table 1 summarizes the list of all variants used in this study.

Table 1. List of all the variants used in our study. The MEN1 missense variants associated with breast cancer are obtained from COSMIC database, whereas the MEN1 pathogenic missense variants associated with Multiple Endocrine Neoplasia are obtained from ClinVar (that is used as a benchmarking set)

Protein Change	Clinical Significance	rs# (dbSNP)	Phenotype List	Source
Breast Cancer Associated Variants (COSMIC)				
S606C	Unknow	.	Breast Cancer	COSMIC
S606Y	Unknow	.	Breast Cancer	COSMIC
L605V	Unknow	.	Breast Cancer	COSMIC
M563I	Unknow	.	Breast Cancer	COSMIC
K562M	Unknow	.	Breast Cancer	COSMIC
P534A	Unknow	.	Breast Cancer	COSMIC
T530I	Unknow	rs750591216	Breast Cancer	COSMIC
R490Q	Unknow	.	Breast Cancer	COSMIC
Q398E	Unknow	.	Breast Cancer	COSMIC
A289E	Unknow	.	Breast Cancer	COSMIC
L272R	Unknow	.	Breast Cancer	COSMIC
S258L	Unknow	rs386134259	Breast Cancer	COSMIC
V220M	Unknow	rs794728621	Breast Cancer	COSMIC
H204Y	Unknow	.	Breast Cancer	COSMIC
E200K	Unknow	.	Breast Cancer	COSMIC
C170Y	Unknow	.	Breast Cancer	COSMIC
R108P	Unknow	.	Breast Cancer	COSMIC
L89R	Unknow	.	Breast Cancer	COSMIC
L36F	Unknow		Breast Cancer	COSMIC
Neoplasia Associated Variants (ClinVar)				
W441R	Pathogenic	rs104894259	Multiple endocrine neoplasia, type 1	ClinVar
A373D	Pathogenic	rs1555164707	Multiple endocrine neoplasia, type 1	ClinVar
E260K	Pathogenic	rs104894268	Multiple endocrine neoplasia, type 1	ClinVar
V189E	Pathogenic	rs104894262	Multiple endocrine neoplasia, type 1	ClinVar
A165P	Pathogenic	-	Multiple endocrine neoplasia, type 1	ClinVar
R280K	Pathogenic	-	Multiple endocrine neoplasia, type 1	ClinVar
D423N	Pathogenic	rs104894264	Hereditary cancer-predisposing syndrome; Multiple endocrine neoplasia, type 1	ClinVar
H139D	Pathogenic	rs104894263	Hereditary cancer-predisposing syndrome; Multiple endocrine neoplasia, type 1	ClinVar

MEN1 Benign Variants (VariSNP)				
P519S	Benign	rs150202288	NA	VariSNP
V555L	Benign	rs562257963	NA	VariSNP
I377T	Benign	rs115859693	NA	VariSNP
E371D	Benign	rs149383809	NA	VariSNP
G508D	Benign	rs375804228	NA	VariSNP

4.2.2 In-house algorithm for pathogenicity prediction

Recently we reported a combinatory in-silico predictor of pathogenicity[99] which combines the 8 best pathogenicity prediction algorithms such as Polyphen2 (Polymorphism Phenotyping V-2)[102], LRT (Likelihood ratio test)[103], MetaSVM[104], VEST3[105], PROVEAN (Protein Variation Effect Analyzer)[106], REVEL[107], Eigen[108], CADD (Combined Annotation Dependent Depletion)[109]. We used this method to characterize the pathogenicity of the missense variant.

4.2.3 Annotating MEN1 missense variants

We also annotated these variants with allele frequency information from 1000 genome[110] and Genome Aggregation Database (gnomAD)[111] to see the distribution of these variants among various population groups. We also annotated the variants with GERP++ scores [112] to evaluate the evolutionary conservation on a functional sequence. We used this score along with another biochemical feature, the folding energy change, to build an in-house classification method (using KNN classification method) to characterize the pathogenicity of missense variants.

4.2.4 3D model of Menin protein and generation of Menin-ER and Menin-MLL complexes

The crystal structure of menin protein was obtained from Protein Data Bank (PDB)[113]. The PDB file (ID: 3U84) contains two chains, chain A and chain B, and has missing residues. The structure was re-modeled using SWISS-model[25] resulting in a full-length protein. (Figure 11A). The structure of ER was modeled with SWISS-modeling using its amino acid sequence taken from Uniprot (ID: Q99527)[114] and a template (PDB ID: 4ZNH) with high sequence similarity (sequence similarity of 0.59) to build the model (Figure 11B). The structure of MLL was taken from PDB file (ID:4GQ6). Then the structures of menin, ER and MLL were used to model the complexes menin-ER and menin-MLL. For this purpose, we used the docking methods ZDOCK[115] and HawkDock[116]. Among the different models given by both methods, one best model is chosen from each method, each with similar position and near to the binding site of the protein. The binding site is predicted using web server meta -PPISP[117] which gives consensus results based on three individual web servers: cons-PPISP[118], PINUP[119] and Promate[120] (Figure 12A-D).

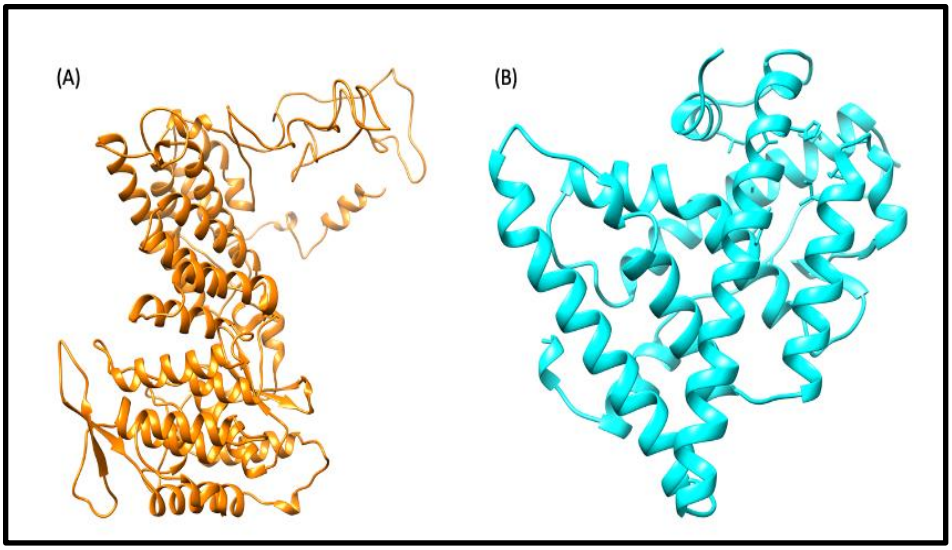

Figure 11. 3D Protein Structure(A) 3D structure of menin; (B) 3D structure of Estrogen Receptor

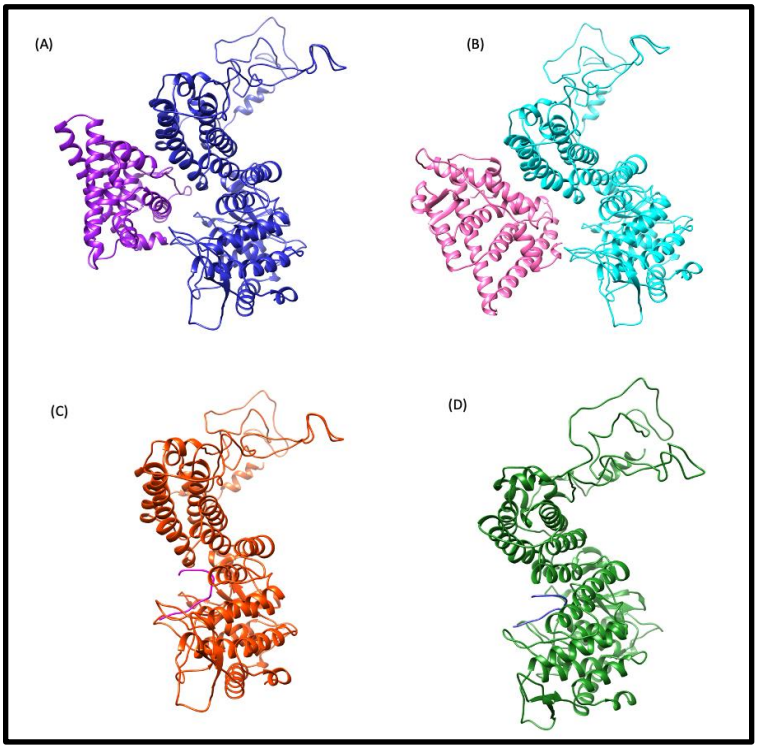

Figure 12. 3D structure of menin –ER and menin –MLL complex (A) 3D structure of menin-ER complex from ZDOCK (blue colored chain menin and purple colored chain ER; (B) 3D structure of menin-ER complex from HawkDock (cyan colored chain menin and pink colored chain ER); (C) 3D structure of menin-MLL complex from ZDOCK (orange colored chain menin and purple colored chain ER); (D) 3D structure of menin-MLL complex from HawkDock (green colored chain menin and blue colored chain ER).

4.2.5 Computing folding free energy change due to mutation

Folding free energy change due to mutation was computed via in-house tool, the SAAFEC[44] algorithm, along with third party webservers as DUET[121], CUPSAT[122], mCSM[123], SDM[124] and I-Mutant 2.0[125].

4.2.6 Computing binding free energy change due to mutation

The change of the binding free energy of menin-ER and menin-MLL complexes was computed with in-house algorithm, the SAAMBE[45] method, along with third party tools as BeAtMuSiC[126], mCSM-PPI2[127] and MutaBind[128].

4.2.7 K-Nearest Neighbors (KNN) classification

K-nearest neighbors algorithm was used to classify the missense mutations using biochemical property along with rate of evolutionary conservation. The dataset includes total of 32 missense mutations. We split this dataset set into two subset (1) train dataset that comprises of 13 (5 Benign and 8 pathogenic) mutations; (2) test dataset consists of all the 19 mutations from COSMIC database (non-classified/unknown effect mutations). The KNN classification was performed using R program and various numbers of K values were tested to obtain the best performance on the training dataset.

4.3 Results & Discussion

4.3.1 Predicting pathogenicity of MEN1 variants

To evaluate the pathogenicity of the 19 MEN1 missense variants taken from COSMIC database we utilized the in-house combinatory approach described in method section[99]. Table 3 (Appendix B) summarizes the score and the assessment from the approach. According to this, 15 out of 19 variants are classified as "***pathogenic***" and 4 variants are categorized as "***benign***". In order to add more evidence to our categorization we mined Uniprot database to see if any of 15 predicted pathogenic variants has reported to have damaging effect on the menin protein, including associations with other disorders. We noticed that some of the variants from COSMIC database classified as pathogenic (by our in-house combinatory approach) such as V220M and A289E are also listed in Uniprot database as causing damage to menin protein and resulting in neoplasia. This provides additional confidence that our predictions are correct and indeed V220M and A289E are pathogenic mutations.

We did not find any allele frequency information associated with 18 variants except for one variant (T530I) categorized as "***benign***" that shows an allele frequency of <5% from both 1000 genome[110] and gnomAD[111] databases. The results from our conservation analysis from GERP++[112] on our dataset, show that all the pathogenic variants have evolutionary conservation rate of larger than 4, while the benign variants have evolutionary conservation rates ranging between 2 and 4.6. Studies have shown that high rates of evolutionary conservation

along with other factors such as biochemical changes could contribute to the disruption of protein function[129].

4.3.2 Mapping the missense mutations onto the Menin-MLL & Menin-ER complex

All the 19 variants taken from the COSMIC database, out of which 15 were predicted to be pathogenic and 4 to be benign, are mapped in the complex structures of menin-ER and menin-MLL complex, along with neoplasia-causing mutations and 5 benign mutations from VariSNP database[101] (Figure 13). It can be seen that all the variants are scattered within the structure and no variants are found near the interface of the protein complexes, both menin-ER and menin-MLL complexes. Neoplasia mutations are found in helixes, beta sheet and loops whereas, COSMIC pathogenic mutations are found in helixes and loops and benign mutations are found in loops except mutation L605Y which is found in helix.

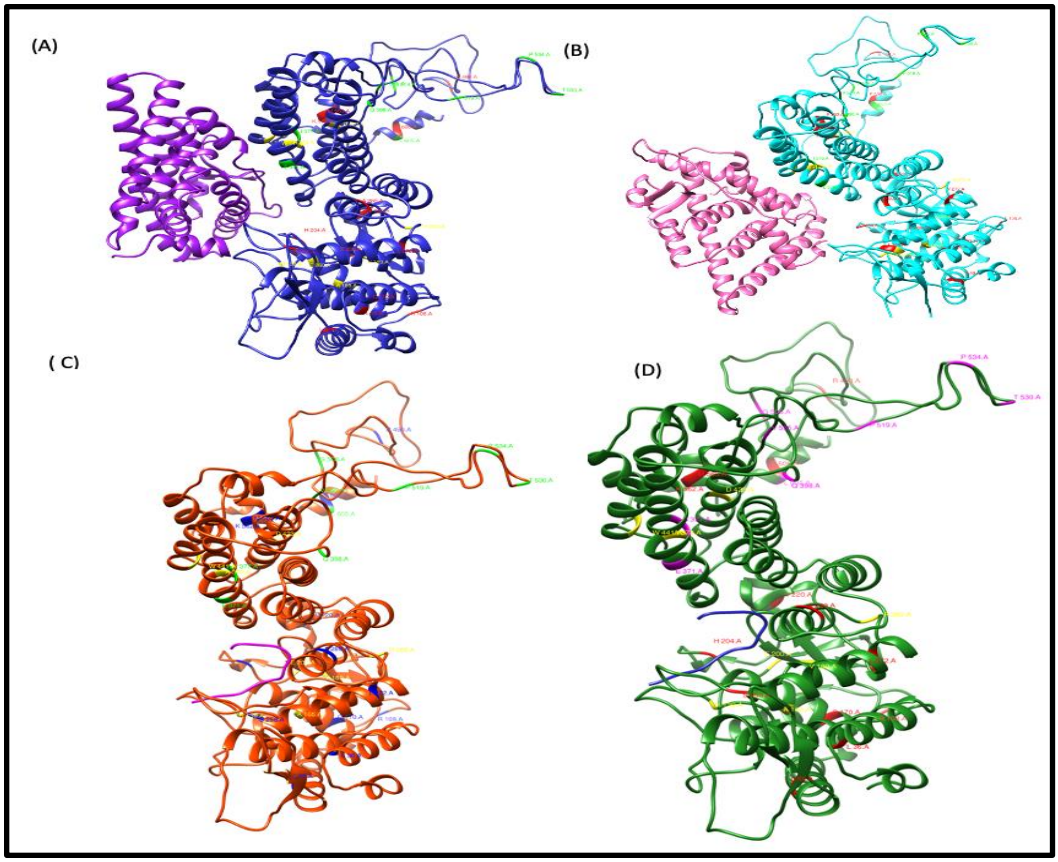

Figure 13. Mapping of pathogenic and benign mutations to menin-ER and menin- MLL complexes (A) mapping of mutations to menin-ER complex from ZDOCK (red colored pathogenic mutations from COSMIC , yellow colored pathogenic mutations associated with Neoplasia from ClinVar, green colored benign mutations from COSMIC and VariSNP); (B) mapping of mutations to menin-ER complex from HawkDock (red colored pathogenic mutations from COSMIC , yellow colored pathogenic mutations associated with Neoplasia from ClinVar, green colored benign mutations from COSMIC and VariSNP); (C) mapping of mutations to menin-MLL complex from ZDOCK (blue colored pathogenic mutations from COSMIC, yellow colored pathogenic mutations associated with Neoplasia from ClinVar , green benign mutations from COSMIC and VariSNP), (D) mapping of mutations to menin-MLL complex from HawkDock (red colored pathogenic mutations from COSMIC, yellow colored pathogenic mutations associated with Neoplasia from ClinVar ,magenta colored benign mutations from COSMIC and VariSNP).

4.3.3 Effect of missense mutation on protein stability

We find that most of the mutations destabilize the menin protein Table 4 (Appendix B). The changes in stability are greater for the pathogenic mutations compared with the benign mutations. Out of fifteen pathogenic mutations, eight are predicted to cause folding free energy change greater than or equal to 1.50 kcal/mol, four to result in folding free energy change less than 1.5 kcal/mol and three to less than 1 kcal/mol. Out of nine benign mutations, only one is predicted to cause change in folding free energy greater than 1.5kcal/mol, six are predicted to cause change in folding free energy less than 1 kcal/mol and two greater than 1kcal/mol and but less than 1.5 kcal/mol. The results indicate that pathogenic mutations tend to cause larger changes of folding free energy compared with predicted for benign mutations. Some of the pathogenic mutations as A289E, R108P and L36F are predicted by all predictors to result in similar change of stability which provides confidence that they are correct. These results can be compared with the change in folding free energy due to pathogenic mutations associated with neoplasia reported to ClinVar (Table 5, Appendix B). It is shown out of eight mutations associated with neoplasia, six have folding free energy change greater than 1.5 kcal/mol and two have less than 1.5 kcal/mol. This shows that pathogenic mutations associated with neoplasia and mutations from COSMIC database predicted to be pathogenic have similar effect on protein stability.

4.3.4 Effect of missense mutation on binding affinity

Table 6 and Table 7 (Appendix B) show the change in binding free energy due to mutations in menin-ER and menin-MLL complexes, respectively. For both complexes, all the mutations

destabilize the protein binding but the changes in binding energy are greater for pathogenic mutation compared with benign mutations. Out of 15 pathogenic mutations, 9 from menin-ER complex are predicted to cause binding free energy to change greater than 0.5 kcal/mol and 6 to less than 0.5 kcal/mol. In menin-MLL complex, 11 mutations are predicted to cause change in binding free energy greater than 0.5 kcal/mol and only four to less than 0.5kcal/mol. Mutations as A289E, C170Y and L272R are predicted to have largest effect on binding free energy change for all methods. Most of the benign mutations of both complexes are predicted to cause change in binding energy less than 0.5kcal/mol which is lower in comparison to pathogenic mutations. These can be compared with the change in binding free energy due to mutations associated with neoplasia reported in ClinVar (Table 8 and Table 9, Appendix B). It is shown that out of eight pathogenic mutations associated with neoplasia in menin_ER complex, four are predicted to cause binding free energy to change greater than or equal to 1.0 kcal/mol, three to cause change in binding free energy greater than 0.5 kcal/mol and less than 1.0 kcal/mol and one to less than 0.5kcal/mol. For menin-MLL complex, five are predicted to cause binding energy change to greater than 1.0 kcal/mol, two to greater than 0.5 kcal mol and less than 1.0 kcal/mol and one to less than 0.5kcal/mol. This shows the similar binding effects of pathogenic mutations associated with neoplasia and mutations predicted to be pathogenic and associated with breast cancer from COSMIC database.

4.3.5 Biophysical properties-based pathogenicity classification using KNN model

Here we use KNN method along with biophysical features such as folding free energy ($\Delta\Delta G$) and conservation score (GERP++) to characterize the 19 MEN1 missense variants from

COSMIC database. The dataset has a total of 32 mutation that encompasses 23 pathogenic mutations and 9 benign mutations. These 32 mutations were partitioned into 13 mutations (composed from ClinVar and VariSNP data sources) and 19 mutations with unknow significance (obtained from COSMIC database) and then subjected to the KNN classifications. The 13 well curated mutations were randomly split into 80% as train dataset and 20% as test data and we preformed 10-fold cross validation. At 100% accuracy we identified the best K value is 5. Using the validated KNN classification model with a K = 5 we performed the categorization of the 19 mutations with unknown significance with both features (GERP++ and ΔΔG) and just folding free energy (ΔΔG). Table 10 (Appendix B) summarizes the predictions results from our KNN classification and compares with the results obtained from our combinatory approach. When we use the both the GERP++ scores and ΔΔG as features, two mutations that were predicted benign, Q398E and L605V, in our combinatory in-silico approach were predicted to be pathogenic. But when only ΔΔG was used as a feature, three mutation, S606Y, H204Yand M563I that was predicted as benign using KNN classification differs from the results our combinatory in-silico method prediction. Previous studies on contribution of biophysical property such as folding free energy change (ΔΔG) has been used characterization of the pathogenicity of the missense variants in *MEN1*[130]. Our results from KNN classification shows that the biophysical property such as folding free energy along with evolutionary conservation rates provide more evidence in characterizing the mutation deleteriousness.

4.4 Conclusions

The study investigated the effects of MEN1 missense mutations and their role in breast cancer. Our utilization of combinatory in-silico predictor approach characterized the pathogenicity of the 19 missense variants with unknown clinical significance taken from COSMIC database. The results from the biophysical analysis using the menin protein along with its complexes (menin-MLL and menin-ER) aided in providing more evidence to the pathogenic impact caused by the variants in breast cancer tissues. The results from our in-house KNN classification method show strong correlation with the results from our in-silico combinatory approach. Furthermore, the benchmarking of the results using the neoplasia exclusive variants indicates that mutations clearly alter the protein stability of menin, and the changes are comparable with predicted pathogenic mutations taken from COSMIC database. Taken together, we provide classification of variants found in breast cancer tissues as pathogenic and benign based on several features.

CHAPTER 5

pH DEPENDANCE OF STABILITY DUE TO MUTATIONS

Computational Investigation of the pH Dependence of Stability of Melanosome Proteins: Implication for Melanosome formation and Disease

5.1 Introduction

The pH of a solution is an important characteristic for many biological processes. On a molecular level, the pH controls macromolecular stability and at extreme pH (acidic or basic extremes) macromolecules unfold. Typically, for every macromolecule, there is a particular pH at which the macromolecule is the most stable and activity is maximum, termed the pH-optimum[68, 131]. Macromolecular interactions are also pH-dependent[31, 132, 133] and typically there is a pH-optimum at which the binding affinity is maximum[134]. Within a cell, subcellular compartments have different pHs, reflecting their function, from low pH in lysosomes to high pH in peroxisomes. Thus, macromolecules tend to have a pH-optimum that is ideal for the pH of the subcellular compartment where they reside[135]. Increasing the scale of this idea, pH plays a crucial role for body organ function and varies form very acidic in the stomach to neutral in the blood. All above examples indicate that the regulation and maintenance of pH is essential for many biological phenomena.

pH is maintained in a given cellular compartment by channels and/or pumps either directly trafficking H^+ or indirectly providing environments that affect local H^+ concentration. These channels and/or pumps can be termed positive (increase pH) or negative (decrease pH)

regulators[136, 137]. Reaching and maintaining the desired pH depends on the balance of H^+ flux controlled by these regulators, including the passive transport across the membrane (Figure 14). One would expect that the positive regulators have activity at acidic pH and almost no activity at basic pH since their role is to increase pH from low to high pH. The converse would be expected for negative regulators; activity increases as the pH rises. At a particular pH, the inward and outward flux of H^+ induced by positive and negative regulators become equal and the pH set-point is established (Figure 14).

