CHAPTER 2

MATERIALS AND METHODS

2.1 PPI PREDICTION ALGORITHM

2.1.1 Data Source for Training Sets

High throughput experimental PPI datasets for three organisms, 
*Saccharomyces cerevisiae* (Uetz et al 2000; Ito et al 2001), *Homo sapiens* (Rual et al 2005) and *Caenorhabditis elegans* (Li et al 2004; Simonis et al 2009) were obtained from published literature. These constitute the five training sets. In the *Homo sapiens* (Rual et al 2005) PPI dataset, ~2800 interactions were detected from the pairwise interactions of ~8100 available Gateway-cloned open reading frames. In *Saccharomyces cerevisiae* (Ito et al 2001), in a comprehensive two-Hybrid screening almost all of the yeast ORFs were cloned individually and as a result 13,754 interaction sequence tags (ISTs) were obtained from the 15,523 clones. The redundant and low-quality ISTs were removed and 4549 two-hybrid interactions were identified from 3,278 proteins. In *Saccharomyces cerevisiae* (Uetz et al 2000) the protein-protein interactions of ORFs predicted from the *Saccharomyces cerevisiae* genome sequence were identified using large-scale yeast two-hybrid screens. Finally 692 interacting protein pairs were identified from 817 yeast ORFs. The *Caenorhabditis elegans* (Simonis et al 2009) protein-protein interactome map was derived by testing a matrix of ~10,000 * ~10,000 proteins using high-throughput yeast two-hybrid system. The WI-2007 dataset consists of 1,816 high-confidence binary, protein-protein interactions from 1496 proteins.
In the *Caenorhabditis elegans* (Li et al 2004) PPI dataset, 1873 baits were used against 14266 preys. Since we require only "all against all" datasets we extracted a subset of the *Caenorhabditis elegans* PPI result set (Li et al 2004) wherein both the interactors in a PPI corresponded to IDs from the bait set (i.e. prey set IDs overlapped with the bait set IDs). This way we ensured that it is an "all against all" dataset.

### 2.1.2 PIR ID Mapping Tool

PIR ID mapping is a web based ID mapping tool available at the PIR web site (http://pir.georgetown.edu/pirwww/search/idmapping.shtml) (Huang et al 2011). This site provides batch mappings and allows transitive mappings between all IDs bridged by UniProtKB. The gene names (in the case of *Saccharomyces cerevisiae* and *Caenorhabditis elegans*) and gene ids (in the case of *Homo sapiens*) were mapped to their corresponding UniProt IDs (UniProt Release 2011-08) (Dimmer et al 2011) using PIR ID mapping tool (PIR release: 2011_08) (Barker et al 2000). PfamA domain information of each protein was retrieved from Pfam database (Finn et al 2010) using PIR ID mapping tool.

Protein-protein interactions are mediated through domain-domain associations. We encode this data in the form of $I$, $P$ and $T$.

### 2.1.3 Construction of Matrices $I$, $P$ and $\bar{I}$

For each of the five training sets the binary matrices, $I$ and $P$ which represent the PPI and PfamA domain composition (of each protein in the training set) respectively were constructed. The $I$ matrix for each training set is a square matrix of dimension $N$ where $N$ is the number of distinct proteins in the training set. In this matrix interactions are represented as 1 and non interactions 0. The $P$ matrix for each training set is of dimensions $N \times M$.
where N is the number of distinct proteins in the set and M the number of
distinct domains. A matrix element in P has a value 1 if the protein contains
that domain and 0 otherwise. A complement of I matrix, \( \bar{I} \) is generated where
interactions are scored 0 and non-interactions are scored 1. All the matrices
were constructed using PERL scripts.

2.1.4 Pruning the Data

In our training set, some proteins have identical domains and
identical PPI. Since these are redundant we chose to retain one representative
from each such set. All instances of the other proteins in each of these sets
were removed from both I and P matrices. We call these pruned I and P
matrices as the original training sets \( I_{\text{org}} \) and \( P_{\text{org}} \). Some of the proteins in the
training set have identical domains but non-identical PPI. The difference in
the interactions may be due to many possible reasons. For the sake of
convenience, in this study, we added an “extra domain” for each of such
proteins in the set. Similarly there are some other proteins in our dataset
without any Pfam domain. For these proteins too we added an “extra
domain”. We call these matrices as the extended training sets \( I_{\text{ext}} \) and \( P_{\text{ext}} \).
Extra domains are added by increasing the dimensions of P matrix to N X (M
+ number of extra domains). The presence of an extra domain for a particular
protein is encoded as 1 for that protein and 0 for other proteins. All the
matrices were constructed using PERL scripts.

