CHAPTER II

REVIEW OF LITERATURE

Starch: Occurrence and function

The major carbon reserve found in the developing storage organs such as seeds, fruits, tubers, and stems of many plants is starch. It is the principal component of seed endosperm, accounting for 56-74% of the available carbohydrates in the grain of most food crops including cereals [Koehler and Weiser, 2013]. Starch has two basic components, amylopectin and amylose, both of which are polymers of α-D-glucose units. Amylose, a linear polymer of several thousand glucose residues assembles with amylopectin (a larger polymer regularly branched with α-1,6-branch points) together form a semi-crystalline starch granule [Manners and Matheson, 1981; Hizukuri, 1995]. This osmotically inert material is readily mobilized as a source of carbon and energy. The exact proportions of these molecules and the size and shape of the granule vary between the species and also between the organs of the same plant. The physico-chemical and biological aspects of the starch granule and its components amylose and amylopectin have been reviewed in some literatures [Morrison and Karkalas, 1990; Hizukuri, 1995].

Animals, fungi, bacteria including the cyanobacteria, and archaebacteria accumulate glycogen; whereas in algae and in plants starch is the polymer synthesized as a carbon and energy reserve. Owing to its role as an important energy and carbon reserve in plants, starch serves as a rich source of nutrition for humans and animals. The diversity of both composition and physical parameters of starches from different biological sources gives rise to their diverse processing properties and applications in both non-food sectors such

as sizing agents in textile and paper industry, adhesive gums, and biodegradable materials, and in food industries, particularly in bakery, thickening, confectionary and emulsification [Burrell, 2003; Mooney, 2009; Slattery et al., 2000]. Starch is also used as a feedstock for first generation bioethanol production [Goldemberg, 2007]. From an industrial perspective, the alternate utilization of starch as a cheap and renewable polymer and clean energy source is becoming increasingly attractive as a consequence of environmental risk concerns about industrial wastes generated from fossil fuels.

Leaf starch

Exposure of the leaf in bright sunlight causes the synthesis of starch granules in the chloroplast organelle naturally and this was demonstrated in the 19th century [Sachs, 1887]. Disappearance of the starch occurs either by exposure of the leaf to low light or by extended exposure in the dark (24-48 h). This can readily be visualized by iodine staining of the tissue [Edwards and Walker, 1983] or by light or electron microscopy [Badenhuizen, 1969]. Carbon fixation during photosynthesis leads to starch accumulation and granule formation. The granules almost degrade when in the dark to products that are in most cases utilized for sucrose biosynthesis. The reason for this is that the accumulated starch is required to synthesize sucrose, which serves as a carbon supply for sink tissues. Mutants of *Arabidopsis thaliana* unable to synthesize starch grow at the same rate as the wild type in a continuous light regime because they are able to synthesize sucrose, but their growth rate is drastically reduced if grown in a day-night regime [Caspar et al., 1985]. Biosynthesis and degradation of starch in the leaf is therefore a dynamic process having diurnal fluctuations in its stored levels.

In leaves, starch also plays an important role in the operation of stomatal guard cells [Outlaw and Manchester, 1979; Zeiger et al., 2002; Ritte and Raschke, 2003]. There, it is degraded during the day while the stomata are open, and it is resynthesized in the late afternoon or evening. Leaf starch is lower in amylose content than what is observed in storage tissues [Matheson, 1996]. The amylose structure is also of a smaller molecular size [Morrison and Karkalas, 1990].

Starch in storage tissues

Biosynthesis of starch occurs in storage organs, fruit or seed etc. during the development and maturation of the tissue [Sivak and Preiss, 1998]. At the time of sprouting or germination of the seed or tuber, or ripening of the fruit, starch is degraded and the derived metabolites are used as a source of both carbon, and energy. The degradative and biosynthetic processes in the storage tissues may therefore be temporally separated. However, there is some possibility that during each phase of starch metabolism, some turnover of the starch molecule do occur. The main site of starch synthesis and accumulation in the cereals is the endosperm, with starch granules being located within the amyloplasts. Starch content in potato tuber, maize endosperm, and in roots of yam, cassava and sweet potato ranges between 65% and 90% of the total dry matter [Sivak and Preiss, 1998]. Patterns of starch accumulation during development of the tissue are specific to the species and are related to the unique pattern of differentiation of the organ. Starch granules in storage tissues can vary in shape, size and composition depending on the plant source. In addition, in each tissue of a plant differences in size and shape are observed within a range. The diameter of the starch granule changes during the

development of the reserve tissue. There are also some fine features, characteristic of each species, for example, the 'growth rings', spaced 4 - 7μm apart, and the fibrillar organization seen in potato starch, which allows one to identify the botanical source of the starch by its microscopic examination [Morrison and Karkalas, 1990; Hizukuri, 1995; Sivak and Preiss, 1998].