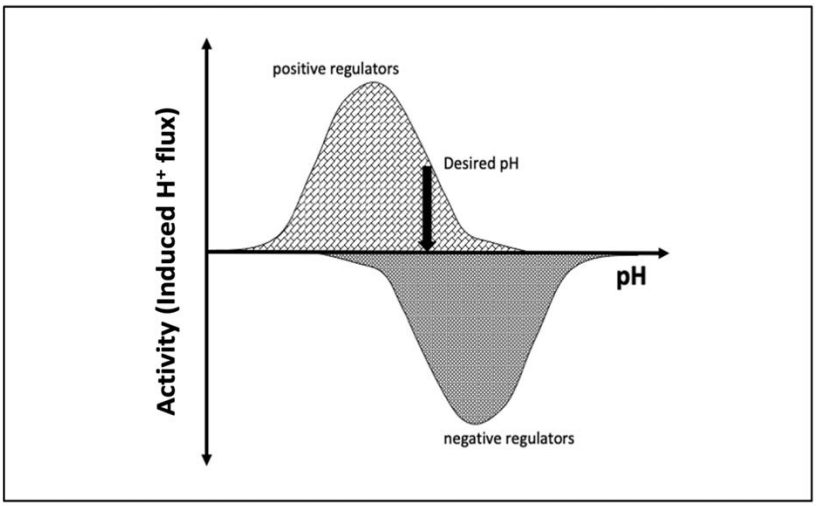

Figure 14. Schematic representation of the induced H^+ flux of positive (increase pH) and negative (decrease pH) regulators. With vertical arrow we indicate the desired pH, at which the total induced H^+ flux is zero.

Melanocytes are a specialized cell type that resides in the skin, eyes, brain, ear, heart, lung and adipose tissue[138]. One of the primary functions of melanocytes is the production of

melanin, a polymer of tyrosine derivatives that has important chemical properties in a wide range of tissues[139]. Melanin is synthesized in a specialized organelle called the melanosome. The pH of this organelle varies during the development of the organelle (a multistage process called maturation) and contributes to common pigmentation variation in human skin, hair, and eye color. Biallelic rare variants in proteins critical for the production of melanin, hair, and eyes, (TYR) or in pH regulation of the melanosome (OCA2 and SLC45A2) lead to a significant reduction in melanin pigmentation in skin and give rise to Oculocutaneous albinism (OCA) (OCA1, OCA2 and OCA4 respectively). Melanin synthesis is critical for the protection of the skin and eyes from ultraviolet radiation and a reduction in melanin synthesis increases the risk of cancer. Furthermore, a dramatic reduction of melanin production in the eye is also correlated with foveal hypoplasia, reduced visual acuity, and photophobia among individuals with OCA[140]. Taken together, the link between altered melanin pigment production and disease is well documented, however it remains poorly understood how the pH set-point of this organelle affects protein function which is critical for melanin synthesis or organelle pH maintenance [141].

Melanosomes originate from the endosome (Figure 15); thus, early melanosomes have a low pH (~ 3–4), whereas in the latter stages the pH reaches a near neutral pH of about 7. The near neutral pH of the mature melanosome is thought to provide a favorable environment for tyrosinase (TYR), the rate-limiting melanin synthesizing enzyme[142-144]. The change in pH during melanosome maturation is thought to be controlled by several membrane proteins[137] (e.g., OCA2, SLC45A2, and TPC2/TPCN2) (Figure 15). OCA2 and SLC45A2 are presumed to be positive pH regulators, while TPC2 is considered to be a negative pH regulator. Based on the

proposed role of positive and negative pH regulators (Figure 14), we anticipated that these proteins have different pH profiles of stability and activity. In addition, there are other melanosome proteins important for melanin synthesis (e.g., the ATP7A protein, which is altered in individuals with Menkes disease, and which supplies Cu^{2+} to the melanosome for TYR catalytic activity) that may exhibit pH-dependent stability and activity. It can be expected that the ATP7A protein, which supplies copper to TYR, should have a similar pH-dependence on activity as compared to TYR[145, 146].

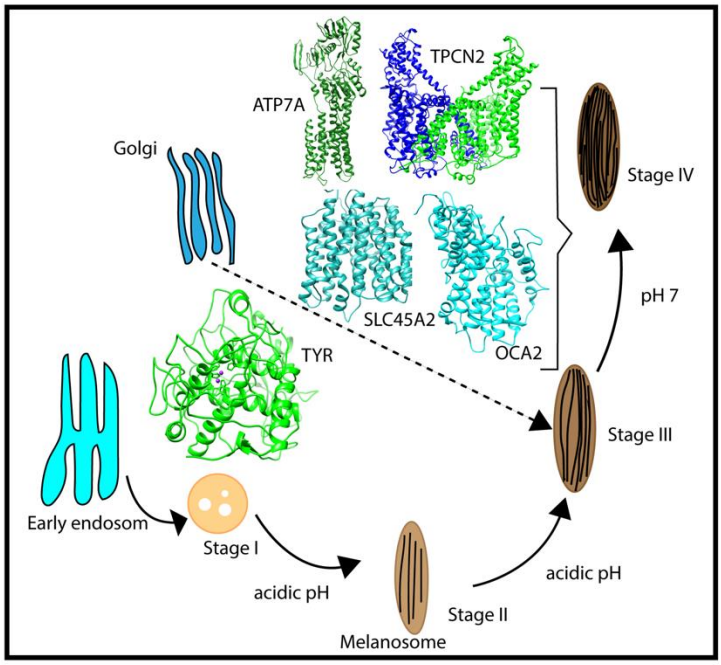

Figure 15. Schematic representation of the multistage processes of melanosome formation and proteins participating in pH regulation and melanin synthesis. The characteristic pH for each melanosome stages are indicated in the figure as well.

We anticipate that OCA2 and SLC45A2 have maximal activity at acidic pH, while TPC2 has maximal activity at basic pH. OCA2 plays a major role in human lighter skin and hair pigmentation and blue vs brown eye color variation[147] and regulates melanosomal pH and

maturation[147-149]. It may also be involved in small molecule transport for the biosynthesis of melanin[150, 151]. SLC45A2 also acts as a melanocyte differentiation agent and participates in the transport of substances required for melanin biosynthesis[149, 152, 153]. TPC2 affects pigmentation by regulating melanosome pH and size by mediating Ca^{2+} release from the organelle[154, 155].

Thus, understanding how melanosomal pH affects the activity of these proteins is essential. Furthermore, these proteins are commonly mutated in disease and those variants may impact the normal pH-optimum of these proteins. Predicting pH-optimum of activity is not an easy task and requires modeling of the details of the corresponding biochemical reactions as a function of pH. Here, we take advantage of the observation that pH-optima of activity and stability typically are the same as indicated in our earlier work[156]. Thus, our goal is to computationally determine the pH-dependence of stability of OCA2, SLC45A2, TPC2, TYR, and ATP7A proteins. Furthermore, we analyze the effects of common pigmentation and disease associated variant alleles in these proteins on the pH-dependence of their stability.

5.2 Results

As pointed earlier, in this work we focus on several proteins participating in melanosome formation, with the goal to contrast their stability pH-dependence and the effect of pathogenic variants. We use the observation made in our previous work that pH-optimum of activity is correlated with the pH optimum of stability[29]. This allows us to speculate that the results obtained on stability can be inferred to activity of these proteins. We present the results according

to the classification of the proteins as “positive” and “negative” regulators and probe our hypothesis that “positive” regulators should have lower pH-optimum as compared with “negative” regulators. Firstly, we present the results of the wild type proteins and sequentially the results about the mutants.

5.2.1 pH dependence of folding free energy on wild type proteins

For each wild type protein, the magnitude of the “constant” in equation (2) is unknown, because there is no experimental data of the folding free energy at a given pH for any of the proteins modeled in this work. Because of that it was set to be zero at the beginning of the simulated pH interval, pH=4.0. Here we present the calculated pH-dependence of the folding free energy using energy minimized structure (Figure 16) and we averaged results over 20 snap shots taken from MD simulations (Figure 25, Appendix C). We do not focus heavily on the results obtained with MD snap shots because DelPhiPKa was developed to calculate pKa’s of ionizable groups using static structures. However, we probe the sensitivity of the results using MD snap shots to investigate the role of plausible conformational changes on the pH-dependence of stability. We see no significant difference of the results obtained with energy minimized structure and averaged results over 20 snap shots which suggests that there are no structural changes contributing to the stability pH-dependence.

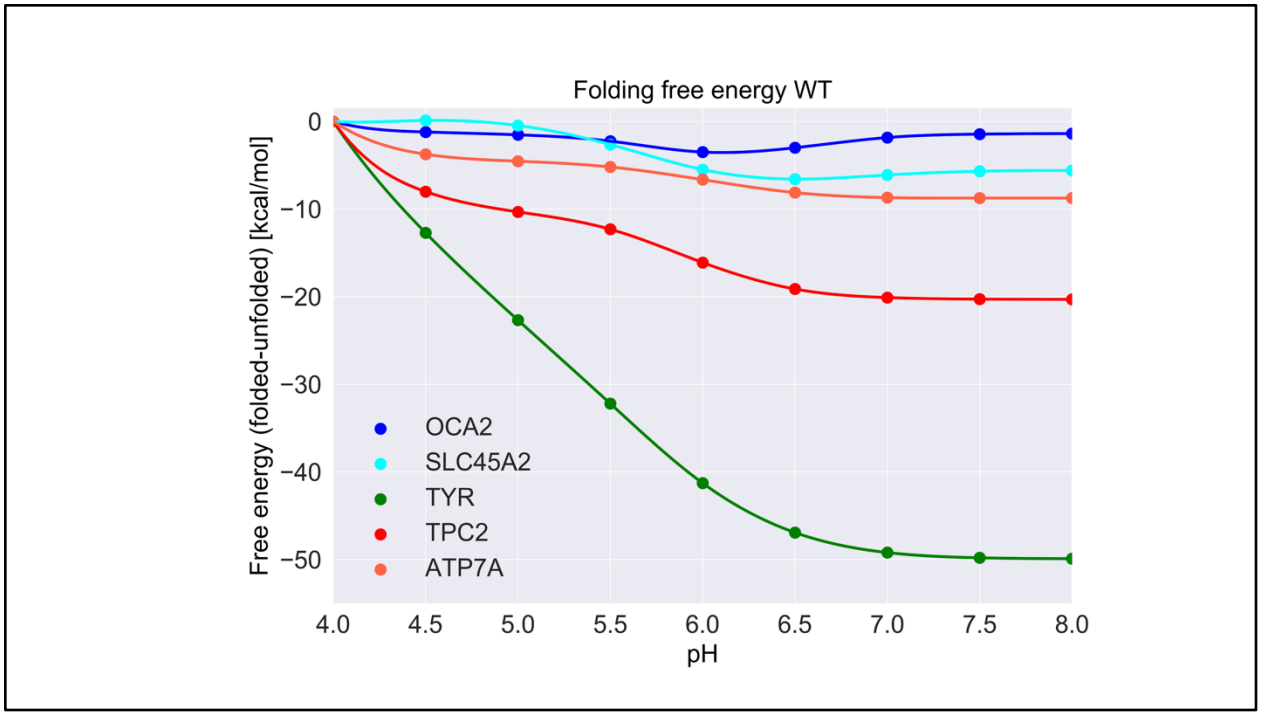

Figure 16. The pH-dependence of the folding free energy of wild type proteins from minimized structures within pH range 4-8.

TYR has the highest pH-optimum of stability, the pH-optimum about 8.0 or higher (Figure 16). In contrast, OCA2 protein (the positive regulator protein) has the lowest pH-optimum of stability, pH-optimum about 5.0 – 6.0. The other positive regulator, the SLC45A2 protein, also has pH-optimum lower that neutral pH, pH-optimum of 6.5. Presumed negative regulator TPC2 and the ATP7A protein which supplies copper to TYR both have pH-optima close to neutral pH. Thus, there is distinctive pH difference of the stability of OCA2, SLC45A2, TYR, ATP7A, and TPC2. Furthermore, the modeling confirms the experimental observation that TYR is most active at neutral pHs and with reduced activity at acidic pHs[144].

The origin of pH-dependence of the folding free energy is the difference of the pKa's of ionizable groups in folded and unfolded state. Thus, if the pKa's in folded and unfolded states are the same, there will be no pH-dependence. Furthermore, even if they are different but are outside the pH region of interest, the pH-dependence of the folding free energy will be affected as well. It is not expected that the pKa's of titratable groups in the unfolded state will be perturbed from standard pKa values[157], and thus most of the pH-dependence of the folding free energy should originate from perturbed pKa's in folded state. However, for completeness, in Claculated PKa's

(Appendix C), we provide the calculated pKa's for both states, folded and unfolded. Indeed, one can see that for "positive" regulators most of perturbed pKa's are for acidic groups, thus resulting in pH-dependence at low pH. In contrast, most of perturbed pKa's for TYR, ATP7 and the "negative" regulator TCP2 are of His residues, resulting in pH-dependence at neutral pH.

5.2.2 Effect of pathogenic variants on protein stability

Error! Reference source not found.. shows the average change of folding free energy due to variants based on predictions made using the methods described above. The low standard deviations reported reflects the consistency of results obtained with different tools. Most of the variants appear to destabilize the proteins by a modest amount. However, some mutants, such as A481T and N489D in OCA2, C1002F and I1264V in ATP7A, are predicted to significantly effect protein stability. In the case of OCA2 A481T and N489D variants, both of which have been observed among individuals with albinism, the predicted large change of the folding free energy can be attributed to the change of the physio-chemical properties of the wild type residues: A→T and N→D. A→T representing a hydrophobic to polar residue change, while N→D represents a polar to charge residue change. In contrast, C1002F (and I1264V in ATP7A are conservative variants but are also predicted to result in a large change of the folding energy. In this case, the

change in folding energy is thought to be caused by the distortion of the residue packing caused by the different geometries of the side chains [158]. The structure of the proteins with variant sites mapped are provided in Figure 17.

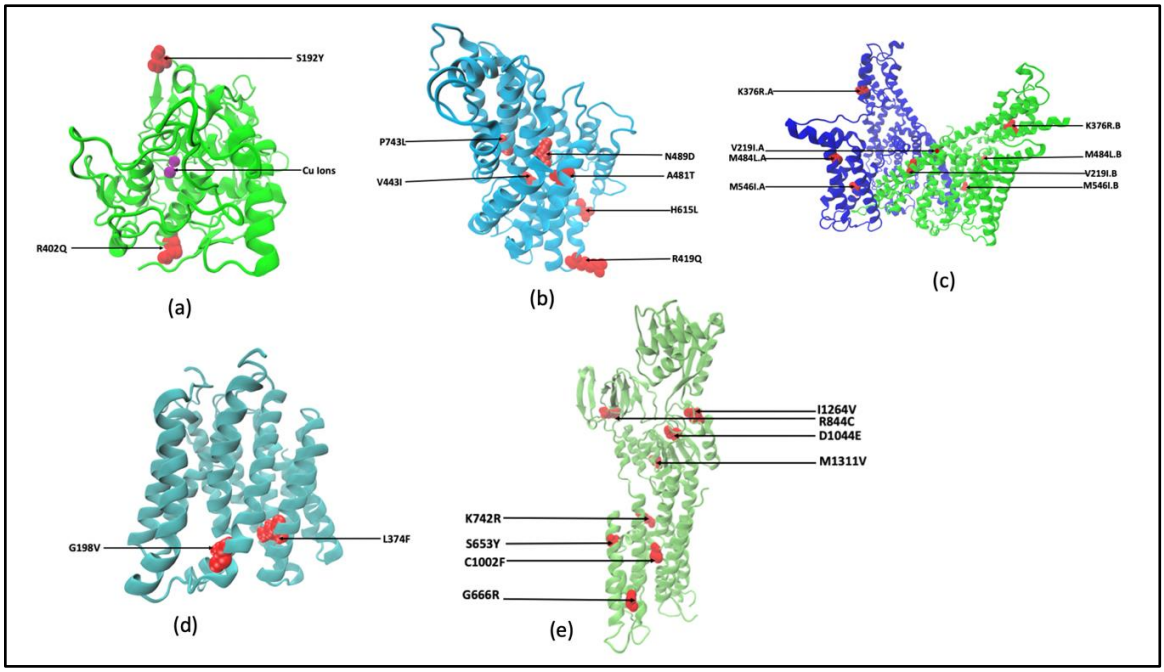

Figure 17. 3D structures with variants (shown in red color): (a) TYR; (b) OCA2; (c) TPC2; (d) SLC45A2; (e) ATP7A

Table 2. Change in folding free energy due to variants

Change in folding free energy (ΔΔG) due to variants(kcal/mol)			
Protein	variant	Avg ΔΔG	SD
TYR	R402Q	-0.5	0.5
	S192Y	-0.27	0.78
	Double MT*	-0.77	1.09

OCA2	A481T	-1.01	0.52
	H615L	0.17	0.39
	N489D	-1.05	1.08
	P743L	-0.9	0.45
	R419Q	-0.54	0.33
	V443I	-0.54	0.48
SLC45A2	G198V	-0.51	0.25
	L374F	-0.84	0.47
TPC2	K376R	-0.49	0.3
	M484L	-0.86	0.33
	M546I	-0.1	0.67
	V219I	-0.11	0.32
ATP7A	C1002F	-1.2	0.74
	G666R	-0.21	0.7
	D1044E	-0.8	0.53
	I1264V	-1.1	0.74
	K742R	0.01	0.35
	M1311V	-0.79	0.35

R844C	-0.48	0.39
S653Y	-0.45	0.54

Note: Positive and negative sign of ΔΔG represents stabilization and destabilization due to variant respectively.

* indicates double mutants (R402Q and S192Y) for TYR and ΔΔG is calculated by taking sum of individual changes.

Overall, the predicted changes of the folding free energy are not extremely large, however, since we do not know the absolute folding free energy of the proteins and how the change in protein stability affects the activity, it is impossible to assess how these moderate changes affect protein activity. However, we can reasonably assume that the activity will decrease when folding free energy changes, even when the variants appear to make the protein more stable (e.g., H615L in OCA2 protein), because in most cases any significant deviation of wild type properties is deleterious for protein function[66, 67].

5.2.3 pH dependence of folding free energy on mutants

We compared the effect of non-synonymous variants on the pH-dependence of protein stability of the wild type and corresponding variant proteins using both free energies minimized structures (Figure 18) and snap shots generated via MD simulations (Figure 26, Appendix C). One can see that there is no significant difference in the results obtained with different protocols. As mentioned in the method section, we considered that the "constant" in equation 2 is the predicted folding free energy change caused by the variants (**Error! Reference source not found.**). The most drastic effects were found for OCA2, whereas the variants in other proteins have moderate effects on the pH-dependence of folding free energy. In the case of OCA2, most

of the variants (except one, H615L) were predicted to alter the pH-dependence of stability suggesting that the variant protein will be less stable at neutral pHs.

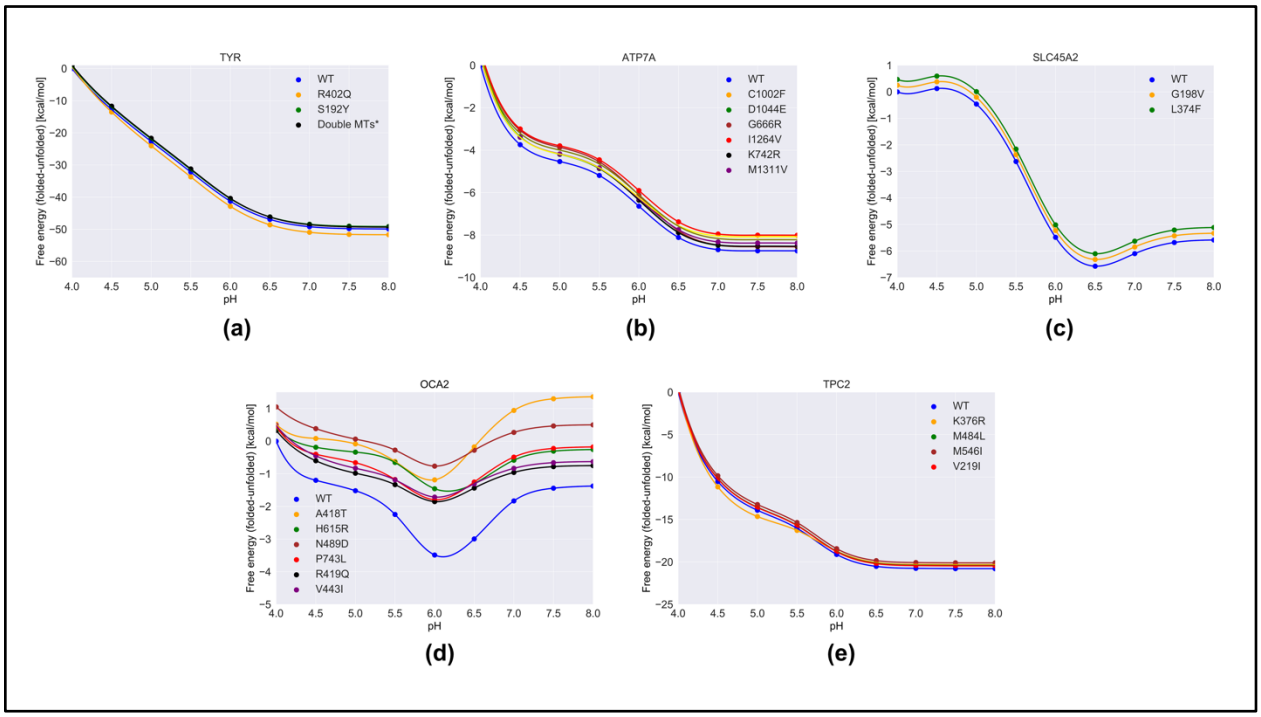

Figure 18. The pH-dependence of the folding free energy of wild type proteins and their mutants from minimized structures within pH range 4-8.