2.1.5 MATLAB

MATLAB is an interactive system whose basic data element is an
array, which allows to solve many technical computing problems with matrix
and vector formulations. The MATLAB mathematical function library is a
collection of computational algorithms with varying functions like simple
sum, sine, cosine, and complex arithmetic, to more sophisticated functions
like matrix inverse, matrix eigenvalues etc. In our algorithm matrix
manipulations were implemented using MATLAB.

2.1.6 Test Sets

For validating our algorithm we used six positive, one negative test
set and a random test set which included both positive and negative entries.

2.1.6.1 Positive test sets

For positive PPI test sets we extracted PPI information confirmed by
multiple experiments from the IntAct database (Aranda et al 2010) and DIP
database (Salwinski et al 2004). *Saccharomyces cerevisiae* test set was
obtained from literature (Huang et al 2007). Duplicates between these three
sets were removed. PPI which were part of our training datasets were also
eliminated from the test sets. The domain association sets were extracted from
an iPfam dataset obtained from literature (Zhao et al 2010) and 3did database
(Stein et al 2011).

2.1.6.2 Negative test set

The negative dataset was a random dataset obtained from literature
wherein they constructed a negative set by selecting an equal sized random set
of non-interacting pairs from a set of 4233 interacting proteins from yeast
retrieved from BIND database (Ben-Hur & Noble 2005). We further filtered
this set to remove instances which are part of our training set and also those
for which we do not have even partial domain information. Our final negative
set had 2745 protein pairs.
2.1.6.3 Random test set

The random test set was generated by 1500 entries randomly from all the six positive test sets and 1500 entries from the negative test set.

2.1.6.4 DIP Species specific dataset

The species specific subsets of DIP database contain all the interactions in which proteins from a particular species participate (Salwinski et al 2004). In our study we used three model organisms *Mus musculus, Drosophila melanogaster* and *Rattus norvegicus*. The PPI dataset of these three organisms were downloaded from DIP database. PIR ID mapping tool was used for mapping Uniprot IDs to their corresponding PfamA domains. We eliminated the proteins which do not have even a single domain. The final dataset for the three organisms *Mus musculus, Drosophila melanogaster* and *Rattus norvegicus* comprised of 1,329, 265 and 272 interactions respectively.

2.1.7 R package

R is a software which provides a wide variety of statistical and graphical techniques. We used R package to generate the ROC curve.

2.1.8 Cytoscape

Cytoscape is a general purpose modelling environment for integrating biomolecular interaction networks from high-throughput expression data which includes protein-protein, protein-nucleic acid and other genetic interactions (Shannon et al 2003). The protein-protein interaction networks and DDA networks were generated using cytoscape.
2.2 DOCKING STUDIES FOR PREDICTED DDAs

DDAs were obtained from the positive DDAs generated using our PPI prediction algorithm. As case studies, we chose three representative DDAs and performed docking studies using FTDock. The UniprotKB IDs and PDB IDs for the corresponding Pfam Id were obtained from UniprotKB. The 3D coordinate files of the proteins were downloaded from PDB database. Homology models were built for proteins which do not have structure using Swiss-Model server (Arnold et al 2006).

2.2.1 FTDock

It is a rigid-body docking used to predict the correct binding geometry of two biomolecules. Shape Complementarity algorithm is used for scoring the docked complexes and further they were rescored using RP score (Residue Level Pair Potential Scoring) (Gabb et al 1997).

2.2.2 Swissmodel

Swiss-model is a fully automated server used for protein homology modelling. The proteins which do not have 3D structure was modelled using swissmodel (Arnold et al 2006). The model that was generated was validated using Procheck from SAVES server. Procheck checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry.

2.2.3 WHAT IF Server

WHAT IF is a versatile molecular modelling package for macromolecular analysis, visualization, modelling and structure validation (Hekkelman et al 2010). WHAT IF server provides many options which includes computation of interatomic contacts, solvent accessibility, hydrogen
bonds and many others. The interatomic contacts of the complex generated using FTDock was found using WHAT IT server.