Biosynthetic pathway of starch synthesis with reference to ADPGlucose Pyrophosphorylase (or AGPase)

The synthesis of starch in plant cells starts with active involvement of the enzyme AGPase. This enzyme in the presence of a divalent metal ion, Mg^{2+} and at the expense of ATP, converts Glc1P into ADPGlc, liberating Pi. This is the first committed step in the biosynthetic pathway leading to starch synthesis in plants. The hydrolysis of ATP into ADP and the expenses of sugar-nucleotides for polysaccharide synthesis essentially make this first step to be rate-limiting and irreversible *in vivo* [Iglesias and Preiss, 1992]. In the later stages of starch synthesis, starch synthase enzymes (plants, E.C. 2.4.1.21) use free ADPGlc, where glucose units are added to the end of a growing polymer chain to construct a starch molecule. Starch branching enzymes (SBEs) (E.C. 2.4.1.18; 1,4 -α-D-glucan: 1,4-α-D-glucan 6-glycosyl-transferase) add branches in the chain by hydrolysing 1,4-glycosidic bonds, and in their place, create 1,6 bonds with neighbouring glucose molecules [Preiss, 1984, 1991; Iglesias and Preiss, 1992; Sivak and Preiss, 1998; Preiss and Sivak, 1998a]. This reaction was first explained from soybean extracts through a classical experiment by Espada (1962). Thereafter, the enzyme was also reported from many other plant tissues and bacterial extracts [Preiss and Sivak, 1998a, b; Preiss, 1999].

Structure of plant AGPase

Plant AGPase is a heterotetramer (α2, β2; M.W.: 200-240 kDa) composed of a pair of SS (AGP-S or SS = α2) and LS (AGP-L or LS = β2) encoded by distinct genes [Smith-White and Preiss, 1992]. The difference in molecular weights between the SS and LS ranges from 1-5 kDa [Ballicora et al., 2004]. Comparative amino acid sequences of different plant AGPases showed that there are (i) about 90% identity among SS, (ii) 50-70% identity among LS, and (iii) 40-50% identity between SS and LS, suggesting that both subunit genes have evolved from a common ancestral gene [Ballicora et al., 2005]. Both subunits have evolved differentially with a different function. Structural validation of potato tuber AGPase indicated that both subunits are required for the optimal enzyme activity but have nonequivalent roles in enzyme function [Iglesias et al., 1993]. The active pocket of AGPase consists of an elongated cleft bounded at each end with a sugar- or adenine binding site. The residues at the active site are conserved in nucleotidyl transferases as well. The entire molecule of ATP fits into the pocket, although the docking pose altogether differs from that seen in common sugar-nucleotide PPase/NTP complexes [Jin et al., 2005]. Two Cys12 residues of the catalytic site in SS form an intermolecular disulfide (S-S) bond that is vital for regulatory activity of plant AGPase. This bond enhances the enzyme activity and provides thermal stability.

Regulation of plant AGPase activity

Plant and bacterial AGPases are allosteric enzymes, the activity of which is regulated by key metabolites from the main carbon assimilation pathway occurring in the organism [Iglesias and Preiss, 1992; Sivak and Preiss, 1998]. Thus AGPase effectors are key

metabolites representing signals of high carbon and energy contents within the cell. On the basis of the affinity for allosteric regulators, AGPase have been classified into nine different classes [Ballicora et al., 2003]. Among these, the enzymes from photosynthetic organisms are included in classes IV, V, VI, VIII, and IX. In most of the plants, 3-PGA and Pi typically regulate the enzyme activity former of which is mainly abundant during the day and scanty at night [Preiss, 1991; Iglesias and Preiss, 1992; Sivak and Preiss, 1998; Preiss and Sivak, 1998b]. Recently, it has been suggested that in higher plants the enzyme activity can also be regulated by its reductive state [Fu et al., 1998b; Ballicora et al., 2000; Tiessen et al., 2002]. Allosteric mutant AGPases from maize endosperm and *C. reinhardtii* and the resultant effects on starch synthesis provide strong evidence that the allosteric effects observed *in vitro* are operative in the *in vivo* situation [Giroux et al., 1996; Van den Koornhuyse et al., 1996; Preiss and Sivak, 1998a, b; Sivak and Preiss, 1998].

Variation of 3-PGA interaction with Pi in AGPases from different plant systems

Activity of most plant AGPases is governed by the abundance of 3-PGA and Pi; however the pattern of regulation may vary from plant to plant. As such, four patterns of interactions between 3-PGA and Pi in the plant enzymes have been worked out till date. First and foremost is the one observed commonly for most enzymes, where Pi and 3-PGA affect the enzyme activity individually. In this pattern, it is observed that increasing concentrations of 3-PGA reverses or overcomes the Pi inhibition.

The second distinct regulatory pattern has been reported in AGPases from storage tissues of some cereals. Purified AGPase from wheat endosperm has been reported to be

regulated by the coordinate action of a series of metabolites [Gomez-Casati and Iglesias, 2002]. The wheat endosperm AGPase is allosterically inhibited by metabolites *viz.*, Pi, ADP, and fructose 1,6-bisphosphate (Fru 6-P). Inhibition is reversed by 3-PGA and Fru 6-P, which individually (in the absence of the inhibitors) have no effect on the enzyme activity [Gomez-Casati and Iglesias, 2002]. This only suggests that the levels of several metabolites also influence the first limiting step in starch biosynthesis.

