HOMOLOGY AND MOLECULAR DOCKING STUDIES OF GLYCOSYLTRANSFERASE6 DOMAIN1(GLT6D1)

Project report submitted

In partial fulfilment of the requirement for the award of

P. G. DIPLOMA IN BIOINFORMATICS

Dissertation submitted by

ZAHERA FATHIMA KHATOON

094224010028

DIVISION OF BIOINFORMATICS

PGRRCDE

OSMANIA UNIVERSITY

HYDERABAD

2021-22

Dr. Someswar R. Sagurthi

Assistant Professor,
Course Coordinator,
G. Diploma in Bioinformatics

P.

Department of Genetics
Osmania University,
Hyderabad-500007. TS, India
Mobile: +91-7729822608
Email:drsomeswar@osmania.ac.in
drsomulab@gmail.com

TO WHOM SO EVER IT MAY CONCERN

This is to certify that the project work entitled "**HOMOLOGY AND MOLECULAR DOCKING STUDIES OF GLYCOSYLTRANSFERASE6 DOMAIN1(GLT6D1)"** submitted for **P. G. Diploma in Bioinformatics, PGRRCDE, Osmania University, Hyderabad** is a Bonafide work carried out by **ZAHERA FATHIMA KHATOON** Roll No. **094224010028** under my guidance **from 2021 to 2022.** As per my knowledge no part of this project work has so far been submitted anywhere for any other degree/diploma. His conduct during this time has been satisfactory.

(Dr. Someswar R Sagurthi)

SELF DECLARATION

The project entitled "**HOMOLOGY AND MOLECULAR DOCKING STUDIES OF GLYCOSYLTRANSFERASE6 DOMAIN1(GLT6D1)**" has been carried out by me under the supervision of **Dr. Someswar R. Sagurthi**, Course Coordinator, Assistant Professor, Department of Genetics, Osmania University, Hyderabad towards the partial fulfilment for the award of P. G. Diploma in Bioinformatics from PGRRCDE, Osmania University. This project work has not been submitted for any degree/diploma or examination in any other university. All the assistance taken during the course of the project work and sources of literature has been duly acknowledged. There is no plagiarism in my entire thesis.

Date:19.08.2022

Place:Osmania university,
(Hyderabad)

Zahera Fathima Khatoon

INDEX:

Abstract:

Glycosyltransferases(GTs) protein are key enzymes in the biosynthesis of ginsenosides, which can catalyse the transfer of glycans from donor molecules to acceptor molecules, and form a variety of biologically active glycoside compounds. Glycosyltransferases are useful synthetic tools for the preparation of natural oligosaccharides, glycoconjugates and their analogues.Glycosylation is an important posttranslational modification of secretory and membrane proteins in all eukaryotes, catalysed by glycosyltransferases.(1) It catalyse the formation of glycosidic bonds by assisting the transfer of a sugar.The biosynthesis of glycosides is generally catalysed by enzymes glycosyltransferases, which transfer monosaccharides from sugar nucleotide donors onto acceptor substrates, usually in a highly selective regio-, stereo-, and chemoselective manner. Glycosyltransferases are mostly specific for both acceptor and donor substrate and hence have more limited substrate ranges than glycosidases and glycosynthases. However, the advantage of using transferases is their very high selectivity and high yields, since the glycosylation is generally irreversible when coupled to phosphate ester hydrolysis.(2)

Introduction:

Glycosyltransferases (GTs) are ubiquitous in nature and are required for the transfer of sugars to a variety of important biomolecules.(3) This essential enzyme family has been a focus of attention

from both the perspective of a potential drug target and a catalyst for the development of vaccines, biopharmaceuticals and small molecule therapeutics. This review attempts to consolidate the emerging lessons from Leloir (nucleotide-dependent) GT structural biology studies and recent applications of these fundamentals toward rational engineering of glycosylation catalysts.(4) Glycosyltransferases are involved in the biosynthesis of cell-wall polysaccharides, the addition of N-linked glycans to glycoproteins, and the attachment of sugar moieties to various small molecules such as hormones and flavonoids. In the past two years, substantial progress has been made in the identification and cloning of genes that encode glycosyltransferases. Moreover, analysis of the recently completed Arabidopsis genome sequence indicates the existence of several hundred additional genes encoding putative glycosyltransferases.(5)