Furthermore, many variants in OCA2 (R419Q, N489D and V443I) result is a shift of pH-optimum to lower pH. This will result in a shift of maximal activity of OCA2 towards the lower pH range and could result in a shift in the balance between positive and negative regulators such that the resulting pH set point will be lower that the wild type melanosome. Low pH in the melanosome in turn will result in reduced TYR activity.

The above observations focused on the shape of the pH-dependence curve of folding free energy without considering the magnitude of the change. It should be mentioned that the changes in the folding free energy of OCA2 are within several kcal/mol, while the changes of the pH-

dependence of folding free energy caused by variants in other proteins are sometimes larger (Figure 18). Despite this, the predicted changes in the stability will likely affect protein activity and alter the balance of pH in melanosome.

The reason why variants in OCA2 have significant effects on the pH-dependence of folding free energy can be found in Table S1. Our study focuses on the pH interval 4.0 to 8.0 and the pH-dependence is induced by titratable groups that have different pKa values in folded versus unfolded state. Such titratable groups are Asp, Glu and His. One can see that in the case of OCA2, variants result in perturbed pKa's of Glu and Asp, while having almost no effect on pKa's of His. This is the reason why the pH-dependence of the folding free energy of OCA2 is mostly affected over acidic pHs.

5.3 Discussions

We studied the effect of coding variants[149] of melanosome proteins on protein stability and the pH-dependence of their folding free energy. In TYR, both variants (R402Q and S192Y) individually do not affect pH-dependence, because they either do not involve titratable groups or involve titratable groups with very high pKa, outside the pH interval of the study. Of note, a single haplotype allele in which these two alleles, R402Q and S192Y, are found together in *cis* has been suggested to present as a pathogenic haplotype for OCA1[159] . Therefore, we examined whether the presence of both variants affected protein stability; TYR modeled with both variants had modest change in protein stability (**Error! Reference source not found.**). In the case of ATP7A, modest changes of the pH-dependence caused by variants are predicted to

occur at neutral and higher pHs. Overall, the changes of stability are quite small. The reason for modest changes in the stability and little effects on the pH-dependence can be attributed to the conservative nature of the variants. In all cases, the physio-chemical properties of wild type sites are preserved. Considering SLC45A2, both variants do not involve titratable groups and do not cause alteration of the wild type pH-dependence of the folding free energy. However, significant change of the protein's stability is predicted which would affect SLC45A2 function. Significant alterations of pH-dependence of the folding free energy caused by variants are predicted for the OCA2 protein. Indeed, most of the variants alter the wild type physio-chemical properties of the protein. As a result, all variants have a shifted pH-optimum shifted from the wild type OCA2 pH-optimum. Lastly, the variants in TPC2 do not cause significant changes of either stability or pH-dependence of the folding free energy.

Our data suggests that variants predicted to be pathogenic (e.g., OCA2: N489D, V443I) likely function by affecting protein stability and/or pH-dependence of folding free energy, our data also identifies known variants with conflicting interpretations/unknown significance (e.g., OCA2: A481T, R419Q) that may affect protein stability and/or pH-dependence of folding free energy. *OCA2**R419Q is thought to modify the penetrance of the OCA2 locus and may affect the risk of melanoma[160]; therefore, our data suggest that melanosomal pH may have a functional role in melanoma genesis. These variants require biological testing to prove this association. However, our analysis failed to find any pH or protein stability effects of variants predicted to be pathogenic (e.g., TYR: R402Q; OCA2: P743L; or ATP7A: G666R). Thus, it is not a foregone conclusion that variants in protein coding sequences affect protein stability and/or pH-dependence of folding free energy suggesting other mechanisms of protein inactivation

occur. The TPC2 variants (K376R, M484I and V219I) are conservative variants and their identification by GWAS may reflect these variants may reside in LD with other variants or structural alleles, that impact expression, protein stability or function. Furthermore, SLC45A2 (rs16891982 = L374F) was a top SNV associated with altered SLC45A2 mRNA expression levels and may be mediating GWAS association via this mechanism [161]. Given the uncertainty of variant associations with function, our data suggests that assessing the function of protein variants on a large scale by structural modeling may be helpful. Perhaps, the addition of a pH polygenic score that takes into consideration the pH impact on the melanosome and all of its channels and enzymes will help in the assignment of variants to predicted functional groups.

5.4 Materials and Methods

The method section consists of four components: (1) obtaining 3D structures of the proteins of interest; (2) generation of mutants in silico; (3) molecular dynamics simulations; and (4) calculating pH-dependence of the folding free energy.

5.4.1 Structures used in the modeling

TYR protein: The 3D structure of TYR was modeled using SWISS-MODEL[162] from an amino acid sequence of length (529 aa) taken from UniProt (ID: P14679)[114]. A template (PDB ID: 5M8P)[163] with percentage identity of 44 percent and covering 81 percent (19-452) of total sequence of TYR was selected. The corresponding model and template are shown in Figure 19.

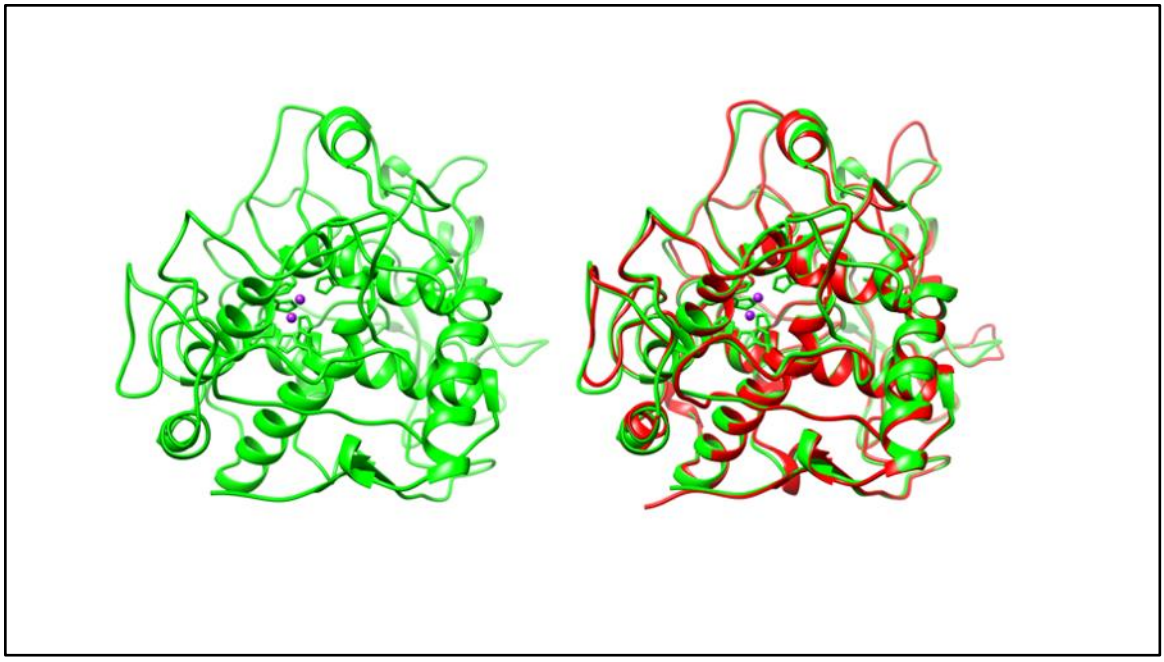

Figure 19**.** 3D model of TYR: (left) 3D structure of TYR; (right) superimposition of the TYR model (green) with its template (red)

OCA2 protein: The 3D structure of OCA2 was modeled using Phyre2 [164]. The full length sequence of OCA2 is 838 aa and is taken from UniProt (ID: Q04671)[114]. A template (PDB ID: 4F35)[165] was selected with percentage identity of 20 percent to query and covering 60 percent of the sequence of OCA2 (Figure 20). The helical content was well preserved between the template and the model (Figure 20).

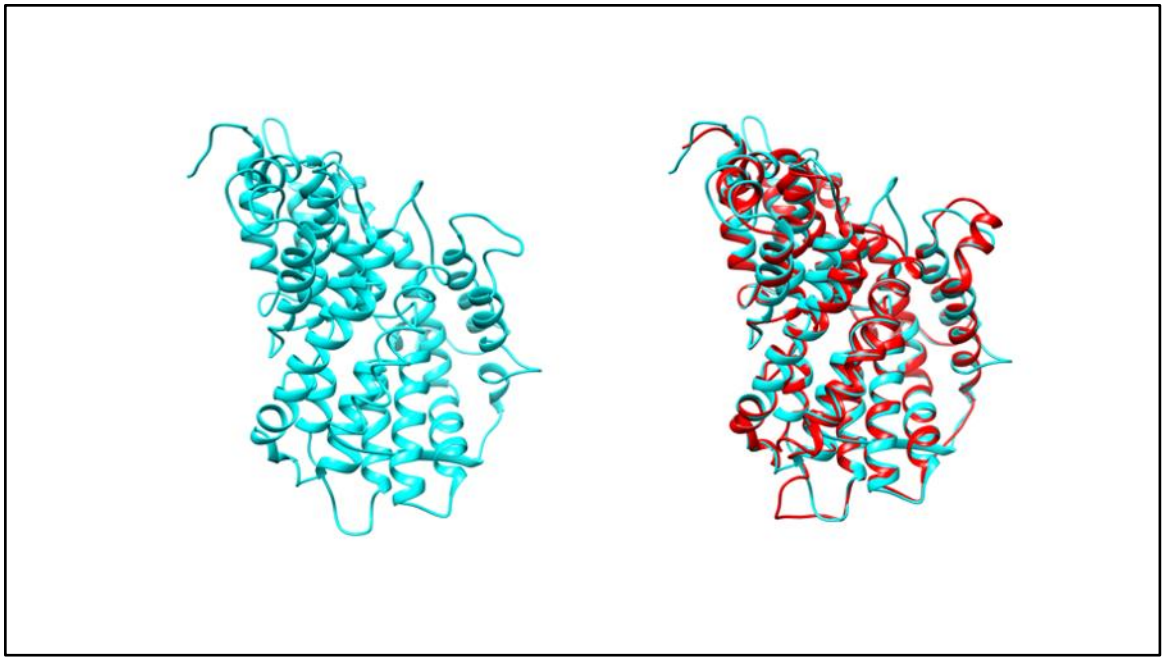

Figure 20. 3D model of OCA2: (left) 3D structure of OCA2; (right) superimposition of the OCA2 model (cyan) with its template (red)

TPC2 protein: A crystal structure for TPC2 is available (PDB ID: 6NQ2)[166] and is a homodimer with 752 residues (Figure 21).

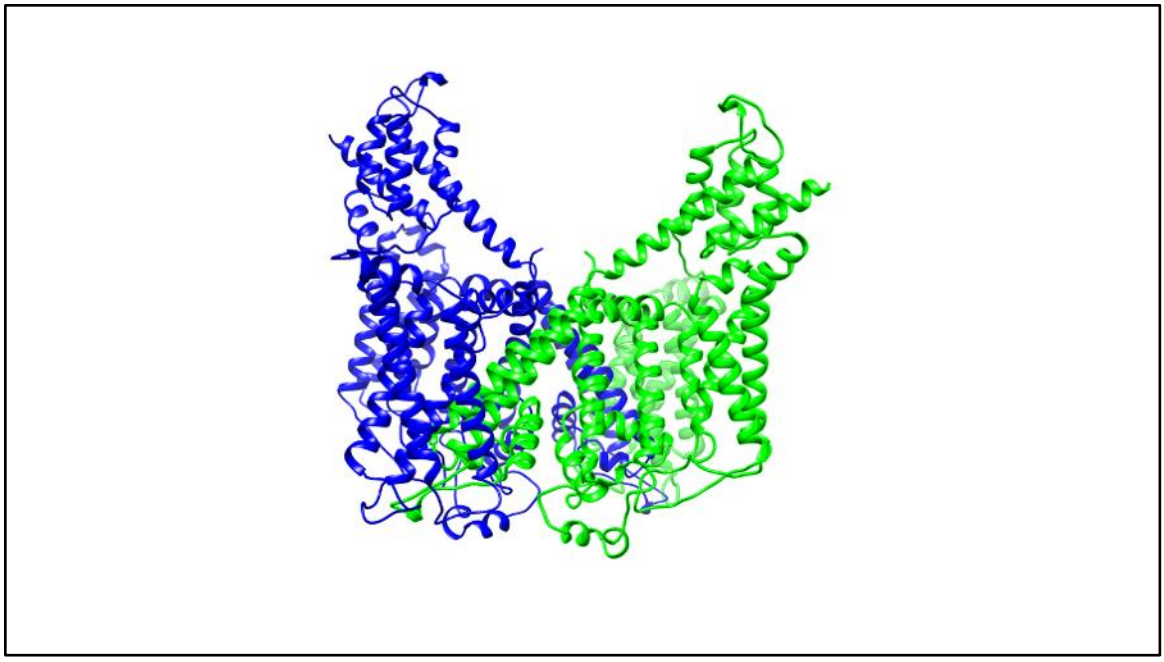

Figure 21. 3D model of TPC2: (blue) monomer A; (green) monomer B

SLC45A2 protein: The 3D structure of SLC45A2 was modeled using Phyre2 [164]. Its sequence was taken from UniProt (ID: Q9UMX9)[114] with a sequence length of 530 amino acids. The chosen template (PDB ID: 4YBQ) [167] covers 94 percent of the sequence with identity of 14 percent (Figure 22).

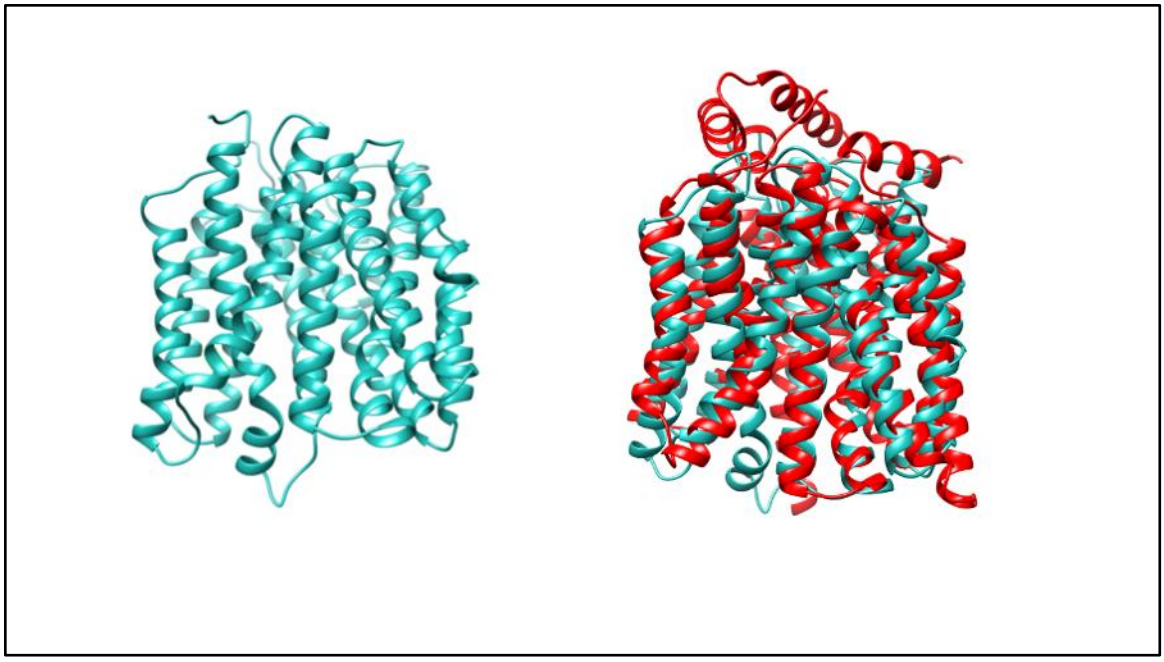

Figure 22. 3D model of SLC45A2: (left) 3D structure of SLC45A2; (right) superimposition of the SLC45A2 model (cyan) with its template (red)

ATP7A protein: The 3D modelling of this protein was also done by using Phyre2 [164] . The sequence was taken from UniProt (ID: Q04656)[114] with a sequence length of 1500 amino acids. The template (PDB ID: 3RFU)[168] covers 57 percent of the sequence (646-1411) with identity of 41 percent (Figure 23).

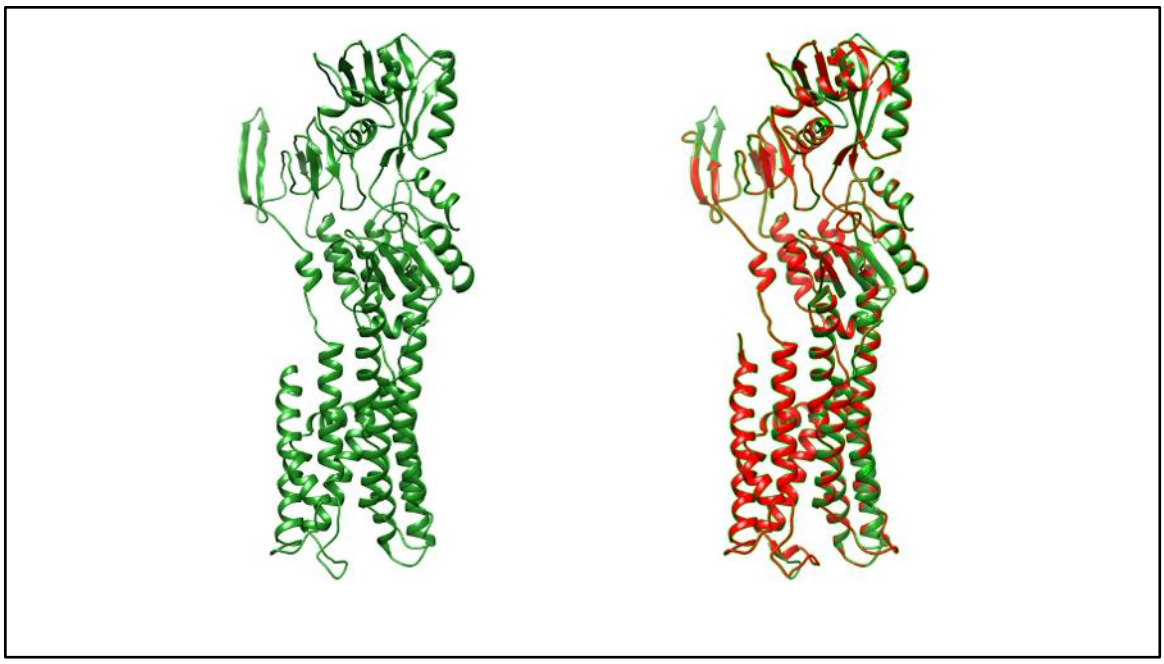

Figure 23. 3D model of ATP7A: (left) 3D structure of ATP7A; (right) superimposition of the ATP7A model (green) with its template (red)

5.4.2 List of non-synonymous GWAS identified pigmentation associated variants

The NHGRI-EBI catalog of human genome-wide association studies (GWAS)[169] was queried April 4th, 2020 to identify all non-synonymous variants in genes TYR, OCA2, SLC45A2 TPCN2 and ATP7A found associated with common human pigmentation variation of skin and hair (see **Error! Reference source not found.**).

Of note the variants identified by GWAS are associations. In the case of non-synonymous coding variants, they may impact protein function or alternatively they may, like those associations identified in non-coding regions of the genome, be in tight linkage disequilibrium (LD) with other variants that may function to impact expression levels or proper splicing. These studies are important as they measure the impact of protein variation on the pH-dependence of

the folding free energy and can help to establish whether a variant directly impacts the protein in question.

5.4.3 Generation of mutants

To generate 3D structure of protein variants while avoiding the introduction of artificial errors, we used the model of the wild type protein and the corresponding residue was mutated using UCSF Chimera[85]. The folded wild type structures and variant sites mapped onto a 3D structure of folded TYR, OCA2, TPC2, and ATP7A are shown in Figure 17. One can see (sequence alignment, Appendix C) that most of the variants are within well preserved structural regions, away from the loops, which reduces the uncertainty of the 3D modeling.

5.4.4 Molecular dynamics (MD) simulations

MD simulations were performed under periodic boundary conditions using NAMD2.9[82] with atomic parameters of the CHARMM forcefield[170]. The protein structures were prepared for the simulations using VMD[86] and TIP3P water molecules were applied to build the explicit water solvated systems. Finally, the neutralized system with NaCl were added wherever neutralization was needed.

Simulations were performed for 20 ns for each protein structure with different initial atomic velocities. In the production stages of the simulations, they were equilibrated under constant volume−temperature (NVT) conditions for 100 ps followed by 2000 ps (= 2 ns) of constant pressure−temperature (NPT) equilibration at 1 atm pressure and 310 K (with the same restraints). The first 15 ns of the simulations were not equilibrated, so they were removed. The structural analysis was sampled from the last 5 ns at every 250ps. This produced 20 snapshots per structure;

all of them were subjected to DelPhiPKa[41, 171, 172] calculations after removing the explicit water molecules.

5.4.5 Modeling pH-dependence of folding free energy

To model the pH-dependent folding free energy, we built a 3D model of the unfolded state[157]. The unfolded structure ensembles of the wild type proteins were generated using the "flexible meccano" approach[173, 174] and among them we selected one representative structure (the structure with no helices and strands) (Figure 24). The unfolded mutants were then generated by using UCSF Chimera[85].