The third variation of 3-PGA activation interaction with Pi inhibition is observed in CAM plant leaf AGPases of *Hoya carnosa* and *Xerosicyos danguyi* [Singh et al., 1984] and of maize endosperm [Plaxton and Preiss, 1987]. A concentration of 2mM 3-PGA upregulates enzyme activity by about 10 to 25 fold; however, presence of 3-PGA makes the enzyme susceptible to Pi inhibition. The enzyme from maize endosperm is only inhibited about 20% by 10mM concentration of Pi and the CAM plant leaf enzymes 50% by 2mM concentration of Pi. Further addition of Pi does not increase their inhibitions. However, at sub-saturating concentrations of 3-PGA (~0.15-0.25mM) the enzyme becomes more sensitive to Pi inhibition and becomes totally inhibited at 0.5-2mM Pi. Higher 3-PGA concentration reverses the Pi inhibition and decreases the affinity of the enzymes for Pi.

The fourth distinct pattern was observed with barley endosperm AGPase, which is poorly activated by 3-PGA or inhibited by Pi. 3-PGA increases the apparent affinity of ATP by lowering up the $S_{0.5}$ of ATP (almost by 3-fold) and the Hill coefficient [Kleczkowski et al., 1993c]. 3-PGA activates upto 4-fold in the presence of ATP at 0.1 mM concentration, and Pi 2.5mM reverses the process. Thus, in barley endosperm, concentration of 3-PGA sand Pi significantly influences the apparent affinity of the substrate, i.e. ATP.

Importance of AGPase in starch biosynthesis and modulation

Many experimental evidences clearly suggest that AGPase is an important regulatory enzyme in the biosynthetic pathway of plant starch synthesis. In *Chlamydomonas reinhardtii*, starch deficient mutants have been isolated and one class of mutant was shown to have an AGPase that could not be activated by 3-PGA [Ball et al., 1991]. Support for the importance of the allosteric regulation by AGPase has also been obtained in *Arabidopsis thaliana* [Lin et al., 1988a, b]. One mutant, TL25, lacked both subunits and accumulated only 2% of the starch seen in the normal plant [Lin et al., 1988a], which would indicate that starch synthesis is almost completely dependent on the synthesis of ADPGlc. The other mutant, TL 46, was starch-deficient and lacked the regulatory 54 kDa subunit [Lin et al., 1988b]. The mutant had only 7% of the wild-type activity and a subsequent study [Neuhaus and Stitt, 1990] showed that in high light (photosynthesis) the rate of starch synthesis of TL46 was only at 9% and at low light, only 26% of the rate of the wild type. This is supporting evidence that the regulation of AGPase is of *in vivo* importance. A maize mutant has also been isolated where the AGPase was less sensitive to the inhibition by Pi than the wild type enzyme; the mutant endosperm had 15% more dry weight and more starch than the normal endosperm [Giroux et al., 1996]. In potato tuber [Stark et al., 1992] and wheat endosperm [Smidansky et al., 2002], genetic manipulation of AGPase activity led to an increase in starch production.

Proteins: The key players in biological function and assembly

Proteins among the most important biological macromolecules in all cells, comprising of one or more polypeptide chains is composed of a sequence of the twenty amino acids

occurring in nature. Understanding protein function, control and interactions is therefore one of the most important goals in structural biology research. The unique 3D structures of proteins which are often formed by their folding and the knowledge of these structures is often essential to understand protein function.

Protein structure prediction

The properties, behaviour and almost all biological phenomena mediated by proteins, including protein-ligand, protein-protein interactions, drug function and protein design are all clarified by the 3D structure of proteins. Experimental efforts have led to deposition of 102863 (as of Aug 26, 2014) experimentally solved structures in the Protein Data Bank (PDB), but there are far more protein sequences reported, with 546000 sequence entries (Release 2014_07; 09-Jul-14 ofUniProtKB/Swiss-Prot). Protein structure prediction remains a significant challenge because the rapidly increasing number of reported protein sequences is not matched by experimental techniques to solve all the associated structures. Structure prediction methods [Ginalski, 2006; Xiang, 2006] represent one promising route to bridge the gap between sequence and structure. This objective is realistic considering that sequences with 50% or more identity can often be modeled within experimental accuracy [Kryshtafovych et al., 2007; Dehury et al., 2013]. Structure prediction methods are classified into homology modeling (HM) also called template-based modeling (TBM) or comparative modeling (CM), and free modeling (FM) [Zhang, 2008]. Proteins with similar sequence fold into similar 3D structures. In HM, the 3D structure of the protein built from a template structure works only if a protein with sufficiently high homology and solved structure is available. In the absence of this

information FM are used, which do not depend on a prior structural information. The success rate of FM methods is presently low [Peng and Xu, 2010].

Comparative modeling

In recent years, many different methods for structure prediction have been proposed [Zaki and Bystroff, 2008], but the most widely used methodology is the comparative structure prediction (also called homology modeling). In HM method, the model of the "target" protein is constructed from the amino acid primary structure and from an experimental 3D structure of a related homologous protein called "template". It is seen that the 3D protein structure is evolutionarily more conserved because of the fact that in nature there is a limited number of folding conformations.