Membrane-associated GT-B glycosyltransferases (GTs) comprise a large family of enzymes that catalyse the transfer of a sugar moiety from nucleotide-sugar donors to a wide range of membrane-associated acceptor substrates, mostly in the form of lipids and proteins.(6) As a consequence, they generate a significant and diverse amount of glycoconjugates in biological membranes, which are particularly important in cell-cell, cell-matrix and host-pathogen recognition events. Membrane-associated GT-B enzymes display two "Rossmann-fold" domains separated by a deep cleft that includes the catalytic centre. They associate permanently or temporarily to the phospholipid bilayer by a combination of hydrophobic and electrostatic interactions.(7) They have the remarkable property to access both hydrophobic and hydrophilic substrates that reside within chemically distinct environments catalysing their enzymatic transformations in an efficient manner. Here, we discuss the considerable progress that has been made in recent years in understanding the molecular mechanism that governs substrate and membrane recognition, and the impact of the conformational transitions undergone by these GTs during the catalytic cycle.(8)

Glycosyltransferases catalyse glycosidic bond formation using sugar donors containing a nucleoside phosphate or a lipid phosphate leaving group. Only two structural folds, GT-A and GT-B, have been identified for the nucleotide sugar-dependent enzymes, but other folds are now appearing for the soluble domains of lipid phosphosugar-dependent glycosyltransferases. Structural and kinetic studies have provided new insights. Inverting glycosyltransferases utilise a direct displacement S(N)2-like mechanism involving an enzymatic base catalyst. Leaving group departure in GT-A fold enzymes is typically facilitated via a coordinated divalent cation, whereas GT-B fold enzymes instead use positively charged side chains and or hydroxyls and helix dipoles. The mechanism of retaining glycosyltransferases is less clear.(32)The expected two-step double-displacement mechanism is rendered less likely by the lack of conserved architecture in the region where a catalytic nucleophile would be expected. A mechanism involving a short-lived oxocarbenium ion intermediate now seems the most likely, with the leaving phosphate serving as the base.(9)

Protein glycosylation is an essential covalent modification involved in protein secretion, stability, binding, folding, and activity.(34,35) One or more sugars may be O-, N-, S-, or C-linked to specific amino acids by glycosyltransferases, which catalyse the transfer of these sugars from a phosphate-containing carrier molecule. Most glycosyltransferases are members of the GT-A, GT-B, or GT-C structural superfamilies. GT-C enzymes are integral membrane proteins that utilise a phospho-isoprenoid carrier for sugar transfer.(36)

To-date, two families of GT-Cs involved in protein glycosylation have been structurally characterised: the family represented by PglB, AglB, and Stt3, which catalyses oligosaccharide transfer to Asn, and the family represented by Pmt1 and Pmt2, which catalyses mannose transfer to Thr or Ser.(10,11)

In eukaryotic protein N-glycosylation, a series of glycosyltransferases catalyse the biosynthesis of a dolichylpyrophosphate-linked oligosaccharide before its transfer onto acceptor proteins. The final seven steps occur in the lumen of the endoplasmic reticulum (ER) and require dolichyl phosphate-activated mannose and glucose as donor substrates. The responsible enzymes-ALG3, ALG9, ALG12, ALG6, ALG8 and ALG10-are glycosyltransferases of the C-superfamily (GT-Cs), which are loosely defined as containing membrane-spanning helices and processing an isoprenoid-linked carbohydrate donor substrate. Here we present the cryo-electron microscopy structure of yeast ALG6 at 3.0 Å resolution, which reveals a previously undescribed transmembrane protein fold. Comparison with reported GT-C structures suggests that GT-C enzymes contain a modular architecture with a conserved module and a variable module, each with distinct functional roles. We used synthetic analogues of dolichyl phosphate-linked and dolichyl pyrophosphate-linked sugars and enzymatic glycan extension to generate donor and acceptor substrates using purified enzymes of the ALG pathway to recapitulate the activity of ALG6 in vitro. A second cryo-electron microscopy structure of ALG6 bound to an analogue of dolichyl phosphate-glucose at 3.9 Å resolution revealed the active site of the enzyme.(39) Functional analysis of ALG6 variants identified a catalytic aspartate residue that probably acts as a general base. This residue is conserved in the GT-C superfamily. Our results define the architecture of ER-luminal GT-C enzymes and provide a structural basis for understanding their catalytic mechanisms.(12)

About disease:

PERIODONTITIS (DISEASE): Periodontal diseases are a group of infectious diseases that mainly include gingivitis and periodontitis. Gingivitis is the most prevalent form of periodontal disease in subjects of all ages, including children and adolescents. Less frequent types of periodontal disease include aggressive periodontitis, acute necrotizing ulcerative gingivitis and various diseases of herpesviral and fungal origin.(61) This review aimed to retrieve relevant information from Latin America on the prevalence of periodontal diseases among children and adolescents of the region. Gingivitis was detected in 35% of young Latin American subjects and showed the highest frequencies in Colombia (77%) and Bolivia (73%) and the lowest frequency in Mexico (23%). The frequency of gingivitis in subjects from other Latin American countries was between 31% and 56%. Periodontitis may affect <10% of the young population in Latin America, but the data are based on only a few studies. A more precise assessment of the distribution and severity of periodontal disease in children and adolescents of Latin America may help policy makers and dentists to institute more effective public health measures to prevent and treat the disease at an early age to avoid major damage to the permanent dentition.(62)

Gingivitis and periodontitis are associated with a negative impact on Oral Health Related Quality of Life , exerting a significant influence on aspects related to the patient's function and esthetics. Periodontitis has been associated with several systemic conditions, including adverse pregnancy

outcomes, cardiovascular diseases, type 2 diabetes mellitus (DM), respiratory disorders, fatal pneumonia in hemodialysis patients, chronic renal disease and metabolic syndrome. The aim of this paper was to review the results of different periodontal treatments and their impacts on patients' Oral Health Related Quality of Life and systemic health. Non-surgical and surgical periodontal treatments are predictable procedures in terms of controlling infection, reducing probing pocket depth and gaining clinical attachment. In addition, the treatment of periodontitis may significantly improve and promote a reduction in the levels of systemic markers of inflammation, including some cytokines associated with cardiovascular diseases.(63)

Physical measurements including the evaluation of probing depth, bleeding on probing, tooth mobility, and inflammation form the basis for most periodontal diagnostics in use today. The interpretation of these observations and the methods available for their measurement, however, have begun to change significantly. The episodic disease activity concept has done much to implement these changes. Observation of episodic attachment loss has been correlated with parallel radiographic changes, alteration in levels of probable pathogens, and changes in inflammatory mediator levels. The failure of pocket depth, suppuration, and bleeding on probing to predict episodic attachment loss has been given plausible explanations and enhanced meanings. Although attachment loss by a continuous process cannot be excluded in some disease conditions, the hypothesis of periodontal disease progression by episodic activity supplements and expands understanding of the disease process(64). Interest in periodontal diagnostics has accelerated in the last decade. As a parallel development, the technology of small computers has decreased in cost and increased in sophistication. The combination of these factors has created an environment for the development of intelligent diagnostic systems. Four commercially available systems and two systems under development are described. The systems, which measure pocket depth, pocket depth or attachment level, tooth mobility, and pocket temperature, all utilise computer processing of measurements. The result is to provide a simplified and more meaningful presentation of diagnostic information. As intelligent diagnostic systems prove themselves, some of these instruments are likely to become common to dental practice. The promise of more accurate identification of areas of the mouth that are diseased can increase both the efficiency and effectiveness of periodontal therapy.(64)

About Protein:

Glycosyltransferases are specific for the type of linkage (α or β), and the linkage position of the glycoside bond formed [e.g. α(1→3) or β(1→4)]. Glycosyltransferases were initially considered to be specific for a single glycosyl donor and acceptor, which led to the "one enzyme-one linkage" concept.[41] Subsequent observations have refuted the theory of absolute enzymatic specificity by describing the transfer of analogs of some nucleoside mono- or diphosphate sugar donors.(41)