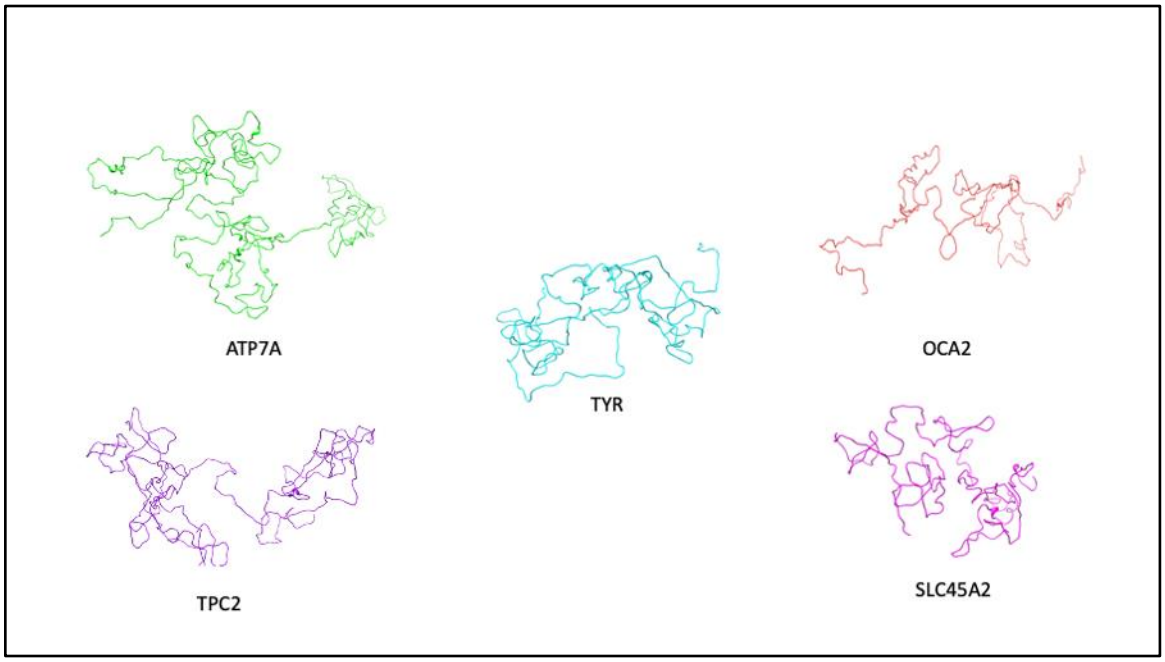

Figure 24. Unfolded structure of wild type proteins

The pKa's and net charge of the wild type protein and mutants, both in the folded and unfolded states, were calculated using DelPhiPKa[41, 171, 172]. We also calculated the pKa's

and net charge for each of the 20 snapshots taken from the MD simulation to obtain the average net charge and its difference with respect to wild type proteins.

The change in folding free energy ($\Delta\Delta G^{folding}$) was calculated from the net charge difference between the folded state and the unfolded state, taking the unfolded state as the initial state. The following equation is used over the pH-range of interest, giving an explicit pH-dependent form of the folding free energy [29, 175].

$$\Delta\Delta G^{folding} = 2.3kT \sum \Delta q dpH + constant \quad (2)$$

where, Δq is the change of net charge from unfolded to folded state and dpH is the pH interval. The constant is the absolute folding free energy at a given pH.

For the analysis of the wild type proteins, the "constant" was considered to be zero at the beginning of the pH interval because there is no information about the absolute folding free energy of the individual proteins and predicting it would introduce significant and unwanted noise. However, for the mutants, there are many algorithms that are benchmarked against experimental data and shown to perform well, and this gives us the opportunity to predict the change of the folding free energy caused by variant with acceptable confidence. Thus, for the mutants, the "constant" was considered to be the free energy difference between wild type and mutant proteins caused by the variant. The folding free energy changes were modeled using an in-house algorithm, the SAAFEC-SEQ[44] method, along with third party tools such as

INPS3D[176], INPS-SEQ[176], mCSM[123], SDM[124], DUET[121], I-Mutant-SEQ[125], MUpro-SEQ[177], iStable-SEQ[178] and DeepDDG[179].

5.5 Conclusions

The importance of melanosomal pH for the regulation of organelle maturation is well studied[150, 154], but how the melanosomal proteins regulate the desired pH remains unknown. The same applies for their genetic variants[154]. Here we suggest a mechanism of competitive pH-dependence of stability and activity of "positive" and "negative" pH regulators. Our data suggests that the "positive" regulators of melanosomal pH should be having maximal activity (and thus maximal stability) at low pH, while the opposite is expected from "negative" regulators. Indeed, we have shown that OCA2 and SLC45A2 have low pH-optimum as compared to TPC2 protein. Furthermore, the TYR and its partner ATP7A are also shown to have a pH-optimum at neutral and higher pHs. We predict that similar mechanisms of pH regulation can be expected for other melanosomal proteins.

APPENDICES

Appendix A

Publications resulting from the dissertation

CHAPTER ONE

Koirala, M., Shashikala, H. B., Jeffries, J., Wu, B., Loftus, S. K., Zippin, J. H., & Alexov, E. (2021). Computational Investigation of the pH Dependence of Stability of Melanosome Proteins: Implication for Melanosome formation and Disease. *International journal of molecular sciences*, *22*(15), 8273.

Ganakammal, S. R.[*], Koirala, M.[*], Wu, B., & Alexov, E. (2020). In-silico analysis to identify the role of MEN1 missense mutations in breast cancer. *Journal of Theoretical and Computational Chemistry*, *19*(06), 2041002. (* equally contributed)

Koirala, M., & Alexov, E. (2020). Ab-initio binding of barnase–barstar with DelPhiForce steered Molecular Dynamics (DFMD) approach. *Journal of Theoretical and Computational Chemistry*, *19*(04), 2050016.

Panday, S. K., Shashikala, M. H., Koirala, M., Pahari, S., Chakravorty, A., Peng, Y., ... & Alexov, E. (2019). Modeling electrostatics in molecular biology: A tutorial of Delphi and associated resources [Article v1. 0]. *Living Journal of Computational Molecular Science*, *1*(2).

Koirala, M., & Alexov, E. (2020). Computational chemistry methods to investigate the effects caused by DNA variants linked with Disease. *Journal of Theoretical and Computational Chemistry*, *19*(06), 1930001.

Li, C., Jia, Z., Chakravorty, A., Pahari, S., Peng, Y., Basu, S., Koirala, M., ... & Alexov, E. (2019). Delphi suite: new developments and review of functionalities. *Journal of computational chemistry*, *40*(28), 2502-2508.

CHAPTER TWO

Li, C., Jia, Z., Chakravorty, A., Pahari, S., Peng, Y., Basu, S., Koirala, M., ... & Alexov, E. (2019). Delphi suite: new developments and review of functionalities. *Journal of computational chemistry*, *40*(28), 2502-2508.

Panday, S. K., Shashikala, M. H., Koirala, M., Pahari, S., Chakrvorty, A., Peng, Y., ... & Alexov, E. (2019). Modeling electrostatics in molecular biology: A tutorial of DelPhi and associated resources [Article v1. 0]. *Living Journal of Computational Molecular Science*, *1*(2).

CHAPTER THREE

Koirala, M., & Alexov, E. (2020). Ab-initio binding of barnase–barstar with DelPhiForce steered Molecular Dynamics (DFMD) approach. *Journal of Theoretical and Computational Chemistry*, *19*(04), 2050016.

CHAPTER FOUR

Ganakammal, S. R.*, Koirala, M.*, Wu, B., & Alexov, E. (2020). In-silico analysis to identify the role of MEN1 missense mutations in breast cancer. *Journal of Theoretical and Computational Chemistry*, *19*(06), 2041002. (* equally contributed)

CHAPTER FIVE

Koirala, M., Shashikala, H. B., Jeffries, J., Wu, B., Loftus, S. K., Zippin, J. H., & Alexov, E. (2021). Computational Investigation of the pH Dependence of Stability of Melanosome Proteins: Implication for Melanosome formation and Disease. *International journal of molecular sciences*, *22*(15), 8273.

Appendix B

Tables for chapter four

Table 3. Summary of annotation of the 19 variants from COSMIC database. This consists of in-silico prediction scores that is used by combinatory approach to make our pathogenicity classification. We also have GERP++ score that shows higher rates of evolutionary conservation for pathogenic variants compared to the benign.

AA Mutation	Polyphen2	LRT	Meta SVM	VEST3	PROVEAN	REVEL	Eigen	CADD	Our Prediction	GREP++	gnomAD	Clnsig
S606C	0.965	0.024	1.062	0.437	-2.320	0.565	4.351	25.9	**Pathogenic**	4.57	.	.
S606Y	0.948	0.024	1.064	0.451	-1.770	0.578	4.324	26.8	**Pathogenic**	4.57	.	.
L605V	0.821	0.000	1.042	0.178	0.070	0.365	2.095	22.8	Benign	4.57	.	.
M563I	0.932	0.000	1.069	0.826	-3.060	0.883	5.017	29.5	**Pathogenic**	4.48	.	.
K562M	0.998	0.000	1.042	0.788	-4.930	0.858	5.063	29	**Pathogenic**	4.48	.	.

P534A	0.849	0.723	0.772	0.144	-0.720	0.318	2.038	6.224	Benign	2.34	.	.
T530I	0.011	0.148	0.551	0.048	-1.800	0.296	0.519	5.558	Benign	2.63	**4.035E-06**	.
R490Q	0.988	0.000	1.045	0.521	-0.340	0.622	4.880	27.1	**Pathogenic**	4.54	.	.
Q398E	0.001	0.052	0.165	0.157	0.520	0.351	1.187	3.064	Benign	3.89	.	.
A289E	0.994	0.000	1.079	0.967	-2.700	0.967	6.548	29.7	**Pathogenic**	4.24	.	.
L272R	1.000	0.000	1.044	0.979	-5.100	0.954	6.031	30	**Pathogenic**	4.39	.	.
S258L	0.979	0.000	1.063	0.974	-3.100	0.913	6.209	27.1	**Pathogenic**	4.80	.	VUS
V220M	0.896	0.002	1.088	0.785	-1.820	0.799	3.983	24.4	**Pathogenic**	4.80	.	VUS
H204Y	0.975	0.000	1.054	0.911	-5.020	0.969	8.195	27.4	**Pathogenic**	4.76	.	.
E200K	0.985	0.000	1.053	0.954	-3.550	0.948	7.484	34	**Pathogenic**	4.76	.	.
C170Y	0.995	0.000	1.029	0.979	-9.330	0.953	7.550	29.3	**Pathogenic**	4.8	.	.
R108P	0.700	0.000	1.081	0.668	-1.280	0.655	4.143	25.4	**Pathogenic**	5.02	.	.
L89R	0.998	0.000	1.036	0.996	-4.640	0.923	6.063	24.5	**Pathogenic**	5.02	.	.
L36F	0.994	0.000	1.048	0.826	-2.540	0.901	5.751	26.8	**Pathogenic**	4.89	.	

*Clnsig: Clinical Significance from ClinVar Database
*VUS: Variant of Unknow Significance

Table 4. Calculation of Folding Free Energy Change (kcal/mol) of variants from COSMIC & VariSNP database. The calculated folding free energy changes in kcal/mol of menin protein due to mutations from COSMIC and VariSNP database. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	**Our Prediction**	**SAAFEC**	**mCSM**	**SDM**	**DUET**	**CUPSAT**	**I-Mutant**	**Average**
Missense Variants from COSMIC DB								
S606C	Pathogenic	2.23	0.27	-0.75	-0.20	-0.73	2.18	1.56
S606Y	Pathogenic	0.29	0.43	-0.54	0.16	0.95	0.51	0.47
L605Y	Benign	-2.53	0.48	0.6	0.52	0.53	-0.03	0.53
M563I	Pathogenic	-1.08	1.10	0.32	0.68	-1.18	-0.48	-0.91
K562M	Pathogenic	5.90	-0.48	0.06	0.56	-0.01	-0.23	2.13
P534A	Benign	2.57	0.27	-0.69	-0.11	1.62	0.98	1.36
T530I	Benign	0.23	0.16	-1.31	-0.41	2.97	0.11	0.87
R490Q	Pathogenic	2.50	0.67	0.32	0.59	-0.61	0.56	0.93
Q398E	Benign	2.07	0.03	-0.97	-0.46	1.25	-1.44	1.12
A289E	Pathogenic	11.39	2.71	2.97	3.04	2.94	0.75	3.96
L272R	Pathogenic	-4.26	2.30	3.02	2.36	5.55	0.51	2.75

S258L	Pathogenic	1.80	0.03	-0.22	-0.18	1.7	0.95	1.12
V220M	Pathogenic	-0.65	0.75	0.14	0.57	2.2	2.15	1.16
H204Y	Pathogenic	-1.42	-1.32	-0.02	-1.35	-1.52	-0.25	-1.14
E200K	Pathogenic	-4.06	2.14	1.15	2.20	-4.5	0.5	1.50
C170Y	Pathogenic	-1.00	1.01	1.9	1.32	4.5	0.78	1.90
R108P	Pathogenic	6.49	0.31	2.23	0.87	0.71	0.81	1.90
L89R	Pathogenic	-2.33	1.54	1.76	1.47	5.16	2.1	2.41
L36F	Pathogenic	1.17	1.51	1.39	1.77	2.02	0.84	1.45
VariSNP Benign Variants								
P519S	Benign	1.00	0.95	-0.72	0.47	-2.93	1.43	0.96
V555L	Benign	-2.58	0.63	0.86	0.45	0.41	0.35	0.54
I377T	Benign	-0.99	2.46	3.46	2.80	4.46	4.18	3.47
E371D	Benign	-4.08	1.04	1.39	1.04	0.31	0.81	0.92
G508D	Benign	-0.88	0.40	-0.26	-0.06	0.53	1.34	0.76

Table 5. Calculation of Folding Free Energy Change (kcal/mol) of variants from ClinVar database. The calculated folding free energy change in kcal/mol of menin protein taken from ClinVar associated with Neoplasia. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	**Conditions**	**Clinical Significance**	**SAAFEC**	**mCSM**	**SDM**	**DUET**	**CUPSAT**	**Average**
A373D	Neoplasia	Pathogenic	6.39	1.93	3.23	2.82	2.25	3.22
E260K	Neoplasia	Pathogenic	-3.32	-0.26	0.77	-0.46	2.13	-1.35
V189E	Neoplasia	Pathogenic	5.73	2.83	3.34	3.21	4.24	3.87
A165P	Neoplasia	Pathogenic	0.42	-0.99	4.45	-0.25	0.17	1.67
R280K	Neoplasia	Pathogenic	1.95	0.66	0.1	0.46	-0.17	0.85
W441R	Neoplasia	Pathogenic	-1.92	2.56	1.11	2.21	5.34	2.81
D423N	Neoplasia	Pathogenic	-0.98	1.66	0.83	1.59	3.7	1.94
H139D	Neoplasia	Pathogenic	5.81	0.32	1.65	0.49	1.25	1.90

Table 6. Calculation of Binding Free Energy change of Menin-ER complex dues to COSMIC and VariSNP mutations. The calculated binding free energy changes in kcal/mol of menin-ER complex due to mutations from COSMIC and VariSNP database. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	Our Prediction	SAAMBE	BeAtMuSiC	mCSM-PPI2	MutaBind	Average
Unclassified mutations from COSMIC Database						
S606C	Pathogenic	0.30	-0.09	0.19	0.60	0.37
S606Y	Pathogenic	0.29	0.05	-0.04	0.55	0.30
L605Y	Benign	0.29	0.01	-0.24	0.62	0.31
M563I	Pathogenic	0.39	0.25	0.49	1.75	0.72
K562M	Pathogenic	0.52	0.15	0.25	1.05	0.50
P534A	Benign	-0.27	0.13	0.59	-0.24	0.36
T530I	Benign	0.20	0.09	-0.01	0.51	0.27
R490Q	Pathogenic	0.37	0.25	0.11	0.75	0.45
Q398E	Benign	0.42	0.07	0.14	0.47	0.28
A289E	Pathogenic	0.57	0.68	-0.55	2.63	1.30
L272R	Pathogenic	0.34	0.92	0.38	2.00	0.91
S258L	Pathogenic	0.40	0.53	0.21	1.53	0.67
V220M	Pathogenic	0.26	0.19	-0.22	0.71	0.39
H204Y	Pathogenic	0.34	0.38	0.32	0.44	0.37
E200K	Pathogenic	-0.24	0.39	0.77	0.90	0.69
C170Y	Pathogenic	0.57	0.23	-0.34	2.75	1.18
R108P	Pathogenic	0.70	0.16	1.08	1.27	0.80
L89R	Pathogenic	0.54	0.48	0.20	1.16	0.60
L36F	Pathogenic	0.29	0.34	0.43	1.55	0.73

Benign mutations from VariSNP Database						
P519S	Benign	-0.22	-0.37	-0.16	-0.43	-0.29
V555L	Benign	0.40	0.22	0.15	0.49	0.32
I377T	Benign	0.27	1.27	0.10	1.53	0.79
E371D	Benign	0.22	0.19	-0.13	1.68	0.70
G508D	Benign	0.70	0.33	-0.27	0.75	0.59

Table 7. Calculation of Binding Free Energy change of Menin-MLL complex dues to COSMIC and VariSNP mutations. The calculated binding free energy changes in kcal/mol of menin-MLL complex due to mutations from COSMIC and VariSNP database. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	**Our Prediction**	**SAAMBE**	**BeAtMuSiC**	**mCSM-PPI2**	**MutaBind**	**Average**
COMSIC Mutations						
S606C	Pathogenic	0.24	-0.09	0.19	0.76	0.40
S606Y	Pathogenic	0.44	0.05	-0.01	0.70	0.41
L605Y	Benign	0.34	0.01	-0.17	0.55	0.30
M563I	Pathogenic	0.39	0.25	0.45	1.59	0.67
K562M	Pathogenic	0.43	0.05	0.24	0.99	0.43
P534A	Benign	0.09	0.2	0.47	-0.19	0.25
T530I	Benign	0.30	0.09	-0.03	0.56	0.32
R490Q	Pathogenic	0.44	0.25	-0.06	0.91	0.53
Q398E	Benign	0.41	0.07	-0.18	0.36	0.28
A289E	Pathogenic	1.00	1.02	-0.70	2.20	1.40
L272R	Pathogenic	0.86	0.88	0.32	2.35	1.10
S258L	Pathogenic	0.18	0.64	0.23	1.32	0.59
V220M	Pathogenic	0.25	0.19	0.23	0.57	0.31
H204Y	Pathogenic	0.73	0.29	-0.05	1.12	0.71
E200K	Pathogenic	0.49	0.39	0.84	1.42	0.78
C170Y	Pathogenic	0.75	0.23	-0.34	2.67	1.22
R108P	Pathogenic	0.78	0.16	1.04	1.16	0.78
L89R	Pathogenic	0.86	0.48	0.18	1.17	0.67
L36F	Pathogenic	0.71	0.34	0.37	1.36	0.71
VariSNP Mutations						
P519S	Benign	0.01	-0.37	-0.14	-0.19	-0.18

V555L	Benign	0.60	0.22	0.16	0.74	0.43
I377T	Benign	0.50	0.99	0.17	1.73	0.85
E371D	Benign	0.29	0.27	0.28	0.77	0.4
G508D	Benign	0.64	0.33	-0.19	1.06	0.68

Table 8. Calculation of Binding Free Energy change of Menin-ER complex dues to ClinVar mutations. The calculated binding free energy changes in kcal/mol of menin-ER complex due to mutations from ClinVar. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	**Conditions**	**Clinical Significance**	**SAAMBE**	**BeAtMuSiC**	**mCSM-PPI2**	**MutaBind**	**Average**
A373D	Neoplasia	Pathogenic	0.58	1.29	-0.61	2.26	1.38
E260K	Neoplasia	Pathogenic	0.08	0.30	0.64	0.90	0.56
V189E	Neoplasia	Pathogenic	0.53	1.92	0.21	2.06	1.18
A165P	Neoplasia	Pathogenic	0.68	1.24	0.18	2.90	1.25
R280K	Neoplasia	Pathogenic	0.08	0.15	0.16	0.31	0.18
W441R	Neoplasia	Pathogenic	-0.74	1.23	1.83	2.33	1.80
D423N	Neoplasia	Pathogenic	0.12	0.16	0.98	1.20	0.62
H139D	Neoplasia	Pathogenic	0.51	0.51	0.34	1.63	0.75

Table 9. Calculation of Binding Free Energy change of Menin-MLL complex dues to ClinVar mutations. The calculated binding free energy changes in kcal/mol of menin-MLL complex due to mutations from ClinVar. The positive value indicates destabilization and negative value indicates stabilization.

AA Mutation	**Conditions**	**Clinical Significance**	**SAAMBE**	**BeAtMuSiC**	**mCSM-PPI2**	**MutaBind**	**Average**
A373D	Neoplasia	Pathogenic	0.21	0.78	-0.65	2.11	1.03
E260K	Neoplasia	Pathogenic	0.40	0.39	0.68	1.20	0.67
V189E	Neoplasia	Pathogenic	0.30	1.92	0.40	2.3	1.23
A165P	Neoplasia	Pathogenic	0.58	1.20	0.17	2.5	1.11
R280K	Neoplasia	Pathogenic	0.25	0.06	0.30	0.30	0.23

W441R	Neoplasia	Pathogenic	0.53	1.01	0.55	2.79	1.22
D423N	Neoplasia	Pathogenic	0.63	0.23	0.74	1.43	0.76
H139D	Neoplasia	Pathogenic	1.70	0.79	0.75	1.84	1.27

Table 10. KNN Classification Results of the 19 mutations from COMSIC database. The training dataset was constructed using well curated mutations from ClinVar and VariSNP. The features used for this supervised learning method is evolutionary conservation score (GERP++) and folding energy change (ΔΔG). And with a K=5 we were able to classify at a higher accuracy.

