Published in final edited form as:

Adv Exp Med Biol. 2016 ; 939: 39–61. doi:10.1007/978-981-10-1503-8_3.

Exploring Human Diseases and Biological Mechanisms by Protein Structure Prediction and Modeling

Juexin Wang^{1,2}, Joseph Luttrell IV³, Ning Zhang^{2,4}, Saad Khan^{2,4}, NianQing Shi⁵, Michael X Wang⁶, Jingqiong Kang⁷, Zheng Wang³, Dong Xu^{1,2,4,*}

¹Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

²Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

³School of Computing, University of Southern Mississippi, 118 College Drive, Hattiesburg, MS, 39406, USA

⁴Informatics Institute, University of Missouri, Columbia, MO, 65211, USA

⁵Department of Medicine, Division of Cardiovascular Medicine, University of Wisconsin, Room 8418, 1111 Highland Ave., Madison, WI 53706, USA

⁶Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, MO 65211, USA

⁷Department of Neurology, Vanderbilt University Medical Center, Nashville, TN 37232, USA.

Abstract

Protein structure prediction and modeling provides a tool for understanding protein functions by computationally constructing protein structures from amino acid sequences and analyzing them. With help from protein prediction tools and web servers, users can obtain the three-dimension protein structural models and gain knowledge of functions from the proteins. In this chapter, we will provide several examples of such studies. As an example, structure modeling methods were used to investigate the relation between mutation-caused misfolding of protein and human diseases including epilepsy and leukemia. Protein structure prediction and modeling were also applied in nucleotide-gated channels and their interaction interfaces to investigate their roles in brain and heart cells. In molecular mechanism studies of plants, rice salinity tolerance mechanism was studied via structure modeling on crucial proteins identified by systems biology analysis; trait-associated protein-protein interactions were modeled, which sheds some light on the roles of mutations in soybean oil/protein content. In the age of precision medicine, we believe protein structure prediction and modeling will play more and more important roles in investigating biomedical mechanism of diseases and drug design.

Keywords

protein structure modeling; protein structure prediction; biological mechanism; protein misfolding; sequence mutation; human disease; GWAS; plant breeding

*Corresponding author: xudong@missouri.edu, Office Phone: 513-882-2299.

3.1 Introduction

As the most versatile macromolecules in living organisms ranging from bacteria to human, proteins serve crucial functions in essentially all biological processes [1]. Folding from an amino acid sequence, three-dimension structures of proteins often gave us informative knowledge of protein functions. However, only limit number of three-dimension structures of proteins were known experimentally (116,258 in Protein Data Bank (PDB)[2] as of Feb. 25th 2016), in contrast to 60,971,489 known protein sequence entries in Release 2016_02 of 17-Feb-2016 of the UniProtKB/TrEMBL database[3]. This huge gap between protein sequence and structure makes protein structure prediction and modeling more and more important, i.e. to computationally predict the protein three-dimension structure from its amino acid sequence and analyze the structural model. With decades of efforts, numerous protein structure prediction methods, software tools and webserver have been developed and deployed. Comparing with the experimental approaches, computational structural prediction and modeling are quicker, cheaper and becoming more and more reliable. Researchers could use these tools to model the target protein that they are interested in, and the structure models could help them obtain further insight on the function of the target protein, and its role and mechanism of the underlying biological process.

This book chapter starts from a brief overview of current protein structure prediction tools and mainstream protein function prediction methods (Section 3.1). Several case studies using structure modeling are presented. The first case study is how to use ligand-binding prediction server to study the proposed protein function (Section 3.2). Then we demonstrate how structure modeling is used in human protein studies on exploring diseases (Section 3.3). Plant studies on abiotic stress and agricultural traits using protein structural models are also presented (Section 3.4).

3.1.1 Protein Structure Prediction Methods

Computational protein structure prediction methods can be generally classified into three categories: *ab initio* prediction[4–6], comparative modeling (CM)[7, 8] and threading[9–11]. Different from CM and threading using other known protein structures as templates, *ab initio* methods predict a protein structure by optimizing some scoring functions based on the physical/statistical properties of proteins. CM methods, based on the fact that evolutionarily related proteins typically share a similar structure, build models for the target protein by aligning the target sequence to evolutionarily related (i.e., homologous) template structures. Threading methods are designed to find and align the target sequence to templates of similar structural folds, where target and template sequences are not required to be evolutionary related. Theoretically, *ab initio* prediction could discover new structural folds with more computational resources, but it has not been consistently successful. Template-based methods often obtain high-resolution models with the available templates and accurate alignments.

To advance the development of protein structure prediction, CASP[12] (Critical Assessment of Techniques for Protein Structure Prediction) was set up in 1994 to provide an objective assessment of the state-of-art in the field. Since then, 11 CASPs have been done on a

biannual basis. Table 3.1 shows several protein structure prediction servers that achieved success in the CASPs.

3.1.2 Protein Function Prediction via Protein-Ligand Binding Comparison

A related problem to protein structure prediction is protein function prediction, which occupies an interesting and challenging area of computational research. In order to understand the process behind predicting a protein's function, it is necessary to understand what is meant by the term "function". For many proteins, this is a computationally challenging problem and can lead to answers that reveal much more than a single biological effect [17, 18]. However, regardless of the difficulty of the question, the goal of studying protein function is typically to simply understand what a given protein "does" in nature [17]. Furthermore, even though the details surrounding the question of a protein's function typically share a stronger relationship with biological research, development of methods for computational prediction of protein function continues to be fueled by the abundance of possible functions and the relative difficulty of determining them experimentally [17].

Just as there are many different functional roles that a protein may have, there are many different ways to predict these functions. One interesting method currently being researched involves protein-ligand binding site comparison. In this kind of prediction, previously known or predicted binding sites are used to infer the functional similarity of two proteins. Essentially, the main argument behind the usefulness of these predictions is the fact that binding sites of proteins with similar biological functions are typically better conserved than many other structural elements in evolutionarily distant proteins [19]. In other words, being able to predict the regions of a protein where binding with a ligand takes place can be a helpful clue in determining the function of that protein by allowing comparison with proteins that have similar binding and classified functions.

