Functional Annotation of Human *Sec24C* protein

By

IMTIAZ NAWAZ

Faculty of Basic and Applied Sciences
Department of Bioinformatics and Biotechnology
International Islamic University Islamabad
(2012)

MS
668-43
IMF

1 - Protein plastics

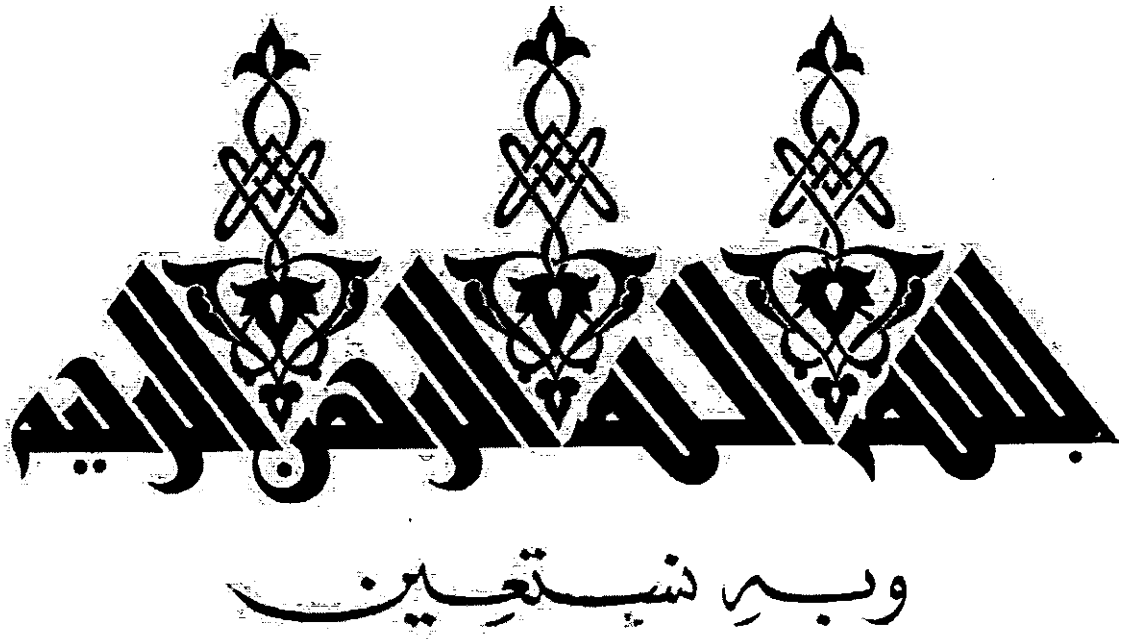
2. Casein - derived plastics

Functional Annotation of Human *Sec24C* protein

Researcher
Imtiaz Nawaz
5-FBAS/MSBI/S11

Supervisor
Dr. Jabar Zaman Khan Khattak
Chairman, Department of Bioinformatics and Biotechnology, International Islamic University Islamabad

Faculty of Basic and Applied Sciences
Department of Bioinformatics and Biotechnology
International Islamic University Islamabad
(2012)


بسم الله الرحمن الرحيم
وبه نستعين

Department of Bioinformatics and Biotechnology
International Islamic University Islamabad

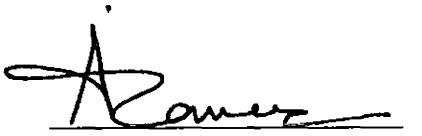
Dated: _____

FINAL APPROVAL

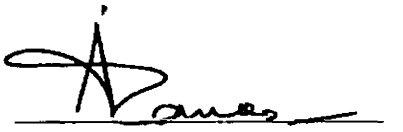
It is certificate that we have read the thesis submitted by Mr. Imtiaz Nawaz and it is our judgment that this project is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for the MS Degree in Bioinformatics

COMMITTEE

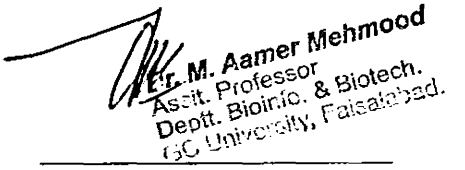
Chairman
Dr. Jabar Zaman Khan Khattak
Department of Bioinformatics and Biotechnology
International Islamic University Islamabad



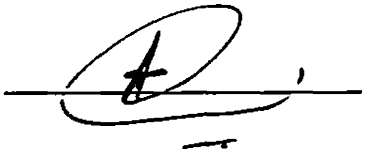
Supervisor
Dr. Jabar Zaman Khan Khattak
Chairman, Department of Bioinformatics and Biotechnology
International Islamic University Islamabad



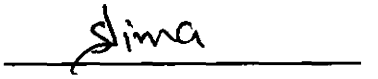
Co-Supervisor
Dr. Muhammad Aamer Mehmood
Assistant Professor
Department of Bioinformatics and Biotechnology
GC University Faisalabad



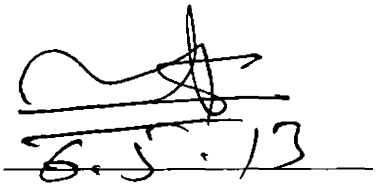
External Examiner
Dr. Amin Ali Abbasi



Internal Examiner
Dr. Saima Umbreen



Dean, FBAS
Dr. Muhammad Sher
International Islamic University Islamabad



A thesis submitted to Department of Bioinformatics and Biotechnology,
International Islamic University, Islamabad as a partial
fulfillment of requirement for the award of the
degree of MS Bioinformatics & Technology

DEDICATION

I dedicate this work to my parents who taught me that the best kind of knowledge to have is that which is learned for its own sake and even the largest task can be accomplished if it is done one step at a time.

DECLARATION

I hereby declare that the work present in the following thesis is my own effort, except where otherwise acknowledged and that the thesis is my own composition. No part of the thesis has been previously presented for any other degree.

Date 04-04-13

Imtiaz Nawaz

CONTENTS

Acknowledgements

All praises, hymns and countless thanks to **Almighty Allah**, omnipotent and the most merciful, who gave me the power and health to accomplish this task and millions of thanks to **The Holy Prophet Muhammad** (Peace Be upon Him) who is forever a model of guidance and knowledge for humanity as a whole.

Sincerest thanks are owned to my research supervisor **Dr. Jabar Zaman Khan Khattak**, chairman, Department of Bioinformatics and Biotechnology, International Islamic University Islamabad, for allocating me the research and his scholastic guidance and support throughout my project. May God enrich his honor and affectionate personality.

I fervently extend my zealous thanks to my co-supervisor **Dr. Muhammad Aamer Mehmood**, Assistant Professor, Department of Bioinformatics and Biotechnology GC University Faisalabad, for his encouragement and affectionate help, keen interest, noble guidance and appreciating attitude that enabled me to complete my research work.

No acknowledgement could ever adequately express my feelings to my **Parents and family** for providing me everything related to this project work, their advice, conviction in my abilities and encouragement to complete this task so that I will not procrastinate in doing it.

Cordial and sincere obligations are rendered to my unforgettable **classmates** and **friends** who helped me in this study and always stood by me throughout my academic period.

Imtiaz Nawaz

LIST OF ABBREVIATIONS

Acronym	Abbreviation
3D	3 Dimensional
BLAST	Basic Local Alignment Search Tool
CADD	Computer Aided Drug Designing
CDD	Conserved Domains Database
COG	Clusters of Orthologous Groups of proteins
COP	Coat Protein Complex
DNA	Deoxyribo Nucleic Acid
EBI	European Bioinformatics Institute
ER	Endoplasmic Reticulum
ExPASy	Expert Protein Analysis System
GRAVY	Grand Average of Hydropathicity
GUI	Graphical User Interface
HMM	Hidden Markov Model
II	Instability Index
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
NMR	Nuclear Magnetic Resonance
Pfam	Protein Families
PSSM	Position Specific Scoring Matrix
RPS BLAST	Reverse PSI-BLAST
SMART	Simple Modular Architecture Research Tool
SPPIDER	Solvent accessibility based Protein-Protein Interface iDEntification and Recognition
TrEMBL	Translated European Molecular Biology Laboratory

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

With the completion of human genome sequencing and impressive progress in the field of proteomics, we have got a huge amount of data that can be utilized for designing drug targets and model the different protein complexes. Bioinformatics and structural biology go hand in hand and assist us to design the drugs computationally. Homology modeling, interaction and docking analysis of proteins has great importance in proteomics as well as Bioinformatics. Homology models and protein interactions have become more accurate and their applicability is becoming increasingly important. In this study, homology model of the human protein Sec24C was created and its docking was carried out with its strong binding partners. *Insilico* study of Sec24C helped in predicting its 3D structure using MODELLER and Swiss Model. After checking its reliability, the docking analysis was carried out by AutoDock 4.2 with its binding partners i.e. Golgi phosphoprotein, Sec31A and Sec23A, which were inferred by using Bioinformatics databases IntAct, STRING and SPPIDER. Flood of sequence and structural information, and improvement in analysis tools and databases has greatly aided the field of computational biology, providing us the valuable information that can help us understand the various biological processes.

INTRODUCTION

INTRODUCTION

Proteins are vital substances which are basically on the scale of nanometers and they still are able to perform complicated biological functions (Lesk, 2001). They comprise a major portion of the cells present in our bodies and all of the other creatures. Proteins help us maintaining the life, replicate the DNA, defense the body against pathogens and help in reproduction. All of the information, that what kind of proteins will be synthesized and what will be its fate is stored in the DNA in the form of nucleotide bases (Ardala *et al.*, 2007). The remarkable progress in the fields of biology, like sequencing of the whole genome, expression of the proteins, Nuclear Magnetic Resonance (NMR), X-Ray crystallography and devising the structures of the proteins has led us to accelerate the process of discovering drugs using computers, and other *insilico* tasks implemented on these structures are also facilitated (Martin and Derewenda, 1999).

20 different types of amino acids are there which are produced by reading the code encoded in DNA and these amino acids help to form polypeptide chains or proteins, each with specific and unique function. Proteins attain the special position in the cells and they are the products of the gene expression. Efforts are being made to synthesize the proteins using unnatural amino acids these days, but there is still a long way to go (Xie *et al.*, 2001).

1.1 Applications of Bioinformatics

Bioinformatics is playing a key role featuring about all the facets that in drug discovery, development and its assessment. Bioinformatics not only handles a huge amount of data easily, but also aids all the processes to analyze, predict and help about clinical findings, including a disease gene, protein domains, macromolecular modeling, genetic interactions or computer aided drug design. With the help of computational biology, we can get a protein structure merely from its amino acid sequence. And in the structure based drug designing, the three dimensional structures of different proteins are used. In the Bioinformatics, computational techniques and expertise about the structures, complexes and their interpretation goes hand in hand which really aids the process of designing the drugs based upon the protein structures.

The completion of human and many other organisms, including pathogenic bacteria genome provided lot of raw material for *insilico* analysis (Dutta *et al.*, 2006). Different databases contain large amount of data and provide us with plenty of information about these genomes. To comprehend the importance of this data, the scientists have to gather the information about these genomes, which proteins they produce and what is their function inside a cell. In spite of this exponential information about the proteins being expressed in different genomes and their structures, we need strategies to functionally annotate these proteins. A basic strategy to know about the function of a specific protein is to carry out similarity search with the other protein sequences in these databases. This type of complex process can still be error prone (Bork and Koonin, 1998), because many of the errors in protein annotation have already been detected (Brenner, 1999).

The study about designing the drugs computationally usually includes protein functional annotation, structure prediction, protein-protein interactions, removal of signal peptides and family and superfamily classification, carried basically on uncharacterized (hypothetical) protein sequences. The *insilico* methods give us a variety of solutions to aforementioned problems. The sequence of amino acids provides us the basis for characterizing the functions of the molecules, and their chemical and physical characteristics. Features of the proteins are characterized with the help of computational power, which ultimately leads us to the information about the protein function and its interaction with the other proteins. This type of information usually includes protein pockets, partner proteins that bind to it and different domains present in its structure.

In the present study, the *insilico* analysis, homology modeling and docking studies of Sec24C protein were performed. Sec24C (a hypothetical protein) is a protein which is encoded by the *SEC24C* gene in humans (Tani *et al*, 1999). This protein is a member of the SEC24 subfamily which belongs to SEC23/SEC24 family. Sec24C is involved in the formation of coat protein complex II (COPII) vesicle coat. This coat facilitates the selective export of membrane proteins arising from the endoplasmic reticulum (ER) (Wendeler *et al.*, 2007).

In the eukaryotic cells, the newly synthesized proteins have to cross through a series of fine membrane bound chambers to reach to the cell surface (Palade, 1975). In the each step of this movement, these newly synthesized proteins are assumed to be packed inside the vesicles

that bud out from the compartment and further again combined to the acceptor compartment. During the transport of vesicles from endoplasmic reticulum (ER) and the Golgi apparatus, numerous vesicles and barriers and reported. First intermediate vesicle stage in the case of movement between ER and Golgi apparatus was devised with the help of electron microscope, which included the budding of vesicles from end region of ER (located in the rough ER) to the Golgi complex (Saraste and Kuismanen, 1984). It is necessary for all the vesicles to be vastly selective to specific proteins, so that an efficient and organized trafficking is ensured (Bonifacino and Glick 2004).

Evidence at protein level is a value which specifies that this protein exists on the level of protein, and it's experimentally proved. The complex of the Sec23/24 is a hetero dimeric protein which is involved in the vesicle coat (COPII) biogenesis (LaPointe and Balch, 2005).