AA Change	**GERP++**	**ΔΔG**	**Combinatory Insilco Prediction**	**KNN Classification Features**	
				GERP++ and ΔΔG	**ΔΔG**
S606C	4.57	1.56	P	P	P
S606Y	4.57	0.47	P	P	***B***
A289E	4.24	3.96	P	P	P
L272R	4.39	2.75	P	P	P
S258L	4.8	1.12	P	P	P
V220M	4.8	1.16	P	P	P
H204Y	4.76	-1.14	P	P	B
Q398E	3.89	1.12	B	***P***	B
L605V	4.57	0.53	B	***P***	B
M563I	4.48	-0.91	P	P	B
K562M	4.48	2.13	P	P	P
P534A	2.34	1.36	B	B	B
T530I	2.63	0.87	B	B	B
R490Q	4.54	0.93	P	P	P
E200K	4.76	1.5	P	P	P
C170Y	4.8	1.9	P	P	P
R108P	5.02	1.9	P	P	P

L89R	5.02	2.41	P	P	P
L36F	4.89	1.45	P	P	P

*P** denotes Pathogenic and *B** denotes Benign
*Highlighted are the difference in prediction between KNN classification and Combinatory insilco prediction approach

Appendix C

Additional materials for chapter five

List of abbreviations

TYR- tyrosinase

OCA2- oculocutaneous albinism 2

SLC45A2- solute carrier 45 member 2

TPC2- two-pore channel 2

ATP7A- copper-transporting ATPase 1

Figures from the result of MD snapshot

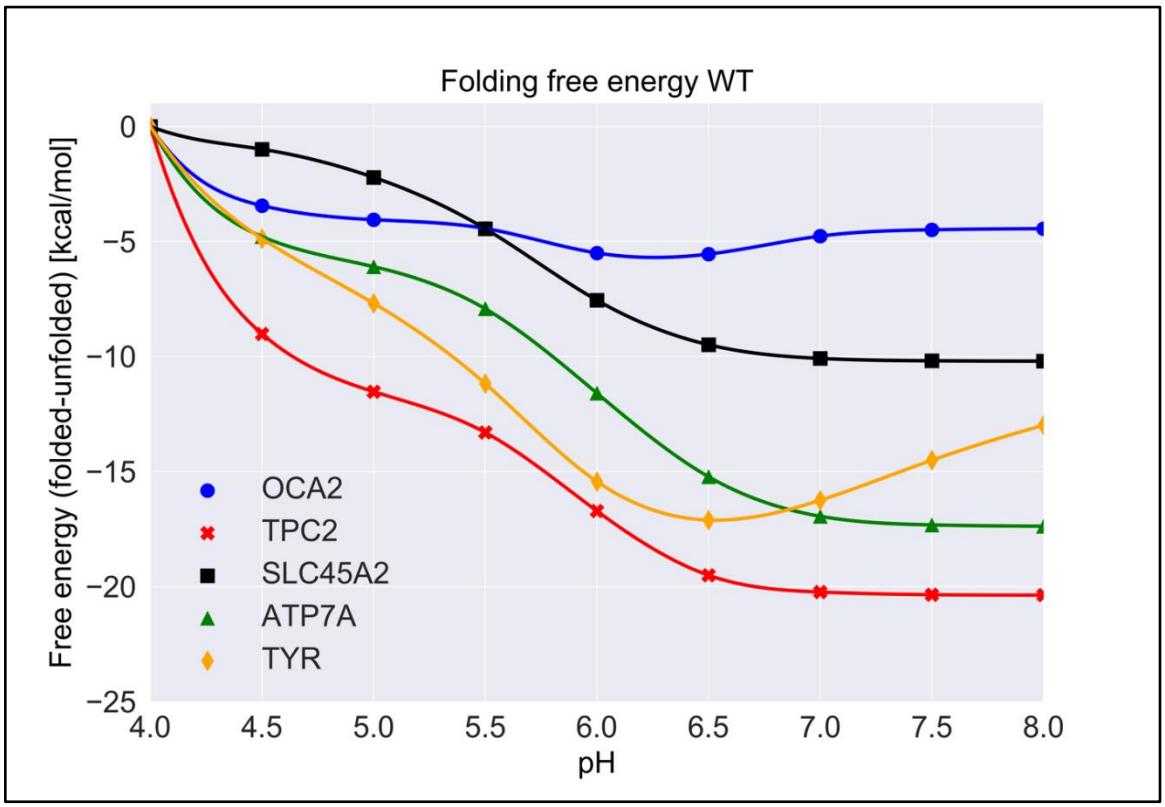

Figure 25. The pH-dependence of the folding free energy of wild type proteins from 20 MD snap shots within pH range 4-8.

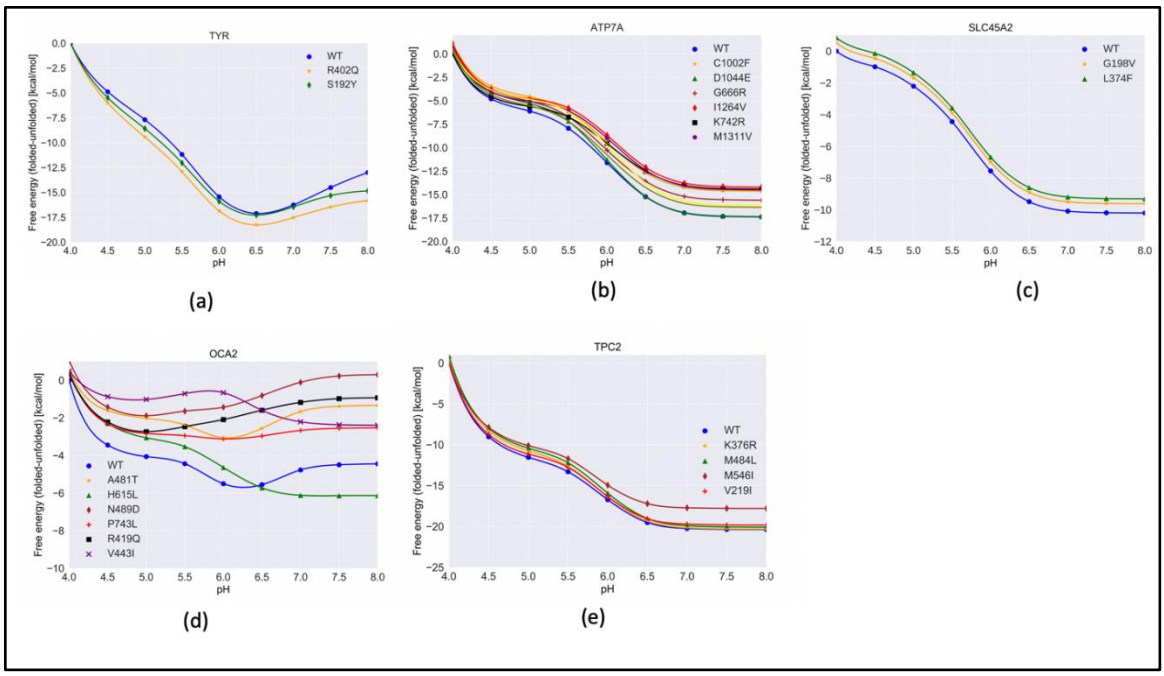

Figure 26. The pH-dependence of the folding free energy of wild type proteins and their mutants from 20 MD snapshots within pH range 4-8.

Sequence alignments :

Sequence alignment of all the models with its template by using T-Coffee web server [180] are shown below. An asterisk (*) indicates positions which have a single, fully conserved residue; a colon (:) indicates conservation between groups of strongly similar properties; a period (.) indicates conservation between group of weakly similar properties.

```
Model.pdb       HFPRACVSSKNLMEKECCPPWSGDRS----PCGQLSGRGSCQNILLSNAP
Template.pdb    QFPRQCATVEALRSGMCCPDLSPVSGPGTDRCGSSSGRGRCEAVTADSRP
                :*** *.: : * .  ***  *   .     **. **** *: :  .. *

Model.pdb       LGPQFPFTGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFWGPNCTE
Template.pdb    HSPQYPHDGRDDREVWPLRFFNRTCHCNGNFSGHNCGTCRPGWRGAACDQ
                 .**:*. * **** **  *:****:*.*** *.***.*: *: *. * :

Model.pdb       RRLLVRRNIFDLSAPEKDKFFAYLTLAKHTISSDYVIPIGTYGQM--KNG
Template.pdb    RVLIVRRNLLDLSKEEKNHFVRALDMAKRTTHPLFVIATRRSEEILGPDG
                * *:****::***  **::*.  * :**:*  . :**.     ::   :*

Model.pdb       STPMFNDINIYDLFVWMHYYVSMDALLG-GSEIWRDIDFAHEAPAFLPWH
Template.pdb    NTPQFENISIYNYFVWTHYYSVKKTFLGVGQESFGEVDFSHEGPAFLTWH
                .** *::*.**: *** ***   .::** *.* : ::**:**.****.**

Model.pdb       RLFLLRWEQEIQKLTGDENFTIPYWDWRDA-EKCDICTDEYMGGQHPTNP
Template.pdb    RYHLLRLEKDMQEMLQEPSFSLPYWNFATGKNVCDICTDDLMGSRSNFDS
                * .*** *:::*::  : .*::***::  . : ******: **.:   :.

Model.pdb       NLLSPASFFSSWQIVCSRLEEYNSHQSLCNGTPEGPLRRNP-GNHDKSRT
Template.pdb    TLISPNSVFSQWRVVCDSLEDYDTLGTLCNSTEDGPIRRNPAGNVARPMV
                .*:** *.**.*::**. **:*::  :***.* :**:**** **  :. .

Model.pdb       PRLPSSADVEFCLSLTQYESGSMDKAANFSFRNTLEGFASPLTGIADASQ
Template.pdb    QRLPEPQDVAQCLEVGLFDTPPFYSNSTNSFRNTVEGFSDP-TGKYDPAV
                 ***.. **  **.:  ::: .: . :. *****:***:.* **  *.:

Model.pdb       SSMHNALHIYMNGTMSQVQGSANDPIFLLHHAFVDSIFEQWLRRHRPLQE
Template.pdb    SSLHNLAHLFLNGTGGQVHLSPNDPIFVLLHTFTDAVFDEWLRRYNADIS
                **:**  *::::** .**: *.*****:* *:*.*::*::****:..  .

Model.pdb       VYPEANAPIGHNRESYMVPFIPLYRNGDFFISSKD-LGYDY------
Template.pdb    TFPLENAPIGHNRQYNMVPFWPPVTNTEMFVTAPDNLGYTYEIQWPS
                .:*  ********:  **** *   * ::*::: * *** *
```

Figure 27. Sequence alignment of TYR with percentage identity of 44 percent with its template, E-value=0 and Score=920.38

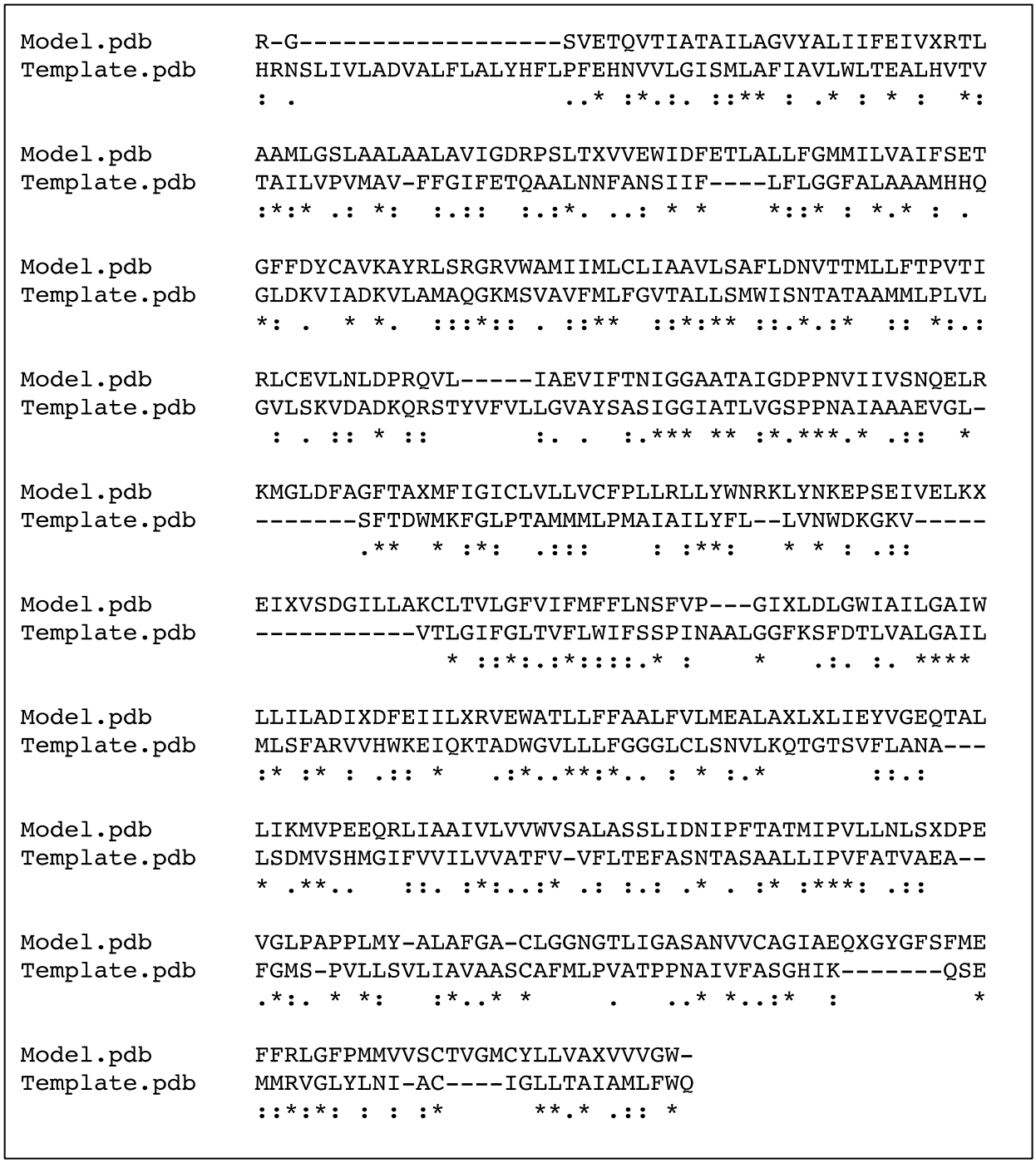

Figure 28. Sequence alignment of ATP7A with percentage identity of 47 percent with its template, E-value=0 and Score=1181.05

```
Model.pdb       R-G------------------SVETQVTIATAILAGVYALIIFEIVXRTL
Template.pdb    HRNSLIVLADVALFLALYHFLPFEHNVVLGISMLAFIAVLWLTEALHVTV
                : .                  ..* :*.:. ::** : .* : * :  *:

Model.pdb       AAMLGSLAALAALAVIGDRPSLTXVVEWIDFETLALLFGMMILVAIFSET
Template.pdb    TAILVPVMAV-FFGIFETQAALNNFANSIIF----LFLGGFALAAAMHHQ
                :*:* .: *:  :.::  :.:*. ..: * *    *::* : *.* : .

Model.pdb       GFFDYCAVKAYRLSRGRVWAMIIMLCLIAAVLSAFLDNVTTMLLFTPVTI
Template.pdb    GLDKVIADKVLAMAQGKMSVAVFMLFGVTALLSMWISNTATAAMMLPLVL
                *: .  * *.  :::*:: . ::**  ::*:** ::.*.:*  :: *:.:

Model.pdb       RLCEVLNLDPRQVL-----IAEVIFTNIGGAATAIGDPPNVIIVSNQELR
Template.pdb    GVLSKVDADKQRSTYVFVLLGVAYSASIGGIATLVGSPPNAIAAAEVGL-
                 : . :: * ::       :. .  :.*** ** :*.***.* .::  *

Model.pdb       KMGLDFAGFTAXMFIGICLVLLVCFPLLRLLYWNRKLYNKEPSEIVELKX
Template.pdb    -------SFTDWMKFGLPTAMMMLPMAIAILYFL--LVNWDKGKV-----
                       .**  * :*:  .:::    : :**:   * * : .::

Model.pdb       EIXVSDGILLAKCLTVLGFVIFMFFLNSFVP---GIXLDLGWIAILGAIW
Template.pdb    -----------VTLGIFGLTVFLWIFSSPINAALGGFKSFDTLVALGAIL
                            * ::*:.:*::::.* :    *   .:. :. ****

Model.pdb       LLILADIXDFEIILXRVEWATLLFFAALFVLMEALAXLXLIEYVGEQTAL
Template.pdb    MLSFARVVHWKEIQKTADWGVLLLFGGGLCLSNVLKQTGTSVFLANA---
                :* :* : .:: *   .:*..**:*.. : * :.*       ::.:

Model.pdb       LIKMVPEEQRLIAAIVLVVWVSALASSLIDNIPFTATMIPVLLNLSXDPE
Template.pdb    LSDMVSHMGIFVVILVVATFV-VFLTEFASNTASAALLIPVFATVAEA--
                * .**..   ::. :*:..:* .: :.: .* . :* :***: .::

Model.pdb       VGLPAPPLMY-ALAFGA-CLGGNGTLIGASANVVCAGIAEQXGYGFSFME
Template.pdb    FGMS-PVLLSVLIAVAASCAFMLPVATPPNAIVFASGHIK-------QSE
                .*:. * *:   :*..* *     .   ..* *..:*  :         *

Model.pdb       FFRLGFPMMVVSCTVGMCYLLVAXVVVGW-
Template.pdb    MMRVGLYLNI-AC----IGLLTAIAMLFWQ
                ::*:*: : : :*      **.* .:: *
```

Figure 29. Sequence alignment of OCA2 with percentage identity of 20 percent with its template, E-value=0 and Score=388.81

```
Model.pdb       -----------------EPPKRPTSRLIMXSMAM----------------
Template.pdb    MGWGGGGGCTPRPPIHQQPPERRVVTVVFLGLLLDLLAFTLLLPLLPGLL
                                 :**:* .  ::: .: :

Model.pdb       --------------------FGREFCYAVEAAYVTPVLLSVGLPSSLYSI
Template.pdb    ESHGRAHDPLYGSWQGGVDWFATAIGMPVEKRYNS--VLFGGLIGSAFS-
                                    *.  :  .**  * :  :*  ** .* :*

Model.pdb       VWFLSPILGFLLQPVVGSASDXCRSRWGRRRPYILTLGVMMLVGMALYLN
Template.pdb    ------VLQFLCAPLTGATSDCLGR-----RPVMLL------CLMGVA-T
                      :* **  *:.*::**         ** :*         *.:  .

Model.pdb       GATVVATMIGVVLFDFAADFIDGP-------IKAYLFDVCSXQDKEKGLX
Template.pdb    SYAVWATSRSFAAFL-ASRLIGGISKGNVSLSTAIVADLGSPLARSQGMA
                . :* **  ... *  *: :*.*         .* : *: *   :.:*:

Model.pdb       YXALFTGFGGALGYLLGAIDWAXLELGRLLGTEFQVMFFFSALVLTLCFT
Template.pdb    VIGVAFSLGFTLGPMLGAS--LPLE----MAPWFALLFAASDLLFIFCFL
                  .:  .:* :** :***     **    :.. * ::*  * *:: :**

Model.pdb       VXLCSIS---------------------EAPLTEVAKGIPPQQTPQDPPL
Template.pdb    PETLPLEKRAPSIALGFRDAADLLSPLALLRFSAVARG-------QDPP-
                    .:.                          :: **:*       ****

Model.pdb       SSDGMYEYGSIEKVKNGYVNPXYRYLCISXLIGWTAFLSNMLFFTDFMGQ
Template.pdb    SGDRLS---SLRR--LGLVYFLYLFL-FSGLEYTLSFLTHQRFQ------
                *.* :    *:.:   * *   * :* :* *    :**::  *

Model.pdb       IVYRGVEVGCWGLCINSVFSSLYSYFQKVLVSYIGLK-----G-------
Template.pdb    ------------------FSSLQQ---GKMFFLIGLTMATIQGAYARRIH
                                 **** .     :.  ***.     *

Model.pdb       -------------LYFTGYLLFGLGTGF--IGLFPNVYSTLVLCSLFGVM
Template.pdb    PGGEVAAVKRALLLLVPAFLLIGWGRSLPVLGLGL------LLYS---FA
                             * ...:**:* * .:  :**        :* *   .

Model.pdb       SSTLYTVPFNLITEYXRVRGKGMDCATLTCMVQLAQILVGGGLGFLVNTA
Template.pdb    AAVVVPCLSSVVAGYGSPGQKGTVMGTLRS------------LGALARAA
                ::.: .   .::: *     **   .** .            ** *..:*

Model.pdb       GTVVVVVITASAVALIGCCFV---------------------
Template.pdb    GPLVAASVYW--LAGAQACFTTWSGLFLLPFFLLQKAENLYFQ
                *.:*.. :    :*   .**.
```

Figure 30. Sequence alignment of SLC45A2 with percentage identity of 14 percent with its template, E-value=4.4e-26 and Score=153.21

Claculated PKa's

Table 11.Calculated pKa's for the folded and unfolded structures.

