CHAPTER III

METHODOLOGY

Computational resources

In this study all the computational analysis were performed on a Xeon(R), 2.13 GHz server equipped with the windows server 2003 environment. Following standalone tools and web servers are employed to carry out the present research work.

Standalone based computational tools

Discovery Studio version 3.5 (for modeling, docking and simulation)

LOOPER

ChiRotor

CDOCKER

Modeller academic version 9.12 (for homology modeling)

Molecular Evolutionary Genetic Analysis (MEGA) version 5.1 (for molecular phylogeny)

PyMOL v1.6 (for molecular visualization and image generation)

Chimera v. 1.5.3 (for image generation)

LigPlot+ v.1.4 (for protein-ligand interaction image analysis)

Web based computational tools

Primary sequence analysis tools

NCBI Database (http://www.ncbi.nlm.nih.gov/)

Conserved Domain Database (CDD) (http://www.ncbi.nlm.nih.gov/cdd/)

The Conserved Domain Architecture Retrieval Tool (CDART) (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi)

Pfam: the protein families database (http://pfam.xfam.org/)

Simple Modular Architecture Research Tool (SMART) (http://smart.embl-heidelberg.de/)

InterProScan sequence search (https://www.ebi.ac.uk/interpro/search/sequence-search)

ProtParam (http://web.expasy.org/protparam/)

CONCORD: Secondary Structure Prediction (http://helios.princeton.edu/CONCORD/)

SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/)

Domain enhanced lookup time accelerated BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROGRAM=blastp&BLAST_PROGRAMS=deltaBlast)

RCSB PDB: RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do)

Protein BLAST: search protein databases using a protein query (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins)

GeneSilico protein structure prediction meta-server (http://genesilico.pl/meta)

Pcons.net: protein structure prediction meta server (http://pcons.net/)

Geno3D: automatic comparative molecular modelling of protein (http://geno3d-pbil.ibcp.fr)

Sequence alignment and structure prediction server

ClustalOmega (http://www.ebi.ac.uk/Tools/msa/clustalo/)

ESPript 3 (http://espript.ibcp.fr/ESPript/ESPript/)

Multialign (http://multalin.toulouse.inra.fr/multalin/)

I-TASSER server for protein 3D structure prediction (http://zhanglab.ccmb.med.umich.edu/I-TASSER/)

PHYRE2 (Protein Fold Recognition Server) (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

Model refinement and validation tools

3Drefine: Protein Structure Refinement Server (http://sysbio.rnet.missouri.edu/3Drefine/)

PROCHECK (http://services.mbi.ucla.edu/PROCHECK/)

ERRAT (http://services.mbi.ucla.edu/ERRAT/)

Verify_3D (http://services.mbi.ucla.edu/Verify_3D/)

ProSA (ProSA-web - Protein Structure Analysis) (https://prosa.services.came.sbg.ac.at/prosa.php)

MolProbity (http://molprobity.biochem.duke.edu/)

ProFunc (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/)

Secondary structure assignment servers

STRIDE (http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py)

Protein-protein docking

HADDOCK (http://haddock.science.uu.nl/)

Computational alanine scanning

ABS-Scan (http://proline.biochem.iisc.ernet.in/abscan/index.php)

Sequence retrieval and primary structure analysis

Fasta formatted amino acid sequences of 37 AGPase SS from five monocots (*Hordeum vulgare, Oryza sativa, Sorghum bicolor, Triticum aestivum* and *Zea mays*) and 17 dicots (*Arabidopsis thaliana, Beta vulgaris, Brassica napus, Brassica rapa subsp. pekinensis, Cicer arietinum, Citrullus lanatus subsp. vulgaris, Citrus unshiu, Cucumis melo, Fragaria* x *ananassa, Ipomoea batatas, Perilla frutescens, Phaseolus vulgaris, Pisum sativum, Solanum lycopersicum, Solanum tuberosum, Spinacia oleracea and Vicia faba var. minor*) were retrieved from the GenBank database of NCBI. Likewise 87 AGPase LS protein sequence from seven monocots (*Oryza sativa, Sorghum bicolor, Triticum aestivum, Zea mays, Hordeum vulgare, Brachypodium distachyon* and *Oncidium hybrid cultivar*) and 21 dicots (*Arabidopsis lyrata subsp. lyrata, Arabidopsis thaliana, Beta vulgaris subsp. vulgaris, Brassica rapa, Cicer arietinum, Citrullus lanatus subsp. vulgaris, Citrus unshiu, Cucumis melo, Fragaria x ananassa, Glycine max, Ipomoea batatas, Medicago truncatula, Perilla frutescens, Phaseolus vulgaris, Pisum sativum, Populus trichocarpa, Ricinus communis, Solanum habrochaites, Solanum lycopersicum, Solanum tuberosum and Vitis vinifera*) were retrieved from the NCBI database.