As a result, a number of tools and databases have been developed using these methods. Even within this closely related group of tools, the methods used vary. For example, ProBis predicts ligand binding sites that may exist on a given query protein by assessing the surface structure of that protein and then comparing it to a database of proteins to find structurally similar binding sites [20, 21]. Some other tools, such as CAST, focus on predicting binding sites through the automatic location and measurement of regions on the input protein known as pockets [22]. Table 3.2 lists a few more of the software packages and projects that are currently utilizing these techniques.

In addition to these servers, a number of databases also contain protein-ligand binding information. These services often distinguish themselves by including information from differing sources and by utilizing differing search algorithms. For example, the PoSSuM database can rapidly compare binding sites among structures with similar or differing global folds [28]. Databases like this may provide users with an idea of the function of similar structures without the need for computationally expensive predictions. In order to make use of the information available from these databases, projects like GIRAF continues to experiment with different data handling approaches [29].

3.1.3 Protein Structure Modeling in the Era of Precision Medicine

Precision medicine often uses invaluable genetics information of many complex diseases. Researchers may explore the link between disease and non-synonymous variation in a large scale, i.e. both population/family scale and individual whole genome-wide scale [30, 31]. Several tools such as SIFT and MutationTaster were developed to evaluate and predict the exon mutation effects on biological functions [31, 32]. However, the potential of using protein structure prediction and modeling for this purpose has been under-explored. There will be increasingly demands in accurately modeling comparative structures between wide-type and mutated proteins. These millions of sequencing data will also advance the method improvement of structure modeling, and the systems biology that incorporates different levels of biological information will expand the usefulness of protein structure information for more comprehensive understanding of biological mechanism. Furthermore, protein design and structure-based drug design will also benefit more and more from integrating the structural information and systems biology data for precision medicine.

3.2 Predicting the Protein Function, A Case Study

One way to start the process of predicting the function of a protein is to obtain protein-ligand binding site predictions from a web server such as ProBis [21]. Using this software simply involves uploading a protein structure file. Also, advanced options include choosing which proteins from the database to compare with the query protein and choosing different limits for the scoring function that judges the results. In order to generate these predictions, ProBis analyzes the solvent accessible surface of the query protein and compares “patches” of this surface with entries in a database of proteins known as the nr-PDB (Non-Redundant Protein Data Bank) [20]. Once the final predictions have been made, the user is sent an email with a link to a results page that contains visualization tools to aid in interpreting the results.

As an example of using ProBis, Figure 3.1 shows images generated with the tools available at the ProBis website and depicts the process of obtaining ligand binding site predictions. In this case, the ProBis web server was given the structure of CASP 9 target protein T0515 as a query for protein-ligand binding site prediction [33]. Using the aforementioned structural comparison process, ProBis determined that the third top scoring similar model was PDB ID 2P3E (chain A). It is important to note that the coloring scheme is consistent across all of the images and was automatically applied by ProBis.

After going through the process to obtain these results, biologists would be more informed in their search to identify the function of T0515. The ProBis server even predicts what type of ligand may bind at each predicted binding site. However, this is just the first step toward predicting the function of a protein. At this point, any predictions made are heavily reliant on inference. Konc et al. performed predictions like these but also took their study a step further into more detailed function prediction using the information that they obtained from protein-ligand binding site comparison [34]. After they had used ProBis to predict the binding sites on their query protein (Tm1631), they also noticed that the server detected binding site similarities between Tm1631 and another protein (PDB ID 2NQJ). Specifically, a similarity was detected between a predicted binding site on Tm1631 and the DNA-binding

site of 2NQJ. Figure 3.2 depicts the structure of Tm1631 as represented in PDB ID 1VPQ and binding site predictions performed with the ProBis server as an example illustration for this passage. By superimposing the structures of Tm1631 and 2NQJ, they were able to formulate a hypothetical model for the interaction of Tm1631 and DNA. In order to further validate this proposed interaction, a molecular dynamics simulation using CHARMM (Chemistry at HARvard Molecular Mechanics) was run on the hypothetical complex[35]. This process essentially tested the stability of the complex and determined that it could reasonably exist in a natural environment. With this as supporting evidence, they were able to draw even more detailed conclusions about Tm1631 and its functional role in binding to DNA. Starting with protein-ligand binding site predictions and making comparisons with other proteins proved to be an effective strategy for this case of function prediction.

While these predictions are a useful starting point, it is important to remember that they do not perfectly reflect the natural conditions that proteins function in. Therefore, it may be beneficial to compare results derived from protein-ligand binding comparison with other computational methods of function prediction. One way to gain further verification of results in this situation is through the use of GO term (Gene Ontology Term) prediction [36]. Essentially, GO terms are identifiers that establish a universal vocabulary for describing the function of a protein [36]. As an example of obtaining GO term predictions through methods other than protein-ligand binding comparison, the sequence for T0515 was submitted to the CombFunc server [37]. Using CombFunc from the web server page is a simple process that only involves entering a protein sequence and an email address for the results to be sent to. Once the sequence has been received, CombFunc utilizes an algorithm that incorporates data from multiple protein function prediction sources. After the relevant data have been gathered, CombFunc produces and ranks the final predictions using an SVM (Support Vector Machine)[38]. In this case, four biological process GO terms were predicted and listed in the order of increasing SVM probability. The first term (GO:0009089) represents the “lysine biosynthetic process via diaminopimelate.” Second, the term GO:0008295 represents the “spermidine biosynthetic process.” Third, the term GO:0006591 represents the “ornithine metabolic process.” Finally, the fourth biological process prediction made by CombFunc was the term representing the “putrescine biosynthetic process.”

Checking the CASP 9 target list revealed that the target T0515 is associated with PDB ID 3MT1. The PDB entry for 3MT1 contains records with two biological process annotations assigned to this protein. These two processes are identified with the GO terms GO:0008295 (described as the “spermidine biosynthetic process”) and GO:0006596 (described as the “polyamine biosynthetic process”). Therefore, the second biological process prediction from CombFunc matched one of the function annotations in the PDB. In some cases, the top scoring predictions can be fairly close to the accepted annotations in terms of semantic similarity. Here, the results from CombFunc were compared with the annotations in the PDB using the “mgoSem” function of the GoSemSim library [39]. In this configuration, GoSemSim takes two lists of GO terms and returns a number (from 0 to 1) that indicates the percentage of similarity between the two lists. Specifically, GoSemSim looks at relationships between “ancestor” terms and the position of GO terms in the graph structure of Gene Ontology data [39]. Simply running the mgoSem function with the two lists (the

two PDB GO terms and the four GO terms predicted by CombFunc) as input resulted in a semantic similarity of 0.723.

