Chapter 3

Materials and methods

3.1. DATASETS

The data on human OCA 1-4 genes were collected from OMIM (Amberger et al., 2009) and Entrez gene on National Center for Biotechnology Information (NCBI) Website. The SNP information of OCA I-IV genes were retrieved from dbSNP (http://www.ncbi.nlm.nih.gov/snp/) (Sherry et al., 2001) and Swissprot databases (Yip et al., 2004; 2008; Boeckmann et al., 2003). The amino acid sequence of TYR, P, TYRP1 and MATP proteins were retrieved from the Uniprot database. The strategy of our investigation is depicted in Fig. 3.1.

Fig. 3.1. The strategy of our investigation
3.2. DISEASE RELATED SNP PREDICTION

3.2.1. SIFT

SIFT (http://sift.jcvi.org/) prediction is based on the sequence homology and the physicochemical properties of amino acids which are dictated by the substituted amino acid. SIFT score is less than equal to 0.05 indicates the amino acid substitution is intolerant or deleterious, whereas the score greater than equal to 0.05 predicted as tolerant (Kumar et al., 2009).

3.2.2. PolyPhen 2.0

PolyPhen 2.0 (http://genetics.bwh.harvard.edu/pph2/) is based on combination of sequence and structure based attributes and uses naive Bayesian classifier for the identification of an amino acid substitution and the impact of mutation. PolyPhen 2.0 identifies homologues of the input sequences through BLAST and calculates the position-specific independent count (PSIC) scores for every variant and evaluates the difference between the variant PSIC scores. The output levels of probably damaging and possibly damaging were classified as functionally significant (greater than equal to 0.5) and the benign level being classified as tolerated (less than equal to 0.5) (Adzhubei et al., 2010).

3.2.3. I-Mutant 3.0

I-Mutant 3.0 (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi) is a support vector machine (SVM) based tool. We used the sequence based version of I-Mutant 3.0 that classifies the prediction into three classes: neutral mutation (−0.5 ≤ DDG ≥ 0.5 kcal/mol), large decrease (≤−0.5 kcal/mol), and a large increase (>0.5 kcal/mol). The free energy change (DDG) predicted by I-Mutant 3.0 is based on the difference between unfolding Gibbs free energy change of mutant and native protein (kcal/mol) (Capriotti et al., 2008).

3.2.4. PANTHER

PANTHER (http://www.pantherdb.org/) program is a protein family and subfamily database which predicts the frequency of occurrence of amino acid at a
particular position in evolutionary related protein sequences. PANTHER is a HMM-based statistical method, and multiple sequence alignment of protein sequence to execute evolutionary analysis of each nsSNP. subPSEC score of PANTHER differ from 0 (neutral) to -10 (more likely to be deleterious). The threshold subPSEC score of -3 has been assigned below which the predictions are considered as deleterious (Thomas et al., 2003).

3.2.5. PhD-SNP

PhD-SNP (http://snps.biofold.org/phd-snp/) is SVM based classifier, trained over the million amino acid polymorphism datasets using supervised training algorithm (Capriotti et al., 2006). The protein sequence and position of native and mutant amino acid residue given as a input to the server. It predicts whether the given amino acid substitution leads to disease associated or neutral along with the reliability index score (Capriotti et al., 2006).

3.2.6. SNP&GO

SNP&GO (http://snps-and-go.biocomp.unibo.it/snps-and-go/) is a Support Vector Machines (SVMs) based method. SNP&GO predicts disease causing mutations from protein sequence, evolutionary information, and functions as encoded in the gene ontology terms (Calabrese et al., 2009). It predicts the given mutation leads to disease associated or neutral with prediction reliability index score. The prediction reliability index score vary from 0 to 10. The probability score greater than 0.5 are predicted to be disease associated.