Four different isoforms of Sec24C protein are expressed in human cells (Wendeler *et al*, 2007). All of these Sec24 isoforms (Sec24A, Sec24B, Sec24C and Sec24D) are expressed in a same cell but it is still a question that whether all off these isoforms performs different functions in the same cell (Pagano *et al*, 1999).

These isoforms are somehow identical on the basis of sequence similarity, as there is 75% similarity between Sec24a and Sec24b, whereas Sec24a and Sec24C are 31% identical to each other (Joseph and Jonathan, 2008). Sec24C protein consists of 1094 amino acids, with molecular weight 118325 Da. The location of *SEC24C* gene is 10q22.2 and this is shown in the figure 1.1.

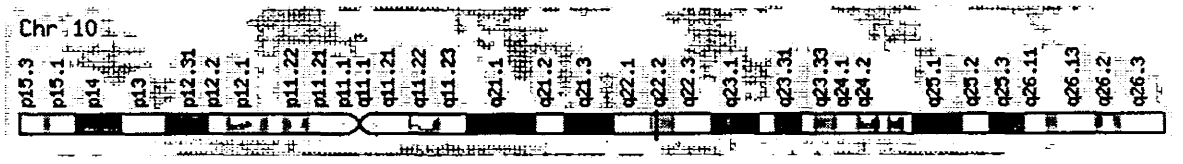


Figure 1.1: Location of SEC24C Gene on Chromosome 10

1.2 Protein structure prediction

Protein structure prediction has tremendous importance because 3D or tertiary structure of a protein delivers essential information about the function of the protein. It also aids to engineer proteins and in understanding of protein-protein interaction as well as protein-ligand interaction.

The protein 3D structure could be predicted by various methods. These methods include experimental and computational approaches. Traditional experimental methods for protein structure prediction include X-Ray crystallography and NMR spectroscopy. While computational methods include prediction of the structure from a scratch (*ab initio* method), fold recognition and homology modeling.

A huge amount of information related to proteins is stored in the databases and the distinction between the experimental and computational areas is based on the fact that to what extent this information is used to infer the structure of a query protein using computers. Different methods use varying amount of information stored in these database.

De novo or *ab initio* prediction is the purest type of protein structure prediction, which involves no use of information present in the databases. In computational biology or Bioinformatics, this method is usually applied for predicting the structure of such proteins for which there is no matching structure already present in the Protein Database (PDB). This method of predicting protein structure uses the structure relationship and local sequence involving short fragments, called motifs (Bradley *et al.*, 2005).

Template based structure prediction or homology modeling is the most reliable and accurate approach of protein structure prediction (Preston and Bianco, 2008). Homology modeling comes into this category in which 3D structure of query protein is predicted based on an optimal alignment of sequence between query protein and template sequence having already predicted structure. This approach allows the users to utilize the generated *insilico* models of the proteins in structure based drug designing, protein function annotation, their interaction with other proteins, and designing rational proteins, having more stability and novel functions in it.

In addition, protein modeling is the only way to obtain structural information if experimental technique fails to devise its structure. Thousands of protein sequences are being discovered in biotechnology labs these days, and we can apply these methods on these sequences to get an insight in their functions (Chen *et al*, 2006).

The three dimensional structure of proteins is basically important to be studied because we generally have to know about their biological activity inside the cell, and their binding with other proteins as in the field of drug designing. When these computational methods of drug designing were not present or advanced to this extent, the drugs were discovered on the basis of probability or random screening process. This thing was totally random and based on trial and error.

Protein structure prediction and protein designing are two different concepts. Predicting the structure of the proteins is the most important goal for Bioinformatics. These structures can bring about lots of improvements in the field of medicine (drug designing) as well as biotechnology (Kotelchuck and Scheraga, 1969).

1.3 Protein docking

All of the activities occurring in the cells are controlled by the interaction between different proteins. Many of the complicated protein interactions have been seen in various organisms using yeast-two hybrid system (Uetz *et al*, 2000) as well as spectrometry (Schwedt and Georg, 2002), which has revealed numerous important interactions between proteins. In spite of these numerous interactions, only a small part of them is characterized experimentally (Gabb *et al*, 1997). So this gap can be filled by showing these interaction with the help of computers which is known as protein docking. These interactions may be between enzymes and substrates, protein and nucleic acids, drugs and proteins and between two or more proteins.

All of these molecular interactions are very important as they perform many important biological processes, including regulation of cell, signal transduction and assemblies of macromolecules. The binding affinities between these molecules and recognition of these

molecules are very important phenomenon in understanding actions involved in designing the drugs.

The computational methods of determining proteins interactions are becoming more popular because the experimental techniques such as X ray crystallography are expensive economically.

Many of the vital processes taking place in the cell e.g. gene expression and transport occur due to protein complexes (Aloy and Russell, 2004). These interactions and complexes are very important to map and characterize and this field is being highlighted now the days (Zhang *et al*, 2005).

Protein tertiary structure is needed if we have to understand the molecular mechanism of the interaction between these proteins (Szilagyi *et al*, 2005), at the same time it is very difficult and time consuming with the traditional experimental methods which are X-ray crystallography and NMR. So the Bioinformatics has to develop the accurate and useful tools to predict these interactions which can be utilized to infer and assess these interactions with the help of machines (Camacho and Vajda, 2008).

In molecular docking, based on the protein structures, thousands of possible poses of association are tried and evaluated; the pose with the lowest energy score is predicted as the "best match", i.e., the binding mode (Wodak and Crombrugghe, 1987). Since Wodak and Crombrugghe's pioneering work, significant progress has been made in docking research to improve the computational speed and accuracy. Among them, protein-ligand docking is a particularly vibrant research area because of its importance to structure-based drug design (Cerqueira *et al*, 2009).

The docking procedures identify the correct poses of the proteins and ligands in the specific pockets present in the proteins. Or this can be described as docking identifies the correct binding positions of the different proteins that bind or fit together in their real location (Schwieters *et al*, 2006).

Drug discovery and protein docking are the two procedures which lie together for many years. Some of the important features are to be considered in the procedure of drug discovery, and they

mainly include the distribution, metabolism, excretion and absorption of the drugs and proteins. These features are helpful in the process of designing the drugs (Rask *et al*, 2011).

Proteins only exhibit their role when they are properly folded in the three dimensional configuration. So we can guess the importance of information about the three dimensional structures of these proteins. As there are many limitations related to the crystallographic studies, the methods which implement computers for showing protein-proteins and protein ligand interactions are becoming more and more acceptable.

Bioinformatics and structural biology both have supported the prediction of protein 3D structures and identifying there hot spots present in the proteins where they perform well established roles; they can now contribute to computer aided drug designing (CADD), determining different pathways and finding protein binding partners, ultimately leading to understanding of numerous diseases.

MATERIALS AND METHODS

MATERIALS AND METHODS

Present study was carried out on a human hypothetical protein (Sec24C), in order to find its 3D structure, finding hotspots and docking of this protein with its ligands implementing various Bioinformatics tools. Figure 2.1 depicts the main steps involved in this study. The tools and databases used to accomplish this study are given below.

2.1 Sequence retrieval

The amino acid sequence of the protein Sec24C was acquired from UniProt http://www.uniprot.org/uniprot/P53992 (Magrane, 2011). UniProt is a complete and authoritative protein information resource which provides a centralized, richly and accurately annotated as well as fully classified protein sequence knowledgebase with extensive cross-references and query interfaces.

2.2 Prediction of signal peptide cleavage site

Signal peptide commonly comprises of a chain of approximately 20 amino acids which describes the protein destiny. These sequences have great importance as they target the translocation of about all the proteins being secreted in eukaryotic and prokaryotic cells (Zheng and Gierasch 1996).

Sec24C is involved in trafficking which involves the vesicle transportation through endoplasmic reticulum ER to Golgi bodies (Tani *et al.*, 1999). Sec24C protein was analyzed to have any possible signal peptide in it by SignalP 4.0 server (http://www.cbs.dtu.dk/services/SignalP/).

The server SignalP 4.0 help predicting the presence and location of any signal peptide cleavage sites in different amino acid sequences. Several artificial neural networks are implemented in order to predict the cleavage sites and signal peptide sequences (Petersen *et al.*, 2011). The removal of signal peptide sequences ensures the protein structure prediction in accordance with its native structure.

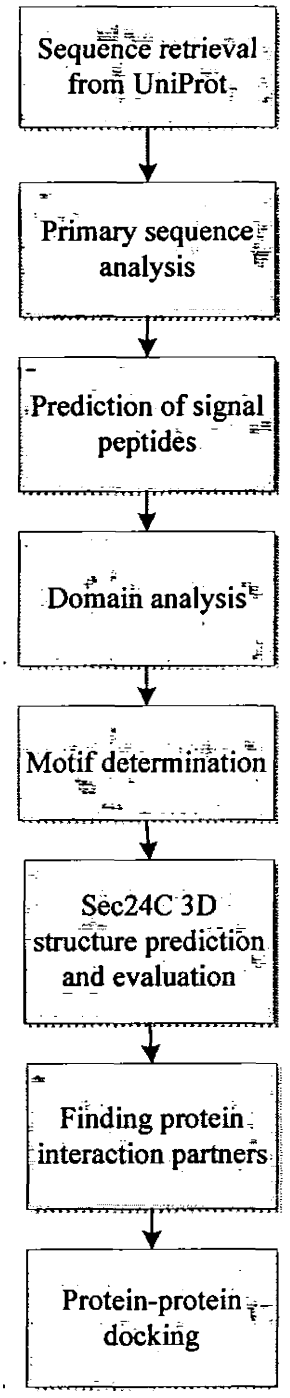


Figure 2.1: Main Steps Involved in the Present Study

2.3 Primary sequence analysis

The various statistics and aspects related to analysis of the proteins are helpful in carrying out this kind of study. A variety of statistical and physic-o chemical properties in our protein sequence were calculated and this task was accomplished by using following tools:

2.3.1 ProtParam

ProtParam, a Bioinformatics tool allows the calculation of numerous chemical and physical parameters for a given protein which is stored in TrEMBL or Swiss Prot, as well as the sequence manually entered. The parameters basically involved in ProtParam are the molecular weight, instability index (II), amino acid composition, the estimated half-life and grand average of hydropathicity (GRAVY) (Gasteiger *et al.*, 2005). All of these mentioned parameters were calculated for our query protein Sec24C by using ProtParam.

2.3.2 SAPS (Statistical Analysis of Protein Sequences)

SAPS computes a variety of properties related to a protein sequence using different statistical approaches. These properties include clusters and charges, types of amino acids lying in the query sequence, compositional biases, presence of periodic motifs and the spacing between the identical residues (Brendel *et al., 1992*).

2.4 Domain analysis

Protein domains or modules are conspicuous structural units in globular proteins. These units were identified in our query protein Sec24C using different databases that are discussed below:

2.4.1 InterPro

InterPro reserves the information related to functional analysis of protein sequences by classifying them into domains and families. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases that make up the InterPro consortium.

InterPro combines signatures from multiple, diverse databases into a single searchable resource, reducing redundancy and helping users interpret their sequence analysis results. InterPro is a composite repository, thus augmenting its importance and use (Hunter *et al.*, 2012).

2.4.2 Pfam database

Pfam is a large collection of information about different protein families. All of the information in the Pfam is either represented by Hidden Markov Models (HMM) or Multiple Sequence Alignments (MSA) (Finn *et al.*, 2010).

2.4.3 CDD: Conserved Domains Database at NCBI

CDD is a valuable store house of the information related to annotation of the proteins and multiple sequence alignment of protein domains and full length sequences. One can run RPS-BLAST for fast identification of the domains, using position specific scoring matrices (PSSMs).

CDD stores the information about NCBI curated domains. This uses the 3D structures of the proteins to define the conserved domains in its amino acid sequence. Further once can explore the structure, function and relationships among different proteins. All of these tasks can be performed on the proteins stored in different repositories like SMART, Pfam and COG (Marchler *et al.*, 2011).

2.5 Motif determination

We can define the motif as "a minute portion of the protein, usually not more than 20 amino acids, which is conserved or similar in the proteins which are homologs to it". The presence of particular motifs reflects basic information about structure and specific functions of the proteins. Motifs were determined in our query sequence Sec24C by PROSITE and PRINTS.

2.5.1 PROSITE

It is a database containing the information about protein domains and families. There are thousands of proteins but still they can be combined together in a few number of families based on the similarity in their amino acid sequences. It is observed that the conserved protein

domains, and proteins sharing the same family also shares common characteristics. This is so because these proteins share a common ancestor (Sigrist *et al.*, 2010).

2.5.2 PRINTS

PRINTS is a database which can be thought as an album of fingerprints of thousands of proteins. A fingerprint can be considered to be a conserved motif in different protein. These fingerprints are helpful to characterize a protein on the basis of its family. PRINTS has high diagnostic power as it carries out an iterative scan of TrEMBL or Swiss Prot databases.

Commonly it is seen that the motifs don't overlap and they are segregated along a sequence length, but they are contiguous in the 3D form. Fingerprints are more flexible and powerful as compared to motifs because they can encode the folds present in proteins and define the functionalities of them more effectively (Attwood *et al.*, 2012).

2.6 Sec24C 3D structure prediction and evaluation

3D structure of Sec24C protein was predicted from sequence of amino acids present in it using two Bioinformatics tools, Swiss-Model (Arnold *et al.*, 2006), and MODELLER (Eswar *et al.*, 2006), using homology modeling approach.

2.6.1 Swiss-Model

The SWISS MODEL (http://swissmodel.expasy.org/) is powerful and automated server for predicting the structure of a protein. This can be accessed via web server at ExPASy, or the Deep View program (Swiss Pdb-Viewer). Swiss Pdb-Viewer is a computer program that provides graphical user interphase (GUI) which is very user friendly, and we can perform analysis of a several proteins in it simultaneously.