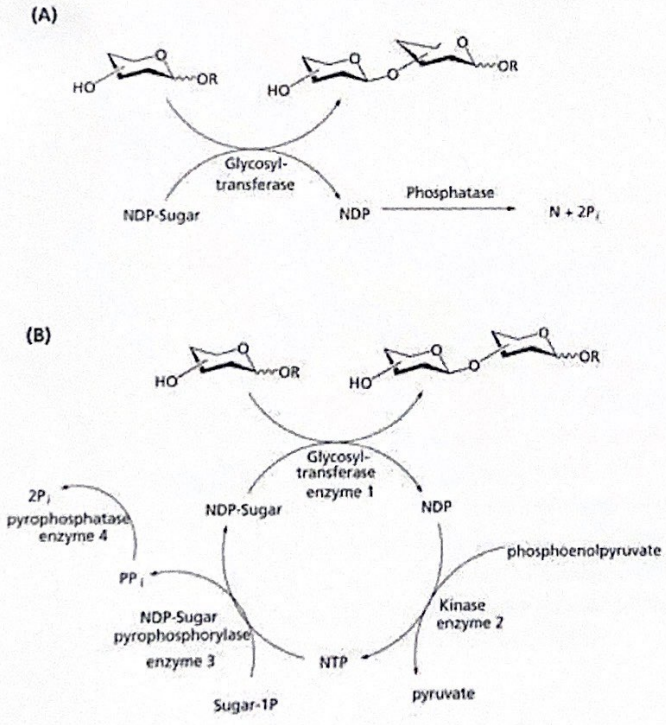
Glycosyltransferases can tolerate modifications to the acceptor sugar, as long as the acceptor meets specific structural requirements, e.g., appropriate stereochemistry and availability of the reactive hydroxyl group involved in the glycosidic bond.(42)

In contrast to organic chemical synthesis, enzymatic glycosylation has potential for application use within biological systems, where the modification of glycosylation sites may be used to investigate the regulation of cell signaling processes.(43)

Various application strategies for glycosyltransferases have employed an assortment of glycosyl donors and reaction conditions for the synthesis of carbohydrates and the glycosylation of natural products.(44)

A major limitation to enzyme-catalyzed glycosylation reactions is the glycosyltransferase inhibition caused by nucleotide diphosphates generated during the reaction. (45)Two strategies have been identified to prevent enzymatic inhibition (**Figure 1**)(46,49)

1. Phosphatase is added to the reaction to degrade the nucleotide diphosphates by removal of the phosphate group (**Figure 1A**).(50)
2. Nucleotide diphosphates are recycled to the appropriate nucleotide triphosphates by employing multi-enzyme regeneration schemes.(51,55) Although several different enzymes and cofactors are involved in these regeneration schemes, the method avoids the use of stoichiometric amounts of sugar nucleotides **(Figure 1B).(56,59)**



Crystal structures of Glycosyltransferases:

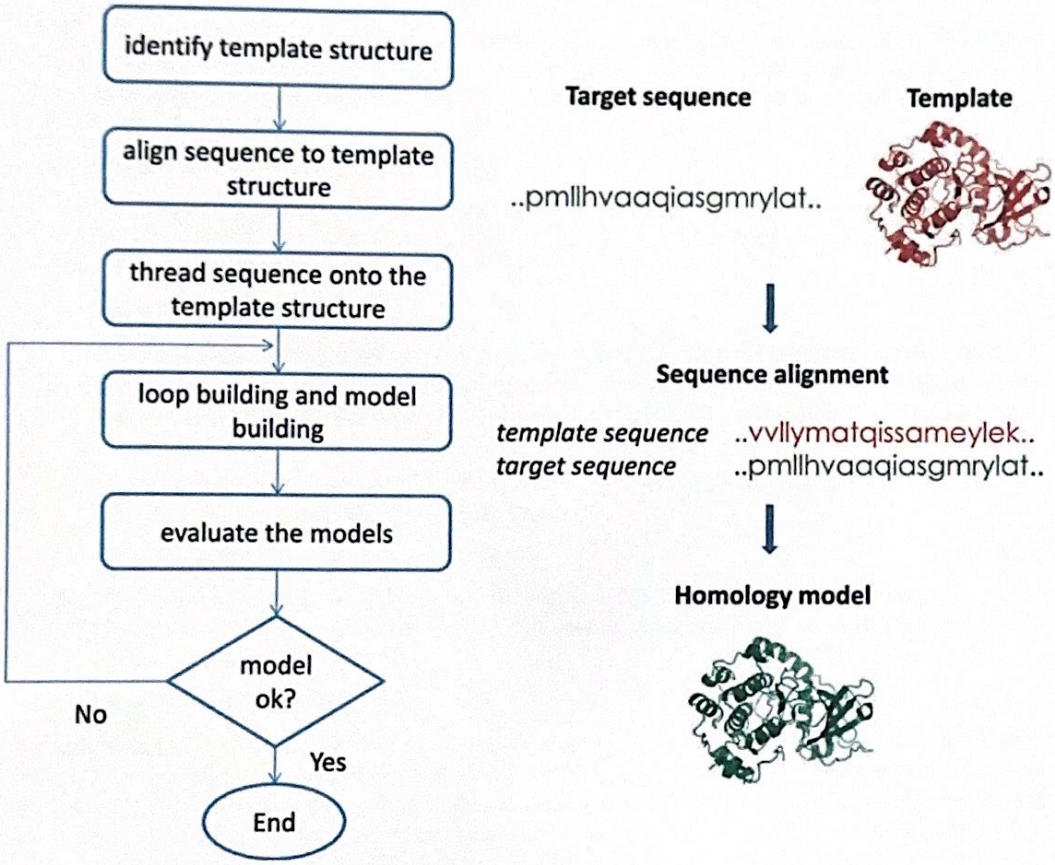
The first X-ray structure was reported in 1994 for bacteriophage T4-glucosyltransferase, an enzyme that transfers glucose from UDP-Glc to phage-modified DNA . Since then, >100 crystal structures have been