3.2.7. PMUT

PMUT (http://mmb2.pcb.ub.es:8080/PMut/) is a neural network based program which is trained on large database of neutral and pathological mutations (Ferrer-Costa et al., 2005). PMUT uses 3 parameters including mutation descriptors, solvent accessibility, and residue and sequence properties to calculate the pathogenicity indexes of given input mutation data ranging from 0 to 1. The mutations with index score greater than 0.5 are predicted to be pathologically significant (Ferrer-Costa et al., 2005).
3.2.8. MutPred

MutPred (http://mutpred.mutdb.org/) is a web based tool, used to predict the molecular changes associated with amino acid variants (Li et al., 2009). It uses SIFT, PSI-BLAST, and Pfam profiles along with some structural disorder prediction algorithms, including TMHMM, MARCOIL, I-Mutant 2.0, B-factor prediction, and DisProt (Li et al., 2009). Functional analysis includes the prediction of DNA-binding site, catalytic domains, calmodulin binding targets, and posttranslational modification sites (Li et al., 2009). Combining the scores of all four servers, the accuracy of prediction rises to a greater extent and finally we filtered the most disease-associated mutation.

3.3. HOMOLOGY MODELLING

The 3-D structure of the OCA (TYR, P, TYRP1 and MATP) proteins structures was not available in PDB database. Hence, we applied threading based approach servers LOMETS and I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) to model the 3-D structure of OCA proteins. LOMETS generated multiple models of which those generated by HHSEARCH and SP3 algorithms. I-TASSER server works by combining the folds and secondary structure by profile-profile alignment threading techniques for non-aligned regions. For the submitted sequences, five 3D models were obtained and the best model was selected based on the lowest energy. Further the native structure was mutated with the most deleterious substitution predicted in this study. In order to build the mutant structures, we made a point mutation in OCA proteins their respective positions using SPDB viewer (Kaplan and Littlejohn, 2001). The native and mutant structures were energetically optimized by GROMACS package 4.5.3 (Hess et al., 2008). During energy minimization both native and mutant structures were solvated in a cubic box with simple point charge (SPC) water molecules at 10Å marginal radius. Initially the solvent molecules were relaxed while all the solute atoms were harmonically restrained to their original positions with a force constant of 100 kcal/mol for 5000 steps. After this, whole molecular system was subjected to energy minimization for 5000 iterations by steepest descent algorithm implementing all atom OPLS force field. The quality of model
structures was verified using the PROCHECK (http://nihserver.mbi.ucla.edu/SAVES/) (Laskowski et al., 1993) and PROSA (https://prosa.services.came.sbg.ac.at/prosa.php) (Wiederstein and Sippl, 2007) programs.

3.4. QUANTITATIVE PROTEIN STRUCTURE ANALYSIS

3.4.1. VADAR

VADAR (Volume Area Dihedral Angle Reporter) is a comprehensive web server for quantitative protein structure evaluation. It accepts 3D coordinates of protein as input and calculates, key structural parameters both for individual residues and for the entire protein. These derived parameters can be used to rapidly identify both general and residue-specific problems within newly determined protein structures (Willard et al., 2003). The VADAR web server is accessible at http://redpoll.pharmacy.ualberta.ca/vadar. 3D coordinates of native and mutant TYR structures were given as input to the server for quantitative analysis.

3.4.2. RESIDUE DEPTH COMPUTATION AND SALT BRIDGES ANALYSIS

Residue depth has been shown earlier to correlate well with hydrogen exchange rates (Chakravarty and Varadarajan, 1999; Pedersen et al., 1991) and thermal stability under mutagenesis (Chakravarty and Varadarajan, 1999). Depth can hence be a useful measure in the prediction of protein–protein interaction hot spots (Chakravarty and Varadarajan, 1999), detecting sites for post-translational modification (Pintar and Pongor, 2005; Pintar et al., 2003a) and predicting the effect of point mutations on protein stability and function (Pintar et al., 2003a; 2003b). Native and mutant structure coordinates were submitted to the DEPTH server (http://mspc.bii.a-star.edu.sg/depth) (Tan et al., 2011) with default parameters to identify depth and surface residues. Salt bridges can play important roles in protein structure and function and have stabilizing and destabilizing effects in protein folding. We observed the number of salt bridges made by core and surface residues. Salt bridges analysis was performed by ESBRI web server (http://
bioinformatica.isa.cnr.it/ESBRI/) (Constantini et al., 2008) by submitting 3D coordinates of modelled native and mutant TYR proteins.