We can compare the active sites and infer the structural alignments with the relevant parts of the protein by superimposing them. The hydrogen bonds, mutations in amino acids, the bond lengths and distances can be easily obtained as this program offers an easy to use and user friendly menu and graphic interface.

2.6.2 MODELLER

The computer program MODELLER can produce the homology models of tertiary and quaternary structures of the proteins by aligning the sequence of the query protein with different templates (Fiser and Marti, 2003). User has to provide a file of alignment of sequences between the sequence which is to be modeled and the sequence with known structures. The MODELLER thus automatically calculates the model for query protein which contains all the non-hydrogen atoms.

MODELLER gives the reliable model by comparing the protein structures considering all the spatial restraints. In addition to structure prediction many other tasks can also be performed by MODELLER such as *de novo* modeling of loops present in protein structures, assess the already predicted model by using flexibility function, sequence database searches and comparing different protein structures.

The latest version of MODELLER (MODELLER 9.10 compatible with Windows 7) was implemented for building 3D structural model of the Sec24C.

2.7 Visualization of the model

Protein models were encoded by atomic coordinate files, having .pdb extension. And these files were visualized from various perspectives using the following visualizing tool:

2.7.1 The PyMOL molecular graphics system. PyMOL

PyMOL was created by DeLano and Warren Lyford, which is a user funded and open source for carrying out molecular visualizations. It is property of a private software company the DeLano Scientific Laboratories.

It can be implemented in order to visualize the high quality images of biological molecules and complexes, for example proteins and their interactions. This is one of the few open source programs of its kind which is frequently used in the structural biology (http://sourceforge.net/projects/pymol/).

2.8 Homology modeling

Protein 3D structure is predicted by a number of methods which include *ab initio* prediction methods and homology modeling techniques. Both of these methods use multiple sequence alignment (MSA) for building the 3D protein structures. These alignments are carried out with the structures already present in the database Protein Data Bank (PDB). BLAST is carried out to select the suitable templates which are further used in the experiment (Arnold *et al.*, 2006).

In this study, we have derived Sec24C protein structure from its amino acid sequence by homology modeling method and this method consisted of 4 different steps. The first step was to find the structurally related sequences (templates) to our query protein sequence, second step involved the alignment of these templates with our query sequence one by one, and third step was to build a model, while fourth step was to assess our model by various techniques. The whole procedure is described in the figure 2.2.

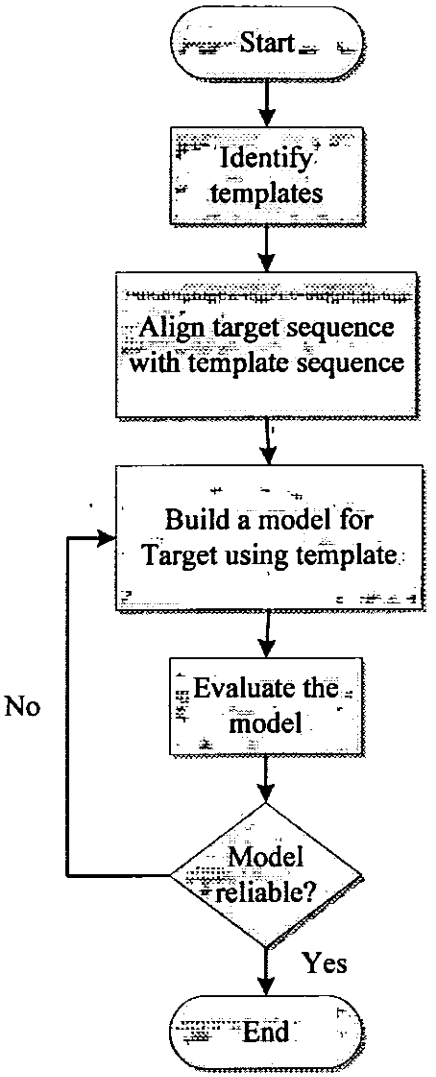


Figure 2.2 Steps Involved in Homology Modeling

2.8.1 Template identification for query sequence

Templates are searched in the first step of homology modeling which are used to model the target sequence. These templates help identifying the structures of target sequences by optimally aligning the query sequence with these templates (Ponder and Richards, 1987).

For this purpose a search was carried out in a database or library of experimentally determined protein structures for identifying the appropriate templates for our query protein (Berman *et al.*, 2009). The similarity between the template and query sequences is obtained by implementing BLAST against PDB (Macey *et al.* 2003). This process was accomplished by using a web based tool Protein-BLAST, found at NCBI (http://blast.ncbi.nlm.nih.gov/), property of Protein Data Bank (PDB) database.

After running BLAST, the PDB files with highest sequence similarity (between query and template) and lowest E-value were downloaded manually. Swiss Model selected the suitable templates by using Template identification tool (http://swissmodel.expasy.org/) and by PDB advance search (http://www.pdb.org/pdb/home/home.do).

2.8.2 Alignment between template and query sequence

A reasonable sequence alignment between query and template is necessary for a reasonable model building. These sequence alignments can be used to carry out the search against database for production of the model. More refined approaches are also there to carry the same process however. One approach gives out sophisticated alignments carried out between the query and template sequences one by one, each with a high level of global similarity (Phuong *et al.*, 2006).

Swiss Model carries out automatic significant sequence alignment under automated mode. While MODELLER uses Align2D command to carry out this process for best selected template. The Align2D implements global dynamic programming with an affine gap penalty function (Renom *et al.*, 2004) and is preferred for aligning a sequence with structures because it tends to place gaps in a better structural context (Eswar *et al.*, 2006).

2.8.3 Building the model

Swiss Model created a personal workspace for the protein Sec24C online and automated model was generated after template selection and significant sequence alignment. The reliability and quality of the models generated by Swiss Model was estimated by the QMEAN4 score available at the QMEAN server (Benkert *et al.*, 2009). It is a composite tool which consists of rating the predicted model between the score 0-1. It is the estimated reliability associated with the model.

After completion of the modeling procedure, the results were received in the email containing work unit ID number. This model was downloaded in the PDB format file to be visualized later.

3D models of the target protein Sec24C were built automatically by MODELLER, using its automodel class. Five models were created by MODELLER of which the model consisting of lowest objective function score, the DOPE (http://swissmodel.expasy.org/qmean/cgi/index.cgi?) (Discrete Optimized Molecule Energy) assessment score (Marti *et al.*, 2006) was selected after carrying out the screening of the log file. The model with the smallest value of the normalized DOPE score is considered to be the best model.

2.8.4 Evaluation and refinement of model

To show our model to be reliable, it was necessary to show the important structural features which were already known about the protein structure in its real place or in general. These criteria included the following important points.

The acceptability of the main chain confirmations as stated in Ramachandran plot, the planes of peptide bonds in polypeptide, confirmations of the side chains as those present in the rotamer repository, H-bonding of all the buried polar atoms, fulfilling the position requirements for hydrophilic and hydrophobic residues, no false contacts between different atoms and ensuring the absence of any holes inside the structure (Bowie *et al.*, 1991).

To check all of these parameters for the model of Sec24C, MolProbity (http://molprobity.biochem.duke.edu/) and NIH server (http://nihserver.mbi.ucla.edu/SAVES/) were employed.

Molprobity gave the Ramachandran plot (Ramachandran *et al.*, 1963), rotamers, C_β deviation, bond angles and bond lengths of our model. NIH server is a composite tool, including PROCHECK, that can be employed in in order to know about the stereo chemical properties with respect to each residue present in the whole protein structure, WHAT_CHECK which is derivative of several protein analyzing tools from the program WHATIF (Vriend, 1990) which checks the chemical and several other characteristics of the residues present in the model, ERRAT which helps analyzing the non-bound interactions of the error functions having the window of 9 residues, VERIFY-3D gives the compatibility of a 3D atomic model against the amino acid sequence of its own, including the loops, beta, alpha and non-polar properties, furthermore, it compares the results of stable structures and PROVE that uses the algorithm to treat the residues as hard components and calculates the volumes of atoms in the macromolecular structures.

2.9 Finding protein interaction partners

The first step to carry out docking procedure is to find the interaction partners for the query protein. The protein interaction partners for Sec24C were found out and validated using IntAct and STRING databases.

2.9.1 IntAct

IntAct is available at Europian Bioinformatics Institute (EBI), which offers a freely available to users the tools and database system which helps analyzing the interactions between the macromolecules. This database stores the information about interactions which is sent by users and available in literature. Search for the protein interaction partners of our protein Sec24C was carried out by entering its name in the search field in IntAct version 4.0.1.

2.9.2 STRING

STRING predicts the supposed associations between proteins which are based upon the genomic neighborhoods which remains conserved in various protein families. It is supposed that the genes which are neighbors to each other in various genomes are also associated functionally. So we can predict these genes to be functionally associated with others by using this tool. Different interacting proteins were determined for Sec24C, using STRING 9.0 by its name, filtering the results only for Homo sapiens.

2.10 Protein binding site prediction

SPPIDER- Solvent accessibility based Protein-Protein Interface iDEntification and Recognition (Porollo and Meller, 2007) was employed to predict the binding sites in our query protein.

The protein interface recognition server – SPPIDER uses the 3D structure of the protein to check whether the residues present in it are at its putative interfaces in each of the chain present in it. Query was processed under prediction of interaction sites using an unbound protein 3D structure by uploading our model. Tradeoff between Sensitivity and Specificity was selected to be at 0.5 (balanced). This tradeoff can be set according to the type of research being carried out. Each version of the SPPIDER works on the basis of different tradeoff values to calculate the knowledge about the residues present in the interface of proteins (Porollo and Meller; Porollo and Meller, 2007 and 2012).

2.11 Protein docking

Protein docking is a computational technique that aims to predict whether and how a particular small molecule, such as protein will stably bind to a target protein (Cases and Mestres, 2009).

Docking or computational modeling of the protein-protein interactions was started with *ab initio* methods, which did not use any template information and totally based on shape complementarity methods (Vakser and Kundrotas, 2008). In the living cells, function of the protein is determined by their ability to interact with other similar molecules, such as proteins, DNA, RNA and small ligands (Russel *et al.*, 2004). Protein complexes control the number of

crucial processing occurring in the cell such as regulation and transport and expression of the genes (Szilagyi *et al.*, 2005).

Many docking procedures have been proposed over about 30 years that range from rigid body docking to flexible docking, these methods are described below.

2.11.1 Rigid body docking and flexible docking

Rigid body docking is the procedure in which the proteins are considered as rigid and no flexibility is taken into account (Gray *et al.*, 2003). A few steps involved in this procedure are illustrated in figure 2.4 (Kohlbacher, 2001), starting from two unbound structures, a lot of docked complex structures are generated by a structure generator. Then these structures are filtered using a scoring function and only a few favorable structures are left for evaluation in more details.

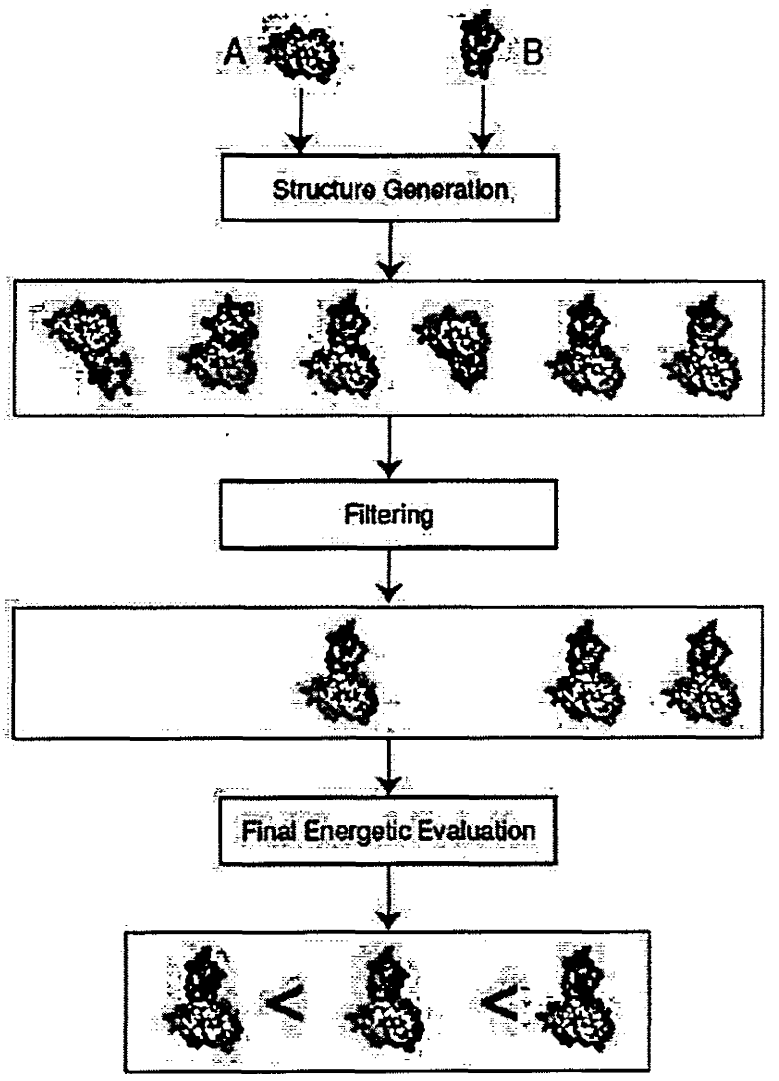


Figure 2.3: The Overall Structure of Rigid-body Docking Algorithm (Kohlbacher, 2001).