2.3 MOLECULAR DYNAMICS SIMULATIONS

The cleaved (PDB ID: 7API), uncleaved (PDB ID: 1QLP) (Berman et al 2000) human alpha 1 antitrypsin and the best complex (AAT-caspase-3) obtained from PPI studies using FTDock was used for MD simulation studies. GROMACS 4.5.5 with G43a1 force field was used to perform all atom Molecular Dynamics Simulation studies (Van Der Spoel et al 2005). A 3 nm cubic box size for cleaved and uncleaved AAT and 3.5 nm cubic simulation box for AAT-caspase-3 complex was made and the protein was solvated with the SPC water model. Counter ions (Sodium ions) were added to the system to neutralize the protein charges. In the case of cleaved AAT 3 chloride ions (Cl-) and for uncleaved AAT and AAT-caspase-3 complex ten and seven sodium ions (Na+) respectively were added. Steepest descents algorithm was used for energy minimization. The system was weakly coupled to an external bath using Berendsen’s method. Simulations were performed for 100 ps under NVT conditions with a constant temperature of 300 K. In a similar fashion, the simulation proceeded for another 100 ps under NPT conditions so that the system reached equilibrium at constant pressure. Then, the full all atom molecular Dynamics Simulation was performed at a constant temperature of 300 K for 10ns (cleaved AAT), 45ns (uncleaved AAT), 50ns (AAT- caspase-3 complex) respectively. The Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuations (RMSF) were computed using GROMACS inbuilt tools. Grace, PyMol and chimera were used for visualization of trajectories.

2.3.1 Protein Protein Interaction studies of Caspase-3 and AAT

The structure of caspase-3 and human AAT determined by X-ray Diffraction was downloaded from PDB database with PDB IDS 1NME and 1QLP respectively (Berman et al 2000). The protein-protein interaction
studies were performed using FT dock (Fourier Transform Dock) (Gabb et al 1997).

2.3.2 Tools used for Structural Analysis and Visualization

WHAT IF server was used for the analysis of interatomic contacts, solvent accessibility and DSSP for secondary structure prediction. Pymol, a molecular visualization system was used for the generation of salt bridges. Chimera, VMD and DS accelyryys Visualizer were used for visualizing the structures and protein-protein complexes.

2.4 PROTEIN DOMAIN COMBINATIONS

2.4.1 UniprotKB

The UniprotKB Ids were downloaded from the Uniprot database (Release 2013_04 of 03-Apr-13). The total number of proteins in the database is 539829. We downloaded the taxonomic classification information, GO ids, etc. from Uniprot database. We also clustered all the proteins using CD-hit algorithm to get a non-redundant dataset. The cutoff used to remove redundancy is 90%. After clustering, the total number of proteins 539829 was reduced to 311159. We used two datasets: one was the original set and the other the clustered dataset. In the original set, out of the processed 539829 proteins, 511817 have atleast one domain. In the clustered set out of the processed 311159 proteins, 21815 proteins did not match any protein domain. These could be the domains obtained from the Uniprot / TrEMBL database, that form a part of the PfamB database.

2.2.2 CD-hit

CD hit is the Cluster Database at High Identity with Tolerance which generates non-redundant dataset for the given fasta sequences (Li and
Godzik 2006). We used Cd-hit to cluster the UniprotKB protein sequences to generate a non-redundant set. The cut-off used for redundancy is 90%.

2.2.3 Pfam

Pfam is a comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation. Each family in Pfam is represented by two multiple sequence alignments and two profile-Hidden Markov Models (profile-HMMs). An area of significant difference in domain definitions between Pfam and the structural databases is due to discontinuous domains defined in SCOP and CATH (Finn et al 2010). The Pfam database was used to retrieve the domain combinations of the proteins. We mapped the Uniprot ids with the Pfam domain combination and retrieved the domain combinations of the corresponding proteins.

2.2.4 Perl Scripts

In house Perl scripts were written to order the domains according to their sequential occurrence in the protein and to find unique domains and domains that occurred in certain combinations with each other. We separated our data into eleven sets such as one domain proteins, two domain proteins etc. up to ten domain proteins. All the proteins with more than 10 domains were placed into one category.