With all of these different pieces of information and methods of prediction, combining them is the first step to uncovering the story behind the function of a protein, which is highly challenging. One place where progress in protein function prediction can be seen is the CAFA (Critical Assessment of protein Function Annotation) experiments [40]. In these experiments, researchers participate to develop the best function prediction methods using various methods including protein-ligand binding comparison and more. As techniques continue to be developed and evaluated, application of these concepts is becoming more feasible.

For example, the medical field can benefit from the ability of predicting ligand-binding sites since many drugs operate by binding to these areas on proteins. This is referred to as druggability prediction and is a feature offered by FPOCKET and many other servers [41]. The general process behind druggability prediction often operates on the same principles of protein-ligand binding prediction. However, druggability prediction tends to focus more on applying knowledge from the field of drug development. Since the process of testing drugs can be very expensive, having the ability to make computational predictions about the interactions of a drug with a potential target protein is very promising for drug development and precision medicine [42]. Given the ability to predict the function of a protein and the types of ligands that it may bind with, it may become easier for healthcare professionals to provide better care for patients on an individual basis. Essentially, this may lead to more efficient treatment because of an increase in the ability to predict the compatibility of a drug with specific cases of a disease [42].

3.3 Protein Structure Modeling and Human disease

In this section, we will show three examples of using protein structure prediction and modeling in studying human diseases: (1) Three truncation mutation of GABA_A Receptor were modelled, which reveals that these mutation caused protein misfolding that links to epilepsy (Section 3.3.1). (2) Structure of hyperpolarization-activated cyclic nucleotide-gated (HCN) channels and caveolin in cardiac tissues were modeled individually, their interactions were studied, and the possible binding interface was predicted associated with pacemakers in heart and brain cells (Section 3.3.2). (3) Whole-exome analysis identified various nonsynonymous mutations in juvenile myelomonocytic leukemia (JMML) patients, and structural modeling on these mutated proteins shed some light on the mechanism of this disease (Section 3.3.3).

3.3.1 Protein Modeling on Truncation Mutation of GABA_A Receptor for Studying Epilepsy

Epilepsy is a central nervous system disorder (neurological disorder) in which the nerve cell activity in the brain is disrupted, causing seizures or periods of unusual behavior, sensations and sometimes loss of consciousness [43]. Genetic epilepsies (GEs) are common neurological disorders that are associated frequently with mutations of ion channel genes. One of them is GABA_A receptor (GABA_AR), which is an ionotropic receptor and ligand-gated ion channel. It has an endogenous ligand, γ -aminobutyric acid (GABA), which is the

major inhibitory neurotransmitter in the central nervous system [44–46]. GABA_AR causes an inhibitory effect on neurotransmission by diminishing the occurrence of a successful action potential. Mutations in GABA_A receptor subunit genes (*GABRs*) are frequently associated with epilepsy, and nonsense mutations in *GABRG2* are associated with various types of epilepsy syndromes including the most severe form epileptic encephalopathy like Dravet syndrome[45]. The molecular basis for the phenotypic heterogeneity of *GABRG2* truncation mutations is still unclear but evidences gathered suggest these mutations caused protein misfolding and abnormal receptor trafficking [47].

The first three-dimension protein structure of GABA_AR was resolved by X-ray diffraction (PDB id: 4COF), and its GABA_AR-β3 homopentamer reveals its role as a pentamer in signal transduction [48]. However, mutants and truncation in different lengths at these subunits still lack structure-based explanation. We applied a structure modeling approach to investigate the structure details on three nonsense mutations in *GABRG2* (*GABRG2(R136X)*, *GABRG2(Q390X)* and *GABRG2(W429X)*) associated with epilepsies of different severities. We mainly used our in-house protein structure prediction tool MUFOLD [13] to construct protein models of mutant GABA_A receptor subunits: (1) γ2 (R136X) subunit, in which all transmembrane regions are deleted and only part of the N-terminal domain remains; (2) γ2 (Q390X) subunit, in which the fourth hydrophobic transmembrane α-helix (YARIFFPTAFCLFNLVYWVSYLEL) is deleted and a new α-helix with many charged amino acids (KDKDKKKKNPAPTIDIRPRSATI) is found to assume its location; and (3) γ2 (W429X) subunit, in which the fourth hydrophobic transmembrane α-helix is truncated. Figure 3.3 presents structure models on wide-type γ2, γ2 (R136X), γ2 (Q390X) and γ2 (W429X) subunits.

Following the MUFOLD protocol, we identified nicotinic acetylcholine receptor (PDB id: 4COF and 2BG9) as the main template for GABA_AR-β3. Then multidimensional scaling (MDS) was used to construct multiple protein decoys based on the template and other minor templates. Then these decoys were clustered and evaluated. With several iterations of model generation and evaluation, one decoy was chosen as the predicted protein model and then refined by Rosetta [49]. For mutant GABA_A receptor subunits, the original input subunits were split into different domains, and each domain was modeled individually and then assembled together.

To further understand the stabilities of these wide-type and mutant subunits, a dimer structure was constructed between two subunits by symmetric docking of SymmDock [50], detailed in Figure 3.4. As classified as transporter protein in the membrane, special filtering on dimer models was applied to make sure the intracellular, transmembrane, and extracellular domains were arranged correspondingly between monomers. Template-free dockings were performed in conjunction with template-based docking [51] between γ2 and α subunits by mapping their corresponding positions to template GABA_AR-β3 homopentamer (PDB id: 4cof). Pentamer and hypothetic homopentamer were also constructed by template-based docking. Chimera [52] and Pymol [53] were used to display the protein structure models.

Along with flow cytometry and biochemical approaches in combination with lifted whole-cell patch clamp recordings, the structural modeling and structure-based analysis indicated that the wild-type $\gamma 2$ subunit surface was naturally hydrophobic, which is suitable to be buried in membrane. The different $\gamma 2$ subunits adopted different conformations, and the mutant $\gamma 2(Q390X)$ subunits formed protein specific or nonspecific stable protein dimers with themselves or other proteins while $\gamma 2(R136X)$ subunits could not form dimers with other partnering subunits but could dimerize with themselves. The $\gamma 2(W429X)$ subunits also dimerized with themselves but the protein conformation was similar to the wild-type $\gamma 2$ subunit protein. Our modeling study provides good hypotheses to understand the mechanisms and effects of the *GABRG2* truncation mutations in epilepsy.