TYR	Folded/Unfolded		
ResName	WT	S192Y	R402Q
ASP0042P	3.38/3.71	3.72/3.71	3.49/3.71
ASP0075P	2.22/3.77	2.42/3.77	2.52/3.77
ASP0076P	2.45/3.99	3.19/3.99	3.1/3.97
ASP0125P	3.57/3.33	3.32/3.33	3.47/3.32
ASP0132P	3.58/3.78	3.71/3.78	3.58/3.77
ASP0148P	3.19/2.97	3.49/2.97	3.57/2.96
ASP0169P	3.76/3.88	3.47/3.88	3.15/3.88
ASP0174P	2.56/3.68	2.84/3.67	2.67/3.72
ASP0186P	2.95/3.49	3.43/3.49	3.09/3.53
ASP0197P	3.54/3.83	3.59/3.83	2.99/3.81
ASP0199P	3.51/3.54	2.32/3.54	1.92/3.54
ASP0228P	3.03/3.41	2.65/3.41	3.45/3.4
ASP0237P	3.36/3.83	2.43/3.83	2.36/3.85
ASP0240P	3.38/3.94	2.27/3.95	2.99/3.93
ASP0245P	2.84/3.35	3.42/3.35	3.58/3.32
ASP0249P	3.54/3.58	3.5/3.58	3.5/3.54

ASP0305P	3.56/3.54	3.49/3.54	3.28/3.52
ASP0317P	3.34/3.86	3.42/3.86	3.34/3.86
ASP0333P	2.59/3.05	3.41/3.06	2.82/3.04
ASP0356P	3.75/3.96	3.44/3.97	3.68/3.96
ASP0383P	2.03/3.76	2.25/3.76	2.27/3.76
ASP0394P	2.19/3.78	2.25/3.78	2.23/3.79
ASP0437P	3.16/3.41	2.52/3.41	2.53/3.41
ASP0444P	3.46/4.03	3.36/4.03	2.71/4.02
ASP0448P	2.83/3.76	3.34/3.76	3.56/3.75
ASP0454P	undet/3.79	3.24/3.79	3.58/3.79
GLU0032P	3.05/3.81	3.99/3.81	3.59/3.8
GLU0034P	3.51/3.99	3.68/3.99	3.19/3.98
GLU0078P	3.23/4.37	2.75/4.37	2.86/4.35
GLU0114P	3.76/3.64	3.79/3.64	3.68/3.64
GLU0130P	3.38/3.95	3.02/3.95	2.63/3.95
GLU0193P	3.83/3.79	4.09/3.8	3.86/3.8
GLU0203P	3.14/3.83	3.28/3.83	3.23/3.84
GLU0219P	2.57/3.39	2.72/3.39	2.7/3.38
GLU0221P	3.1/4.01	2.83/4.01	3.02/4
GLU0229P	3.8/3.9	3.05/3.9	3.43/3.9
GLU0242P	3.51/3.77	3.54/3.78	3.49/3.76

GLU0250P	3.92/3.96	3.76/3.96	3.64/3.96
GLU0280P	3.74/3.45	3.61/3.45	3.54/3.42
GLU0281P	3.28/3.82	3.02/3.82	3.45/3.82
GLU0294P	2.97/3.05	3.44/3.05	2.94/3.05
GLU0319P	3.74/4	2.71/4	3.89/4
GLU0328P	2.02/3.59	2.89/3.58	2.03/3.56
GLU0345P	1.55/3.57	2.37/3.57	1.82/3.56
GLU0398P	2.09/3.85	2.08/3.85	2.53/3.98
GLU0409P	3.82/3.53	3.74/3.53	3.91/3.6
GLU0413P	3.73/3.77	3.59/3.77	3.85/3.77
GLU0423P	3.55/3.86	3.46/3.92	3.55/3.86
HIS0019P	6.33/6.5	6.38/6.5	6.37/6.5
HIS0143P	6.47/6.36	6.49/6.36	6.46/6.37
HIS0180P	5.38/6.32	5.9/6.32	7.04/6.35
HIS0202P	6.64/7.24	6.87/7.24	6.42/7.27
HIS0211P	5.9/6.38	5.95/6.38	5.71/6.38
HIS0256P	6.55/6.48	6.27/6.48	6.33/6.49
HIS0285P	6.68/6.67	6.83/6.67	6.66/6.67
HIS0304P	6.66/6.63	6.04/6.62	6.54/6.62
HIS0363P	6.04/6.5	6.39/6.5	6.35/6.51
HIS0367P	5.98/6.48	6.12/6.49	5.76/6.48

HIS0389P	5.64/6.52	6.08/6.52	5.87/6.52
HIS0390P	3.59/6.5	5.29/6.5	4.35/6.51
HIS0404P	6.56/6.3	6.68/6.3	6.58/6.46
HIS0420P	6.06/6.85	6.64/6.85	6.97/6.86

TPC2	Folded/Unfolded				
ResName	WT	K376R	V219I	M546I	M484L
ASP0043A	3.93/3.84	3.55/3.84	3.81/3.84	3.3/3.83	3.52/3.84
ASP0047A	3.37/3.86	3.05/3.86	3.27/3.86	2.71/3.83	3.53/3.86
ASP0055A	2.57/3.89	2.82/3.89	2.3/3.89	2.88/3.87	2.41/3.89
ASP0067A	3.04/3.75	2.95/3.75	2.84/3.75	3.04/3.75	3.25/3.75
ASP0110A	2.18/3.92	3.34/3.92	2.96/3.92	2.45/3.85	2.76/3.92
ASP0139A	2.99/3.06	2.71/3.06	3.26/3.06	2.88/2.69	3.15/3.06
ASP0171A	2.11/3.54	2.54/3.54	2.4/3.54	2.72/3.53	2.3/3.54
ASP0244A	3.54/3.63	3.53/3.63	3.92/3.63	3.48/3.58	2.83/3.63
ASP0245A	3.82/3.95	3.74/3.95	3.04/3.95	3.68/3.95	3.8/3.95
ASP0248A	3.76/3.46	3.53/3.46	3.62/3.46	3.77/3.12	3.67/3.46
ASP0276A	3.32/3.95	3.37/3.95	3.43/3.95	3.48/3.93	3.49/3.95
ASP0372A	3.14/3.24	2.99/3.22	3.23/3.24	2.99/3.23	3.14/3.24
ASP0404A	2.78/3.77	3.46/3.77	2.78/3.77	3.18/3.58	3.26/3.77

ASP0435A	3.23/3.89	3.49/3.89	3.03/3.89	3.11/3.87	2.71/3.89
ASP0456A	3.92/3.56	3.72/3.56	3.6/3.56	3.68/3.55	3.48/3.56
ASP0458A	3.83/3.31	3.76/3.31	3.63/3.31	2.79/3.06	3.12/3.31
ASP0465A	3.65/4	3.39/4	3.65/4	3.76/3.94	3.9/4
ASP0466A	3.03/4.09	3.49/4.09	3.51/4.09	3.53/3.63	3.49/4.09
ASP0505A	2.48/3.72	3.03/3.72	2.52/3.72	2.47/3.7	2.6/3.72
ASP0542A	3.28/3.79	3.38/3.79	2.43/3.79	3.16/3.77	2.05/3.79
ASP0637A	2.35/3.9	2.87/3.9	2.85/3.9	2.82/3.84	2.84/3.9
ASP0638A	1.95/3.97	2.35/3.97	2.77/3.97	2.2/3.97	3.33/3.97
ASP0660A	3.21/3.76	3.01/3.76	2.36/3.76	3.3/3.72	3.39/3.76
GLU0054A	2.17/3.85	2.97/3.85	2.44/3.85	3.42/3.82	2.32/3.85
GLU0100A	1.32/3.94	1.91/3.94	2.25/3.94	2.49/3.93	1.84/3.94
GLU0119A	3.85/3.54	4.02/3.54	3.95/3.54	3.92/3.5	3.97/3.54
GLU0126A	2.74/3.83	2.44/3.84	2.42/3.83	3.05/3.8	3.06/3.83
GLU0129A	1.75/3.9	1.79/3.9	2/3.9	1.98/3.87	1.93/3.9
GLU0182A	3.12/3.85	3.44/3.85	3.38/3.85	2.76/3.85	3.54/3.85
GLU0215A	3.84/3.97	3.8/3.97	3.75/3.97	3.65/3.93	3.82/3.97
GLU0250A	2.75/2.83	2.87/2.83	3.4/2.83	1.96/2.87	3.38/2.83
GLU0260A	3.7/4.1	3.59/4.1	3.74/4.1	2.95/4.07	3.72/4.1
GLU0339A	3.66/3.94	3.78/3.94	3.49/3.94	3.78/3.93	3.5/3.94
GLU0347A	3.75/3.91	4.04/3.91	3.87/3.91	3.86/3.91	3.83/3.91

GLU0381A	3.68/3.56	2.76/3.44	3.64/3.56	3.43/3.48	2.82/3.56
GLU0394A	3.55/3.85	3.44/3.85	3.74/3.85	2.68/3.83	3.72/3.85
GLU0395A	3.7/3.46	3.53/3.46	2.45/3.46	3.67/3.43	2.94/3.46
GLU0402A	3.44/4.01	2.14/4.01	2.83/4.01	2.66/3.97	3.01/4.01
GLU0410A	3.18/3.93	3.83/3.93	2.36/3.93	3.41/3.88	3.76/3.93
GLU0416A	3.8/3.53	3.78/3.53	3.35/3.53	3.55/3.53	3.37/3.53
GLU0463A	3.87/3.62	3.89/3.62	3.68/3.62	3.74/3.44	3.8/3.62
GLU0483A	2.02/3.96	2.28/3.96	2.53/3.96	3.1/3.94	3.5/3.96
GLU0516A	3.51/3.97	3.95/3.97	3.95/3.97	3.93/3.97	3.85/3.97
GLU0533A	3.58/3.45	3.5/3.45	2.91/3.45	3.78/3.2	3.49/3.45
GLU0627A	2.8/3.97	3.44/3.97	3.49/3.97	3.02/3.96	3.78/3.97
GLU0630A	3.41/3.99	4.01/3.99	3.77/3.99	4.11/3.97	4.01/3.99
GLU0695A	2.55/4	2.66/4	2.82/4	2.65/4.01	2.47/4
HIS0064A	6.5/6.46	6.05/6.46	6.3/6.46	6.15/6.47	6.28/6.46
HIS0151A	6/6.48	5.81/6.47	5.98/6.48	5.92/6.47	5.69/6.48
HIS0181A	6.82/6.5	6.68/6.5	6.71/6.5	6.74/6.49	6.73/6.5
HIS0226A	6.22/6.73	6.22/6.73	6.02/6.73	6.17/6.73	6.28/6.73
HIS0375A	6.18/6.55	6.52/6.53	6.35/6.55	6.51/6.53	6.46/6.55
HIS0411A	6.04/6.49	5.96/6.49	5.96/6.49	6.26/6.49	6.64/6.49
HIS0431A	6.61/6.25	6.5/6.25	6/6.25	6.46/6.24	6.21/6.25
HIS0527A	6.65/6.49	5.82/6.49	6.66/6.49	6.62/6.49	6.47/6.49

HIS0699A	5.99/6.88	6.26/6.88	6.33/6.88	5.86/6.86	5.76/6.88

SLC45A2	Folded/Unfolded		
ResName	WT	G198V	I374F
ASP0093X	1.73/3.31	2.19/3.31	2.19/3.31
ASP0153X	4.16/3.82	4.2/3.82	4.2/3.82
ASP0157X	3.92/3.94	4.07/3.94	4.07/3.94
ASP0160X	2.05/3.78	2.62/3.78	2.62/3.78
ASP0169X	3.08/3.86	3.49/3.86	3.49/3.86
ASP0175X	2.87/3.19	3.25/3.19	3.25/3.19
ASP0201X	3.99/3.99	4.02/3.99	4.02/3.99
ASP0257X	3.79/3.98	3.83/3.98	3.83/3.98
ASP0263X	3.66/3.84	3.84/3.84	3.84/3.84
ASP0340X	3.69/3.7	3.7/3.7	3.7/3.7
ASP0475X	2.64/3.72	3.49/3.72	3.49/3.72
GLU0026X	3.85/3.81	3.97/3.81	3.97/3.81
GLU0046X	3.43/3.83	3.6/3.83	3.6/3.83
GLU0052X	4.41/3.74	4.46/3.74	4.46/3.74
GLU0177X	3.04/3.45	3.72/3.45	3.72/3.45

GLU0206X	3.65/2.88	4.04/2.88	4.04/2.88
GLU0214X	4.04/3.84	4.19/3.84	4.19/3.84
GLU0239X	3.73/3.77	3.78/3.77	3.78/3.77
GLU0244X	3.62/3.82	3.79/3.82	3.79/3.82
GLU0267X	3.72/3.99	4.01/3.99	4.01/3.99
GLU0272X	2.96/3.66	3.07/3.68	3.07/3.68
GLU0368X	3.96/3.82	3.98/3.82	3.98/3.82
GLU0448X	3.2/3.64	3.68/3.64	3.68/3.64
HIS0038X	5.55/6.48	6.24/6.48	5.55/6.48
HIS0094X	6.24/6.31	6.54/6.31	6.03/6.31
HIS0173X	6.24/6.13	6.35/6.13	6.17/6.13
HIS0181X	6.03/6.26	6.36/6.26	6.3/6.26
HIS0183X	6.22/5.69	6.69/5.69	6.03/5.69
HIS0204X	6.76/6.72	6.69/6.72	6.97/6.72
HIS0233X	4.84/6.46	5.63/6.46	6.07/6.46
HIS0316X	6.22/6.33	6.73/6.33	6.25/6.33
HIS0324X	6.32/6.68	6.51/6.68	6.12/6.68
HIS0450X	6.6/6.34	5.66 /6.34	6.61/6.34

OCA2	Folded/Unfolded						
ResName	WT	V433I	R419Q	P743L	N489D	H615L	A481T
ASP0372U	3.66/3.85	3.63/3.56	3.05/3.94	3.64/3.83	3.48/3.94	3.02/3.86	3.73/3.88
ASP0384U	4.06/4.46	3.92/4.01	4.14/3.94	4.04/3.99	3.98/3.94	3.91/4.02	3.98/4.08
ASP0408U	3.54/4.03	3.19/3.53	3.56/3.93	3.44/3.79	3.36/3.95	3.74/3.84	3.56/3.77
ASP0441U	3.05/4.16	2.74/2.91	1.97/3.94	2.68/3.88	3.37/3.94	3.47/3.79	2.94/3.83
ASP0463U	2.51/3.85	2.96/2.81	2.35/3.93	2.98/3.97	2.69/3.94	2.51/4.16	2.73/3.71
ASP0486U	3.43/4.01	3.51/3.73	3.47/3.95	3.38/3.98	2.52/3.94	3.59/3.87	3/3.99
ASP0504U	3.48/4.13	3.46/3.29	3.26/4.04	3.49/3.76	3.46/3.94	3.69/3.68	3.52/3.75
ASP0601U	2.83/4.09	3.42/2.77	3.11/3.79	2.54/3.53	2.7/3.96	3.25/3.76	2.46/3.32
ASP0619U	2.99/3.73	2.2/2.81	2.86/3.89	2.98/3.89	3.22/3.94	2.87/3.64	2.35/3.83
ASP0649U	3.91/3.79	3.5/3.85	3.56/3.76	3.95/3.75	3.8/3.93	3.18/3.71	3.7/3.63
ASP0666U	4.03/4	3.79/4.1	3.98/3.93	3.97/4.3	4.04/3.94	4.07/3.46	4.02/3.44
ASP0669U	3.57/3.64	3.42/3.71	3.63/3.94	3.75/3.99	3.5/3.94	3.8/3.87	3.54/4.04
ASP0740U	3.49/3.99	3.05/4.05	3.6/4.06	2.33/3.87	3.4/3.93	2.53/3.97	3.5/3.87
ASP0758U	3.93/4.05	3.5/3.99	2.98/4.02	3.9/3.67	4.11/3.96	3.53/3.77	3.88/3.85
GLU0328U	3.77/4.19	3.9/3.54	3.06/4.01	3.5/3.87	3.57/3.92	3.57/4.01	3.48/4.03
GLU0348U	3.32/3.8	3.7/3.2	3.53/4.01	3.2/3.79	3.3/4.02	3.87/3.52	3.01/3.97
GLU0381U	3.81/4.03	3.98/3.48	3.59/4.01	3.91/4.02	3.8/4.04	3.9/3.93	3.89/4.08
GLU0386U	3.31/4.68	3.63/3.86	4.25/4.01	3.55/4.07	3.43/4	3.84/3.91	3.69/4.05
GLU0403U	3.95/4.26	3.54/3.76	3.78/4	3.62/4.17	3.92/4.02	3.65/3.85	3.6/3.96

GLU0458U	3.35/4.09	3.73/3.96	3.4/4	3.85/3.87	3.61/4.03	3.94/4.04	3.75/3.79
GLU0471U	2.5/4.21	3.99/4.01	4.03/4.02	3.92/3.96	3.87/4	4.13/3.98	2.67/4.19
GLU0497U	3.7/3.96	3.6/3.23	3.61/4	3.24/3.97	3.81/4	3.42/3.74	3.59/3.95
GLU0540U	3.76/4.4	3.72/3.51	3.68/4.03	3.91/4	3.44/4.02	3.89/4.05	3.43/4.01
GLU0543U	3.55/4.38	3.75/3.82	2.31/4.02	3.75/3.85	3.98/4	3.78/3.84	3.76/3.26
GLU0546U	2.78/3.92	3.26/3.2	3.55/4	2.83/3.94	3.08/4	3.28/3.98	2.94/3.96
GLU0550U	3.16/3.63	2.92/3.8	3.77/4.03	3.52/3.65	2.65/4	3.62/3.78	3.56/3.71
GLU0567U	2.27/4.05	3.27/3.58	3.02/4.01	3.09/3.95	3.74/4	3.86/3.66	3.93/3.72
GLU0568U	2.25/4	4.03/3.6	3.41/4.03	3.34/3.77	2.34/4	1.89/3.43	3.94/3.49
GLU0583U	3.83/4.12	3.73/3.7	3.08/4.07	3.75/3.78	3.39/4.02	3.47/3.54	3.63/3.56
GLU0600U	3.82/4.05	2.8/3.98	2.71/4.01	3.78/4.06	3.07/4	3.79/3.42	3.73/3.55
GLU0605U	2.5/4.01	2.46/3.22	3.77/3.95	2.33/3.68	2.46/4.04	2.85/3.2	2.82/3.72
GLU0610U	3.19/3.98	3.87/3.56	3.82/3.97	2.4/3.97	2.01/4	2.87/3.8	2.43/3.84
GLU0671U	3.98/3.41	3.96/3.89	3.79/3.86	3.88/3.04	3.98/4.02	3.91/3.96	3.98/4.05
GLU0678U	3.22/4.15	2.49/3.31	3.62/3.97	3.41/3.85	3.72/4	2.99/3.82	3.55/3.94
GLU0693U	3.07/4	3.42/2.81	2.54/3.63	3.07/3.82	2.54/3.84	3/3.62	2.81/4.04
GLU0702U	3.96/4.21	4.03/3.84	3.32/3.99	3.6/4.03	4.15/3.68	3.88/4.01	3.78/3.93
GLU0706U	3.22/4.29	3.89/4.08	4.15/3.77	3.97/3.83	4.12/3.96	4.05/4.12	4.08/3.78
GLU0717U	4.05/3.56	3.91/4.05	4.08/3.92	4.06/3.83	4.15/4.04	4.01/3.95	4.01/3.82
GLU0718U	3.45/3.5	3.97/4.07	4.04/4.33	4.03/3.93	4.24/3.97	4.02/4.15	4.01/3.53
GLU0760U	3.29/3.8	3.98/4.08	3.92/3.98	3.92/4.01	3.94/3.8	4/3.99	3.8/4

GLU0798U	4.07/4.56	3.65/3.51	3.96/3.96	3.81/3.82	3.7/3.74	3.02/3.97	3.78/3.89
GLU0808U	1.92/3.53	1.98/1.74	2.46/6.48	1.97/3.65	1.79/3.92	2.35/3.83	2.3/3.95
HIS0351	6.4/6.64	6.2/5.97	6.1/6.48	6.1/6.62	6.3/3.76	6.2/6.82	6.5/6.5
HIS0378	6/6.85	6.9/6.62	6/6.49	6.7/6.51	6.5/6.49	6.1/6.33	6.8/6.37
HIS0511	6.8/6.53	6.3/6.66	6.4/6.49	6.5/6.55	6.6/6.49	6.9/6.39	6.6/6.26
HIS0549	6.5/6.33	6.2/6.54	6.3/6.49	6.3/6.44	6.2/6.48	6.4/6.68	6.2/6.29
HIS0552	6.2/7.12	6.3/6.44	6.3/6.5	5.8/6.81	6.5/6.49	6.3/6.37	6.4/6.8
HIS0584	6.1/6.38	6/5.6	6.3/6.49	6.3/6.37	6.1/6.49	6.5/6.67	5.7/6.52
HIS0591	6.2/6.58	6.2/5.8	6.2/6.49	5.7/5.98	6.1/6.49	6.1/5.95	6.1/6.15
HIS0594	5.9/5.82	6.2/6.01	6.1/6.27	6.3/6.51	6.2/6.49	6.1/5.83	6.6/6.58
HIS0647	6.2/6.76	6.6/7.02	6.6/7.01	6.3/6.54	6.7/6.5	6.8/6.52	6.2/6.17
HIS0668	6.8/6.9	6.9/6.76	7.1/6.72	6.6/6.72	6.5/6.49	6.8/6.38	6.4/6.54
HIS0675	6.6/6.64	6.7/6.72	6.8/6.86	6.7/6.29	6.7/6.49	6.7/6.5	6.7/6.52
HIS0697	6.6/6.59	6.6/6.76	6.7/6.55	6.6/6.21	6.8/6.18	6.7/6.64	6.7/6.26
HIS0699	6.5/6.55	6.1/7.03	6.6/6.78	6.9/6.4	7.1/6.9	6.5/6.84	6.4/6.29
HIS0757	6.7/7.22	7.1/6.59	6.7/6.67	6.8/7	7/6.51	6.7/6.53	6.9/6.54
HIS0800	6.5/6.77	6.4/6.46	6.7/10.77	6.6/6.63	6.6/6.76	6.5/10.66	6.4/6.82