A detailed sequence analysis of the protein was done to have a wide spectrum on the primary structure of the protein. Domain analysis of both SS and LS proteins were done with CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) [Marchler-Bauer et al., 2011], conserved domain architecture retrieval tool (CDART) (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi) [Geer et al., 2002], Pfam (http://pfam.janelia.org/) [Finn et al., 2014] and SMART (http://smart.embl-heidelberg.de/) [Letunic et al., 2012]. InterProScan tool

(http://www.ebi.ac.uk/Tools/pfa/iprscan/) [Joneset al., 2014] was used to predict the protein family, super family and the domain arrangement within the protein. ProtParam tool (http://web.expasy.org/protparam/) [Gasteiger et al., 2005] was used to obtain detailed information on various physico-chemical properties of both SS and LS of AGPase. CONCORD server (http://helios.princeton.edu/CONCORD/) [Wei et al., 2012] which uses consensus-based method for protein secondary structure prediction from its primary amino acid sequence integrating several popular tools, such as PSIPRED, DSC, GOR IV, Predator, Prof, PROFphd, and SSpro was used for predicting the secondary structure of all the AGPase protein sequences. Signal peptide sequences were predicted through SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/) [Petersen et al., 2011].

Homology modeling and model quality assessment

Full length amino acid sequences of both SS and LS of AGPase were subjected to DELTA-BLAST (domain enhanced lookup time accelerated BLAST) (http://blast.ncbi.nlm.nih.gov) against PDB (http://www.rcsb.org/) in order to find suitable templates for the comparative modeling and furthermore functional prediction. DELTA-BLAST searches a database of pre-constructed position specific score matrices (PSSMs) before searching a protein-sequence database to yield better homology detection. DELTA-BLAST was preferred against normal BLASTP because of its retrieval accuracy and sensitivity towards protein analysis. To ensure the sensitivity and accuracy of template selection, various meta-servers like GeneSilico (http://genesilico.pl/meta) [Kurowski and Bujnicki, 2003], Pcons.net (http://pcons.net/) [Wallner, 2007] and Geno3D (http://geno3d-pbil.ibcp.fr) [Combet, 2002] were also used

to find reliable templates with conserved domains. In addition, the protein threading approach implemented by I-TASSER (http://zhanglab.ccmb.med.umich.edu/ITASSER/) [Roy, 2010] and protein fold recognition server Phyre Version 2.0 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) [Kelley, 2009] were also used to determine the best templates in terms of fold recognition. In all cases the template was selected based on high level of sequence identity, query coverage and alignment quality which promises a more reliable and good quality model.

All the template selection methods suggested 1YP3 (C-chain) to be the most appropriate template for SS as well as LS model building having highest sequence identity, query coverage and less E-value.

Different model building algorithms i.e. MODELLER 9.11 [Sali et al., 1995], Discovery Studio version 3.5 (DS3.5; Accelrys Inc. San Diego, CA, USA) followed by Phyre2 and I-TASSER were used for model building of both SS and LS of AGPase. As comparative modelling relies on a sequence alignment between target and the template sequence, the target-template alignment was performed using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) [Sievers, 2011] and rendered using The Easy Sequencing in Postscript 2.2 (ESPript) (http://espript.ibcp.fr/ESPript/ESPript/) server. Based on the target-template alignment 20 different 3D models for each SS and LS of AGPase were generated by MODELLER9.11 [Sali et al., 1995] and were ranked based on their normalized discrete optimized protein energy (DOPE) scores (a statistical potential to assess the quality of the models) and the model with the lowest DOPE score was selected for further validation. MODELLER uses both homology and CHARMm force field derived spatial restraints for model building.