In HM, one or more known structures that are similar to the structure of the query sequence were identified and then an alignment is created between the residues of the query sequence and the residues in the template sequence [Tang et al., 2003; Ohlson et al., 2004; Marti-Renom et al., 2004]. The target model is created from the sequence alignment and the template structure. The higher the quality of the alignment between the sequence and the template, the better the model that is produced. Gaps in the sequence alignment or the template structure can decrease the quality of the model. Loops region are normally the most involved in this kind of errors, where the target and template proteins may be completely different. When there is no high homology template for the whole sequence, we have to use loop modeling techniques to fill the gaps in the structure, which typically result in less accurate model structures. Low quality models should not be used for studies such as drug design and protein-protein interaction prediction; but

sometimes they can still be used to obtain some interesting information of the biochemistry of the query sequence, for example by making hypotheses about the conservation of certain residues, information that could also be useful for experimentalist.

Model building

There are four major methods for model building which includes, the spatial restraint method (SSR) [Sali and Blundell, 1993], the segment matching method (SMM) (Levitt, 1992), the multiple template method (MTM) [Chothia and Lesk, 1986; Blundell et al., 1987] and artificial evolution (AE) [Petrey et al., 2003]. SSR assumes that while comparing equivalent positions several geometrical features such as distances and angles are conserved in homologous proteins. This method involves two main steps which includes extraction of spatial restraints based on alignment and construction of the target 3D model by fulfilling the spatial restraints [Sali, 1995]. One of the most frequently used homology modeling programs is MODELLER that uses SMM to divide the target into a series of short segments, each matched to its own template fitted from the PDB. The alignment of the sequence is done over segments rather than over the entire protein. An extension to the SMM program SegMod/ENCAD called Pfrag was proposed by Larsson and co-workers which can use multiple templates [Larsson et al., 2008].

In MTM, several solved protein 3D structures are used to build the target-protein model. The multiple templates that are based on sequences and structures are used to align with each other and the target is optimally aligned with the multiple templates. Loops are present between the conserved regions where they are usually exposed at the surface of

the proteins. MTM has been implemented in several packages such as SWISS-MODEL [Schwede et al., 2003] and MOE [Boyd et al., 2005], 3D-JIGSAW [Bates et al., 2001].

In AE, the model of the target protein is built by editing the template structure which is based on the alignment of the sequences of the template and target achieved using evolutionary concepts like mutations, insertions and deletions. The aligned residues are mutated and this causes changes in scoring function (energy) which is minimized in the procedure. Algorithms for modeling mutations are followed by procedures of modeling the deletions and then insertions. When there is no significant energy penalty, the operation is considered successful [Petrey et al., 2003]. Some automated web servers such as I-TASSER [Roy et al., 2010] and ROBETTA [Raman et al., 2009] are implemented in consensus servers, such as Pcons [Wallner et al., 2003] and HHPred3 [Soding. 2005], which have been successful in accurate prediction of the protein target structures. 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 prediction by ROBETTA makes use of extensive and computationally expensive conformational sampling and all-atom energy refinement [Chivian et al., 2003]. Other web servers for model building are M4T (Multiple Mapping Method with Multiple Templates) [Fernandez-Fuentes et al., 2007] and PROTEUS2 [Montgomerie et al., 2008]; LOMETS (Local Meta-Threading-Server) [Wu and Zhang, 2007], Phyre2 (Protein Homology/analogY Recognition Engine V 2.0) [Kelley and Sternberg, 2009].

Loop modeling

Loop modeling methods can be classified into two major approaches: (i) knowledge based and (ii) energy based methods. A few methods have also been reported which combine the two approaches.

Knowledge based methods are limited by the availability of relevant loop structures from known protein structures [Peng and Yang, 2007] because it is difficult to find a suitable loop segment that fits between the two stem regions of the loop from a database of structures. A database of structural motifs, ArchDB, was developed [Espadaler et al., 2004] and evaluated using two different sequence profiles. Other methods use Monte-Carlo simulation of the loop, ranking the fragment database for the loop prediction using the DFIRE potential.

Energy based methods use a de novo energy function for conformational search of the loops to test their quality [Soto et al., 2008]. *De novo* loop modeling is a good method for loops not longer than seven residues [Jacobson et al., 2004]. Loop conformational search can be carried out using tools such as local move Monte-Carlo (LMMC) [Cui et al., 2008], torsion angle conformational search [Felts et al., 2008], Loop Builder [Xiang etal., 2002], replica exchange [Olson et al., 2008] or a dihedral angle-based build up procedure in hierarchical loop prediction (HLP) [Jacobson et al., 2004]. The conformers generated are scored using a force field or other physics-based energy calculations [Zhu et al., 2006] usually including solvation effects. LOOPER allows a systematic and efficient sampling strategy, searching for loop conformers with optimal interactions of the loop backbone with the rest of the protein atoms. For the final ranking the CHARMm energy scoring function with a generalized Born solvation term is used [Spassov et al., 2008].