On the other hand the flexibility in the protein structures is an important phenomenon to understand molecular interactions because most of the proteins exhibit significant changes in their structures when binding to ligands or other proteins (Koshland 1963), named "flexible docking" (Gabb *et al.*, 1997).

The structure based designing of drugs includes the docking of macromolecules such as DNA or proteins and receptors which is based on the binding capacity in the protein complex under consideration (Daniel and Bert, 2010).

To completely understand the phenomenon of binding of the proteins to form a complex and their affinity with each other, we need accurate and fast programs to use structures and visualize their complex formation. In the present study we have met the goals to find the interacting proteins, protein binding site prediction and implementing the docking procedure on our query protein Sec24C with its strong interacting partners. Docking experiments on Sec24C with selected protein partners were performed with AutoDock 4.2 (Morris *et al.*, 2009) - http://autodock.scripps.edu/. Graphical User Interphase – GUI and AutoDock Tools version 1.5.4 was applied for this purpose.

AutoDock 4.2 is freely downloadable using a proper license at the web server WWW site: http://autodock.scripps.edu. AutoDock Tools – ADT is also available free of charge in the MGLTools package on the website http://mgltools.scripps.edu/downloads. The complexes obtained by the docking procedures were visualized by using PyMOL (DeLano, 2002) - http://www.pymol.org.

In the Autodock the receptor and binding proteins were represented in .pdbqt format, which is an altered form of protein data bank format. It contains significant information, like types of atoms present in protein, and rotatable bonds. This file was created with the help of Autodock tools package. In this file the binding sites were defined and proteins to be docked were prepared.

Prior to running the docking procedure, the interaction maps were calculated by using the grid function, which specifies the interaction energy between the proteins to be docked. The main steps involved in the docking procedure are given in the figure 2.5. The files created by this method were stored in a library containing proteins to be docked and this file was launched with the help of command line for seeing the results of docking procedure.

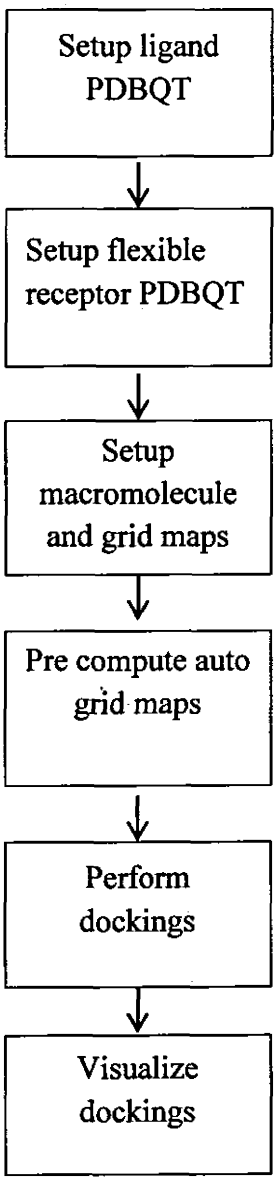


Figure 2.4: Main Steps Involved in the Docking Procedure

RESULTS

RESULTS

The results of this study are as follows for each step:

3.1 Query protein sequence

The amino acid sequence of Sec24C, the hypothetical protein was acquired from UniProt, which is given below in the FASTA format.

```
>sp|P53992|SC24C_HUMAN Protein transport protein Sec24C OS=Homo sapiens
GN=SEC24C PE=1 SV=3
MNVNQSVPPVPPFGQPQPIYPGYHQSSYGGQSGSTAPAIPYGAYNGPVPGYQQTPPQGMS
RAPPSSGAPPASTAQAPCGQAAYGQFGQGDVQNGPSSTVQMQRLPGSQPFGSPLAPVGNQ
PPVLQPYGPPPTSAQVATQLSGMQISGAVAPAPPSSGLGFGPPTSLASASGSFPNSGLYG
SYPQGQAPPLSQAQGHPGIQTPQRSAPSQASSFTPPASGGPRLPSMTGPLLPGQSFGGPS
VSQPNHVSSPPQALPPGTQMTGPLGPLPPMHSPQQPGYQPQQNGSFGPARGPQSNYGGPY
PAAPTFGSQPGPPQPLPPKRLDPDAIPSPIQVIEDDRNNRGTEPFVTGVRGQVPPLVTTN
FLVKDQGNASPRYIRCTSYNIPCTSDMAKQAQVPLAAVIKPLARLPPEEASPYVVDHGES
GPLRCNRCKAYMCPFMQFIEGGRRFQCCFCSCINDVPPQYFQHLDHTGKRVDAYDRPELS
LGSYEFLATVDYCKNNKFPSPPAFIFMIDVSYNAIRTGLVRLLCEELKSLLDFLPREGGA
EESAIRVGFVTYNKVLHFYNVKSSLAQPQMMVVSDVADMFVPLLDGFLVNVNESRAVITS
LLDQIPEMFADTRETETVFVPVIQAGMEALKAAECAGKLFLFHTSLPIAEAPGKLKNRDD
RKLINTDKEKTLFQPQTGAYQTLAKECVAQGCCVDLFLFPNQYVDVATLSVVPQLTGGSV
YKYASFQVENDQERFLSDLRRDVQKVVGFDAVMRVRTSTGIRAVDFFGAFYMSNTTDVEL
AGLDGDKTVTVEFKHDDRLNEESGALLQCALLYTSCAGQRRLRIHNLALNCCTQLADLYR
NCETDTLINYMAKFAYRGVLNSPVKAVRDTLITQCAQILACYRKNCASPSSAGQLILPEC
MKLLPVYLNCVLKSDVLQPGAEVTTDDRAYVRQLVTSMDVTETNVFFYPRLLPLTKSPVE
STTEPPAVRASEERLSNGDIYLLENGLNLFLWVGASVQQGVVQSLFSVSSFSQITSGLSV
LPVLDNPLSKKVRGLIDSLRAQRSRYMKLTVVKQEDKMEMLFKHFLVEDKSLSGGASYVD
FLCHMHKEIRQLLS
```

3.2 Signal peptide cleavage sites

A signal peptide usually consists of a short fragment of the residues (about 3 to 60) in the peptide chain that decides the final destination of a protein to a specific region in a cell (Rapoport T. 2007). The graphical output from SignalP, (available at The Center for Biological Sequence Analysis at the Technical University of Denmark - DTU) consists of three different scores, C, S and Y. The score C represents the score of the "cleavage site" which is higher on the region of the cleavage as compared to the other regions. Y is combined with the C score for the better prediction of the cleavage site. The Y score corresponds the cleavage site, in which the S score slope is steep and C score is most significant. Average of the S score is represented by the S

mean. It starts from the N terminal amino acids and goes up to the amino acids having the maximum Y score. It means that the S mean is calculated for the whole length of the signal peptide.

Figure 3.1 shows the maximum C-score calculated by SignalP for our query protein Sec24C on the position 48, with cutoff 0.450. This shows the absence of any signal peptide in our query protein.

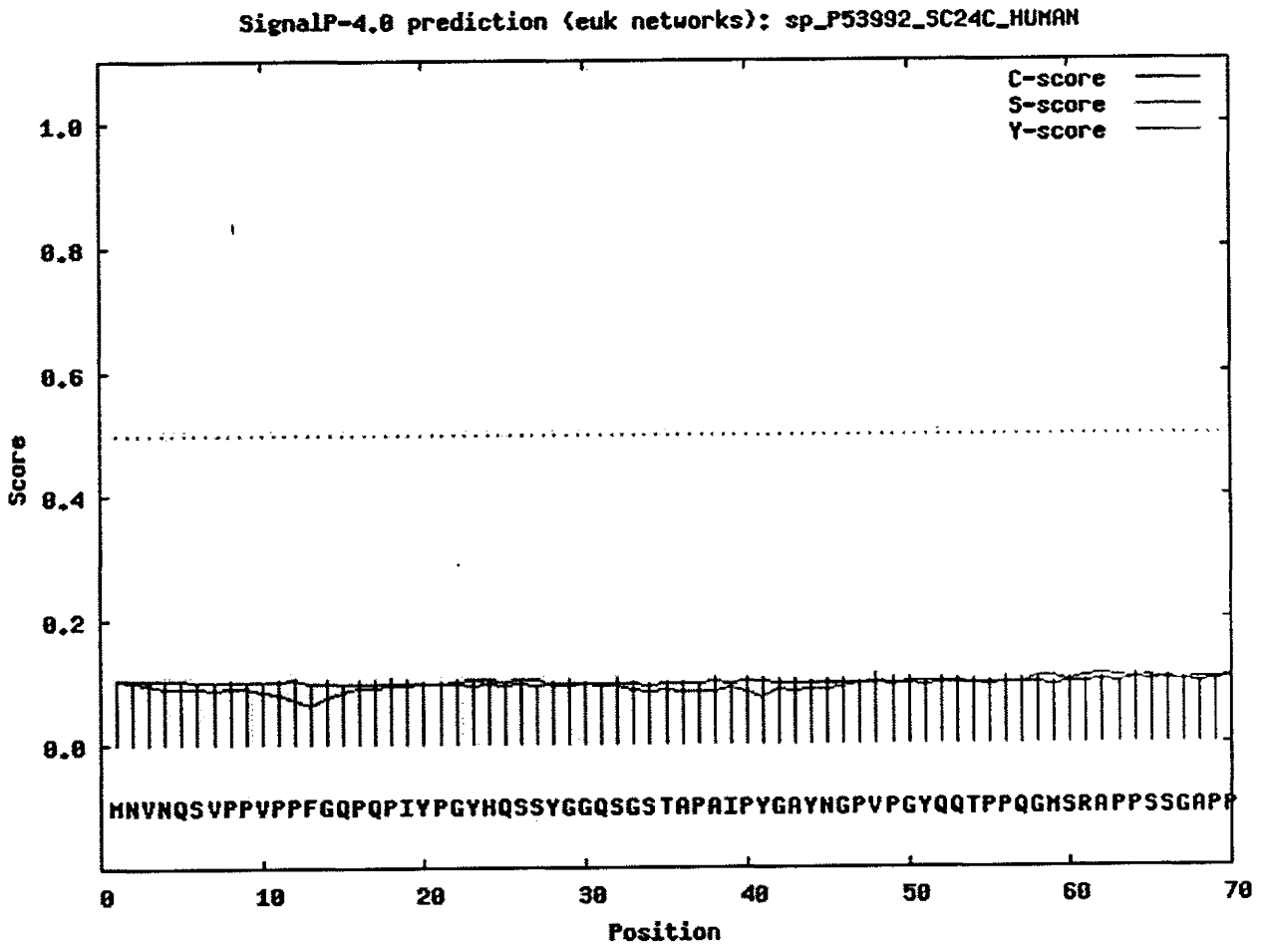


Figure 3.1: SignalP Results for Sec24C. Max. C-score is present on position 48 with cutoff value 0.450, showing absence of signal peptide in this sequence.

Mostly the cleavage takes place on the carboxyl side present in the polypeptides. Coordinates of signal peptides for Sec24C were also predicted by using version 4.0 of SignalP. Different parameters like the measure, position and values of these signal peptides were calculated using SiganlP, with the cutoff 0.450. This analysis showed the absence of any signal peptide in our query sequence. Table 3.1 describes the values of various parameters calculated by SignalP.

Table 3.1: Measeure, Position and Values Generated by SignalP for Sec24C, with Cutoff 0.450.

Measure	Position	Value
Max. C	48	0.113
Max. Y	70	0.105
Max. S	62	0.109
Mean S	1-69	0.094

3.3 Primary sequence analysis

Various physic-o chemical parameters like number of amino acids, molecular weight and theoretical isoelectric point were calculated for Sec24C using ProtParam and SAPS. Table 3.2 shows the number of amino acid, molecular weight and theoretical isoelectric point for our query protein Sec24C, while table 3.3 shows the aligned matching blocks in the query protein sequence.

Table 3.2: ProtParam result for the Primary Sequences Analysis of Sec24C

No of Amino Acids	Molecular Weight	Theoretical pI
1094	118324.7	6.71

Table3.3: Aligned Matching Blocks in Sec24C

Block	Sequence
[62 - 67]	APPSSG
[152- 157]	APPSSG
[226 - 232]	MTGPL__LP
[260 - 268]	MTGPLGPLP

Moreover, ProtParam showed that the target protein has overall 89 negatively charged residues and 87 positively charged residues, while the instability index (II) is computed to be 51.85 which classify the protein as unstable.

Charge distribution analysis in SAPS showed that there are no high scoring positive and negative charge segments as well any high scoring hydrophobic and trans-membrane segments.

3.4 Protein domains in Sec24C

InterPro, Pfam and CDD databases were employed in order to analyze the domains present in our query protein Sec24C. This analysis shows that there are five different domains present in this protein and their positional distribution is shown over the entire sequence.

First domain, named zinc finger starts at the position 408, similarly further domains present in this protein are shown named trunk domain, beta-sandwich domain, helical domain and gelsolin domain (ending at position 1,034) respectively. These results are shown in the figure 3.2 (a & b). While table 3.4 contains the domain names present in Sec24C, their respective positions and E-values.