3.3.2 Exploring the Interaction of Caveolin-3 with Hyperpolarization-Activated Cyclic Nucleotide-Gated Channels 2 and 4

Hyperpolarization-activated cyclic nucleotide-gated (HCN) channels are a group of cation channel proteins serving as pacemakers in heart and brain cells [54, 55]. They play essential roles in regulating cardiac and neuronal rhythmicity [56, 57]. Currently four types of HCN (HCN1 to HCN4) channels have been discovered. Among them, HCN4 and HCN2 are the main isoforms expressed in cardiac tissues. A type of related proteins to HCN is the caveolin family of integral membrane proteins, which are the building blocks of caveolae, a type of lipid rafts on cell membrane [58, 59]. In addition, caveolins act as scaffolding proteins and interact with a variety proteins to form macromolecular complexes [60]. Three types of caveolin proteins have been identified so far (Cav1 to Cav3). Cav3 is reported to be specifically expressed in skeletal, smooth, and cardiac and muscle cells [61]. Interestingly, studies have shown that Cav3 is associated with HCN4, and affects its function [62]. However, the detailed interaction information is still largely unknown. In this work, we explored the interaction of Cav3 with HCN2/HCN4 and predicted the possible binding interface of Cav3 with HCN2/HCN4 using computational methods.

We started our analysis by searching for caveolin-binding motif on HCN2 and HCN4 protein sequences, respectively. The existence of caveolin-binding motif was not a definitive evidence of the binding interface but it served as a reasonable starting point. We used ScanProsite [63] and targeted on two known motif patterns, i.e., [FWY]X[FWY]XXXX[FWY] and [FWY]XXXX[FWY]XX[FWY] [64]. We found two hits on HCN2 and one hit on HCN4 (Table 3.3). One hit (214 – 222) on HCN2 was located in the transmembrane region, making it less likely to be binding interface. The other hit on HCN2 (202 – 210) had the exactly the same motif sequence with HCN4 hit. Both hits were in the N-terminal cytoplasmic domain.

Meanwhile, to determine the possible binding interfaces of HCN2/HCN4 and Cav3, we performed correlated mutation analysis using i-COMS webserver [65]. Correlated mutation aims to discover those co-evolved pairs of amino acid residues between two proteins. Based on our motif search, we limited our search in the N-terminus of HCN2/HCN4. We predicted possible correlated amino acid residues between the ion transport protein N-terminal domain (PF08412) of HCN2/HCN4 and the caveolin domain (PF01146) of Cav3. We combined the prediction results from three algorithms of computing correlated mutations, including

mutual information, pseudo likelihood maximization direct coupling analysis and mean field direct coupling analysis.

We found several amino acid pairs whose prediction scores ranked top 100 inter-protein links for all three methods: D209 in HCN2 vs. T66 in Cav3, D260 in HCN4 vs. T66 in Cav3, and S230 in HCN4 vs. F65 in Cav3. For the first two pairs (HCN2 D209 vs. Cav3 T66 and HCN4 D260 vs. Cav3 T66), the amino acid residues were in caveolin-binding motif regions previously determined. This further inferred that the possible binding sites on HCN2/HCN4 were located within the caveolin-binding motif.

Next we built a structural model showing the interaction between HCN2/HCN4 and Cav3. Results from protein disorder analysis indicated that most of the N-terminal sequences of HCN2/HCN4 were disordered regions. Therefore, the structure of the N-terminal domains could be highly variable, and it might be difficult to obtain a reliable and consistent structure. We selected the ordered region of HCN2 (186 – 225) and HCN4 (236 – 275), and used I-TASSER [11] to predict their 3D structures. On the other hand, since Cav3 is a transmembrane protein, we predicted the 3D structures using its cytoplasmic N-terminal sequences (1 – 85) based on I-TASSER. Next, the best predicted structures were docked using PatchDock [50] and FireDock [66] to determine and refine the possible binding conformations of the N-terminal structures of HCN2/4 and Cav3 (Figure 3.5). Among the top docking solutions, we could find interaction interface via caveolin binding motif in HCN2/HCN4 with the N-terminus of Cav3.

1.3.4 Point Mutations Identified in Juvenile Myelomonocytic Leukemia (JMML) Patient's Exome and Their Effects on Protein Structure

Juvenile myelomonocytic leukemia (JMML) is a rare and chronic leukemia found to occur in 1.2 cases per million. JMML affects children in the age group of four and below. It is thought that JMML is a congenital disorder. The majority of mutations that have been found in JMML patients so far belong to RAS/MAPK signalling pathway these include NRAS, KRAS, NF and PTPN11 mutations. It has been established that JMML is fundamentally a disease of hyperactive RAS signalling but targeted chemotherapy of this pathway has not been successful. [67, 68]

We did whole exome analysis of a 2-year old JMML patient's bone marrow specimen using whole exome sequencing and verified it using cancer panel sequencing. We were able to identify several novel mutations in NTRK1, HMGA2, MLH3, MYH9 and AKT1 genes. We were also able to confirm the already identified mutation of PTPN11 (exon 3 181G>T)[69] in JMML. Here we discuss the methods we used to elucidate whether any of these novel/ already identified mutations at DNA level affects the respective protein structure of their proteins.

Whole exome analysis identified various nonsynonymous mutations, which were then confirmed by cancer panel sequencing. These include ITPR3 (chr6:33651070,G>A,Exon35,A1562>T), PTPN11(chr12:112888165,G>T,Exon3, D61>Y), AKT1(chr14:105239869,G>A,Exon9,R251>C), MLH3(chr14:75514537,A>T, Exon2,F608>I), and MYH9(chr22:36685257,C>A,Exon32,K1477>N). In order to identify

whether these mutations have any effect on their protein sequence, SIFT and Provean predictions were used [32]. The already identified mutations in PTPN11 and the novel mutations in AKT1, MLH3 and MYH9 were found to be deleterious and damaging. In order to identify whether the particular mutation is likely to be associated with the disease, we used SuSPect webserver (<http://www.sbg.bio.ic.ac.uk/~suspect/>). SuSPect webserver indicated that out of all the mutations, two mutations namely AKT1 and already known PTPN11 were most likely disease causing mutations. Protein sequences of the two mutations were further extracted from the vcf file using customProDB [70]. Mutation taster [31] was used to detect that the AKT1 mutation occurs at an evolutionary conserved site by comparison against different species. A homology model of the AKT1 mutation was constructed using SWISS-MODEL[71] based on the template (PDB: 3O96) with 97% identity as shown in Figure 3.6A (superimposed with the native structure). The 3D structure shows that mutation R251>C occurs at the surface of the 3D protein as indicated by ARG and CYS residues in the structure created visualized using UCSF Chimera[52].