Side chain modeling

Regarding side chain prediction most of the methods use rotamer libraries which are constructed using statistical knowledge of protein 3D structures such as the Grow-to-Fit molecular dynamics method (G2FMD) [Zhang and Duan, 2006], statistica lmachine learning methods [Yan et al., 2007] and IRECS that selects more than one rotamer in order to have a representation of the conformational space flexibility of the side-chain. The ranking is calculated by a knowledge based statistical potential, ROTA [Hartmann et al., 2007]. To improve side-chain modeling modifications to the ROSETTA, energy functions with softer van der Waals terms and extended rotamer libraries have been implemented [Dantas et al., 2007].

Model quality assessment

Quality assessment (QA) methods attempt to rate the quality of the model based on statistical evidence. The assessment methods include statistical as well as physico-chemical methods, which are based on alignment to a single template or multiple templates or on meta server results. The QA method gives a local score as a function of residue or residue window [Wiederstein and Sippl, 2007; McGuffin, 2008; Mereghetti et al., 2008] or a global score [Eramian et al., 2006; Benkert et al., 2008; Qiu et al., 2008; Randall and Baldi2008] which may be based on single or multiple assessment criteria.

Quality assessment programs include ModFOLD server [McGuffin, 2008] that combines ModSSEA [McGuffin, 2007], MODCHECK [Pettitt et al., 2005] and ProQ [Wallner et al., 2003]. AIDE [Mereghetti et al., 2008] scores with secondary structure information. The local quality of a structure can be quantified using ProQres, which relies on 3D

information and quantifies structural qualities such as secondary structure, solvent accessibility, atom-atom and residue-residue contacts to have a measure of local quality. Furthermore it can be quantified by ProQprof, using a model generated from sequence alignment [Wallnerand Elofsson 2003] of both target and template (the sum of the two scores is arranged as ProQlocal). Structural Analysis and Verification Server v4.0 (SAVES) hosted at National Institute of Health also checks the quality of modeled protein integrating five different tools i.e., PROCHECK, WHATCHECK, ERRAT, Verify_3D and PROVE.

Applications of homology modeling

Various applications of HM include:

- Identification of active sites of protein.
- Modeling substrate specificity.
- Protein-protein docking.
- Structure guided design of mutagenesis experiments.
- Design of mutants to test hypothesis about a protein's function.
- Design of *in vitro* test assays.
- Homology model based ligand design.
- Tool compound design for probing biological function.
- Structure based prediction of drug metabolism and toxicity.
- Structure based assessment of target drugability etc.

Molecular docking

Docking is a computational method which predicts the preferred orientation of a molecule to a second one [Lengauer and Rarey 1996]. Such methods have been applied in a number of contexts: e.g. protein-protein, protein-DNA and protein-ligand docking.

In protein-ligand docking, a small molecule binds to a protein receptor. In the past, two complementary approaches have been investigated: one, where the protein and the ligand are described as complementary surfaces [Jorgensen, 1991; Kitchen et al., 2004]; and the second approach where the protein and the ligand conformational spaces are explored with a search algorithm to optimize the scoring function which evaluates an approximation of the binding energy (scoring function) for every conformation of the complex [Wei et al., 2004]. Typical moves in such algorithms includes translational and rotational (rigid body transformations), and torsion angle rotations (internal changes of the ligand). Finally the scoring function for the predicted complex is calculated, yielding an estimate of the affinity. This method has some advantages, as for example that the process is very similar to the docking process in reality, but the scoring functions used today often give poor approximations of the affinity [Warren et al., 2006].

Depending on the docking speed, the method can be employed for screening large compound databases in the search for drug compounds. One major complication in docking methods is treatment of the conformational flexibility of the receptor and treatment of induced fit effects [Kokh and Wenzel, 2008]. Generally docking method involves generation of a set of poses which are ranked based on a scoring function for a ligand that can fit the binding site [Mobley and Dill, 2009]. There are different scoring

functions like regression-based or interaction-based scoring functions that can be used depending on the desired approach.

Interaction-based scoring functions are similar to molecular mechanics force fields, for example the CHARMm [Brooks et al., 2009] and AMBER [Cornell et al., 1995], which have been used to calculate the enthalpy of binding. Normally the values of non-bonded energy terms (van der Waals, electrostatic and internal energy related to bond angles, bond lengths and torsional angles) are pre-calculated on a grid which is used to compute the energy contributions for atoms placed in the receptor pocket. Moreover an approximation of solvent effects can be added. Using such scoring functions or force fields, the ligand deformation is treated in the same way as the interaction between the ligand and the protein. Knowledge-based potentials improve modeling of protein complexes by taking advantage of the rapidly increasing amount of experimentally derived information on protein-protein association. An essential element of knowledge-based potentials is defining the reference state for the optimal description of residue-residue pairs in the non-interaction state. Studies in scoring functions showed that XSCORE [Obiol-Pardo and Rubio-Martinez, 2007] is one of the most used programs to assess ligand-binding affinity as it performs better than the other empirical scoring functions [Wang et al., 2003]. Efforts have been made to develop techniques to speedup the search for the best conformation. For example, Monte-Carlo methods [Huey et al., 2007] or its generalizations [Wenzel and Hamacher, 1999] are often used, while other methods use genetic algorithms, e.g. Gold [Jones et al., 1997]. Fragment based approaches use a different technique to sample the large conformational space of a protein-ligand complex. Different fragments of the ligand are created and then