Moreover, to have a better confidence on model building, model generated by MODELLER were compared with the best models generated by DS3.5, Phyre2 and I-TASSER server. Phyre2 constructs a non-redundant fold library of known protein sequences which are mined from the PDB and Structural Classification of Proteins (SCOP) database. All protein sequences stored in this fold library are annotated with known and predicted secondary structures. A non-redundant sequence database and a hidden Markov model (HMM) is iteratively scanned by each sequence in the fold library for each known structure generated. Similarly, the non-redundant sequence database is scanned upon submission of a new protein sequence and a profile HMM is created. Both close and remote sequence homologues are retrieved and an alignment is constructed by PSI-Blast and subsequently secondary structure is predicted. Using HMM-HMM, the fold library is scanned by the profile HMM and the secondary structure. All alignments are ranked based upon a score generated by the alignment and an E-value is generated. The top twenty scoring matches are then used to build up the 3D models of each sequence.

I-TASSER screens the whole PDB library to find appropriate protein fragments from which the global structure is assembled by combining aligned fragments. For portions for which no alignment matches are found, the 3D structure is built using *de novo* simulations. The final refinement of the model is made with a search of the lowest energy conformation [Zhang 2007, 2008].

Model refinement

To increase the compatibility score of each residue, the target model was further refined by loop modeling and side chain refinement using Accelrys *ab initio* loop prediction

algorithm LOOPER [Spassov, 2008] and ChiRotor [Spassov, 2007] for refining protein side chain conformations. Looper generates a set of low energy conformations for the specified loop region and ChiRotor systematically search side chain conformation and scores based on their CHARMm [Wu, 2003] energy. Out of five different models generated by Looper and ChiRotor, the best model was selected based on the lowest DOPE score generated by DS3.5.

Furthermore 3Drefine and its iterative implementation i3Drefine (http://sysbio.rnet.missouri.edu/3Drefine/) [Bhattacharya and Cheng, 2013 a,b] were used to have a reliable refinement of the predicted structures in atomic-level both in terms of global and local measures of structural qualities. It uses two steps of minimization, wherein as the first iteration it optimizes the hydrogen bond network followed by energy minimization using a composite physics and knowledge based force field. This webserver shows its potency over other refinement methods while testing extensively in Critical Assessment of Techniques for Protein Structure Prediction (CASP) data sets.

Quality assessment and validation

Generated models were tested for quality both by geometric and energetic means. PROCHECK (http://services.mbi.ucla.edu/PROCHECK/) [Laskowski, 1993], ERRAT (http://services.mbi.ucla.edu/ERRAT/) [Colovos and Yeates,1993] and Verify_3D (http://services.mbi.ucla.edu/Verify_3D/) [Luthy, 1992] tools which are embedded in structure analysis and validation server (SAVES) (http://nihserver.mbi.ucla.edu/SAVES/) were used for validation of the modeled proteins. The PROCHECK provides an idea of the stereo-chemical quality of the protein. It analyzes the Ramachandran plot quality,

peptide bond planarity, non-bonded interactions, main chain H-bond energy, Cα chiralities and overall G factor. ERRAT tool which finds the overall quality factor of the protein was used to check the statistics of non-bonded interactions between different atom types. Verify_3D was used to access the compatibility of the atomic models with its own amino acid sequence. A high Verify_3D profile score indicates the high quality of protein model. Subsequently the ProSA (https://prosa.services.came.sbg.ac.at/prosa.php) [Wiederstein, 2007] tool was employed in the refinement and validation of the modeled structure to check the native protein folding energy of the model by comparing the energy of the model with the potential mean force derived from a large set of known protein structures. Furthermore, MolProbity (http://molprobity.biochem.duke.edu) [Chen et al., 2010] tool was used for the quality estimation of the predicted 3D models which provides detailed information about the atomic contact and steric problems within the molecules as well as updated dihedral-angle diagnostics, if any.

After each loop and side chain refinement step, the above mentioned model quality assessment programs were employed to check the error at each residue in the protein. This process was repeated iteratively until the most geometrically and energetically stable structural conformation was attained.

To investigate how well the modeled structure matches the X-ray data of template protein, pairwise 3D structural superimposition of the predicted models of both SS and LS of AGPase was carried out with their respective template protein to compute the root mean square deviation (RMSD) between the Cα-atoms and all atoms between target and template. RMSD between equivalent Cα and backbone atom pairs (target and template) was calculated by superimpose module of DS 3.5. STRIDE

(http://webclu.bio.wzw.tum.de/stride/) [Frishman, 1995] server was used to recognize secondary structural elements of the predicted 3D models from their atomic coordinates.

Detailed structural analysis

Knowing the 3D structure of a protein opens up the possibility of ascertaining its function by various analysis of that structure. The theoretical 3D models of both SS and LS of AGPases, belonging to different monocot and dicot species were analyzed extensively to have a wide spectrum on the 3D structure and the role of key residues responsible for catalytic and inhibitory function.