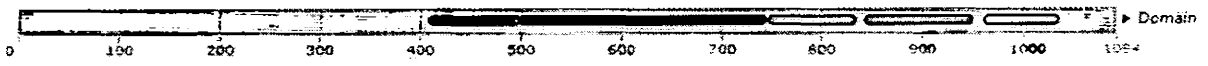


Figure 3.2 a: InterPro Domain Analysis of Sec24C

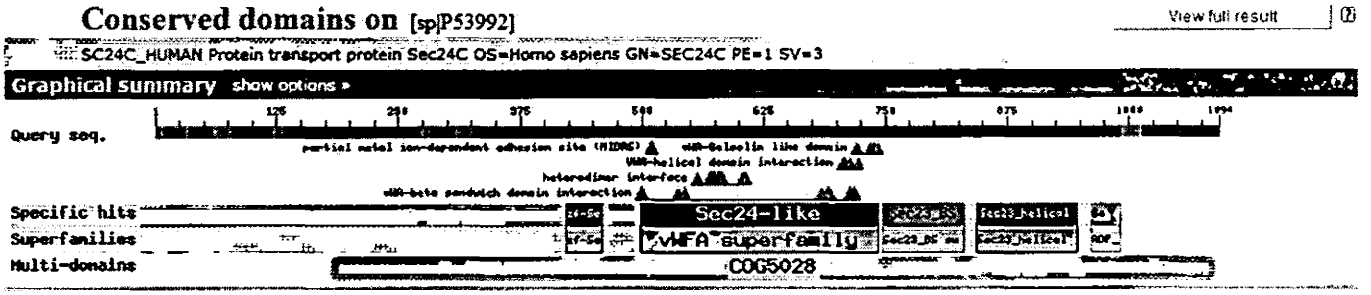


Figure 3.2 b: CDD-Conserved Domain Database analysis

Table 3.4: Domains, Respective Positions and Their E-values in the Sec24C

Domain Name	**Position In The Sequence**	**E-value**
Zinc finger	408-496	5.94e-17
Trunk domain	499-743	4.11e-108
Beta-sandwich	748-831	4.74e-27
Helical domain	845-945	1.44e-31
Gelsolin domain	962-1,043	4.11e-06

3.5 Motifs in Sec24C

PROSITE found one motif in the sequence of Sec24C at the position of 980-988, the glycosyl hydrolase and "Nucleophile" by similarity. Table 3.5 shows this motif with its related information and figure 3.3 describes the location of this motif with respect to a ruler.

Table 3.5: Motif and its Description Found on PROSITE

Motif	**Position**	**PROSITE ID**	**Description**	**Sequence tag**
ACT_SITE	980-988	PS00572	Glycosyl hydrolases family 1 active site	IYLLENGLN

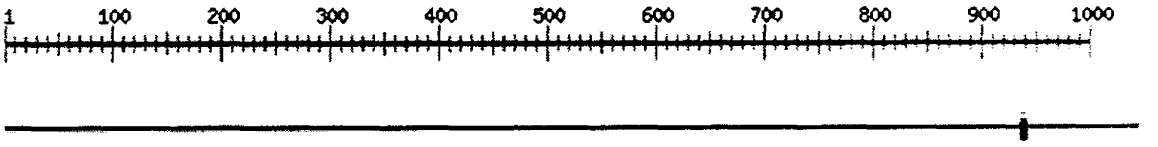


Figure 3.3: Motif Hit with Relative Position and Ruler, found on PROSITE

9 possible hits were reposted by PRINTS in our protein with cutoff block expected value 1 (Henikoff and Henikoff, 1994). These results along with their descriptions are given in the table 3.6.

Table 3.6: Motifs Found in the Query Sequence by PRINTS and their Description.

Family	Location	Description	PRINTS E-value
IPB006895	423-450, 501-519	Sec23/Sec24 zinc finger	5.2e-27
IPB006895	272-312	Sec23/Sec24 helical domain	6.1e-27
IPB006900	445-457	Tumour susceptibility gene 101	0.00016
IPB008883	14-58, 265-309	Sec23/Sec24 trunk domain	0.00067
IPB008883C	927-994	Synaptophysin/synaptoporin	0.0018
IPB006896	147-166	Gelsolin family signature	0.013
IPB006896A	274-315	Gelsolin region	0.013
IPB001285F	265-305	Nucleoporin Nup153-like	0.21
IPB007122	259-299	Prion protein	0.99

3.6 Template Identification for Sec24C

Four different templates for Sec24C were selected through BLAST. These templates were selected on the basis of lowest E-value and maximum similarity. These templates had similarity with our query protein 83%, 76%, 74% and 66% respectively. These files were downloaded in PDB file format to use further in MODELLER. Swiss Model automatically selected best template to model our query protein.

3.7 Alignment of template with query sequence

3.7.1 Swiss Model template-query alignment

Following template-query alignment was accomplished in Swiss Model.

```
Query: 331  QVIEDDRNNRGTEPFVTGVRGQVPPLVTTNFLVKDQGNASPRYIRCTSYNIPCTSDMAKQ 390
            QVIEDDRNNRGTEPFVTGVRGQVPPLVTTNFLVKDQGNASPRYIRCTSYNIPCTSDMAKQ
Sbjct: 1    QVIEDDRNNRGTEPFVTGVRGQVPPLVTTNFLVKDQGNASPRYIRCTSYNIPCTSDMAKQ 60

Query: 391  AQVPLAAVIKPLARLPPEEASPYVVDHGESGPLRCNRCKAYMCPFMQFIEGGRRFQCCFC 450
            AQVPLAAVIKPLARLPPEEASPYVVDHGESGPLRCNRCKAYMCPFMQFIEGGRRFQCCFC
Sbjct: 61   AQVPLAAVIKPLARLPPEEASPYVVDHGESGPLRCNRCKAYMCPFMQFIEGGRRFQCCFC 120

Query: 451  SCINDVPPQYFQHLDHTGKRVDAYDRPELSLGSYEFLATVDYCKNNKFPSPPAFIFMIDV 510
            SCINDVPPQYFQHLDHTGKRVDAYDRPELSLGSYEFLATVDYCKNNKFPSPPAFIFMIDV
Sbjct: 121  SCINDVPPQYFQHLDHTGKRVDAYDRPELSLGSYEFLATVDYCKNNKFPSPPAFIFMIDV 180

Query: 511  SYNAIRTGLVRLLCEELKSLLDFLPREGGAEESAIRVGFVTYNKVLHFYNVKSSLAQPQM 570
            SYNAIRTGLVRLLCEELKSLLDFLPREGGAEESAIRVGFVTYNKVLHFYNVKSSLAQPQM
Sbjct: 181  SYNAIRTGLVRLLCEELKSLLDFLPREGGAEESAIRVGFVTYNKVLHFYNVKSSLAQPQM 240

Query: 571  MVVSDVADMFVPLLDGFLVNVNESRAVITSLLDQIPEMFADTRETETVFVPVIQAGMEAL 630
            MVVSDVADMFVPLLDGFLVNVNESRAVITSLLDQIPEMF      TETVFVPVIQAGMEAL
Sbjct: 241  MVVSDVADMFVPLLDGFLVNVNESRAVITSLLDQIPEMF-----TETVFVPVIQAGMEAL 295

Query: 631  KAAECAGKLFLFHTSLPIAEAPGKLKNRDDRKLINTDKEKTLFQPQTGAYQTLAKECVAQ 690
            KAAECAGKLFLFHTSLPIAEAPGKLKNRDDRKLINTDKEKTLFQPQTGAYQTLAKECVAQ
Sbjct: 296  KAAECAGKLFLFHTSLPIAEAPGKLKNRDDRKLINTDKEKTLFQPQTGAYQTLAKECVAQ 355

Query: 691  GCCVDLFLFPNQYVDVATLSVVPQLTGGSVYKYASFQVENDQERFLSDLRRDVQKVVGFD 750
            GCCVDLFLFPNQYVDVATLSVVPQLTGGSVYKYASFQVENDQERFLSDLRRDVQKVVGFD
Sbjct: 356  GCCVDLFLFPNQYVDVATLSVVPQLTGGSVYKYASFQVENDQERFLSDLRRDVQKVVGFD 415

Query: 751  AVMRVRTSTGIRAVDFFGAFYMSNTTDVELAGLDGDKTVTVEFKHDDRLNEESGALLQCA 810
            AVMRVRTSTGIRAVDFFGAFYMSNTTDVELAGLDGDKTVTVEFKHDDRLNEESGALLQCA
Sbjct: 416  AVMRVRTSTGIRAVDFFGAFYMSNTTDVELAGLDGDKTVTVEFKHDDRLNEESGALLQCA 475

Query: 811  LLYTSCAGQRRLRIHNLALNCCTQLADLYRNCETDTLINYMAKFAYRGVLNSPVKAVRDT 870
            LLYTSCAGQRRLRIHNLALNCCTQLADLYRNCETDTLINYMAKFAYRGVLNSPVKAVRDT
Sbjct: 476  LLYTSCAGQRRLRIHNLALNCCTQLADLYRNCETDTLINYMAKFAYRGVLNSPVKAVRDT 535
```

```
Query: 871  LITQCAQILACYRKNCASPSSAGQLILPECMKLLPVYLNCVLKSDVLQPGAEVTTDDRAY 930
            LITQCAQILACYRKNC      GQLILPECMKLLPVYLNCVLKSDVLQPGAEVTTDDRAY
Sbjct: 536  LITQCAQILACYRKNC------GQLILPECMKLLPVYLNCVLKSDVLQPGAEVTTDDRAY 589

Query: 931  VRQLVTSMDVTETNVFFYPRLLPLTKSPVESTTEPPAVRASEERLSNGDIYLLENGLNLF 990
            VRQLVTSMDVTETNVFFYPRLLPLT    ESTTEPPAVRASEERLSNGDIYLLENGLNLF
Sbjct: 590  VRQLVTSMDVTETNVFFYPRLLPLT----ESTTEPPAVRASEERLSNGDIYLLENGLNLF 645

Query: 991  LWVGASVQQGVVQSLFSVSSFSQITSGLSVLPVLDNPLSKKVRGLIDSLRAQRSRYMKLT 1050
            LWVGASVQQGVV          QITSGLSVLPVLDNPLSKKVRGLIDSLRAQRSRYMKLT
Sbjct: 646  LWVGASVQQGVV----------QITSGLSVLPVLDNPLSKKVRGLIDSLRAQRSRYMKLT 695

Query: 1051 VVKQEDKMEMLFKHFLVEDKSLSGGASYVDFLCHMHKEIRQLLS 1094
            VVKQEDKMEMLFKHFLVEDKSLSGGASYVDFLCHMHKEIRQLLS
Sbjct: 696  VVKQEDKMEMLFKHFLVEDKSLSGGASYVDFLCHMHKEIRQLLS 739
```

3.7.2 MODELLER template-query alignment

The alignment bellows was carried out by MODELLER with best selected template 3EH2 for our query protein Sec24C.

```
>P1;Sec24C
sequence:Sec24C:     : :      : :::-1.00:-1.00
NCNPELFRCTLTSIPQTQALLNKAKLPLGLLLHPFKDLVQLPVVTSSTIV---------R
CRSCRTYINPFVSFLDQ-RRWKCNLCYRVNDVPEEF----------LEPHRRPEVQNATI
EFMAPSEYML--RPPQPPVYLFVFDVSHNAVETGYLNSVCQSLLDNLDLLP-----GNTR
TKIGFITFDSTIHFYGLQESLSQPQMLIVSDIEDVFIPMPENLLVNLNESKELVQDLLKT
LPQMFTKTLETQSALGPALQAAFKLM--SPTGGRMSVFQTQLPTL-GVGALKPREEPNHR
SSAK----MTPSTDFYKKLALDCSGQQVAVDLFLLSGQYSDLASLGCISRYSAGSVYYYP
SYHHQHNPVQVQKLQKELQRYLTRKIGFEAVMRIRCTKGLSIHTFHGNFFVRSTDLLSLP
NVNPDAGYAVQMSVEESLTDTQLVSFQSALLYTSSKGERRIRVHTLCLPVVSTLNDVFLG
ADVQAISGLLANMAVDRSMTASLSDARDALVNAVIDSLSAYR------SSVPGLMVPFSL
RLFPLFVLALLKQKSFQTGTNARLDERIFAMCQVKNQPLVYLMLTTHPSLYRVDNLSDEG
ALNISDRTIPQPPILQLSVEKLSRDGAFLMDAGSVLMLWVGKNCTQNFLSQVLGVQNYAS
IPQPMTDLPELDTPESARIIAFISWLREQRPFFPILYVIADESPMKANFLQNMIEDRTES
-ALSYYEFLLHIQQQVNK*
```

3.8 Model of the Sec24C

Full length 3D models of the protein Sec24C were created by Swiss Model and MODELLER. The reliable models, fulfilling the requirements of the evaluation procedures were selected and visualized by PyMol version 1.3. Figure 3.4 (a & b) shows both of these models, predicted by Swiss Model and MODELLER respectively.