independently docked to the protein to be finally assembled again [Taylor et al., 2002]. Glide uses a series of hierarchical filters to search for possible locations of the ligand in the active-site region of the receptor [Friesner et al., 2004]. For the reconnection of the broken bonds, the most popular approach is the incremental construction algorithm implemented in FlexX [Rarey et al., 1996]. The *de novo* ligand design methodology is linked to this approach where a totally new ligand is constructed by docking fragments of a database to the protein instead of screening a database of ligands [Sándor et al., 2010].

Affinity estimation

Estimating the affinity of protein-ligand interactions with reliable methodologies has become imperative in drug discovery to detect new lead compounds, and chemical genomics to search for inhibitors to elucidate gene function [Gilson and Zhou, 2007]. To upsurge the impact of these computational techniques it is essential to compare/correlate the experimentally determined affinities with the predicted affinities. The experimentally determined affinities are used as guidelines to calibrate these empirical scoring functions. In recent years, free energy studies have been computed using relative or absolute binding free energy calculations [Guimarães et al., 2005], explicit or implicit solvent models [Fujitani et al., 2005; Michel et al., 2006].

Protein-ligand interactions

Protein-ligand interactions have a central role in enzeme activity, metabolism, host-pathogen interactions and many other processes of living systems. Understanding the protein interactions with small molecules is of great interest as it allows understanding

and influencing protein function, and also for therapeutic applications. The conformational changes of proteins, including folding, as well as many biological functions are mediated by intra- and inter-molecular interactions. Models for such interactions are required to predict the preferred orientation of the molecules and the strength of association or binding affinity between two molecules. Among the most important non-covalent inter-atomic interactions mediating the binding of small molecules with protein are, electrostatic and van der Waals interactions. Other important factors that contribute to protein-ligand affinity, include entropy changes in the solvent and intra-molecular contributions arising from the flexibility of the receptor in the binding site [Gohlke and Klebe, 2002; Bissantz et al., 2010]. Hydrogen bonding, salt bridge and metal interactions [Gohlke and Klebe, 2002] are often treated in the framework of the electrostatic model, but in particular for hydrogen bonding, one of the most important interactions in biological macromolecules, this may not be fully justified. Hydrophobic interactions involve contacts between non-polar parts of the molecule and have been shown to play a crucial role in ligand binding [Bissantz et al., 2010]. Such models may be used to model the process of receptor ligand binding in order to determine the binding pose and the binding affinity.

Protein-protein interactions

Techniques to obtain the atomic structures of single proteins are maturing. The size limit of macromolecules that can be modeled accurately is continually expanding and the number of structures deposited in the PDB has followed a nearly exponential growth over the past two decades. One of the remaining challenges, however, is obtaining assemblies

of two or more macromolecules (proteins, DNA, RNA). On average, it is believed that each protein interacts with about 8-10 other macromolecules, but only 2990 complexes are present in the PDB. For X-ray diffraction (XRD), the difficulty in co-crystallizing a complex is much greater than for individual proteins. For Nuclear magnetic resonance (NMR) spectroscopy, the large molecular weight of complexes presents a problem, making it more difficult to obtain and analyze data. Furthermore, inter-molecular nuclear overhauser effects (NOEs), which provide the most useful information, often involve amino acid side chains, the resonances of which are much harder to assign tham the backbone resonances of the protein. Currently, the only method that can systematically give insights into the large number of protein complexes encountered in biological processes is protein docking *in silico*. This consists of predicting the binding mode of protein complexes starting from their free-form (unbound) experimental or modeled individual 3D structure. Several software packages have been developed for this purpose. The majority of them try to predict protein-protein complexes using solely geometrical and/or energetic considerations. HADDOCK [de Vries et al., 2010] distinguishes itself by including experimental, notably NMR data and/or bioinformatics information to efficiently drive the docking process.

Compared to methods studying protein-ligand interactions for small molecule (ligands), a lot of work has been done to develop models to predict conformation affinity of protein-protein complexes [Stoddard and Koshland, 1992; Smith and Sternberg, 2002; Wiehe et al., 2008], which are important regulators of biological function. These methods are either based on empirical bioinformatics approaches, such as functional matrices, [Marrero-Ponce et al., 2005] or on molecular modeling techniques, such as molecular

mechanics MM Generalized Born/Surface Area GBSA [Chong et al., 1999]. Protein-protein interactions are indispensable to understand the function of biological systems, and their depiction has become a vital task for both experimental and computational approaches in systems biology. Experimental methods including yeast two-hybrid systems [Uetz et al., 2000; Ito et al., 2001], mass spectrometry [Ewing et al., 2007] and protein chips can also be used to study protein-protein interactions apart from computational approaches.