described for proteins corresponding to 23 different GTs, from prokaryotes and eukaryotes. Structural information is now available for 17 distinct GT families, including both retaining and inverting enzymes. In contrast to glycosylhydrolases that adopt a large variety of folds, including all α, all β, or mixed α/β structures, GT folds have been observed to consist primarily of α/β/α sandwiches similar or very close to the Rossmann-type fold, a classical structural motif (six-stranded parallel β-sheet with 321456 topology) found in many nucleotide-binding proteins. Until recently, only two structural superfamilies have been described for GTs, named GT-A and GT-B, and which were first observed in the original SpsA and β-glucosyltransferase (BGT) structures. A third family has recently emerged which comprises the bacterial sialyltransferase (CstII) . This protein displays a similar type of fold than GT-A, but with some differences, so it can be considered as a new fold. The GT-A and GT-B folds are also shared by non-GT enzymes, such as nucleotidyl transferases and sugar epimerases(60)

HOMOLOGY MODELLING:

Homology modelling, an important technique has matured in structural biology, significantly contributing to narrowing the gap between known protein sequences and experimentally determined structures. Fully automated workflows and servers simplify and streamline the homology modelling process, also allowing users without a specific computational expertise to generate reliable protein models of modelling results and have easy to understand results, their visualisation and interpretation. Homology modelling is based on the observation that related protein sequences adopt similar three-dimensional structures. Therefore, the homology model of a protein can be derived using related protein structures as modelling templates.

Homology modelling is one of the computational structure prediction methods that are used to determine protein 3D structure from its amino acid sequence. It is considered to be the most accurate of the computational structure prediction methods. It consists of multiple steps that are straightforward and easy to apply. There are many tools and servers that are used for homology modelling. There is no single modelling program or server which is superior in every aspect to others. Since the functionality of the model depends on the quality of the generated protein 3D structure, maximising the quality of homology modelling is crucial. Homology modelling has many applications in the drug discovery process. Since drugs interact with receptors that consist mainly of proteins, protein 3D structure determination, and thus homology modelling is important in drug discovery. Accordingly, there has been the clarification of protein interactions using 3D structures of proteins that are built with homology modelling. It depends on programs such as BLAST to search for similar proteins in various databases, structural or otherwise, such as the Protein Data Bank (PDB). Homology modelling is also called "comparative modelling," because you're comparing the model structure with known template structures as you build it.(1)



1. Target Sequence Selection

The protein sequence we wish to model is termed the "target sequence.In some cases, we need to model an entire protein. And in such cases, sticking with the essential protein sequence/domains will save you work and speed things up.(9)

2. Template Protein Recognition

The sequence of similarity can be searched using BLAST or Psi blast or fold recognition methods and align with the known structures in PDB. PDB, which is the largest database, contains only experimentally resolved structures. BLAST allows comparing a query sequence with a database such as PDB and identifying the best sequence which shares a high degree of similarity. (10)The sequence of similarity of each line is summarised with its E-value (Expected value) which is closer to zero, and has a high degree of similarity. The E-value describes the number of hits one can "expect" when searching through a database of a particular size. The sequences which fall under the safe zone are expected to be getting a better structure than the twilight zone and midnight zone. After identifying one or more possible templates, alignment correction is performed. Sometimes it is difficult to align two sequences that have percentage identity. Such cases, one can use other sequences from homologous proteins to solve this problem. Multiple Sequence Alignment programs such as CLUSTALX align sequences by insertions and deletions. Alignment correction is the critical step in homology modelling, otherwise

which in turn creates a defective model. The backbone generation from the aligned regions can be done using modelling tools such as Modeller. The actual experimentally determined structures contain manual errors due to poor electron density. Therefore a good model has to be chosen with less number of errors.(12)

3. Preparation of Template Protein

To trim back the template proteins because the experimental structures will contain extraneous matter. For example, symmetry equivalent protein chains, water molecules, ligands, and solvent of crystallisation.