Both Motif Scan and Uniprot indicate that the R251>C lies near or within multiple protein domains which can potentially act as active sites in the protein. In order to see if any active site lies near the mutation which can potentially have a damaging effect on overall structure of the protein HOPE[72] webserver was used. Results from HOPE server point towards a structural disruption in an Interpro [73] domain i.e., the protein kinase domain (IPR000719) [74].

In order to find if there is a change in the polar contacts of the neighboring residues of the mutated residue we compared the polar contact maps of native ARG251 and the mutated CYS189 residues and their immediate neighbors using Ligplot+ [75]. As indicated by Figure 3.6, there is a change in the electron density due to mutation as shown by change in positions of the hydrophobic contacts and loss of hydrogen bonding between mutated site and ASP248 and PHE407.

3.4 Protein Modeling and Plant Analysis

In this section, we present two examples of integrating protein structure prediction/modeling with other omics data for understanding plant protein mechanism. Although they represent agriculture research problems, applications in medicine can work in a similar fashion. The first example is to combine rice microarray and structure modeling to study proteins that may have important roles in the rice salinity resistance process (Section 3.4.1). The second example is to conduct protein modeling on trait associated protein-protein interactions identified from soybean genome wide association study (GWAS) (Section 3.4.2).

3.4.1 Structure Modeling in Exploring Rice Tolerance Mechanism

Under the pressure of global climate change and global population explosion, soil salinity causes rice production reduction in about 30% of the rice-growing area worldwide. The study in exploring the mechanism of salt tolerance starts from selecting differential expressed genes in the whole gene set, then a putative mechanism network was built and several network modules were identified upon information from protein-protein interaction. These modules were annotated and assessed by quantitative trait loci (QTL) [76], co-

expression and regulatory binding motif analysis. The topological hub genes in these modules are considered as the most important genes dominating the inherent function [77].

Among all the genes in the module, one of the most interesting genes is LOC_Os01g52640.3, which is a hub gene in the largest module and overlaps with a QTL region. This gene corresponds to a hypothetical protein Os01g0725800, which interacts with 32 of the 51 proteins in the module. It contains four InterPro domains, namely, IPR000719, IPR001680, IPR011046, and IPR011009. IPR011009 domains can also be found in RIO kinase (IPR018935), a SPA1-related, serine/threonine-specific and tyrosine-specific protein kinase. This protein also has an ortholog in *Arabidopsis thaliana* as SPA4 (SPA1-RELATED 4), which is a binding protein and a signal transducer. MUFOLD [13, 14] was applied to predict the structure for LOC_Os01g52640.3. Using the identified templates of 2GNQ, 3EMH, and 3DM0 in PDB, the model for the protein region of 196–627 for the protein with the length of 432 was constructed, as shown in Figure 3.7. The protein structural model contains the WD40 structure motif repeats, each with a tryptophan-aspartic acid (W-D) dipeptide termination. As WD40 proteins often play important roles in signal transduction and transcription regulation [78], the structure prediction suggests that this protein may be related to signal transduction in the salt resistance process.

3.4.2 Structure Modeling for Soybean Trait Improvement

Soybeans represent one of the most important agricultural crops providing nutrition and sustenance to humans and household animals, and become an increasingly valuable feedstock for industrial applications [79]. Among hundreds of agricultural trials, seed oil content and seed protein content are both polygenic traits controlled by several gene loci in soybeans. Many of the QTL alleles with positive and negative effects on oil content are often dispersed among genotypes [80], which suggests that accumulation of the positive alleles from different genetic backgrounds could eventually lead to the development of genotypes with higher seed oil content or protein content [81]. To address the “Missing Heritability” problem in complex traits by the original GWAS analysis under the hypothesis of single SNP association with the phenotype [82], BHIT (Bayesian High-order Interaction Toolset) [83] was applied to explore the SNP-interactions associated with the phenotypes. The most interesting interactions identified were 4 loci across two chromosomes located in position 20,897,627; 20,954,490 of Chromosome 8, and 8,642,446; 12,051,017 of Chromosome 19 in soybean genome. Among them, protein Glyma08g26580.1 containing first SNP (SNP293) and protein Glyma19g07330.1 containing third SNP (SNP792) were computationally predicted to interact by PropInt [84].

The first SNP (named as SNP293) in the results is located in gene Glyma08g26580.1, which has an Arabidopsis homology AT3G0140 (EC/6.3.2.19) and an ubiquitin-protein ligase. At the sequence level, this polymorphism makes the minor allele nucleotide adenine (A) replaced the major allele nucleotide guanine (G), which causes the 73th amino acid of the protein change from glycine (G) to arginine (R). The added positive charged arginine may have significant impact on the protein conformation and function. The third SNP (named as SNP792) in the results is located in gene Glyma19g07330.1, which also causes amino acid change from glycine (G) to arginine (R). This gene has the Arabidopsis homolog

AT3G48990.1, which encodes an oxalyl-CoA synthetase and is required for oxalate degradation and normal seed development processes.

MUFOLD [13, 14] was applied to predicted protein structures of gene Glyma08g26580.1 and gene Glyma19g07330.1. The two predicted structures were docked together using GRAMMX [85]. Interestingly, the distance between the residue containing SNP293 and the residue containing SNP792 in the docking complex was 1.17Å, shorter than 0.0052% of all the paired distances between the two structures, as shown in Fig. 3.8. This result suggests that the epistatic interaction between the two SNPs may play a role in the interaction between the two proteins. And this interaction caused by amino acid changes may shed some light on the mechanism in controlling oil/protein contents in soybean.