3.4.3. NORMAL MODE ANALYSIS

Normal mode analysis (NMA) is a powerful tool for predicting the possible movements of a given macromolecule. It has been shown recently that half of the known protein movements can be modelled by using two low-frequency normal modes. NMA provides an alternative to molecular dynamics for the study of motions of macromolecules. A quantitative measure of the atomic motions in proteins can be obtained from the mean square fluctuations of the atoms relative to their average positions. Understanding structural dynamics of proteins is essential for gaining greater insights into their biological functions (Purohit and Sethumadhavan, 2009; Purohit et al., 2008; 2011a; 2011b; Rajendran et al., 2012). Since protein flexibility is important for protein function and for rational drug design. Therefore flexibility of certain amino acids in protein is useful for various types of interactions which can be analyzed by B factor which are computed from the mean square displacement. We used WEBnm (http://www.bioinfo.no/tools/ normalmodes) (Hollup et al., 2005) to calculate the slowest modes and Elnemo (http://igs-server.cnrs-mrs.fr/elnemo/index.html) (Suhre and Sanejouand, 2004) to calculate normal mode analysis by observing mean square fluctuations c-alpha carbon atoms. 3D coordinates of modelled native and mutants TYR structures were given as input to the server.

3.5. MOLECULAR DYNAMICS SIMULATION

3.5.1. ALL ATOM MOLECULAR DYNAMICS SIMULATION

Molecular dynamics simulation was performed by using GROMACS 4.5.3 package (Hess et al., 2008) running on a single Intel Core2Duo machine with 3 GB RAM and running Ubuntu 11.10 Linux package. Structure of native and mutant OCA (TYR, P and TYRP1) proteins was used as starting point for MD simulations. Systems were solvated in a cubic box with simple point charge (SPC) water molecules at 10 Å marginal radius. At physiological pH the structures were found to be either positively or
negatively charged; thus in order to make the simulation system electrically neutral, we added sodium (Na+) and Cholride (Cl-) ions to the simulation box using the “genion” tool that accompanies with GROMACS package. Initially the solvent molecules were relaxed while all the solute atoms were harmonically restrained to their original positions with a force constant of 100 kcal/moL for 5000 steps. After this, whole molecular system was subjected to energy minimization for 5000 iterations by steepest descent algorithm implementing GROMOS96 43a1 force field. Berendsen temperature coupling method (Berendsen et al., 1984) was used to regulate the temperature inside the box. Electrostatic interactions were computed using the Particle Mesh Ewald method (Cheatham et al., 1995). The ionization states of the residues were set appropriate to pH 7 with all histidines assumed neutral. The pressure was maintained at 1 atm with the allowed compressibility range of 4.5e⁻⁵ atm. SHAKE algorithm was used to constrain bond lengths involving hydrogen, permitting a time step of 2 fs. Van der Waals and coulomb interactions were truncated at 1.0 nm. The non-bonded pair list was updated every 10 steps and conformations were stored every 0.5 ps. Position restraint simulation for 500 ps was implemented to allow solvent molecules to enter the cavity region of structure. Finally, systems were subjected to MD simulation. The general workflow of molecular dynamics simulation was showed in Fig. 3.2.

3.5.2. MEMBRANE SIMULATION

Molecular dynamics (MD) simulations were performed on MATP native and mutant (Y317C) model structures in lipid and water environment using the GROMACS 4.5.3 package (Hess et al., 2008) running on a Intel Core i7 Dell machine with 8 GB RAM under Ubuntu 12.04 operating system. The lipid bilayer consisted of a pre-equilibrated layer of 288 molecules of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC), generously gifted by X. Periole. All the Berger lipid parameters including those for POPC were obtained from P. Tieleman’s site at http://moose.bio.ucalgary.ca/Downloads. The modified GROMOS 43a1 force field with lipid parameters was applied on the system. The MATP protein was inserted carefully into the lipid bilayer using the InflateGro program (Kandt et al., 2007). The entire lipid
bilayer was inflated and then slowly compressed around the protein until an area per lipid value of 69 Å\(^2\) was reached, which is just above the experimental value of 65 Å\(^2\) for pure POPC. Each compression step was followed by a round of steepest descent energy minimization to relax the lipid molecules, keeping the protein restrained. The entire system (native & mutant) was then solvated with a single-point charge (spc) water model and neutralized with Cl\(^-\) ions. After this the systems were subjected to energy minimization for 1000 steps by steepest descent. Further, the minimized systems were then subjected to MD simulations in two steps (Van Gunsteren et al., 1996). Initially we performed under an NVT ensemble (constant number of particles, volume, and temperature) for 1000 ps, followed by an NPT ensemble (constant number of particles, volume, and temperature) for 1000 ps each at 300 K with positions restrained for the entire system, except the water molecules, in order to make sure a balance of the solvent molecules around the residues of the protein. The temperature was kept constant at 300 K with a Berendsen thermostat throughout the process of simulations. LINCS constraint algorithm (Hess, 2008) was used to maintain the geometry of the molecules. Long-range electrostatic interactions were calculated using the particle-mesh Ewald (PME) method. Van der Waal’s interactions and Coulomb interactions were cut off at 12 Å with updates every five steps. Finally a well-equilibrated system was then subjected to MD simulations for 50 ns each at 300 K without any constraint. The steps of membrane simulation were showed in Fig. 3.3.
Fig. 3.2. Workflow of molecular dynamics simulation approach
Fig 3.3. Flowchart of membrane simulation
3.5.3. ESSENTIAL DYNAMICS (ED)