4. Sequence Alignment

Align the target and template protein sequences using a sequence alignment tool such as CLUSTALX. This is a very important step in homology modelling because using an appropriate alignment algorithm is necessary for bagging the most valid template structures.

The alignment compares all of the proteins, target, and templates, and it tells us which parts have completely conserved amino acid sequences.

5. Prediction of Secondary Structure

The secondary structure of the model is built. It compares the proposed secondary structures of your target to those of the template proteins and ranks them to iteratively build up the model.

6. Side Chain Modelling

Proteins that are structurally similar, have similar torsion angle about Ca-Cb bond (psi angle) when comparing with side chain conformations. In such cases, copying conserved residues entirely from the template to the model will result in higher accuracy than copying the backbone or re-predicting side chains. Side chain conformations are partially knowledge based which uses libraries of rotamers extracted from high resolution X ray structures. To build a position-specific rotamer library, one can take high-resolution protein structures and collect all stretches of three to seven residues (method dependent) with a given amino acid at the centre. Prediction accuracy is usually quite high for residues in the hydrophobic core, where more than 90% of all psi angles fall with 20o of experimental values, it is much lower for surface residues, where the percentage is often lower than 50%.(13)

There are two reasons for this:

1. Flexible side chains on the surface tend to adopt multiple conformations, which are additionally influenced by crystal contacts.
2. Energy functions used to score rotamers can easily handle hydrophobic packing in the core (Van der Waals interactions), but are not accurate enough to get complicated electrostatic interactions on the surface.

7. Loop Modeling

Regions of the target sequence that are not aligned to a template are modelled by loop modelling; they are the most susceptible to major modelling errors and occur with higher

frequency when the target and template have low sequence identity. The coordinates of unmatched sections determined by loop modelling programs are generally much less accurate than those obtained from simply copying the coordinates of a known structure, particularly if the loop is longer than 10 residues. The first two sidechain dihedral angles (χ_1 and χ_2) can usually be estimated within 30° for an accurate backbone structure; however, the later dihedral angles found in longer side chains such as lysine and arginine are notoriously difficult to predict. Moreover, small errors in χ_1 (and, to a lesser extent, in χ_2) can cause relatively large errors in the positions of the atoms at the terminus of side chain; such atoms often have a functional importance, particularly when located near the active site.(14)

In most cases, alignment between model and template sequence contains gaps. By means of insertions and deletions with some conformational changes to the backbone it can be modelled, although it rarely happens to secondary structures. So it is safe to shift the insertion and deletions of the alignment, out of helices or strands and placing them in loops or coils. But this loop conformational change is difficult to predict due to many reasons like, Surface loops tend to be involved in crystal contacts, leading to a significant conformational change between template and target.(15)

The interchange of the side chains can lead to change in the orientation and spatial arrangement especially when it is an interchange between small and a bulky group.

Proline and glycine are an exception when a "**Ramachandran plot"** is considered. Proline has a restriction in the plot due to its 5 membered ring whereas glycine has a hydrogen atom as its side chain which is very difficult to predict from the plot. This makes it difficult for detect mutations that have happened to loop residue from/to either glycine or proline.(16)
There are two main ways to overcome this and model the loop region:

1. Knowledge based:
Users can search PDB for known loops with endpoints that match the residues between loops that have to be inserted and simply copy the loop conformation.

2. Energy based:
The quality of a loop is determined with energy function and minimizes the function using Monte Carlo or molecular dynamics to find the best loop conformation.

8. Model Optimization and Validation

Finally, once we have an almost complete model, we need to improve it to attain a near-native confirmation via energy minimization. Such validation tests show whether the protein model is energetically satisfactory based on the spread of conformations observed in experimental structures for any given fold or feature. Sometimes the rotamers are predicted based on incorrect backbone or incorrect prediction.(17) Such cases modelling programs either restrain the atom positions,or apply only a few hundred steps of energy minimization to get an accurate value. This accuracy can be achieved in 2 ways.