Till date a number of methods have been developed for predicting protein function from their structure. They tend to match the protein's fold against other proteins of experimentally determined 3D structure. Moreover identification of more local features, such as active site residues or DNA/RNA/ligand-binding motifs is also exploited to search for their functions [Watson, 2005]. However, none of these structure-based methods are successful in all cases. For example, methods that are capable to detecting catalytic sites in a protein 3D structure will fail to provide useful information if the query protein is not an enzyme. Therefore, a discrete method which provides both structure-based and sequence-based information not only to increase the chances of obtaining a helpful match, but also to benefit from cases where several methods equivocally comes to the same conclusion is always important. Based on this fact, we used ProFunc server (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) [Laskowski, 2005] for predicting the likely function and various key features of both SS and LS of AGPase. ProFunc uses both existing and novel sequence and structure based methods to provide a convenient

summary about the sequence, structural motif or their close relationships to functionally characterize proteins. It is a user friendly server where the user needs to submit only the coordinates of the protein structure in PDB format to have a detailed sequence and structural information. The whole methodology of ProFunc is shown below in Figure 2. Image is adopted from ProFunc website.

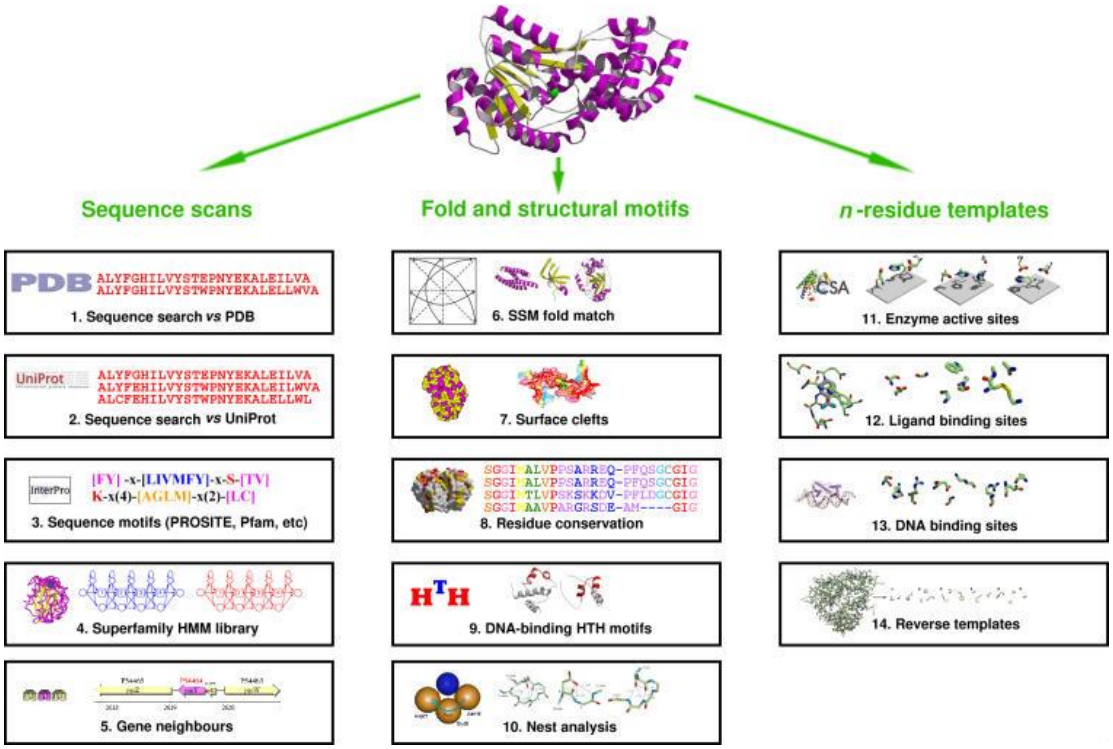

Figure 2: Step by step workflow of the ProFunc server.