Protein-protein docking

Protein docking is generally applied to individual pairs of proteins that are known to interact with each other. This problem can be addressed by two ways: a scoring/energy function must be developed that can discriminate correctly or near-correctly docked orientations from incorrectly docked ones, and a search method must be implemented to find a near-correctly docked orientation. The problem is much more complicated than small-molecule ligand docking simulations, because in protein-protein docking, the docking site is generally not known. Therefore all respective orientations of the molecules with respect to each other must be sampled, which is difficult with a standard dynamic method. Instead, methods exploiting surface shape complementarity (viz., simplest scoring function) are used as an initial step, which is defined later. This may be done by discretizing the molecule in to a grid space and considering which cells are occupied, or by using some sort of surfacing algorithm which calculates the solvent-accessible or solvent-excluded surface, and a point set that triangulates. Most docking programs comprises of two standard steps: generation of thousands of alternative poses to

sample all possible interaction modes, followed by scoring these poses using an energy function. In many cases, conformations very similar to the natural one are generated by the first step, but scoring functions often fail to rank them properly [Lensink et al., 2007]. The fast Fourier transform (FFT) method is the most used technique for the first stage of docking. This approach is used in various programs like GRAMM [Vakser, 1997], 3D-Dock [Sternberg et al., 1998], DOT [Mandell et al., 2001], HADDOCK [de Vries et al., 2010] and ClusPro2.0 [Kozakov et al., 2013].

Phylogenetics

The term "phylogenetics" has been derived from the Greek terms "phyle" and "phylon" meaning "tribe" and "race"; and the term "genetikos" which imply "relative to birth", from "genesis" i.e. "birth" [Roy et al., 2014]. Phylogenetics is the study of evolutionary relatedness among groups of organisms (e.g. species, populations). In other words, phylogenetic analysis of a family is to determine how the family might have been derived during evolution. Evolution can be defined in a number of ways in different contexts. From the biologists' point of view evolution can be defined as the development of a biological form from other pre-existing forms or its origin to the current existing form through natural selections and modifications i.e. change across successive generations. Natural selection is said to be the driving force behind evolution in which "unfit" forms are eliminated through changes of environmental conditions or sexual selection so that only the fittest are selected (Darwinism) to survive. The underlying mechanism of evolution is genetic mutation that occur spontaneously on the genetic material. These mutations help in providing the variability of the individuals within a population to

survive successfully in a given environment. Genetic diversity thus provides the source of raw material for the natural selection to act on.

Phylogenetic analysis

Computational phylogenetic analysis, an important bioinformatics tool is considered to be a highly reliable tool for studying the molecular evolution. Its importance lies in its simple manifestation and easy handling of data. Representation of the evolution by a simple tree makes the phylogenetic analysis easy to represent and comprehend as well. Its various applications makes this tool an important necessity in different fields of biology. It is a line drawing which provides a visual means of representation to a group of species or sequences and indicates their time of origin. It is a two dimensional representation of relatedness among a varied number of biological species. Phylogenetic tree is represented in three forms: phylogram, dendrogram, and cladogram [Roy et al., 2014].

Merits and demerits of tree building methods

For building up a phylogenetic tree, mainly two methods are used which include distance based methods and character based methods. The most commonly used distance based methods include UPGMA (unweighted paired group method with arithmetic mean) [Murtagh, 1984], NJ (neighbor joining) [Saitou and Nei, 1987], ME (minimum evolution method) [Rzhetsky and Nei, 1993], and FM (Fitch-Margoliash method) [Fitch and Margoliash, 1967]. The most commonly used character based method includes MP (Maximum Parsimony) method [Sober, 1983] and ML (Maximum Likelihood) method [Felsenstein, 1981]. The character based methods derive trees that optimize the

distribution of the actual data pattern for each character. The different established tree building methods can be compared based on some important criteria such as computational speed, consistency of estimated topology, statistical consistency of phylogenetic tree, probability of obtaining the correct topology, and reliability of estimated branch length. Different algorithms are designed for the tree building methods on which various criterions become dependent, e.g. the computational speed. According to this criterion (i.e., computational speed), the NJ method is superior over other tree-building methods which are currently in use as this method can handle a large number of sequences with bootstrap tests with ease unlike MP, ME, and ML methods that examine all possible topologies searching for the individual trees. When the number of input sequences is high, the possible number of topologies increases sharply which becomes cumbersome in using these methods. A tree-building method is considered as a "consistent estimator" if the method tends to give the correct topology as the number of experimental sequence tends to infinity. Simplified different algorithms have been developed for different methods. In the case of ME, these simplified advanced algorithms become efficient in the frame of timescale for obtaining the correct tree and also for MP methods the branch and bound method is often used when number of sequences is relatively high. If no bias is applied during the estimation of distance through substitution NJ and ME methods have been found to be consistent for estimating trees but MP is often inconsistent. On the other hand the ML methods have the additional advantage of being more flexible in choosing the evolutionary model but this method is lengthy and time consuming [Tamura et al., 2013]