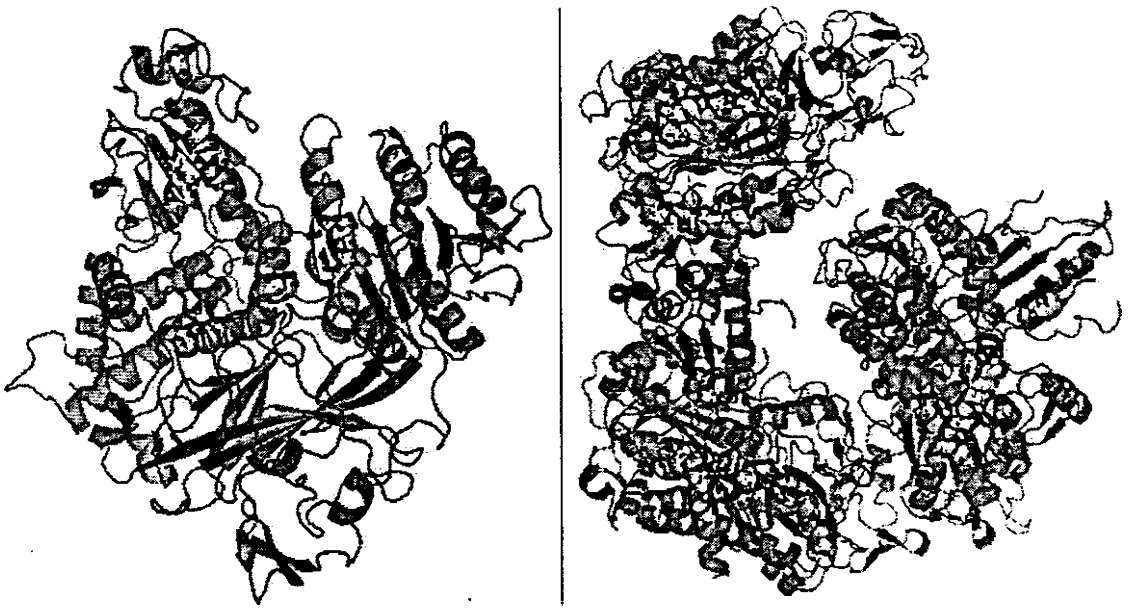


Figure 3.4: 3D Models of Sec24C Visualized using PyMOL Version 1.3, (a) Swiss Model Result, (b) MODELLER Result.

3.8.1 Evaluation of the model

MolProbity was used to draw the Ramachandran plot against our query protein Sec24C which showed that 94.6% (723/764) of all residues were in favored (98%) regions and 98.4% (752/764) of all residues were in allowed (>99.8%) regions. A model with these parameters is considered satisfactory and reliable. This Ramachandran plot is given in the figure 3.5.

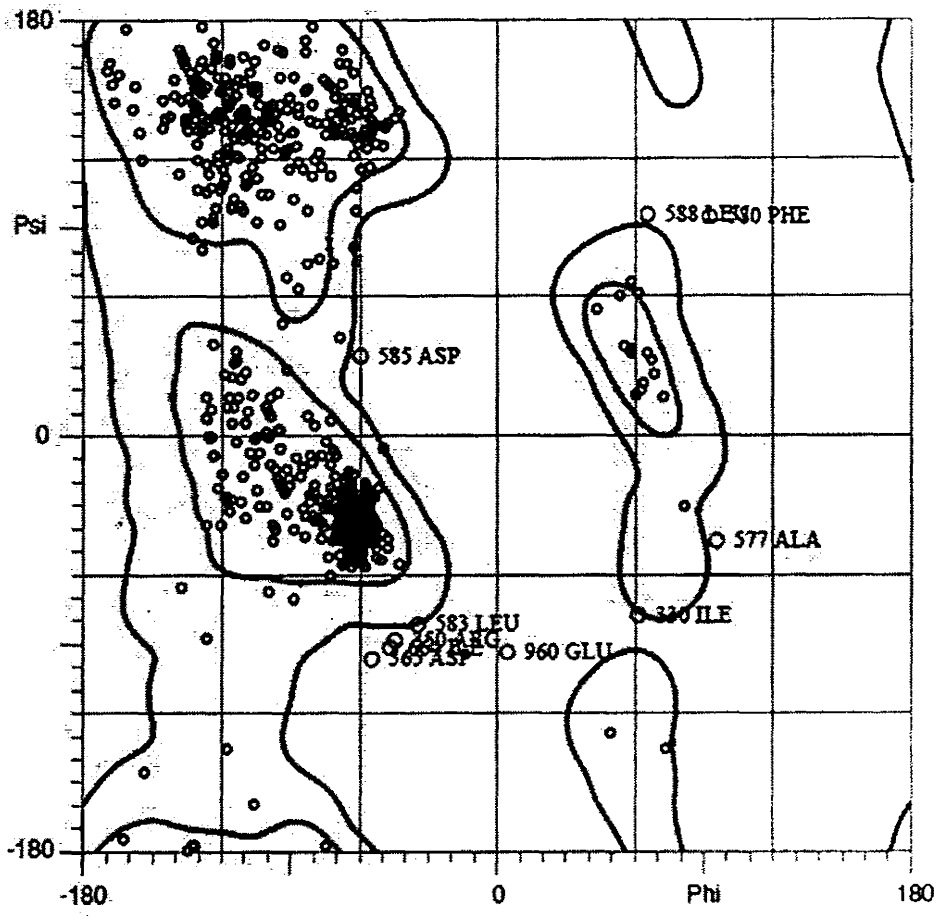


Figure 3.5: MolProbity Ramachandran analysis (general) of Sec24C model showing 12 (phi, psi) outliers i.e 330 ILE, (61.1, -77.6), 350 ARG (-45.6, -88.2), 365 ASP (-55.4, -96.5), 577 ALA (95.7, -45.7), 580 PHE (92.8, 95.7), 583 LEU (-35.9, -81.5), 585 ASP (-60.1, 35.4), 588 LEU (66.0, 96.5), 664 ILE (-47.0, -91.1), 960 GLU (4.1, -93.0), 964 GLU (-111.6, -27.6) and 965 PRO (-4.9, 128.6). (Lovell *et al.*, 2003), http://kinemage.biochem.duke.edu/

SAVES suite was also employed to evaluate our model using several parameters including analysis of the entire residues, calculating the Z score and outliers. These outliers were found to be 144 in number and 4.900 in percent. SAVES analysis further showed that 97.65% of the residues present in our model was found to have an averaged 3D-1D score > 0.2 and this showed our model to be passed.

QMEAN4 global scores were also calculated in order to ensure the quality of our model. QMEAN4 combines different scores based on linear combination of four different statistical terms where estimated reliability of the model is 0 to 1.

Table 3.7 describes various parameters calculated by the SAVES suite where pseudo energies of various terms is given with their respective Z score as compared to the significant scores obtained by other experimental method, the X ray crystallography (Benkert *et al.*, 2011).

Table 3.7: SAVES Suite Result for the Query Protein Sec24C

Scoring function term	**Raw score**	**Z-score**
C_beta interaction energy	-159.89	-1.05
All-atom pairwise energy	-20205.38	-0.35
Solvation energy	-70.25	0.06
Torsion angle energy	-145.79	-1.36
QMEAN4 score	0.676	-1.36

3.9 Docking of the Sec24C

3.9.1 Protein interaction partners

The proteins interacting with our query protein Sec24C were determined by using IntAct and STRING 9.0. The results of both of these databases are given below.

3.9.1.1 IntAct results

20 binary interactions were found in IntAct database for Sec24C. We selected four different interaction partners with highest confidence value i.e. IntAct micro score. Interaction type was also considered to be whether direct interaction or physical interaction. Table 3.8 shows the results of selected interaction partners along with their specific details.

Table 3.8: Interaction Partners Determined by IntAct Database version 4.0.1

Molecule A	Molecule B	Interaction type	Confidence value
Sec24C	Sec23A	Direct interaction	0.68
Sec24C	Sec23B	Physical association	0.37
Sec24C	TMED10	Physical association	0.40
Sec24C	Sec31 and Sec 13	Direct association	0.68

3.9.1.2 STRING 9.0 results

STRING 9.0 was used to find interaction partners for our query protein Sec24C. Its result is a confidence view in which stronger associations are shown with thicker lines. These results were filtered by using Homo sapiens as a target organism. Different results are shown by different nodes showing strong association of Sec24C with Sec13 and Sec23A, as represented in the figure 3.6.

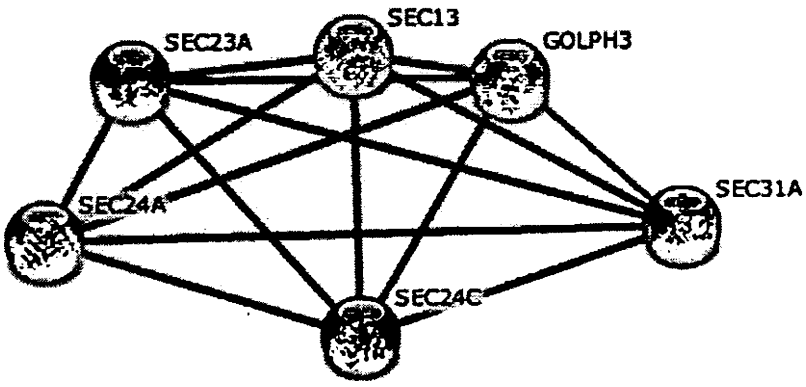


Figure 3.6: Protein interaction partners for Sec24C determined by STRING 9.0.

3.10 Sec24C binding sites

The information about the binding sites residing in the target protein Sec24C was derived and this task was performed using SPPIDER. The results of this analysis were calculated in POLYVIEW-2D (Porollo and Adamczak, 2004) and POLYVIEW-3D (Porollo and Meller, 2007).

POLYVIEW-2D and POLYVIEW-3 both are macromolecular structure visualization tools which are available on the web. These tools offer a wide variety of tools and options to carry out analysis about the structural and functional analysis on the proteins and their complexes. The results obtained by these tools are given in the figures 3.7 and 3.8 respectively.

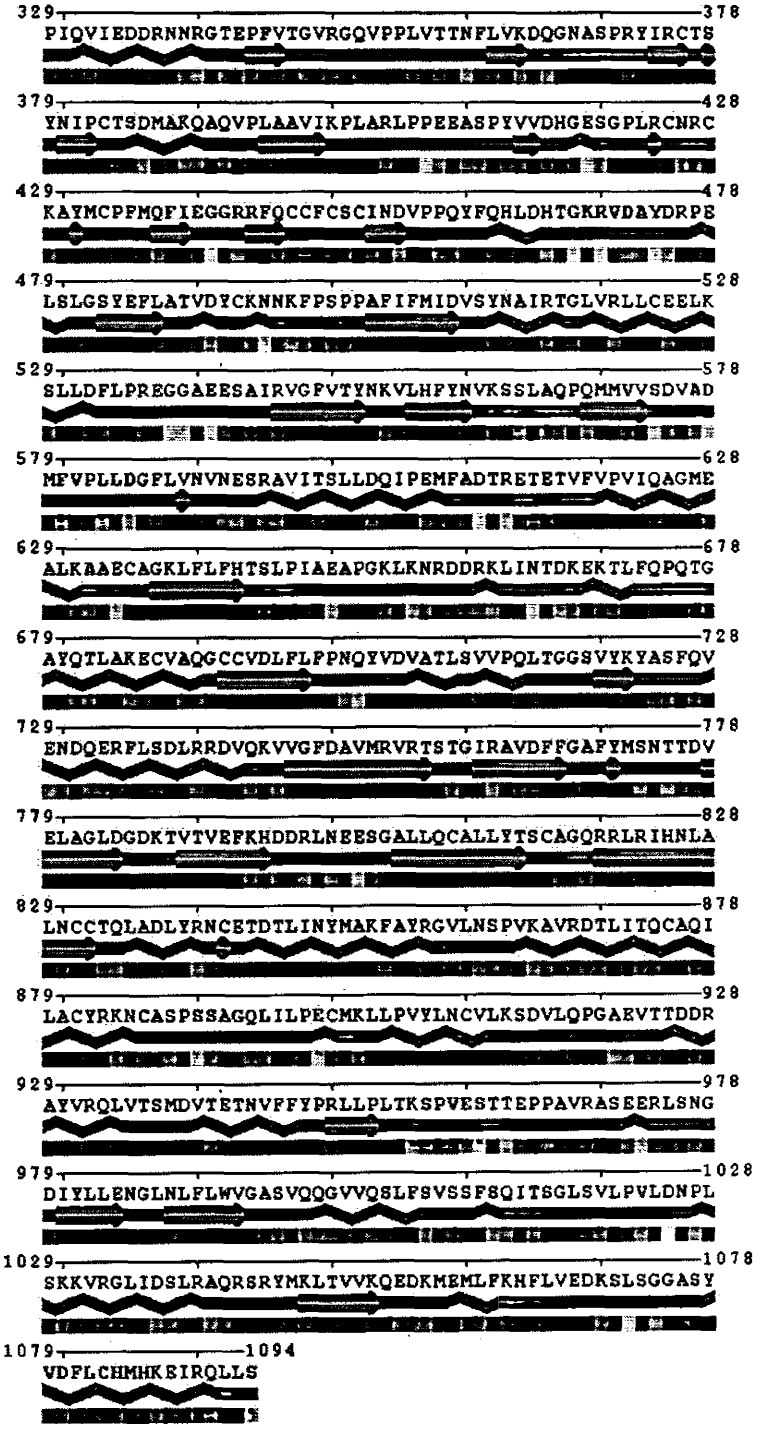


Figure 3.7: POLYVIEW-2D view of protein Sec24C showing the distribution of H - α and other helices, E - β-strand or bridge, C-coil and 0 - completely buried (0-9% RSA) residues. Magenta colored residues are identified as sites of binding to other chains.