ED is a widely applied technique based on the principal component analysis of conformational ensembles that allows the identification of the most relevant or correlated motions of groups of residues of a protein along a trajectory generated by MDS (Amadei et al., 1993; Stepanova, 2007). The covariance matrix of C-α atoms was constructed and then diagonalized by using g_covar and g_anaeig within the GROMACS package (Hess et al., 2008; Van Der Spoel et al., 2005) yielding a set of eigenvectors and their respective eigenvalues. Each eigenvector represents one single direction of collective motion, whereas the corresponding eigenvalue represents the amplitude of motion along that vector. The deviations of the MATP protein C-α atom from their time-averaged positions were obtained as averages over the eigenvectors representing 80% of the total mobility of the system (i.e. the total sum of eigenvalues for the system).

3.5.4. ROOT MEAN SQUARE CUT-OFF OR DEVIATION (RMSD) CLUSTERING

In order to select a reduced set of representative models of the native and mutant MATP protein, RMSD conformational clustering was performed using the gromos method (Daura et al., 1999) implemented in GROMACS (g_cluster). In the gromos clustering algorithm, the conformation with the highest number of neighbors, identified within the chosen RMSD cutoff, is chosen as the center of the first cluster. All the neighbors of this conformation are removed from the ensemble of conformations. The center of the second cluster is then determined in the same way, and the procedure is repeated until each structure has been assigned to a cluster (Daura et al., 1999).

3.5.5. CROSS-CORRELATION ANALYSIS

The dynamic cross-correlation matrix (DCCM) \( C_{ij} \) that reflects the fluctuations of the coordinates of Cα atoms was calculated to analyze the collective motions of native and mutant MATP. The cross-correlation coefficient \( C_{ij} \), between atoms i and j, is a measure of the correlated nature of their atomic fluctuations and is computed as follows: 

\[
C_{ij} = \frac{(\Delta r_i \times \Delta r_j)}{(\Delta r_i^2 \times \Delta r_j^2)^{1/2}}
\]

where \( \Delta r_i \) and \( \Delta r_j \) correspond to the atomic displacement vectors for atoms i and j, respectively, and the angle brackets
indicate time averages (Laberge and Yonetani, 2008). The elements $C(i,j)$ can be collected in matrix form and displayed as a 3D dynamical cross-correlation map (Swaminathan et al., 1991).

RMSD, RMSF, Rg and SASA plot analysis were carried out by using $g_{\text{rms}}$, $g_{\text{rmsf}}$, $g_{\text{gyrate}}$ and $g_{\text{sas}}$ tool, respectively. Number of distinct hydrogen bonds formed by specific residues to other amino acids within the protein during the simulation (NH bond) was calculated using $g_{\text{hbond}}$. NH bond determined on the basis of donor-acceptor distance smaller than 0.35 nm and of donor-hydrogen-acceptor. All the graphs were plotted using XMGRACE (Turner, 2005) program. The trajectories were visualized using VMD (Humphrey et al., 1996) and all the images were rendered using both VMD (Humphrey et al., 1996) and PyMol (DeLano, 2003).