3.5 Conclusions

Protein structure modeling provides a tool to explore the mechanism of a biological process and the function of a protein. In the studies reviewed in this book chapter, we showed various use cases of combining with docking, protein-protein interaction prediction, GWAS analysis, systems biology, and other analysis, protein structure modeling in studying protein conformation, function, mutation, and disease/phenotype effects. The structure-based prediction and analysis expand our knowledge in biological mechanism and human disease and help design drug treatment in the age of precision medicine.

References

1. Rossmann MG, Moras D, and Olsen KW, Chemical and biological evolution of nucleotide-binding protein. *Nature*, 1974 250(463): p. 194–9. [PubMed: 4368490]
2. Berman HM, et al., The Protein Data Bank. *Nucleic Acids Research*, 2000 28(1): p. 235–242. [PubMed: 10592235]
3. Bairoch A, et al., The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, 2009 37: p. D169–D174. [PubMed: 18836194]
4. Li Z and Scheraga HA, Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci U S A*, 1987 84(19): p. 6611–5. [PubMed: 3477791]
5. Xu D and Zhang Y, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, 2012 80(7): p. 1715–35. [PubMed: 22411565]
6. Kim DE, Chivian D, and Baker D, Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 2004 32: p. W526–W531. [PubMed: 15215442]
7. Bowie JU, Luthy R, and Eisenberg D, A Method to Identify Protein Sequences That Fold into a Known 3-Dimensional Structure. *Science*, 1991 253(5016): p. 164–170. [PubMed: 1853201]
8. Soding J, Biegert A, and Lupas AN, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 2005 33: p. W244–W248. [PubMed: 15980461]
9. Simons KT, et al., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 1997 268(1): p. 209–225. [PubMed: 9149153]
10. Xu Y and Xu D, Protein threading using PROSPECT: Design and evaluation. *Proteins-Structure Function and Genetics*, 2000 40(3): p. 343–354.
11. Zhang Y, I-TASSER server for protein 3D structure prediction. *Bmc Bioinformatics*, 2008 9.
12. Kryshchukovych A, Fidelis K, and Moult J, CASP10 results compared to those of previous CASP experiments. *Proteins-Structure Function and Bioinformatics*, 2014 82: p. 164–174.
13. Zhang JF, et al., MUFOLD: A new solution for protein 3D structure prediction. *Proteins-Structure Function and Bioinformatics*, 2010 78(5): p. 1137–1152.

14. Zhang J, et al., Prediction of protein tertiary structures using MUFOLD, in *Functional Genomics*. 2012, Springer p. 3–13.
15. Wang Z, Eickholt J, and Cheng J, MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics*, 2010 26(7): p. 882–8. [PubMed: 20150411]
16. Kallberg M, et al., Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, 2012 7(8): p. 1511–1522. [PubMed: 22814390]
17. Friedberg I, Automated protein function prediction - the genomic challenge. *Briefings in Bioinformatics*, 2006 7(3): p. 225–242. [PubMed: 16772267]
18. Borgwardt KM, et al., Protein function prediction via graph kernels. *Bioinformatics*, 2005 21: p. I47–I56. [PubMed: 15961493]
19. Konc J and Janezic D, Binding site comparison for function prediction and pharmaceutical discovery. *Current Opinion in Structural Biology*, 2014 25: p. 34–39. [PubMed: 24878342]
20. Konc J and Janezic D, ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 2010 26(9): p. 1160–1168. [PubMed: 20305268]
21. Konc J, et al., ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in Protein Binding Sites. *Journal of Chemical Information and Modeling*, 2015 55(11): p. 2308–2314. [PubMed: 26509288]
22. Liang J, Edelsbrunner H, and Woodward C, Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, 1998 7(9): p. 1884–1897. [PubMed: 9761470]
23. Nilmeier JP, et al., Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure. *Plos One*, 2013 8(5).
24. Kirshner DA, Nilmeier JP, and Lightstone FC, Catalytic site identification-a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Research*, 2013 41(W1): p. W256–W265. [PubMed: 23680785]
25. Roy A, Yang JY, and Zhang Y, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*, 2012 40(W1): p. W471–W477. [PubMed: 22570420]
26. Volkamer A, et al., Combining global and local measures for structure-based druggability predictions. *Abstracts of Papers of the American Chemical Society*, 2012 243.
27. Gherardini PF, et al., Modular architecture of nucleotide-binding pockets. *Nucleic Acids Research*, 2010 38(11): p. 3809–3816. [PubMed: 20185567]
28. Ito JI, et al., PoSSuM: a database of similar protein-ligand binding and putative pockets. *Nucleic Acids Research*, 2012 40(D1): p. D541–D548. [PubMed: 22135290]
29. Nagarajan N and Kingsford C, GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Res*, 2011 39(6): p. e34. [PubMed: 21177643]
30. Gao M, Zhou HY, and Skolnick J, Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure*, 2015 23(7): p. 1362–1369. [PubMed: 26027735]
31. Schwarz JM, et al., MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth*, 2010 7(8): p. 575–576.
32. Kumar P, Henikoff S, and Ng PC, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 2009 4(7): p. 1073–1082. [PubMed: 19561590]
33. Moult J, et al., Critical assessment of methods of protein structure prediction (CASP)-Round IX. *Proteins-Structure Function and Bioinformatics*, 2011 79: p. 1–5.
34. Konc J, et al., Structure-Based Function Prediction of Uncharacterized Protein Using Binding Sites Comparison. *Plos Computational Biology*, 2013 9(11).
35. Brooks BR, et al., CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry*, 2009 30(10): p. 1545–1614. [PubMed: 19444816]
36. Ashburner M, et al., Gene Ontology: tool for the unification of biology. *Nature Genetics*, 2000 25(1): p. 25–29. [PubMed: 10802651]