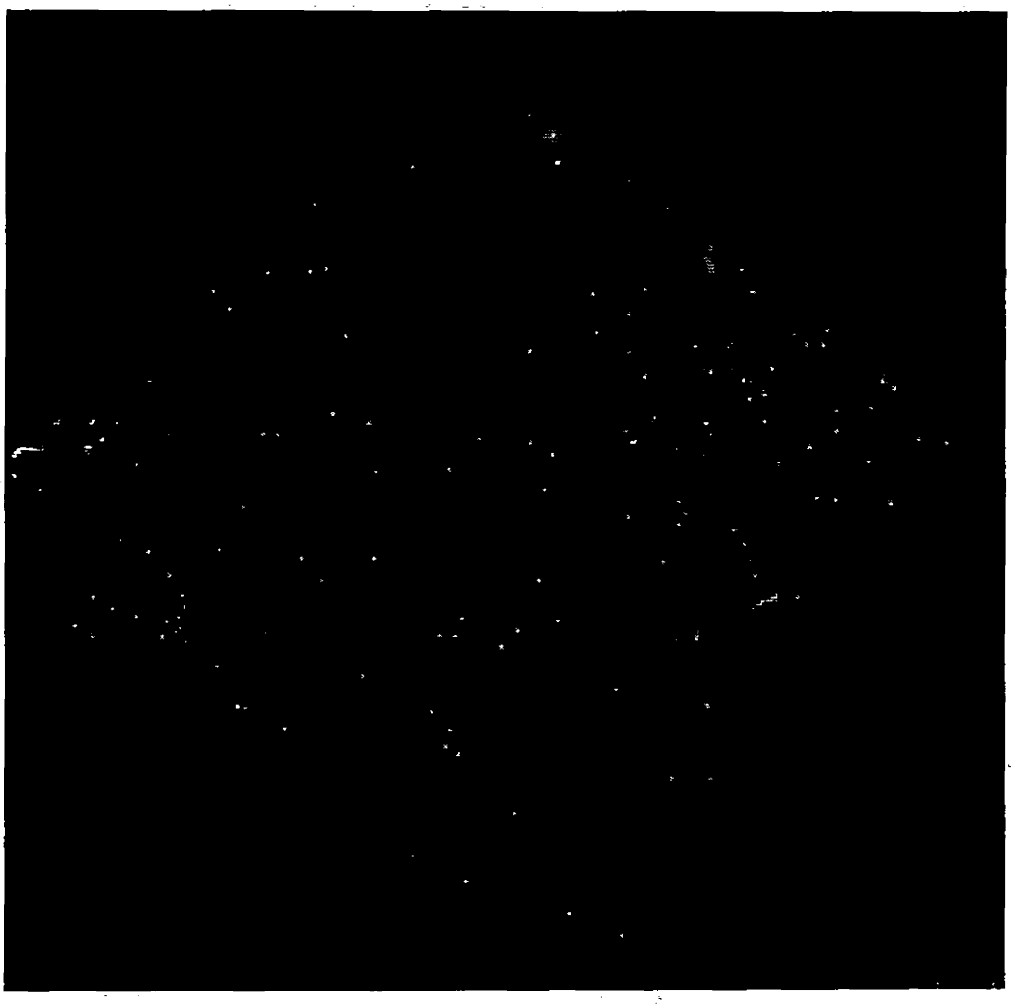


Figure 3.8: POLYVIEW-3D view of the Sec24C protein. Positively inferred residues are shown on the surface in red color, which take part in the complex formation. The residues shown in blue color are not recognized as the interfacial residues. The image was viewed by JMol (Chen and Jim, 2008) java-applet.

3.11 Binding partners for Sec24C

After predicting the binding sites and binding partners for Sec24C, we selected three proteins i.e. 2PM6 (Fath and Mancias, 2007), 2YRC (Nagashima and Hayashi, 2009), and 3KN1 (Wood *et al.*, 2009) to dock with Sec24C.

All of these proteins are involved in forming the COPII (coat protein complex II) vesicle coat (Wendeler *et al.*, 2007). Structures of these proteins were downloaded in PDB format to be used in docking analysis. Table 3.9 contains different characteristics of these proteins.

Table 3.9: Proteins to be Docked with Sec24C, and there Relevant Information.

Binding protein	**Protein function**	**PDB ID**	**Length (amino acids)**
2PM6	Golgi phosphoprotein	3KN1	249
Sec31A	Forms vesicle coat	2PM6	399
Sec23A	Forms vesicle coat	2YRC	59

The first step of docking was to prepare our protein Sec24C for docking analysis. Water molecules present in the structure hinder the docking analysis, so these molecules were removed and hydrogen was added to the protein. AutoDock has to compute the interaction points on the protein before docking and this task was performed by using the grid option offered by AutoDock tools. We setup the grid for our protein in the using different values for X, Y and Z centers, while keeping the spacing (Angstrom) 0.778. The values to make a grid along the each axis on our protein are represented in the table 3.10.

Rigid as well as flexible molecule files were created. Flexible molecule was created by adding arginine and assigning the Gasteiger charges (Gasteiger *et al.*, 2003) to the protein and this file was saved in .gpf format.

The next step was to prepare the binding proteins for docking. The number of torsions was set for each of the protein i.e. 3KN1, 2PM6 and 2YRC with the help of AutoDock tools 4.2 and each file was saved with the extension PDBQT.

Docking of each of these binding proteins was carried out one by one. And both the flexible and rigid protein molecules were selected for each of the docking run.

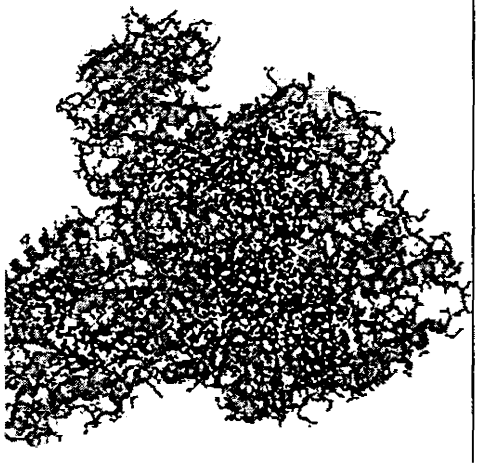
Table 3.10: Values for Different Centers used to make the Grid in the Sec24C Protein by AutoDock 4.2

Centers	Values	Offset
X center	12.1	-6.250
Y center	-10.43	3.361
Z center	-7.958	-2.306

Parameters for GA - genetic algorithm (Eibn *et al.*, 1994) and docking were set to default values for each docking run. The Lamarckian genetic algorithm conformational search was performed using Lamarckian GA (Morris *et al.*, 1998) 4.2 and the file was stored with .dpf extension. The confirmations were found to bury inside hydrophobic portions and tend to make internal hydrogen bonds.

In the prompt window, the process was run to completion and results were visualized with the help of PyMol. The results are given in the figure 3.9 a, b and c.

a



b

c

Figure 3.9: AutoDock 4.2 Results Visualized in PyMol viwer. Receptor protein is shown in green color and binding proteins in orange and blue colors. (a) The complex of Sec24C with Sec31A. (b) Sec24C complex with 2PM6, and (c) Sec24C complex with Sec23A.

DISCUSSION

DISCUSSION

The proteins mediate virtually all biological processes and protein structure prediction using its amino acid sequence is a vast challenge in Bioinformatics. To infer the function and interaction of a protein, its structure is needed and luckily, Bioinformatics has the luck in predicting three dimensional structures of proteins from merely its amino acid sequence using various tools and databases. Since protein structure defines its function so this type of information is of great importance and has vast range of applications like refining NMR structures, supporting the detection of side-directed mutagenesis, molecular replacement in X-ray crystallography, defining antibody epitopes, visual screening and docking of small ligands, drug design, prediction of protein partners by docking of macromolecular structures, signal transduction mechanism and studying catalytic mechanisms. Different tools use various approaches for predicting the protein 3D structures computationally; one of them is homology modeling, which builds a 3D model of a protein using structures of evolutionary related proteins. Docking of these proteins can be carried out if we are familiar with their structures.

The present study is carried out on a human hypothetical protein Sec24C, whose sequence was retrieved from UniProt database. Literature review showed that this protein is involved in the formation of COPII (coat protein complex II) vesicle coat, which mediates the selective export of membrane proteins from the endoplasmic reticulum (ER) to Golgi apparatus. Signal peptide in the proteins is a significant characteristic that decides the fate of the protein. The presence of any signal peptide in our protein was also predicted using SignalP 4.0 server, which reposted absence of any signal peptide in its sequence. Furthermore, ProtParam and SAPS were employed in order to know the nature of our protein, its amino acid composition, check its hydrophobicity and charges present on the amino acids in our sequence. This type of information is valuable because it decides the final residual positions and configurations of each of the residue in the structure of the protein.

Domains and motifs are the main features which are directly associated with the protein structure and give us the insight into its final structure and confirmation. They can evolve

independently of the remaining protein and perform specific functions. Using InterPro, Conserved Domain Database - CDD and Pfam databases, we came to know that our protein has five domains in it, named Zinc finger, beta-sandwich, helical and gelsolin domain. While motifs search using PROSITE and PRINTS posted one motif in Sec24C which was present at the position 980-988. This information was necessary and a key for predicting the model of the protein, as it gave us the hint that our model will have multiple domains in it, so the model can be categorized reliable and similar to the crystal structure of the protein.

Two different softwares were used to predict the 3D structure of our protein i.e. Swiss Model and MODELLER. These both servers follow template based approach for predicting the structure of the protein, which is known as homology modeling. Swiss model is a web based structure prediction server while MODELLER is desktop based server. BLAST was used for searching the templates for our query protein; these templates and query protein sequence are to be aligned with each other prior to build the model. Ramachandran plot was created for the model which showed the psi and phi angles for each of the residue present in our query protein. And this plot confirmed our model to be stable. Our model was found to be reliable as it was also compared with the crystal structure of the protein determined by X-ray crystallography. Our model showed to have five domains in it as it was already predicted using the databases.

Most of the proteins are able to perform their function when they are bound to some other proteins in nature. Our protein also binds with a several other proteins to perform its function *invivo.* To show this binding computationally, it was necessary to have the evidence about the proteins which bind to it and their specific binding positions. These goals were met by implementing the IntAct, STRING and SPPIDER databases. Their results showed a few proteins which can bind to Sec24C, and the position where they are bound. We selected the proteins Sec23A, Sec23B and Sec31 to dock with Sec24C, all of which are involved in forming the COPII (coat protein complex II) vesicle coat.

Docking is the computational modeling of the protein complexes. It shows whether how and, to what position the specific proteins bind to other protein. These protein complexes must be stable so that they can perform there function.

The selected binding proteins were docked with Sec24C using AutoDock 4.2. This step showed the formation of a complex of our query protein with its binding proteins which were making complex together *invivo*. We came to know that these proteins bind to their specific location and at the different domains which were already predicted in the Sec24C.

A little information about these proteins and the function of this complex was known. This study reveals the complete information about Sec24C that what is the nature of the protein, where it is found inside the cell; and its binding partners that how they join together and what is their final configuration in the bound state.

This type of study has great importance as is can answer many important questions. Interaction between the biomolecules (DNA, proteins and small molecules) drive biological systems. Interaction of proteins with each other or ligands is involved in virtually all the critical processes i.e. signaling, development and metabolism. Given the post genomic era the knowledge about complexes and bio molecular complexes is crucial. These tasks have become easier to perform with the ever increasing knowledge about the proteomics, development of the databases containing information about the bio molecules, genomics and advancement in the computational power. There are many other proteins *invivo* which are present in the databases, but the information about their structures, complex formation with other proteins and function is still unknown.

Conclusion and future prospects

In the present study, homology model of the human hypothetical protein Sec24C was created and its docking was carried out with its strong binding partners. *Insilico* study of Sec24C helped in predicting its 3D structure, findings its pockets and binding partners and carrying out the docking procedure. Flood of sequence and structural information, and improvement in analysis tools and databases has greatly aided the field of computational biology, providing us the valuable information that can help us understand the various biological processes and this type of information is valuable in the computer aided drug designing.

Bioinformatics provides us with extensive range of tools, databases and scoring functions that can be employed on proteins in order to find their function, interaction with each other and how they exactly bind with each other. In spite of this pivotal progress, structural ligand designing and flexible protein docking still faces a few major limitations and challenges, which include lack of availability of numerous important receptors and restrictions with rigid body protein docking. So ample computational approaches, including algorithms, and scoring functions still need to be developed and utilized to deal with these problems.