Molecular docking

Protein-ligand docking

The interaction between the enzyme and its substrate provides a detailed and accurate picture of the interacting amino acid residues between the substrate and the active site. Different binding site prediction methods were employed for finding the binding site amino acid residues of both SS and LS of AGPases. DS3.5 binding site prediction

module was also employed to predict the binding sites amino acid residues and functional residues which identify the binding sites based on eraser and flood-filling algorithm. MetaPocket2.0 server (http://projects.biotec.tu-dresden.de/metapocket/) [Zhang et al., 2011] was also used for predicting binding sites of all the models. It employs a consensus method by combining the results of four different methods (LIGSITEcs, PASS, Q-SiteFinder, and SURFNET) to improve the prediction success rate. The potential ligand binding sites were generated using a probe of radius 5.0Å and the binding site having highest z-score was considered for further investigation. Moreover, the co- crystallized ligands were used to identify the binding site of the regulator and substrate molecule. In the present study, a binding site sphere was used around the co-crystallized ligand which is large enough to cover the ligand binding residues within the active site. Subsequently the coordinates of the binding site sphere was used as the binding site for identifying regulator and substrate binding with AGPases.

For ligand-protein interaction, both protein and ligand were optimized using the "prepare protein and ligands tool" of DS3.5 that adds hydrogen ions to the protein and add charges, and applies force field to the ligand based on the CHARMm force field. A high temperature simulated annealing dynamics scheme was used for searching the random conformations of the ligand. Ten random conformations were generated by heating the ligand to 700 K in 2000 steps, followed by annealing at 300 K in 5000 steps.

After delineating the binding sites and preparing the protein and ligand molecules, CDOCKER [Wu et al., 2003] module of DS 3.5 was used to carry out the docking analysis. It is a grid based docking which uses CHARMm molecular simulation program to dock ligands within the active site of receptors. A set of random ligand conformations

were generated using a high-temperature molecular dynamics and successively the produced conformations were translated into the binding site. The generated poses were further created using random rigid-body rotations followed by simulated annealing. Finally the ligand poses were refined by minimization.

The prepared ligand molecules were docked in to the active site of the AGPase to elucidate its binding affinity towards the regulator and substrate molecules which in turn reflects the insight into the allosteric regulation and substrate binding specificity of the protein. During the docking process different numbers of ligand conformations (poses) were prepared based on the top orientation of the molecules in the active site of the protein. The binding affinity of the ligand molecule into the active site of the protein was calculated based on the consensus scoring scheme of CDOCKER ENERGY, CDOCKER_Interaction Energy, Ligscore1_Dreiding, LigScore2_Dreiding, PLP1, PLP2, Jain, PMF and PMF4 implemented in the protein-ligand interaction module of DS3.5.

Protein-protein docking

For studying the protein-protein interaction HADDOCK (High AmbiguityDriven protein-protein Docking) web server was used. There are two main reasons for using the server. First, HADDOCK uses the known information about the contact regions of the interacting molecules to accelerate and accurate the docking procedure and secondly both the main and side chain flexibility can be assimilated into the docking process.

It is an information based, data driven flexible docking method for bio-molecular complex formation between protein-protein, protein-nucleic acids, protein-peptides, protein-ligand etc. [de Vries et al., 2010]. It uses the biochemical and/or biophysical

information of the interface regions between the molecules and their relative orientations in the stable conformation to carry out the docking study [de Vries et al., 2010; Dominguez et al., 2003]. In the absence of experimental information, HADDOCK accepts the data of the interface regions generated by various computational interface prediction tools [de Vries et al., 2010]. In contrast to many other flexible docking programs, it allows the conformational changes in the side chains of the interacting proteins, as well as in the respective backbones. Data regarding the interface residues are provided in the form of active and passive residues to HADDOCK, which are then converted to ambiguous interaction restraints (AIR). Active residues are crucial as they remain in the interface between the two molecules (ligand and receptor) and plays active role in complex formation, whereas passive residues are absent in the interface when the complex is formed. The passive residues can also be automatically defined in HADDOCK which select them automatically around the active residues.

HADDOCK uses a three step docking protocol to perform the task. In the first step a rigid body energy minimization is computed and subsequently a semi-flexible refinement in torsion angle space is performed. The final step is optional where flexible refinement in Cartesian space with explicit solvent, e.g. water, can be done. After execution of these three stages, final docked conformations are scored and ranked based on a scoring function to obtain the best docked conformation.The ranking of the docked complex is done based on the score calculated from van der Waals (E_{vdW}), electrostatic (E_{elec}), restraint violation energies ($E_{AIR)}$, empirical desolvation energy (E_{desolv}) and buried surface area (BSA). Finally scores for all these energies for a given cluster of the docked complex is reported along with the RMSD, buried surface area and HADDOCK score.