37. Wass MN, Barton G, and Sternberg MJE, CombFunc: predicting protein function using heterogeneous data sources. *Nucleic Acids Research*, 2012 40(W1): p. W466–W470. [PubMed: 22641853]
38. Vapnik VN, An overview of statistical learning theory. *Ieee Transactions on Neural Networks*, 1999 10(5): p. 988–999. [PubMed: 18252602]
39. Yu GC, et al., GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 2010 26(7): p. 976–978. [PubMed: 20179076]
40. Radivojac P, et al., A large-scale evaluation of computational protein function prediction. *Nature Methods*, 2013 10(3): p. 221–227. [PubMed: 23353650]
41. Schmidtke P and Barril X, Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. *Journal of Medicinal Chemistry*, 2010 53(15): p. 5858–5867. [PubMed: 20684613]
42. Khoury MJ, et al., A Population Approach to Precision Medicine. *American Journal of Preventive Medicine*, 2012 42(6): p. 639–645. [PubMed: 22608383]
43. Noebels JL, Exploring new gene discoveries in idiopathic generalized epilepsy. *Epilepsia*, 2003 44: p. 16–21.
44. Ishii A, et al., Association of nonsense mutation in GABRG2 with abnormal trafficking of GABAA receptors in severe epilepsy. *Epilepsy Res*, 2014 108(3): p. 420–32. [PubMed: 24480790]
45. Harkin LA, et al., Truncation of the GABA(A)-receptor gamma2 subunit in a family with generalized epilepsy with febrile seizures plus. *Am J Hum Genet*, 2002 70(2): p. 530–6. [PubMed: 11748509]
46. Kang J-Q, et al., The human epilepsy mutation GABRG2 (Q390X) causes chronic subunit accumulation and neurodegeneration. *Nature neuroscience*, 2015.
47. Kang JQ, et al., Slow degradation and aggregation in vitro of mutant GABAA receptor gamma2(Q351X) subunits associated with epilepsy. *J Neurosci*, 2010 30(41): p. 13895–905. [PubMed: 20943930]
48. Miller PS and Aricescu AR, Crystal structure of a human GABAA receptor. *Nature*, 2014.
49. Leaver-Fay A, et al., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 2011 487: p. 545. [PubMed: 21187238]
50. Schneidman-Duhovny D, et al., PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 2005 33(Web Server issue): p. W363–7. [PubMed: 15980490]
51. Szilagyi A and Zhang Y, Template-based structure modeling of protein–protein interactions. *Current opinion in structural biology*, 2014 24: p. 10–23. [PubMed: 24721449]
52. Pettersen EF, et al., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 2004 25(13): p. 1605–1612. [PubMed: 15264254]
53. DeLano WL, The PyMOL molecular graphics system. 2002.
54. DiFrancesco D, Pacemaker mechanisms in cardiac tissue. *Annu Rev Physiol*, 1993 55: p. 455–72. [PubMed: 7682045]
55. Pape HC, Queer current and pacemaker: the hyperpolarization-activated cation current in neurons. *Annu Rev Physiol*, 1996 58: p. 299–327. [PubMed: 8815797]
56. Santoro B, et al., Identification of a gene encoding a hyperpolarization-activated pacemaker channel of brain. *Cell*, 1998 93(5): p. 717–29. [PubMed: 9630217]
57. Ludwig A, et al., Two pacemaker channels from human heart with profoundly different activation kinetics. *EMBO J*, 1999 18(9): p. 2323–9. [PubMed: 10228147]
58. Philippova MP, et al., T-cadherin and signal-transducing molecules co-localize in caveolin-rich membrane domains of vascular smooth muscle cells. *FEBS Lett*, 1998 429(2): p. 207–10. [PubMed: 9650591]
59. Simons K and Toomre D, Lipid rafts and signal transduction. *Nat Rev Mol Cell Biol*, 2000 1(1): p. 31–9. [PubMed: 11413487]
60. Boscher C and Nabi IR, Caveolin-1: role in cell signaling. *Adv Exp Med Biol*, 2012 729: p. 29–50. [PubMed: 22411312]
61. Tang Z, et al., Molecular cloning of caveolin-3, a novel member of the caveolin gene family expressed predominantly in muscle. *J Biol Chem*, 1996 271(4): p. 2255–61. [PubMed: 8567687]

62. Ye B, et al., Caveolin-3 associates with and affects the function of hyperpolarization-activated cyclic nucleotide-gated channel 4. *Biochemistry*, 2008 47(47): p. 12312–8. [PubMed: 19238754]
63. de Castro E, et al., ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*, 2006 34(Web Server issue): p. W362–5. [PubMed: 16845026]
64. Couet J, et al., Identification of peptide and protein ligands for the caveolin-scaffolding domain. Implications for the interaction of caveolin with caveolae-associated proteins. *J Biol Chem*, 1997 272(10): p. 6525–33. [PubMed: 9045678]
65. Iserte J, et al., I-COMS: Interprotein-CORrelated Mutations Server. *Nucleic Acids Res*, 2015 43(W1): p. W320–5. [PubMed: 26032772]
66. Mashiah E, et al., FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res*, 2008 36(Web Server issue): p. W229–32. [PubMed: 18424796]
67. Baumann I, M.B. J, Niemeyer CM, Thiele J, Shannon K, Juvenile Myelomonocytic Leukemia (JMML) WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues, ed. Swerdlow SH, I.A.f.R.o. Cancer, and W.H. Organization. 2008: International Agency for Research on Cancer.
68. Lauchle JH and Braun B, Targeting RAS Signaling Pathways in Juvenile Myelomonocytic Leukemia (JMML), in *Molecularly Targeted Therapy for Childhood Cancer*, Houghton PJ and Arceci RJ, Editors. 2010, Springer New York p. 123–138.
69. Loh ML, Vattikuti S, Schubert S, Reynolds MG, Carlson E, Lieu KH, ... Ptpn T, Mutations in PTPN11 implicate the SHP-2 phosphatase in leukemogenesis. *Blood*, 2004 103 (6): p. 2325–2332. [PubMed: 14644997]
70. Wang X and Zhang B, customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*, 2013.
71. Biasini M, et al., SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research*, 2014: p. gku340.
72. Venselaar H, et al., Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, 2010 11: p. 548–548. [PubMed: 21059217]
73. Mitchell A, et al., The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, 2015 43(D1): p. D213–D221. [PubMed: 25428371]
74. Hanks SK, Quinn AM, and Hunter T, The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science (New York, N.Y.)*, 1988 241(4861): p. 42–52.
75. Laskowski RA and Swindells MB, LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *Journal of Chemical Information and Modeling*, 2011 51(10): p. 2778–2786. [PubMed: 21919503]
76. Seaton G, et al., QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics*, 2002 18(2): p. 339–40. [PubMed: 11847090]
77. Wang J, et al., A computational systems biology study for understanding salt tolerance mechanism in rice. *PloS one*, 2013 8(6): p. e64929. [PubMed: 23762267]
78. Neer EJ, et al., The ancient regulatory-protein family of WD-repeat proteins. *Nature*, 1994 371(6495): p. 297–300. [PubMed: 8090199]
79. Snyder CL, et al., Acyltransferase action in the modification of seed oil biosynthesis. *N Biotechnol*, 2009 26(1–2): p. 11–6. [PubMed: 19501678]
80. Zhao JY, et al., Oil content in a European x Chinese rapeseed population: QTL with additive and epistatic effects and their genotype-environment interactions. *Crop Science*, 2005 45(1): p. 51–59.
81. Weselake RJ, et al., Increasing the flow of carbon into seed oil. *Biotechnology Advances*, 2009 27(6): p. 866–878. [PubMed: 19625012]
82. Mackay TFC, Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 2014 15(1): p. 22–33.
83. Wang J, et al., A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genomics*, 2015 16(1): p. 1011. [PubMed: 26607428]