ATP7A	Folded/Unfolded								
ResName	WT	C1002F	D1044E	G666R	I1264V	K742R	M1311V	R844C	S653Y
ASP0675P	3.52/3.47	3.23/3.53	3.19/3.35	3.52/3.47	3.73/3.35	3.36/3.2	3.65/3.35	undet/3.55	3.54/3.47
ASP0747P	2.46/2.23	2.75/2.7	2.99/2.68	2.91/3.1	3.07/3.78	3.18/2.81	3.15/3.78	2.52/2.28	2.58/2.7
ASP0782P	3.37/3.65	3.53/3.35	3.62/3.69	3.46/3.48	3.32/3.91	3.51/3.49	3.7/3.91	3.55/3.62	3.75/3.77
ASP0826P	3.38/3.31	3.48/3.13	3.58/3.89	3.43/3.84	3.79/3.93	3.52/3.9	3.89/3.93	3.01/3.05	4/3.23
ASP0828P	3.95/4	3.94/3.99	3.98/3.96	4.1/3.96	3.86/3.98	3.93/3.96	3.57/3.98	3.79/3.73	3.99/4
ASP0838P	3.95/3.98	3.91/3.9	3.95/3.84	3.87/3.95	3.95/4.11	3.82/3.94	3.93/4.11	3.93/3.91	3.97/3.88
ASP0846P	3.58/3.76	3.72/3.67	3.81/3.71	3.59/3.77	3.68/2.98	3.7/3.71	3.74/2.98	3.84/3.85	3.77/3.69
ASP0859P	1.79/1.5	2.31/2.37	3.15/2.71	2.05/2.72	2.64/3.28	2.78/2.51	2.55/3.28	2.51/2.19	2.44/2.45
ASP0870P	3.74/3.54	3.48/3.49	3.66/3.68	3.93/3.53	3.5/3.95	3.41/3.59	3.57/3.95	2.73/2.83	3.89/3.91
ASP0910P	3.51/3.56	3.13/3.39	3.56/3.92	3.64/3.46	3.38/3.78	3.73/3.71	3.75/3.78	3.58/3.47	3.43/3.52
ASP0935P	3.35/3.44	2.66/2.93	2.53/2.32	3.11/3.08	3.28/3.77	3.35/3.08	3.08/3.77	2.23/2.19	2.65/2.57
ASP1044P	2.75/2.4	2.45/2.46	1.56/3.66	2.55/3.57	2.69/3.54	3.14/3.25	2.34/3.54	2.51/2.49	2.32/2.79
ASP1101P	3.74/3.84	3.73/3.58	3.57/3.67	3.44/3.53	3.64/3.74	3.76/3.76	3.67/3.74	3.85/3.84	3.7/3.7
ASP1110P	3.4/3.31	3.57/3.55	3.66/3	3.7/3.53	3.22/3.8	3.13/3.2	3.36/3.8	3.49/3.55	3.3/3.67
ASP1139P	3.34/3.84	3.7/3.69	3.75/3.65	3.67/3.48	3.58/3.86	3.78/3.85	3.39/3.86	3.73/3.7	3.77/3.57
ASP1151P	3.65/3.72	3.44/3.49	3.81/3.41	2.88/3.69	3.75/3.88	3.81/3.94	3.8/3.88	3.83/3.85	3.54/3.84
ASP1166P	3.81/3.82	3.81/3.85	3.49/3.59	3.48/3.78	3.56/4.26	3.71/3.46	3.75/4.26	3.82/3.76	3.93/3.84
ASP1198P	3.78/3.42	3.76/3.78	3.64/3.72	3.75/3.7	3.47/3.98	3.72/3.6	3.57/3.98	3.62/3.57	3.37/3.63
ASP1201P	3.74/3.71	3.76/3.76	3.6/2.16	3.65/3.15	3.62/3.82	3.7/3.72	3.69/3.82	3.77/3.69	3.7/3.81
ASP1219P	3.65/3.47	3.34/3.44	2.05/3.59	3.18/3.93	2.33/3.82	3.53/3.35	2.59/3.82	3.52/3.5	3.19/3.71
ASP1220P	3.04/2.7	3.35/3.62	3.58/2.69	3.97/2.96	3.95/3.77	2.97/3.41	2.99/3.77	3.33/3.35	3.13/3.4
ASP1230P	2.87/3.2	2.93/3.31	2.66/3.25	2.46/3.17	1.97/3.96	2.51/2.27	2.72/3.96	2.19/2.24	3.01/3.05
ASP1256P	2.72/2.62	2.97/2.94	3.61/3.62	2.98/2.68	2.53/3.67	3.05/2.81	2.72/3.67	2.82/2.92	3.01/2.99
ASP1301P	3.52/3.33	3.38/3.52	3.61/2.66	3.36/3.01	3.95/3.83	3.43/3.41	3.49/3.83	3.15/3.19	3.02/3.01
ASP1305P	2.42/2.65	2.89/2.65	2.74/3.31	2.55/3.56	2.3/3.56	2.46/2.23	2.74/3.55	2.46/1.94	2.11/1.98

ASP1323P	3.73/3.79	3.74/3.78	3.53/3.47	3.58/3.41	3.73/3.76	3.6/3.58	3.6/3.76	3.5/3.57	3.56/3.55
ASP1330P	2.71/2.8	2.67/2.38	3.56/3.16	3.46/3.68	3.44/3.8	3.29/3.24	2.65/3.8	2.88/2.79	3.41/3.49
ASP1337P	3.57/3.65	3.46/3.44	3.17/2.91	3.18/3.54	2.89/3.93	3.26/3.49	3.21/3.93	3.51/3.56	3.18/3.49
ASP1340P	2.71/2.98	2.31/2.43	2.94/2.71	3.67/2.7	3.05/3.7	3.67/3.84	2.66/3.7	2.9/3.09	3.47/3.41
ASP1346P	2.38/2.39	2.54/2.84	2.8/3.49	2.46/3.54	2.63/3.85	2.78/2.85	2.78/3.85	2.47/2.53	2.33/2.47
GLU0646P	3.37/2.39	3.83/3.66	3.9/3.87	3.56/3.84	3.78/4	3.13/2.72	3.72/4	3.72/3.73	3.21/3.56
GLU0690P	3.73/3.86	3.59/3.81	3.97/2.22	3.65/3.91	3.9/3.98	3.75/3.81	3.74/3.98	3.28/3.14	3.95/3.95
GLU0691P	3.83/3.96	3.75/4.01	2.18/3.44	3.75/2.72	3.75/3.99	2.78/3.69	3.94/3.99	3.88/3.95	3.77/3.94
GLU0702P	3.83/3.49	3.8/3.96	3.77/3.05	1.53/3.15	3.85/4.01	3.71/3.8	3.46/4.01	3.71/3.66	3.43/3.48
GLU0771P	3/3.46	3.42/3.48	3.26/3.03	3.62/2.69	2.86/3.65	3.12/2.76	3.42/3.65	3.45/3.64	3.23/3.05
GLU0798P	3.23/3.15	2.56/2.93	2.49/3.93	2.89/3.24	3.02/3.74	2.41/3.14	3.38/3.74	2.61/2.63	3.07/2.49
GLU0807P	3.71/3.83	4.01/3.93	4.14/4.01	3.84/4.02	3.24/3.81	4.05/3.92	3.95/3.81	2.84/3.13	3.69/3.29
GLU0819P	4.09/4.02	4.04/3.94	3.95/3.69	4.01/3.44	4.09/4.51	3.9/3.85	3.98/4.51	4.14/4.11	3.87/3.69
GLU0834P	3.47/3.52	3.26/3.45	3.52/4.13	3.12/4.04	2.97/3.94	3.06/3.48	3.27/3.94	3.49/3.47	3.29/3.21
GLU0835P	4.11/4.2	4.24/4.13	4.27/4.16	4.28/3.97	4.07/4.15	4.24/4.06	4.13/4.15	4.07/4.09	4.32/4.1
GLU0840P	4.06/4.05	4.02/4.04	4.01/3.74	4.02/3.74	3.94/4.01	3.96/4.08	4.01/4.01	4/3.98	3.98/3.98
GLU0864P	3.64/3.74	3.52/3.78	3.64/3.03	3.8/2.96	3.54/3.94	3.81/3.82	3.48/3.94	3.55/3.64	3.82/3.66
GLU0871P	3.33/2.81	3.17/3.23	2.92/4.06	2.83/3.76	3.3/3.81	3.34/3.43	3.45/3.81	3.67/3.86	3.49/3.46
GLU0877P	3.53/3.75	3.78/3.73	3.95/3.95	3.19/3.86	3.59/3.76	3.54/3.61	3.66/3.76	3.38/3.16	3.88/3.94
GLU0921P	3.76/3.76	3.8/3.91	3.57/3.55	3.32/4	3.74/3.73	3.97/4.02	3.45/3.73	3.72/3.87	3.75/3.13
GLU0922P	3.81/3.84	2.84/3.76	3.73/3.96	3.57/3.87	3.38/3.49	3.54/2.96	3.65/3.49	3.87/3.24	3.8/3.93
GLU0965P	3.26/4	4/3.78	4.15/3.96	3.87/4	2.75/3.97	4.01/3.73	1.99/3.97	4.01/3.88	2.69/2.77
GLU0968P	3.97/4.02	4.02/3.99	3.97/3.72	3.82/2.84	3.97/4.01	4/4.01	3.98/4.01	3.96/4.01	4.03/3.98

GLU0982P	3.76/3.14	3.54/3.51	3.7/3.14	3.51/3.54	2.7/3.77	2.88/2.77	3.02/3.77	3.52/3.54	3.69/3.82
GLU1030P	3.06/2.7	3.36/3.48	3.05/1.12	3.77/3.07	3.63/4.07	2.45/2.48	3.12/4.07	3.09/3.31	3.03/3.3
GLU1033P	3.1/2.99	3.05/3.08	2.23/2.1	3.03/2.39	2.94/3.83	3.12/3.1	3.23/3.83	2.76/2.78	2.83/2.93
GLU1064P	3.56/3.72	3.54/3.82	3.73/3.72	3.76/3.72	3.11/3.87	3.91/3.88	3.58/3.87	3.81/3.9	3.87/3.48
GLU1081P	2.27/3.29	2.85/3.18	3.5/3.49	3.33/3.34	2.63/4.01	3.39/3.53	3.47/4.01	3.47/3.46	2.36/2.3
GLU1085P	2.99/2.85	3.85/3.94	3.91/3.96	3.83/3.98	3.98/3.68	3.68/3.9	3.97/3.68	3.79/4.01	3.84/3.85
GLU1099P	3.85/3.79	3.56/3.67	3.85/4	3.97/3.97	3.97/3.23	3.96/3.97	3.93/3.23	3.56/3.58	3.87/3.98
GLU1103P	3.84/3.73	3.3/3.86	3.3/2.64	3.57/3.16	2.97/3.92	3.13/3.86	3.27/3.92	3.81/3.75	3.54/2.65
GLU1127P	3.16/3.73	3.86/3.6	3.84/3.87	3.99/3.83	3.84/3.87	3.73/3.67	3.83/3.87	3.79/3.72	2.75/3.64
GLU1138P	3.77/3.93	3.74/3.71	3.78/3.97	4/3.94	3.74/3.95	3.99/4.04	3.98/3.95	3.82/3.78	3.82/3.94
GLU1155P	3.74/3.85	4.01/4	3.96/3.85	3.75/3.66	4.07/4.01	3.96/3.68	3.02/4.01	3.69/3.85	4.04/3.97
GLU1186P	3.4/2.45	2.25/3.19	2.98/2.8	2.36/2.85	2.48/3.84	3.41/3.15	3.46/3.84	3.54/3.43	3.49/3.47
GLU1205P	3.52/3.42	3.27/3.22	2.86/2.93	3.23/3.03	2.98/3.22	2.98/3.18	3.25/3.22	3.05/2.79	3.47/3.06
GLU1207P	2.73/2.68	3.07/3.38	2.93/3.49	2.5/2.46	3.34/3.47	3.12/3.46	3.22/3.47	2.74/3.33	3.39/3.41
GLU1221P	2.91/3.67	4.12/4.1	2.94/3.54	4.07/4.07	3.15/3.55	4.06/3.71	4.09/3.55	4.11/4.12	3.53/3.68
GLU1235P	3.92/3.82	3.49/3.38	2.58/2.48	3.52/3.4	3.81/4.04	3.59/3.56	3.83/4.04	3.53/3.61	3.54/3.74
GLU1237P	3.92/3.92	3.59/3.76	3.8/3.93	3.77/3.83	3.78/4	3.85/3.67	3.56/4	3.75/3.75	3.78/3.84
GLU1249P	2.35/3.11	2.89/2.74	2.98/2.62	3.04/2.99	3.15/3.32	3.15/2.85	3.66/3.32	2.52/2.8	2.86/2.72
GLU1276P	3.42/3.52	3.51/3.58	3.61/3.6	3.49/3.57	3.46/3.96	3.6/3.67	3.36/3.96	3.34/3.08	3.48/3.55
GLU1291P	3.73/3.79	3.88/3.92	3.51/3.49	4.02/3.21	3.34/3.95	3.68/3.78	3.89/3.95	3.3/3.55	3.42/3.02
GLU1292P	3.55/3.81	3.93/3.8	3.94/3.63	3.97/3.82	3.83/3.77	3.77/3.6	3.82/3.77	3.55/3.94	3.56/3.72
GLU1327P	3.53/2.95	3.72/3.63	3.52/3.77	3.18/3.51	3.7/3.7	3.09/2.92	3.36/3.7	4/4	3.06/2.55
HIS0676P	6.05/6.51	6.65/6.41	6.33/6.49	6.66/6.57	6.71/6.51	6.66/6.69	6.46/6.51	6.46/6.52	6.29/6.48

HIS0677P	6.48/6.42	6.46/6.34	5.95/4.17	6.32/6.52	6.53/6.95	6.04/6.33	6.19/6.95	6.2/6.29	6.19/6.33
HIS0682P	5.54/5.97	6.46/6.56	6.46/6.33	6.42/6.5	6.41/6.85	6.45/6.34	6.45/6.85	6.48/6.43	6.46/6.48
HIS0683P	6.49/6.46	6.51/6.28	6.04/6.17	6.34/6.54	6.04/6.54	6.49/6.45	6.2/6.54	6.34/6.33	6.25/6.28
HIS0696P	6.64/6.51	6.45/6.52	6.57/6.55	6.67/6.51	6.66/6.47	6.66/6.5	6.71/6.47	6.5/6.49	6.53/6.54
HIS0741P	6.37/6.46	6.32/6.46	6.14/6.18	6.51/6.48	6.17/6.49	6.14/6.18	6.3/6.49	6.17/6.13	6.21/5.62
HIS0799P	6.3/6.38	6.29/6.35	6.35/6.26	5.99/6.54	6.47/6.33	6.32/6.16	6.38/6.33	6.3/6.24	6.34/6.52
HIS0866P	6.46/6.45	5.85/5.81	5.53/6.25	5.63/6.16	5.99/6.47	6.28/6.36	6.18/6.47	6.29/6.31	6.29/6.34
HIS0906P	6.47/6.51	6.54/6.25	6.32/6.66	6.49/6.3	6.78/6.06	6.62/6.64	6.36/6.06	6.56/6.53	6.47/6.67
HIS1036P	5.67/5.81	5.76/5.78	5.5/5.78	5.77/5.7	5.99/6.79	5.96/6.11	5.88/6.79	5.71/5.77	5.75/5.37
HIS1051P	6.37/6.46	6.44/6.29	6.3/6.22	6.34/6.34	6.47/6.24	6.48/6.44	6.33/6.24	6.27/6.2	6.05/6.34
HIS1070P	6.17/5.93	5.94/6.29	5.76/6.06	6.14/6.17	5.98/6.44	6.04/6.06	6.18/6.44	6.07/6.14	5.48/5.57
HIS1071P	6.31/6.68	6.52/6.56	6.6/6.53	6.36/6.18	6.5/6.48	6.16/6.22	6.51/6.48	6.22/6.47	6.46/6.35
HIS1086P	6.53/6.34	6.47/6.65	6.64/6.56	6.2/6.33	6.01/6.27	6.57/6.27	6.15/6.27	6.38/6.45	6.55/6.65
HIS1131P	6.49/6.48	5.95/6.45	6.56/6.54	6.65/6.68	6.61/6.46	6.74/6.53	6.52/6.46	6.35/6.26	6.48/6.24
HIS1206P	6.25/6.15	6.06/5.96	6.06/6.01	6.13/5.93	5.93/6.71	5.99/6.06	6.18/6.71	6.18/6.17	6.16/6.02
HIS1241P	6.49/6.51	6.5/6.54	6.32/6.46	6.49/6.45	6.46/6.51	6.49/6.46	6.54/6.51	6.34/6.35	6.28/6.46
HIS1281P	6.12/5.52	5.83/6.11	6.22/6.26	5.76/6.45	6.05/6.47	6.33/6.29	6.05/6.47	5.36/5.98	6.2/6.27

* undet = Indetermined

pKa (Folded-Unfolded)	Color code
$0.5 \leq \Delta pKa < 1.0$	
$\Delta pKa \geq 1.0$	
$-1.0 < \Delta pKa \leq -0.5$	
$\Delta pKa \leq -1.0$	

REFERENCES

1. Cisneros, G.A., et al., *Classical electrostatics for biomolecular simulations.* Chemical reviews, 2014. **114**(1): p. 779-814.
2. Shashikala, H.B.M., A. Chakravorty, and E. Alexov, *Modeling Electrostatic Force in Protein-Protein Recognition.* Frontiers in Molecular Biosciences, 2019. **6**.
3. Pinak, M., *CHAPTER 10 - Enzymatic recognition of radiation-produced oxidative DNA lesion. Molecular dynamics approach*, in *Modern Methods for Theoretical Physical Chemistry of Biopolymers*, E.B. Starikov, J.P. Lewis, and S. Tanaka, Editors. 2006, Elsevier Science: Amsterdam. p. 191-210.
4. Li, L., et al., *Forces and Disease: Electrostatic force differences caused by mutations in kinesin motor domains can distinguish between disease-causing and non-disease-causing mutations.* Scientific reports, 2017. **7**(1): p. 1-12.
5. Koirala, M. and E. Alexov, *Computational chemistry methods to investigate the effects caused by DNA variants linked with disease.* Journal of Theoretical and Computational Chemistry, 2020. **19**(06): p. 1930001.
6. Kucukkal, T.G., et al., *Impact of rett syndrome mutations on MeCP2 MBD stability.* Biochemistry, 2015. **54**(41): p. 6357-6368.
7. Kim, S. and A. Misra, *SNP genotyping: technologies and biomedical applications.* Annu. Rev. Biomed. Eng., 2007. **9**: p. 289-320.
8. Chanock, S., *Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease*. Disease markers, 2001. **17**(2): p. 89-98.
9. Stalin, A., et al., *Computational analysis of single nucleotide polymorphisms (SNPs) in PPAR gamma associated with obesity, diabetes and cancer*. Journal of Biomolecular Structure and Dynamics, 2022. **40**(4): p. 1843-1857.
10. Sukhumsirichart, W., *Polymorphisms*. 2018: IntechOpen.
11. Li, G., S.K. Panday, and E. Alexov, *SAAFEC-SEQ: a sequence-based method for predicting the effect of single point mutations on protein thermodynamic stability.* International journal of molecular sciences, 2021. **22**(2): p. 606.
12. Pahari, S., et al., *SAAMBE-3D: predicting effect of mutations on protein–protein interactions.* International journal of molecular sciences, 2020. **21**(7): p. 2563.
13. Li, G., et al., *SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity.* Bioinformatics, 2021. **37**(7): p. 992-999.
14. Koirala, M., et al., *Computational Investigation of the pH Dependence of Stability of Melanosome Proteins: Implication for Melanosome formation and Disease.* International journal of molecular sciences, 2021. **22**(15): p. 8273.
15. Ganakammal, S.R., et al., *In-silico analysis to identify the role of MEN1 missense mutations in breast cancer.* Journal of Theoretical and Computational Chemistry, 2020. **19**(06): p. 2041002.
16. Jankauskaitė, J., et al., *SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation.* Bioinformatics, 2019. **35**(3): p. 462-469.

17. Kucukkal, T.G. and E. Alexov, *Structural, dynamical, and energetical consequences of Rett syndrome mutation R133C in MeCP2.* Computational and mathematical methods in medicine, 2015. **2015**.
18. Liu, Z., et al., *PDB-wide collection of binding data: current status of the PDBbind database.* Bioinformatics, 2015. **31**(3): p. 405-412.
19. Zhang, N., et al., *PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions.* PLoS computational biology, 2018. **14**(12): p. e1006615.
20. De Las Rivas, J. and C. Fontanillo, *Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.* PLoS computational biology, 2010. **6**(6): p. e1000807-e1000807.
21. Koirala, M. and E. Alexov, *Ab-initio binding of barnase–barstar with DelPhiForce steered Molecular Dynamics (DFMD) approach.* Journal of Theoretical and Computational Chemistry, 2020. **19**(04): p. 2050016.
22. Peng, X., et al., *Protein–protein interactions: detection, reliability assessment and applications.* Briefings in Bioinformatics, 2016. **18**(5): p. 798-819.
23. Kuzmanov, U. and A. Emili, *Protein-protein interaction networks: probing disease mechanisms using model systems.* Genome Medicine, 2013. **5**(4): p. 37.
24. Lu, H., et al., *Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials.* Signal Transduction and Targeted Therapy, 2020. **5**(1): p. 213.
25. Schwede, T., et al., *SWISS-MODEL: an automated protein homology-modeling server.* Nucleic acids research, 2003. **31**(13): p. 3381-3385.
26. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Current opinion in structural biology, 1996. **6**(3): p. 402-406.
27. Abagyan, R., M. Totrov, and D. Kuznetsov, *ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation.* Journal of computational chemistry, 1994. **15**(5): p. 488-506.
28. Bottegoni, G., et al., *A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE).* Journal of Computer-Aided Molecular Design, 2008. **22**(5): p. 311-325.
29. Alexov, E., *Numerical calculations of the pH of maximal protein stability: The effect of the sequence composition and three-dimensional structure.* European Journal of Biochemistry, 2004. **271**(1): p. 173-185.
30. Peng, Y. and E. Alexov, *Computational investigation of proton transfer, p K a shifts and p H-optimum of protein–DNA and protein–RNA complexes.* Proteins: Structure, Function, and Bioinformatics, 2017. **85**(2): p. 282-295.
31. Mitra, R.C., Z. Zhang, and E. Alexov, *In silico modeling of pH-optimum of protein–protein binding.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(3): p. 925-936.
32. Peng, Y., et al., *pH-dependent interactions of Apolipophorin-III with a lipid disk.* Journal of Computational Biophysics and Chemistry, 2021. **20**(02): p. 153-164.
33. Xiao, L., et al., *A continuum Poisson–Boltzmann model for membrane channel proteins.* Journal of chemical theory and computation, 2017. **13**(7): p. 3398-3412.

34. Li, C., et al., *Progress in developing Poisson-Boltzmann equation solvers.* Computational and Mathematical Biophysics, 2013. **1**(2013): p. 42-62.
35. Li, L., et al., *DelPhi: a comprehensive suite for DelPhi software and associated resources.* BMC biophysics, 2012. **5**(1): p. 1-11.
36. Li, C., et al., *DelPhi suite: New developments and review of functionalities.* Journal of computational chemistry, 2019. **40**(28): p. 2502-2508.
37. Li, L., et al., *On the dielectric "constant" of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi.* Journal of chemical theory and computation, 2013. **9**(4): p. 2126-2136.
38. Jia, Z., et al., *Treating ion distribution with G aussian-based smooth dielectric function in DelPhi*. 2017, Wiley Online Library.
39. Chakravorty, A., et al., *Gaussian-based smooth dielectric function: a surface-free approach for modeling macromolecular binding in solvents.* Frontiers in molecular biosciences, 2018. **5**.
40. Panday, S.K., et al., *Modeling electrostatics in molecular biology: A tutorial of DelPhi and associated resources [Article v1. 0]*. Living Journal of Computational Molecular Science, 2019. **1**(2): p. 10841-10841.
41. Wang, L., L. Li, and E. Alexov, *pKa predictions for proteins, RNA s, and DNA s with the Gaussian dielectric function using DelPhi pKa.* Proteins: Structure, Function, and Bioinformatics, 2015. **83**(12): p. 2186-2197.
42. Li, L., A. Chakravorty, and E. Alexov, *DelPhiForce, a tool for electrostatic force calculations: Applications to macromolecular binding.* Journal of computational chemistry, 2017. **38**(9): p. 584-593.
43. Li, L., et al., *DelPhiForce web server: electrostatic forces and energy calculations and visualization.* Bioinformatics, 2017. **33**(22): p. 3661-3663.
44. Getov, I., M. Petukh, and E. Alexov, *SAAFEC: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified MM/PBSA approach.* International journal of molecular sciences, 2016. **17**(4): p. 512.
45. Petukh, M., L. Dai, and E. Alexov, *SAAMBE: webserver to predict the charge of binding free energy caused by amino acids mutations.* International journal of molecular sciences, 2016. **17**(4): p. 547.
46. Peng, Y., et al., *Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver.* Bioinformatics, 2018. **34**(5): p. 779-786.
47. Szklarczyk, D., et al., *STRING v10: protein–protein interaction networks, integrated over the tree of life.* Nucleic acids research, 2015. **43**(D1): p. D447-D452.
48. Li, L., et al., *Forces and Disease: Electrostatic force differences caused by mutations in kinesin motor domains can distinguish between disease-causing and non-disease-causing mutations.* Sci Rep, 2017. **7**(1): p. 8237.
49. Peng, Y., et al., *Revealing the Effects of Missense Mutations Causing Snyder-Robinson Syndrome on the Stability and Dimerization of Spermine Synthase.* Int J Mol Sci, 2016. **17**(1).