Haddock score is a linear combination of the above listed four types of energies along with the buried surface area [de Vries et al., 2010].

$$HADDOCK_{score} = 0.1\ E_{AIR} + 1.0\ E_{vdw} + 0.2\ E_{elec} + 1.0\ E_{desol}$$

Result with the lowest HADDOCK score and Z-Score were considered as the best docked complex and were used for further analysis.

Phylogenetic analysis

In order to gain new insights into the AGPase evolutionary history and to discriminate between different models of evolution and sub-functionalization, molecular evolutionary genetic tree of both SS and LS of AGPase was construced. Full-length AGPase sequences from 37 AGPase SS along with two out-group sequences and 87 AGPase LS sequences were selected. Multiple sequences alignment was performed by ClustalW using the MEGA software [Tamura et al., 2013] with BLOSUM62 matrix with a gap opening penalty of 10, gap extension penalty of 0.1, and gap separation distance of 4 followed by manual refinement of the resulting alignments. Model testing for the phylogenetic tree construction of both the SS and LS of AGPase was done in MEGA5. Neighbor-Joining (NJ) method [Saitou and Nei, 1987] was used to construct the phylogenetic tree for both SS and LS of AGPase separately using MEGA5. The reliability of topology was tested by bootstrap analysis using 1000 iterations.

Alanine scanning mutagenesis (ASM)

In order to identify the key residues responsible for regulator binding with AGPase, ASM was performed using ABS-Scan web server which carries out the ASM for a given

protein-ligand complex. It allows the user to submit a protein-ligand complex of their interest in PDB format. Thereafter binding site residues were selected based on a user-defined distance cut-off threshold and the ligand was selected. Subsequently, Modeller was used to perform a site-specific mutagenesis on all selected residues, followed by energy minimization. The mutated protein-ligand energetics was then evaluated based on the Autodock score and the scores obtained are compared with the wild type to evaluate the energetic contributions made by each residue in the binding site towards ligand recognition. The whole methodology of ABS-Scan is shown in Figure 3. The image is adopted from ABS-Scan website.

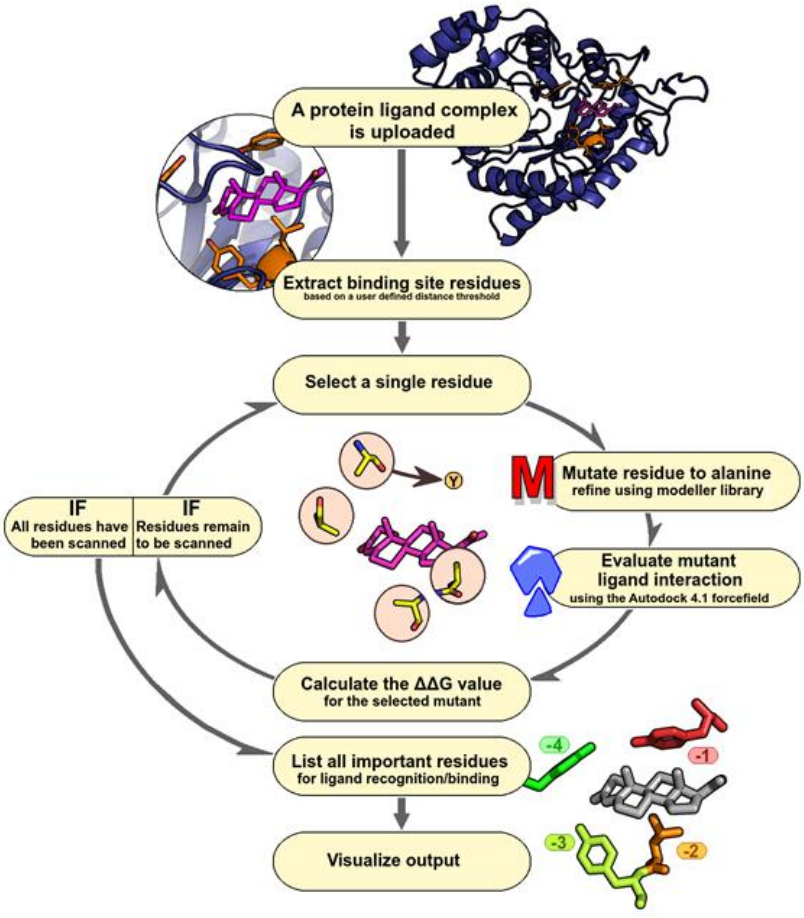

Figure 3: Step by step workflow of the ABS-Scan server.