84. Rashid M, Ramasamy S, and Raghava GPS, A Simple Approach for Predicting Protein-Protein Interactions. *Current Protein & Peptide Science*, 2010 11(7): p. 589–600. [PubMed: 20887258]
85. Tovchigrechko A and Vakser IA, GRAMM-X public web server for protein-protein docking. *Nucleic Acids Research*, 2006 34: p. W310–W314. [PubMed: 16845016]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.1: An example of using the ProBis web server to perform protein-ligand binding site prediction.

Section A represents the template protein 2P3E and depicts its binding region (represented by red spheres) as identified by ProBis. Section B shows an enlarged view of this binding region surrounded by a simplified view of the tertiary structure of 2P3E. Section C depicts the structure of T0515 with its predicted binding region and the aligned region of 2P3E. Section D provides a closer view of the predicted binding region of T0515 without the aligned region of 2P3E and surrounded by a simplified representation of the tertiary structure of T0515. Section E depicts the same view as section C but also includes the aligned region of 2P3E (represented as a mixture of blue and red spheres). Section F offers a more detailed view of the predicted binding region of T0515 and the aligned region of 2P3E.

Figure 3.2: ProBis results depicting the tertiary structure of PDB ID 1VPQ (Tm1631) and its predicted binding region (red spheres).
The red spheres represent the predicted binding region and the rest of the image follows the same coloring scheme as Figure 3.1.

Figure 3.3: Predicted protein structural modeling of the wild-type $\gamma 2$ and the mutant $\gamma 2$ (R136X), $\gamma 2$ (Q390X) and $\gamma 2$ (W429X) subunits.

Figure 3.4: Docking models for potential mutant $\gamma 2$ subunit homodimers by SymmDock.

In each panel, the two $\gamma 2$ subunits are shown in red and green; (A) wide-type $\gamma 2$ dimer; (B) $\gamma 2$ (R136X) mutant dimer; (C) $\gamma 2$ (Q390X) mutant dimer; (D) $\gamma 2$ (W429X) mutant dimer.

Figure 3.5: A schematic representation of the interaction between HCN2/4 and Cav3. HCN2 was shown in left and HCN4 in right. Cav3 N-terminal domain was shown in blue, and HCN2/HCN4 N-terminal ordered region was shown in red. The side chains of the three hydrophobic residuals (W, Y, F) in caveolin-binding motif were labeled in orange.

Figure 3.7: Predicted structural model of protein Os01g0725800.

Figure 3.8: Protein-protein interaction on predicted protein structures containing SNP locations. SNP293 is located in the protein Glyma08g26580.1 (upper, green) and SNP 792 is located in the protein Glyma19g07330.1 (lower, cyan). The polymorphism sites (red) are located at the interface of the interaction

Table 3.1

A sample of current tools in the field of protein structure prediction

Method / Server	URL	Brief Description
HHpred[8]	http://toolkit.tuebingen.mpg.de/hhpred	Structure prediction by sensitive HMM- HMM search of template
I-TASSER[11]	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	A hierarchical approach using multiple threading results and conformation sampling
MUFOLD[13, 14]	http://mufold.org/	Graph-based model generation using MDS and comprehensive model quality assessment
MULTICOM[15]	http://sysbio.rnet.missouri.edu/multicom_cluster/	Model generation using multi-template comparative modeling and refinement
RaptorX[16]	http://raptorx.uchicago.edu/StructurePrediction/predict/	Highly sensitive method for remote homolog identification and alignment
ROBETTA/ROSETTA[6]	http://www.rosetta.org/submit.jsp	Model generation using both <i>ab initio</i> and comparative models of protein domains

Table 3.2 -

A sample collection of current tools in the field of protein function prediction that are based on protein-ligand binding comparison.

Method / Server	URL	Summary
CAST [22]	http://sts.bioe.uic.edu/castp/	Predicting binding sites through shape matching
CatSId [23, 24]	http://catsid.llnl.gov/	Predicting catalytic sites by using a structure matching algorithm.
COFACTOR [25]	http://zhanglab.ccmb.med.umich.edu/COFACTOR	Structure-based function predictions (ligand binding sites, GO, and enzyme commission).
DoGSiteScorer [26]	http://dogsite.zbh.uni-hamburg.de/	Pocket detection on protein surface and druggability prediction.
Nucleos [17, 27]	http://nucleos.bio.uniroma2.it/nucleos/	Binding site prediction for different types of nucleotides.
SA-Mot [17]	http://sa-mot.mti.univ-paris-diderot.fr/main/SA_Mot_Method	Using HMM-SA to extract and describe structural motifs from protein loop structures.
SiteBinder	http://webchem.ncbr.muni.cz/Platform/App/SiteBinder	Building models of binding sites and superimposing an arbitrary number of small protein fragments.
ProBis [20, 21]	http://probis.nih.gov/	Ligand binding site prediction by searching for protein surface with known binding of ligand.

Table 3.3.

Predicted caveolin-binding motif for HCN2 and HCN4.

	Sequence	Pattern Matched	Topology
HCN2-human-wt	WiihpYsdF (202 – 210)	[FWY]XXXX[FWY]XX[FWY]	Cytoplasmic (1 – 215)
HCN2-human-wt	WdFtmllF (214 – 221)	[FWY]X[FWY]XXXX[FWY]	Transmembrane (216 – 236)
HCN4-human-wt	WiihpYsdF (253 – 261)	[FWY]XXXX[FWY]XX[FWY]	Cytoplasmic (1 – 266)