Modern trends in phylogenetics

The acquisition of large multilocus sequence data sets have left the researchers with an unprecedented amount of information in resolving these difficult problems. The modern trends in phylogenetics focus specifically on the topics of multiple sequence alignment and methods of tree reconstruction. More sophisticated algorithms for analysis of the heterogeneous data sets have become necessary as the traditional methods are inadequate for analysis. With the third generation sequencing technology rapidly approaching, it will become more feasible to large multilocus data sets to infer evolutionary relationships. Following the construction of an MSA for the traditional 2-step MSA phylogeny estimation procedure, the researcher is left with the decision of how to handle the gaps inserted into the data set by the MSA algorithm to account for INDEL events. Again for the most traditional MP analyses, gaps have been either coded as missing data (most cases) or coded as a fifth character state. Both of these methods are potentially problematic in that the former completely discards relevant evolutionary information, whereas the latter assumes that gaps represent independent evolutionary events which is a highly unlikely scenario. These issues also extend into probabilistic phylogenetic inference where in parameters are estimated without taking indel events into account. An alternative to constructing an MSA prior to phylogenetic inference is to use direct optimization (DO)procedures that are different from other approaches where in the alignment and the phylogenetic tree are estimated simultaneously. Optimization can be performed either under parsimony or under a probabilistic framework.

The program POY [Wheeler et al., 2014], for example, estimates both the phylogenetic tree and the best alignment based on the MP criterion. Previous versions of POY were

also able to implement DO in a likelihood framework. Newer programs such as Stat-Align, BAli-Phy, and BEAST incorporate models of sequence evolution to estimate the posterior distribution of a set of trees and alignments based on Bayesian inference (BI). The Bali-Phy software allows for nested or overlapping indel events, whereas other methods utilize the more common TKF1 and TKF2 indel models. The multi-locus data sets become the norm across laboratories. As highly heterogeneous data sets become available, testing the accuracy of both modern alignment algorithms and DO methods through simulation will become even more important. Furthermore, model-based concatenation methods using mixture models in Bayes Phylogenies seem promising for multi-locus data sets. For traditional phylogenetic inference, MP analysis will no doubt continue to play a role. In this regard, TNT (Tree Analysis Using New Technology) is showing promise for dealing with difficult phylogenetic problems. However, there have been few simulations to quantify the accuracy of the model compared with other methods including direct species tree inference [Holder and Lewis, 2003].

Computational alanine scanning mutagenesis (CASM)

The interactions in protein interfaces can be inhibited if knowledge about affinity and specificity in protein interfaces is known. A powerful tool for analyzing interactions in protein interfaces is experimental alanine scanning mutagenesis (ASM) [De Genst et al., 2002] that helps to identify hotspots of protein-protein interactions. As it is well known that the protein-protein complex formation depends, in most cases, on only a few interface residues, called “hotspots“ [Bogan and Thorn, 1998] that account for the highest contributions to the binding free energy [Clackson and Wells, 1995; Ofran and Rost,

2007]. Even though ASM cannot be applied easily to high throughput screening of protein-protein interfaces, it still represents a very important experimental effort. For this reason, CASM have been developed, which calculate the change in binding free energy ($\Delta\Delta G$) of a protein-protein complex after mutation of an amino acid residue with alanine. To complement the experiments and to enhance the understanding of protein stability, computational prediction methods are used. Several methods of alanine scanning such as Robetta [Kortemme and Baker, 2002], FoldEF [Guerois et al., 2002] and knowledge-based methods (K-FADE, K-CON) [Darnell et al., 2007; Darnell et al., 2008] have been proposed. Thermodynamic simulations are mostly used to estimate the free energy of association. There is still a large discrepancy between the predicted values and experimentally measured free energy changes, although these methods include energy terms which are important for protein stability. Recently, a knowledge-based model was introduced to predict binding "hot spots" [Darnell et al., 2007; Darnell et al., 2008], but the prediction accuracy was relatively low. The Rosetta fragment based approach has also been extended to study protein-protein ininteractions based on sequence information and homology to proteins of known structure combined with a *de novo* protocol for non-homologous portions of the protein. Rosetta creates protein structures as well as energies for wild-type proteins and alanine screens of protein complexes with available sequence information [Simons et al., 1997; Chivian et al., 2003]. The empirical algorithm FoldEF carries out free-energy calculations based on 3D structural data using experimental terms for interactions and weighting factors for energy terms derived from experimental data sets that can be used for computational alanine screening (CAS) by comparing interaction energies of wild-type and mutated complexes [Guerois et al., 2002; Schymkowitz et al.,

2005]. In the MM-GBSA approach interaction free energies of individual residues in a protein-protein complex are estimated by calculating gas-phase energies, solvation free energies, and entropic contributions for the free proteins and the complex derived from selected snap shots of the trajectories [Huo et al., 2002; Simonson et al., 2002; Gohlke et al., 2003; Benedix et al., 2009]. As conserved residues have generally restricted flexibility in the unbound state, molecular dynamics simulations have been used to detect and characterize hotspots [Rajamani et al., 2004; Yogurtcu et al., 2008]. Further efforts has been made to identify correlations between binding hot spots and protein structure along with the sequence information [Hu et al., 2000; Ma et al., 2003; Halperin et al., 2004]. These methods consider that structurally conserved residues are strongly correlated with interaction hot spots.