50. Peng, Y. and E. Alexov, *Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding.* Proteins, 2016. **84**(2): p. 232-9.
51. Petukh, M., T.G. Kucukkal, and E. Alexov, *On human disease-causing amino acid variants: statistical study of sequence and structural patterns.* Hum Mutat, 2015. **36**(5): p. 524-534.
52. Kucukkal, T.G., et al., *Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins.* Curr Opin Struct Biol, 2015. **32**: p. 18-24.
53. Petukh, M., et al., *Chronic Beryllium Disease: revealing the role of beryllium ion and small peptides binding to HLA-DP2.* PLoS One, 2014. **9**(11): p. e111604.
54. Nishi, H., et al., *Cancer missense mutations alter binding properties of proteins and their interaction networks.* PLoS One, 2013. **8**(6): p. e66273.
55. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proceedings of the National Academy of Sciences, 1996. **93**(1): p. 13-20.
56. Chothia, C., *The nature of the accessible and buried surfaces in proteins.* Journal of molecular biology, 1976. **105**(1): p. 1-12.
57. Li, Y., X. Zhang, and D. Cao, *The Role of Shape Complementarity in the Protein-Protein Interactions.* Scientific Reports, 2013. **3**(1): p. 3271.
58. Skrabanek, L., et al., *Computational prediction of protein–protein interactions.* Molecular biotechnology, 2008. **38**(1): p. 1-17.
59. Yan, C., et al., *Characterization of protein–protein interfaces.* The protein journal, 2008. **27**(1): p. 59-70.
60. Jones, S. and J.M. Thornton, *Analysis of protein-protein interaction sites using surface patches.* Journal of molecular biology, 1997. **272**(1): p. 121-132.
61. Teng, S., et al., *Modeling effects of human single nucleotide polymorphisms on protein-protein interactions.* Biophysical journal, 2009. **96**(6): p. 2178-2188.
62. Jochim, A.L. and P.S. Arora, *Assessment of helical interfaces in protein–protein interactions.* Molecular BioSystems, 2009. **5**(9): p. 924-926.
63. Petukh, M., M. Zhang, and E. Alexov, *Statistical investigation of surface bound ions and further development of BION server to include pH and salt dependence.* J Comput Chem, 2015. **36**(32): p. 2381-93.
64. Petukh, M., T. Kimmet, and E. Alexov, *BION web server: predicting non-specifically bound surface ions.* Bioinformatics, 2013. **29**(6): p. 805-6.
65. Petukh, M., et al., *Predicting nonspecific ion binding using DelPhi.* Biophys J, 2012. **102**(12): p. 2885-93.
66. Takano, K., et al., *An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity.* Hum Mol Genet, 2012. **21**(20): p. 4497-507.
67. Witham, S., et al., *A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics.* Proteins, 2011. **79**(8): p. 2444-54.
68. Peng, Y. and E. Alexov, *Computational investigation of proton transfer, pKa shifts and pH-optimum of protein-DNA and protein-RNA complexes.* Proteins, 2017. **85**(2): p. 282-295.

69. Yang, Y., et al., *Binding Analysis of Methyl-CpG Binding Domain of MeCP2 and Rett Syndrome Mutations.* ACS Chem Biol, 2016. **11**(10): p. 2706-2715.
70. Kucukkal, T.G., et al., *Impact of Rett Syndrome Mutations on MeCP2 MBD Stability.* Biochemistry, 2015. **54**(41): p. 6357-68.
71. Li, L., J. Alper, and E. Alexov, *Cytoplasmic dynein binding, run length, and velocity are guided by long-range electrostatic interactions.* Sci Rep, 2016. **6**: p. 31523.
72. Li, L., J. Alper, and E. Alexov, *Multiscale method for modeling binding phenomena involving large objects: application to kinesin motor domains motion along microtubules.* Sci Rep, 2016. **6**: p. 23249.
73. Tajielyato, N., et al., *E-hooks provide guidance and a soft landing for the microtubule binding domain of dynein.* Sci Rep, 2018. **8**(1): p. 13266.
74. Salsbury Jr, F.R., *Molecular dynamics simulations of protein dynamics and their relevance to drug discovery.* Current opinion in pharmacology, 2010. **10**(6): p. 738-744.
75. Duan, L., et al., *Accelerated Molecular Dynamics Simulation for Helical Proteins Folding in Explicit Water.* Front Chem, 2019. **7**: p. 540.
76. Zhang, L.Y., et al., *Solvent models for protein–ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations.* Journal of Computational Chemistry, 2001. **22**(6): p. 591-607.
77. Mongan, J., et al., *Generalized Born model with a simple, robust molecular volume correction.* Journal of chemical theory and computation, 2007. **3**(1): p. 156-169.
78. Peng, Y., et al., *DFMD: Fast and Effective DelPhiForce Steered Molecular Dynamics Approach to Model Ligand Approach Toward a Receptor: Application to Spermine Synthase Enzyme.* Frontiers in Molecular Biosciences, 2019. **6**(74).
79. Li, L., et al., *DelPhi: a comprehensive suite for DelPhi software and associated resources.* BMC biophysics, 2012. **5**(1): p. 9.
80. Panday, S.K., et al., *Modeling electrostatics in molecular biology: A tutorial of DelPhi and associated resources [Article v1. 0].* Living Journal of Computational Molecular Science, 2019. **1**(2): p. 10841.
81. Wang, J., et al., *Development and testing of a general amber force field.* Journal of computational chemistry, 2004. **25**(9): p. 1157-1174.
82. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of computational chemistry, 2005. **26**(16): p. 1781-1802.
83. Tanner, D.E., et al., *Parallel generalized Born implicit solvent calculations with NAMD.* Journal of chemical theory and computation, 2011. **7**(11): p. 3635-3642.
84. Buckle, A.M., G. Schreiber, and A.R. Fersht, *Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-. ANG. resolution.* Biochemistry, 1994. **33**(30): p. 8878-8889.
85. Pettersen, E.F., et al., *UCSF Chimera—a visualization system for exploratory research and analysis.* Journal of computational chemistry, 2004. **25**(13): p. 1605-1612.
86. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38.
87. Janin, J., et al., *CAPRI: a critical assessment of predicted interactions.* Proteins: Structure, Function, and Bioinformatics, 2003. **52**(1): p. 2-9.

88. Moreira, I.S., et al., *SpotOn: high accuracy identification of protein-protein interface hot-spots.* Scientific reports, 2017. **7**(1): p. 1-11.
89. Agrawal, P., et al., *Benchmarking of different molecular docking methods for protein-peptide docking.* BMC bioinformatics, 2019. **19**(13): p. 426.
90. Imachi, H., et al., *Menin, a product of the MENI gene, binds to estrogen receptor to enhance its activity in breast cancer cells: possibility of a novel predictive factor for tamoxifen resistance.* Breast cancer research and treatment, 2010. **122**(2): p. 395-407.
91. Chandrasekharappa, S.C., et al., *Positional cloning of the gene for multiple endocrine neoplasia-type 1.* Science, 1997. **276**(5311): p. 404-407.
92. Lemos, M.C. and R.V. Thakker, *Multiple endocrine neoplasia type 1 (MEN1): analysis of 1336 mutations reported in the first decade following identification of the gene.* Human mutation, 2008. **29**(1): p. 22-32.
93. Matkar, S., A. Thiel, and X. Hua, *Menin: a scaffold protein that controls gene expression and cell signaling.* Trends in biochemical sciences, 2013. **38**(8): p. 394-402.
94. Imachi, H., et al., *Raloxifene inhibits menin-dependent estrogen receptor activation in breast cancer cells.* Journal of Endocrinological Investigation, 2011. **34**(11): p. 813-815.
95. Dreijerink, K.M., et al., *Enhancer-mediated oncogenic function of the menin tumor suppressor in breast cancer.* Cell reports, 2017. **18**(10): p. 2359-2372.
96. Dreijerink, K.M., et al., *Menin links estrogen receptor activation to histone H3K4 trimethylation.* Cancer research, 2006. **66**(9): p. 4929-4935.
97. Dreijerink, K.M., et al., *Breast-cancer predisposition in multiple endocrine neoplasia type 1.* The New England journal of medicine, 2014. **371**(6): p. 583.
98. Tate, J.G., et al., *COSMIC: the catalogue of somatic mutations in cancer.* Nucleic acids research, 2019. **47**(D1): p. D941-D947.
99. Ganakammal, S.R. and E. Alexov, *Evaluation of performance of leading algorithms for variant pathogenicity predictions and designing a combinatory predictor method: application to Rett syndrome variants.* PeerJ, 2019. **7**: p. e8106.
100. Landrum, M.J., et al., *ClinVar: improvements to accessing data.* Nucleic acids research, 2020. **48**(D1): p. D835-D844.
101. Schaafsma, G.C. and M. Vihinen, *V ari SNP, A Benchmark Database for Variations From db SNP*. Human Mutation, 2015. **36**(2): p. 161-166.
102. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Current protocols in human genetics, 2013. **76**(1): p. 7.20. 1-7.20. 41.
103. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome research, 2009. **19**(9): p. 1553-1561.
104. Dong, C., et al., *Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.* Human molecular genetics, 2015. **24**(8): p. 2125-2137.
105. Carter, H., et al., *Identifying Mendelian disease genes with the variant effect scoring tool.* BMC genomics, 2013. **14**(3): p. 1-16.
106. Choi, Y., et al., *Predicting the functional effect of amino acid substitutions and indels.* 2012.

107. Ioannidis, N.M., et al., *REVEL: an ensemble method for predicting the pathogenicity of rare missense variants.* The American Journal of Human Genetics, 2016. **99**(4): p. 877-885.
108. Ionita-Laza, I., et al., *A spectral approach integrating functional genomic annotations for coding and noncoding variants.* Nature genetics, 2016. **48**(2): p. 214-220.
109. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome.* Nucleic acids research, 2019. **47**(D1): p. D886-D894.
110. Clarke, L., et al., *The 1000 Genomes Project: data management and community access.* Nature methods, 2012. **9**(5): p. 459-462.
111. Karczewski, K. and L. Francioli, *The genome aggregation database (gnomAD).* MacArthur Lab, 2017.
112. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS computational biology, 2010. **6**(12): p. e1001025.
113. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
114. Consortium, U., *UniProt: a hub for protein information.* Nucleic acids research, 2015. **43**(D1): p. D204-D212.
115. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-1773.
116. Weng, G., et al., *HawkDock: a web server to predict and analyze the protein–protein complex based on computational docking and MM/GBSA.* Nucleic acids research, 2019. **47**(W1): p. W322-W330.
117. Qin, S. and H.-X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction.* Bioinformatics, 2007. **23**(24): p. 3386-3387.
118. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data.* Proteins: Structure, Function, and Bioinformatics, 2005. **61**(1): p. 21-35.
119. Liang, S., et al., *Protein binding site prediction using an empirical scoring function.* Nucleic acids research, 2006. **34**(13): p. 3698-3707.
120. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein–protein binding sites.* Journal of molecular biology, 2004. **338**(1): p. 181-199.
121. Pires, D.E., D.B. Ascher, and T.L. Blundell, *DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach.* Nucleic acids research, 2014. **42**(W1): p. W314-W319.
122. Parthiban, V., M.M. Gromiha, and D. Schomburg, *CUPSAT: prediction of protein stability upon point mutations.* Nucleic acids research, 2006. **34**(suppl_2): p. W239-W242.
123. Pires, D.E., D.B. Ascher, and T.L. Blundell, *mCSM: predicting the effects of mutations in proteins using graph-based signatures.* Bioinformatics, 2014. **30**(3): p. 335-342.
124. Pandurangan, A.P., et al., *SDM: a server for predicting effects of mutations on protein stability.* Nucleic acids research, 2017. **45**(W1): p. W229-W235.

125. Capriotti, E., P. Fariselli, and R. Casadio, *I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure.* Nucleic acids research, 2005. **33**(suppl_2): p. W306-W310.
126. Dehouck, Y., et al., *BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations.* Nucleic acids research, 2013. **41**(W1): p. W333-W339.
127. Rodrigues, C.H., et al., *mCSM-PPI2: predicting the effects of mutations on protein–protein interactions.* Nucleic acids research, 2019. **47**(W1): p. W338-W344.
128. Li, M., et al., *MutaBind estimates and interprets the effects of sequence variants on protein–protein interactions.* Nucleic acids research, 2016. **44**(W1): p. W494-W501.
129. Shu, Y., et al., *Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1.* Proceedings of the National Academy of Sciences, 2003. **100**(10): p. 5902-5907.
130. Caswell, R.C., et al., *Using structural analysis in silico to assess the impact of missense variants in MEN1.* Journal of the Endocrine Society, 2019. **3**(12): p. 2258-2275.
131. Alexov, E., *Numerical calculations of the pH of maximal protein stability. The effect of the sequence composition and three-dimensional structure.* Eur J Biochem, 2004. **271**(1): p. 173-85.
132. Garcia-Moreno, B., *Adaptations of proteins to cellular and subcellular pH.* Journal of biology, 2009. **8**(11): p. 1-4.
133. Peng, Y., et al., *pH-dependent interactions of Apolipophorin-III with a lipid disk.* Journal of Computational Biophysics and Chemistry, 2020: p. 1-12.
134. Mitra, R.C., Z. Zhang, and E. Alexov, *In silico modeling of pH-optimum of protein-protein binding.* Proteins, 2011. **79**(3): p. 925-36.
135. Garcia-Moreno, B., *Adaptations of proteins to cellular and subcellular pH.* J Biol, 2009. **8**(11): p. 98.
136. Luo, Z., et al., *Bbmsn2 acts as a pH-dependent negative regulator of secondary metabolite production in the entomopathogenic fungus Beauveria bassiana.* Environ Microbiol, 2015. **17**(4): p. 1189-202.
137. Wiriyasermkul, P., S. Moriyama, and S. Nagamori, *Membrane transport proteins in melanosomes: Regulation of ions for pigmentation.* Biochim Biophys Acta Biomembr, 2020. **1862**(12): p. 183318.
138. Yamaguchi, Y. and V.J. Hearing, *Melanocytes and their diseases.* Cold Spring Harbor perspectives in medicine, 2014. **4**(5): p. a017046.
139. Pavan, W.J. and R.A. Sturm, *The genetics of human skin and hair pigmentation.* Annual review of genomics and human genetics, 2019. **20**: p. 41-72.
140. Grønskov, K., J. Ek, and K. Brondum-Nielsen, *Oculocutaneous albinism.* Orphanet journal of rare diseases, 2007. **2**(1): p. 1-8.
141. Shain, A.H. and B.C. Bastian, *From melanocytes to melanomas.* Nat Rev Cancer, 2016. **16**(6): p. 345-58.
142. Talley, K. and E. Alexov, *On the pH-optimum of activity and stability of proteins.* Proteins: Structure, Function, and Bioinformatics, 2010. **78**(12): p. 2699-2706.
143. Ikehata, K. and J.A. Nicell, *Characterization of tyrosinase for the treatment of aqueous phenols.* Bioresource Technology, 2000. **74**(3): p. 191-199.

144. Zaidi, K.U., A.S. Ali, and S.A. Ali, *Purification and characterization of melanogenic enzyme tyrosinase from button mushroom.* Enzyme research, 2014. **2014**.
145. White, C., et al., *A role for the ATP7A copper-transporting ATPase in macrophage bactericidal activity.* Journal of Biological Chemistry, 2009. **284**(49): p. 33949-33956.
146. Cobbold, C., et al., *Novel membrane traffic steps regulate the exocytosis of the Menkes disease ATPase.* Human molecular genetics, 2002. **11**(23): p. 2855-2866.
147. Duffy, D.L., et al., *A three–single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation.* The American Journal of Human Genetics, 2007. **80**(2): p. 241-252.
148. H. BRILLIANT, M., *The mouse p (pink-eyed dilution) and human P genes, oculocutaneous albinism type 2 (OCA2), and melanosomal pH.* Pigment cell research, 2001. **14**(2): p. 86-93.
149. Le, L., et al., *SLC45A2 protein stability and regulation of melanosome pH determine melanocyte pigmentation.* Molecular biology of the cell, 2020. **31**(24): p. 2687-2702.
150. Ancans, J., et al., *Melanosomal pH controls rate of melanogenesis, eumelanin/phaeomelanin ratio and melanosome maturation in melanocytes and melanoma cells.* Experimental cell research, 2001. **268**(1): p. 26-35.
151. Eiberg, H., et al., *Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression.* Human genetics, 2008. **123**(2): p. 177.
152. Branicki, W., et al., *Association of the SLC45A2 gene with physiological human hair colour variation.* Journal of human genetics, 2008. **53**(11-12): p. 966-971.
153. Fernandez, L., et al., *SLC45A2: a novel malignant melanoma-associated gene.* Human mutation, 2008. **29**(9): p. 1161-1167.
154. Ambrosio, A.L., et al., *TPC2 controls pigmentation by regulating melanosome pH and size.* Proceedings of the National Academy of Sciences, 2016. **113**(20): p. 5622-5627.
155. Chao, Y.-K., et al., *TPC2 polymorphisms associated with a hair pigmentation phenotype in humans result in gain of channel function by independent mechanisms.* Proceedings of the National Academy of Sciences, 2017. **114**(41): p. E8595-E8602.
156. Talley, K. and E. Alexov, *On the pH-optimum of activity and stability of proteins.* Proteins, 2010. **78**(12): p. 2699-706.
157. Tajielyato, N. and E. Alexov, *Modeling pKas of unfolded proteins to probe structural models of unfolded state.* Journal of Theoretical and Computational Chemistry, 2019. **18**(04): p. 1950020.
158. Skjørringe, T., et al., *Characterization of ATP7A missense mutants suggests a correlation between intracellular trafficking and severity of Menkes disease.* Scientific reports, 2017. **7**(1): p. 1-18.
159. Grønskov, K., et al., *A pathogenic haplotype, common in Europeans, causes autosomal recessive albinism and uncovers missing heritability in OCA1.* Scientific reports, 2019. **9**(1): p. 1-7.
160. Sturm, R.A., et al., *A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color.* The American Journal of Human Genetics, 2008. **82**(2): p. 424-431.

161. Zhang, T., et al., *Cell-type–specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes.* Genome research, 2018. **28**(11): p. 1621-1635.
162. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling.* electrophoresis, 1997. **18**(15): p. 2714-2723.
163. Lai, X., et al., *Structure of human tyrosinase related protein 1 reveals a binuclear zinc active site important for melanogenesis.* Angewandte Chemie International Edition, 2017. **56**(33): p. 9812-9815.
164. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nature protocols, 2015. **10**(6): p. 845-858.
165. Mancusso, R., et al., *Structure and mechanism of a bacterial sodium-dependent dicarboxylate transporter.* Nature, 2012. **491**(7425): p. 622-626.
166. She, J., et al., *Structural mechanisms of phospholipid activation of the human TPC2 channel.* Elife, 2019. **8**: p. e45222.
167. Nomura, N., et al., *Structure and mechanism of the mammalian fructose transporter GLUT5.* Nature, 2015. **526**(7573): p. 397-401.
168. Gourdon, P., et al., *Crystal structure of a copper-transporting PIB-type ATPase.* Nature, 2011. **475**(7354): p. 59-64.
169. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic acids research, 2019. **47**(D1): p. D1005-D1012.
170. Brooks, B.R., et al., *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.* Journal of computational chemistry, 1983. **4**(2): p. 187-217.
171. Panday, S.K., et al., *Modeling electrostatics in molecular biology: A tutorial of DelPhi and associated resources [Article v1. 0].* Living Journal of Computational Molecular Science, 2019. **1**(2): p. 10841.
172. Wang, L., M. Zhang, and E. Alexov, *DelPhiPKa web server: predicting p K a of proteins, RNAs and DNAs.* Bioinformatics, 2016. **32**(4): p. 614-615.
173. Ozenne, V., et al., *Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables.* Bioinformatics, 2012. **28**(11): p. 1463-1470.
174. Estrada, J., et al., *ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble.* BMC bioinformatics, 2009. **10**(1): p. 104.
175. Schaefer, M., M. Sommer, and M. Karplus, *pH-dependence of protein stability: absolute electrostatic free energy differences between conformations.* The Journal of Physical Chemistry B, 1997. **101**(9): p. 1663-1683.
176. Savojardo, C., et al., *INPS-MD: a web server to predict stability of protein variants from sequence and structure.* Bioinformatics, 2016. **32**(16): p. 2542-2544.
177. Cheng, J., A. Randall, and P. Baldi, *Prediction of protein stability changes for single-site mutations using support vector machines.* Proteins: Structure, Function, and Bioinformatics, 2006. **62**(4): p. 1125-1132.
178. Chen, C.-W., J. Lin, and Y.-W. Chu. *iStable: off-the-shelf predictor integration for predicting protein stability changes*. in *BMC bioinformatics*. 2013. Springer.

179. Cao, H., et al., *DeepDDG: predicting the stability change of protein point mutations using neural networks.* Journal of chemical information and modeling, 2019. **59**(4): p. 1508-1514.
180. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* Journal of molecular biology, 2000. **302**(1): p. 205-217.