REFERENCES

REFERENCES

Ardala B., Napoleão F., Osmar N., And Richard C., (2007) A Protein Structure, Modelling and Applications, 29, p. 413-23

Ardala B., Napoleão F., Osmar N., And Richard G., (2006) Protein structure, modeling and applications, Bookshelf ID p548462

Arnold K., Bordoli L., Kopp J., And Schwede T., (2006) The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling, Bioinformatics, 22, p. 195-201

Attwood T., Coletta A., Muirhead G., Pavlopoulou A., Philippou P., Popov I., Roma C., Theodosiou A., And Mitchell A., (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource - its status in 2012, Database, 10, p. 1093-96

Benkert P., Biasini M., And Schwede T., (2011) Toward the estimation of the absolute quality of individual protein structure models, Bioinformatics, 27, p. 343-50

Benkert P., Künzli M., And Schwede T., (2009) QMEAN Server for Protein Model Quality Estimation, Nucleic Acids Research, p. 510-4

Benkert P., Schwede T., And Tosatto S., (2009) QMEANclust: Estimation of protein model quality by combining a composite scoring function with structural density information, BMC Struct Biol, 20, p. 9-35

Bermanh M., Westbrook D., Gabanyi J., Tao w., Shah R., Bordoli L., Kopp J., Podvinec M., Carter L., Minor W., Nair R., And Baer J., (2009) The protein structure initiative structural genomics knowledgebase, Nucleic Acids Res, 37, P. 365-8, Biology, cell bio, 6, p. 5-7

Blundell T., Sibanda B., Montalvão R., Brewerton S., Chelliah V., Worth C., Harmer N., Davies O., And Burke D., (2006) *Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery*, Department of Biochemistry, University of Cambridge 80 Tennis Court Road, Cambridge, UK

Bonifacino S., And Glick S (2004) The mechanisms of vesicle budding and fusion, Cell, 116, p. 153-166

Bork P., And Koonin V., (1998) Predicting functions from protein sequences-/where are the bottlenecks, Nat. Genet, 18, p. 313-318

Bowie J., Luthy R., And Eisenberg D., (1991) A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure, Science, 253, p. 164-170

Bowie J., Lüthy R., And Eisenberg D., (1991) A method to identify protein sequences that fold into a known three-dimensional structure, Molecular Biology Institute, University of California, Los Angeles, 356, p. 5-8

Bradley P., Malmstrom L., Qian B., Schonbrun J., Chivian D., Kim E., Meiler J., Misura M., And Baker D., (2005) Free modeling with Rosetta in CASP6, Proteins, 7, p. 128-34

Brendel V., Bucher P., Nourbakhsh I., Blaisdell B., And Karlin S., (1992) Methods and algorithms for statistical analysis of protein sequences, Proc. Natl. Acad. Sci. USA, 89, p. 2002-2006

Brenner E., (1999) Errors in genome annotation, Trends Genet, 15, p. 132-133

Brinda K., Jaroslaw P., And Ron E., (2008) A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins, Proteins, 72, p. 910-928

Camacho J., And Vajda S., (2008) Protein docking along smooth association pathways, Proceedings of the National Academy of Sciences, 98 (19), p. 10636–10641

Cases M., And Mestres J., (2009) Achemogenomic approach to drug discovery: focus on cardiovascular diseases, Drug Discovery, 14, p. 479–485

Cerqueira N., Fernandes A., Eriksson A., And Ramos M., (2009) MADAMM: A multistaged docking with an automated molecular modeling protocol, Proteins: Structure, Function, and Bioinformatics, 74 (1), p. 192–206

Chen C., Zhou X., Tian Y., Zou X.,And Cai P., (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal. Biochem, 357 (1), p. 116–21

Chen J, And Jim X., (2008) Guide to Graphics Software Tools, Springer, ed., p. 471

Daniel S., Bert L., And Groot J., (2010) Ligand docking and binding site analysis with PyMOL and Autodock/Vina, Comput Aided Mol Des, 24, p. 417–422

DeLano W., (2002) The PyMOL molecular graphics system, http://www.pymol.org

Dutta A., Singh S., Ghosh P., Mukherjee R., Mitter S. And Bandyopadhyay D., (2006) In Silico sequence analysis of proteins, Springer, ed. p78

Eiben A., et al (1994) Genetic algorithms with multi-parent recombination PPSN III: Proceedings of the International Conference on Evolutionary Computation, The Third Conference on Parallel Problem Solving from, Nature, p. 78–87

Eswar N., Marti M., Webb B., Madhusudhan M., Eramian D., Shen M., Pieper U., And Sali A., (2006) Comparative Protein Structure Modeling With MODELLER, Current Protocols in Bioinformatics, John Wiley & Sons, Inc., 15, p. 51-56

Fath S., Mancias J., Bi, X., And Goldberg, J (2007)Structure and organization of coat proteins in the COPII cage, Cell(Cambridge,Mass.), 129, p. 1325-1336

Finn J., Mistry J., Tate P., Coggill A., Heger J., Pollington O., Gavin P., Gunesekaran G., Ceric K., Forslund L., Holm E., Sonnhammer S., Eddy A., And Bateman., (2010) The Pfam protein families database, Nucleic Acids Research, Database Issue, 38, D. 211-222

Gabb H., Jackson M., And Sternberg J., (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information, J. Mol. Biol. 272 (1), p. 106–120

Gabb H., Jackson M., And Sternberg M., (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information, Mol. Biol, 272, p.106–120

Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R., And Bairoch A., (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res, 31, p. 37848

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M., Appel R., And Bairoch A., (2005) Protein Identification and Analysis Tools on the ExPASy Server, The Proteomics Protocols Handbook, Humana Press, p. 571-607

Gray J., Moughon S., Wang C., Schueler O., Kuhlman B., Rohl C., And Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, Mol. Biol., 331, p. 281–299

Henikoff S., And Henikoff J., (1994) Protein family classification based on searching a database of blocks, Genomics, 19, p. 97-107

Herbert J., And Bernstein B., (2000) Recent changes to RasMol, recombining the variants, Trends in Biochemical Sciences (TIBS), 25, p. 453-455

Hunter S., Jones P., And Wilkipn L., (2012) InterPro in 2011: New developments in the family and domain prediction database, Nucleic Acids Res. 40 (Database issue), D306-12

Jensen L., Kuhn M., Stark M., Chaffron S., Creevey C., Muller J., Doerks T., Julien P., Roth A., Simonovic M., Bork P., And Mering C., (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, Nucleic Acids Res, 37, p. 412–6

Joseph D., And Jonathan G., (2008) Structural basis of cargo membrane protein discrimination by the human COPII coat machinery, EMBO J, 27, p. 2917-2928

Kerrien S., et al (2012) The IntAct molecular interaction database in 2012, Nucl. Acids Res, doi: 10.1093/nar/gkr1088

Kohlbacher O., And Lenhof H., (2000) Ball - rapid software prototyping in structural and functional analysis, Prototyping, 12, p. 56-57

Koshland D. (1963) Correlation of structure and function in enzyme action, Sci, 142 p. 1533–1541

Kotelchuck D., And Scheraga H., (1969) The Influence of Short-Range Interactions on Protein Conformation, II. A Model for Predicting the α-Helical Regions of Proteins, Proc Natl Acad Sci USA 62 (1), p. 14–21

LaPointe P., And Balch W (2005) Purification and properties of mammalian Sec23/24 from insect cells, Methods Enzymol, 404, p. 66-74

Lensink M., Mendez R., And Wodak S., (2007) Docking and scoring protein complexes: CAPRI 3rd edn. Proteins, 69, p. 704–718

Lesk M., Irving A., And Whisstock C., (2001) Protein structural alignments and functional genomics, Proteins, 42, p. 378–382

Luthy, R., Bowie, J.U., And Eisenberg, D. (1992) Assessment of Protein Models with Three-Dimensional Profiles. Nature, 356, p. 83-85

Macey M., Jenny J., Williams R., Thibodeaux K., Beal M., Almeida J., Cunningham C., Mancia A., Warr W., Burge J., Holland F., And Chapman R., (2003) Modelling interactions of acid-base balance and respiratory status in the toxicity of metal mixtures in the American oyster Crassostrea virginica, Comp Biochem Physiol A Mol Integr Physiol, 155(3), p. 341-9

Magrane M., And the UniProt consortium (2011) Publications on UniProt databases, UniProt Knowledgebase: a hub of integrated protein data, Bioinformatics, 8, p. 11-16

Marchler A., Lu S., Anderson J., Chitsaz F., Derbyshire M., DeWeese C., Fong J., Geer L., Geer R., Gonzales N., Gwadz M., Thanki N., Yamashita R., Zhang D., Zhang N., And Zheng C., (2011) CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Res, 39, p. 225–9

Marti M., Eramian D., Shen D., Devos F., Melo A., And Sali., (2006) A composite score for predicting errors in protein structure models, Protein Science, 15, p. 1653–1666

Martin W., And Derewenda Z., (1999) The name is Bond—H bond, Nature Structural Biology 6(5), p. 403–6

Morris G., Goodsell S., Halliday S., Huey R., Hart W., Belew K., And Olson A., (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, Journal of Computational Chemistry, 19, p. 1639–1662

Morris M., Huey R., Lindstrom W., Sanner F., Belew K., Goodsell S. And Olson J., (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity, J. Computational, Chemistry 2009, 16, p. 2785-91

Muckstein U., Hofacker I., And Stadler P., (2002) Stochastic pairwise alignments, Bioinformatics, 18, p 153 60,

Nagashima T., Hayashi F., And Yokoyama S (2007) Solution structure of the zf-Sec23_Sec24 from human Sec23A

Pagano A., Letourneur F., Garcia-Estefania D., Carpentier J., Orci L., And Paccaud P (1999) Sec24 proteins and sorting at the endoplasmic reticulum, J Biol Chem, 274, p. 7833–7840

Palade G., (1975) Intracellular aspects of the process of protein secretion, Science, 189, p. 347–358

Petersen T., Brunak S., von Heijne G., And Nielsen H., (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions, Nature Methods, 8, p. 785-786

Pettersen F., Goddard T., Huang C., Couch S., Greenblatt D., Meng E., And Ferrin T., (2004) UCSF Chimera--a visualization system for exploratory research and analysis, Computational Chemistry 1605-12

Phuong T., Edgar R., And Batzoglou S., (2006) Multiple alignment of protein sequences with repeats and rearrangements, Nucleic Acids Res, 34(20), p. 5932-42

Ponder W., And Richards F., (1987) Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes, J. Mol. Biol, 193 (4), p. 775–791

Porollo A., Adamczak R., And Meller J., (2004) POLYVIEW: A Flexible Visualization Tool for Structural and Functional Annotations of Proteins, Bioinformatics, 20, p. 2460-2462

Porollo A., And Meller J., (2007) Versatile Annotation and Publication Quality Visualization of Protein Complexes Using POLYVIEW-3D, BMC Bioinformatics, 8, p. 316

Porollo J., And Meller J., (2007) Prediction-based Fingerprints of Protein-Protein Interactions, Proteins: Structure, Function and Bioinformatics, 66, p. 630-45

Preston G., Bianco E., Wu Y., And (2008) Chitinase is stored and secreted from the inner body of microfilariae and has a role in exsheathment in the parasitic nematode Brugia malayi, Mol Biochem Parasitol, 161, p. 55-62

Ramachandran G., Ramakrishnan C., And Sasisekharan V., (1963) Stereochemistry of polypeptide chain configurations, Journal of Molecular Biology, 7, p. 9-11

Rapoport T., (2007) Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes, Nature, 663-9

Rask M., Almén M., And Schiöth H., (2011) Trends in the exploitation of novel drug targets, Nat. Rev. Drug Disc, 8 (10), p. 549–90

Renom A., Madhusudhan M., Eramian D., Shen Y., Pieper U., And Sali, A., (2006) Comparative protein structure modeling using Modeller, Curr Protoc Bioinformatics, Chapter 5, Unit 5-6

Roger S., And James M., (2000) RasMol: Biomolecular graphics for all, Trends in Biochemical Sciences (TIBS), 20, p. 374

Russell R., Aloy P., Alber F., Davis F., Korkin D., Pichaud,M., Topf,M. And Sali A., (2004) A structural perspective on protein–protein interactions, Curr. Opin. Struct Biol, 14, p. 313–324

Sarah H., Philip J., Alex M., Rolf A., Teresa K., Alex B., Thomas B., David B., Peer B., Sara B., Conor M., Huaiyu M., Prudence M., Nicola M., Darren N., Christine O., And Paul D., (2011) InterPro in 2011: new developments in the family and domain prediction database, Nucleic Acids Research, doi: 10.1093/nar/gkr948

Saraste J., And Kuismanen E., (1984) Pre-and post-Golgi vacuoles operate in the transport of Semilki forest virus membrane glycoprotein to the cell surface, Cell, 38, p. 535-549

Schwedt Z., And Georg I., (2002) The Essential Guide to Analytical Chemistry (Brooks Haderlie, trans.), Chichester, NY: Wiley, p. 16-17

Schwieters J., Kuszewski H., And Clore G., (2006) Using Xplor-NIH for NMR molecular structure determination, Progr. NMR Spectroscopy, 48, p. 47-62

Sievers F., Wilm A., Dineen D., Gibson T., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson D., And Higgins D., (2011) Scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology, 539, p. 44-53

Sigrist C., Cerutti L., Castro E., Langendijk P., Bulliard V., Bairoch A., And Hulo N., (2010) PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Res, p. 38161-6

Szilagyi A., Grimm V., Arakaki A., And Skolnick J., (2005) Prediction of physical protein protein interactions, Phys Biol, 2, p. 1-16

Tani K., Oyama Y., Hatsuzawa K., And Tagaya M., (1999) Hypothetical protein KIAA0079 is a mammalian homologue of yeast Sec24p, FEBS Lett, 447, p. 247-250

Uetz P., Schwikowski B., And Fields S., (2000) A network of protein-protein interactions in yeast, Nature Biotechnology, 18, p. 1257–61

Vakser I., And Kundrotas P., (2008) Predicting 3D structures of protein-protein complexes. Curr. Pharm. Biotech, 9, p. 57–66

Vriend G., (1990) WHAT IF: a molecular modeling and drug design program, J Mol Graph 8, p. 6-29

Wendeler M., Paccaud J., And Hauri H., (2007) Role of Sec24 isoforms in selective export of membrane proteins from the endoplasmic reticulum, EMBO Rep, 8, p. 256-64

Wodak J., Crombrugghe M., And Janin J., (1987) Computer Studies of Interactions between Macromolecules, Progress in Biophysics and Molecular Biology, 49, p. 29–63

Wood C., Schmitz R., Bessman J., Setty G., Ferguson M., And Burd G., (2009) PtdIns4P recognition by Vps74/GOLPH3 links PtdIns kinase signaling to retrograde Golgi traffic king, J.Cell Biol, 187, p. 967-975

Xie W., Huang X., And Gong Z., (2000) Characteristics and antifungal activity of a chitin binding protein from Ginkgo biloba, FEBS Lett, 478, p. 123-6

Zhang C, Liu S, Zhu Q, Zhou Y., (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. Journal of Medicinal chemistry 7 (48), p. 2325–2335

Zheng N., And Gierasch L., (1996) Signal sequences: the same yet different, Cell, 86, p. 849-852