1. Quantum force field: To handle large molecules efficiently force field can be used, energies are therefore normally expressed as a function of the positions of the atomic nuclei only. Van der Waals forces are, for example, so difficult to treat, that they must often be completely omitted. While providing more accurate electrostatics, the overall precision achieved is still about the same as in the classical force fields.(18)

2. Self-parametrizing force fields: The precision of a force field depends to a large extent on its parameters (e.g., Van der Waals radii, atomic charges). These parameters are usually obtained from quantum chemical calculations on small molecules and fitting to experimental data, following elaborate rules (Wang, Cieplak, and Kollman,2000). By applying the force field to proteins, one implicitly assumes that a peptide chain is just the sum of its individual small molecule building blocks—the amino acids. To increase the precision of the force field, the following steps can be used. Take initial parameters (for example, from an existing force field), change a parameter randomly, energy minimise models, see if the result improved, keep the new force field if yes, otherwise go back to the previous force field.(19)

Validation:

The models we obtain may contain errors. These errors mainly depend upon two values.

1. The percentage identity between the template and the target.

If the value is > 90% then accuracy can be compared to crystallography, except for a few individual side chains. If its value ranges between 50-90 % r.m.s.d. error can be as large as 1.5 Å, with considerably more errors. If the value is <25% the alignment turns out to be difficult for homology modelling, often leading to quite larger errors.(20)

2. The number of errors in the template.

Errors in a model become less of a problem if they can be localised. Therefore, an essential step in the homology modelling process is the verification of the model. The errors can be estimated by calculating the model's energy based on a force field.(21)This method checks to see if the bond lengths and angles are in a normal range. However, this method cannot judge if the model is correctly folded. The 3D distribution functions can also easily identify misfolded proteins and are good indicators of local model building problems.(22)

DRUG DESIGNING:

Drug design is an integrated developing discipline which portends an era of 'tailored drug'. It involves the study of effects of biologically active compounds on the basis of molecular interactions in terms of molecular structure or its physico-chemical properties involved. It studies the processes by which the drug produces their effects, how they react with the protoplasm to elicit a particular pharmacological effect or response, how they are modified or detoxified, metabolised or eliminated by the organism.

Disposition of drugs in individual regions of biosystems is one of the main factors determining the place , mode and intensity of their action . The biological activity may be "positive" as in drug design or "negative" as in toxicology. Thus drug design involves either total innovation of lead or an optimization of already available lead. These concepts are the building stones up on which the edifice of drug design is built up.

The drug is most commonly an organic small molecule that activates or inhibits the function of a bio molecule such as a protein, which in turn results in a therapeutic benefit to the patient. In the most basic sense, drug design involves the design of small molecules that are complementary in shape and charge to the biomolecular target with which they interact and therefore will bind to it. Drug design frequently but not necessarily relies on computer modelling techniques. This type of modelling is often referred to as computer-aided drug design. Finally, drug design that relies on the knowledge of the three-dimensional structure of the biomolecular target is known as structure-based drug design.

APPROACHES FOR DRUG DESIGNING:

The various approaches used in drug design include the following..

1) Random screening of synthetic compounds or chemicals and natural products by bioassay procedures.

2) Novel compound preparation based on the known structures of biologically active, natural substances of plant and animal origin, i.e., lead skeleton.

3) Preparation of structural analogs of lead with increasing biological activity and

4) Application of bioisosteric principle.

The current trend in drug design is to develop new clinically effective agents through the structural modification of lead nucleus. The lead is a prototype compound that has the desired biological or pharmacological activity but may have many undesirable characteristics, like high toxicity, other biological activity, insolubility or metabolism problems. Such organic leads, once identified, are easy to exploit. This process is rather straightforward. The real test resides with the identification of such lead real test resides with the identification of such lead bioactive positions on the basic skeleton of such leads.

Computer-aided drug design

Computer-aided drug design uses computational chemistry to discover, enhance, or study drugs and related biologically active molecules. The most fundamental goal is to predict whether a given molecule will bind to a target and if so how strongly. Molecular mechanics or molecular dynamics are most often used to predict the conformation of the small molecule and to model conformational changes in the biological target that may occur when the small molecule binds to it.
Semi-empirical, ab initio quantum chemistry methods, or density functional theory are often used to provide optimised parameters for the molecular mechanics calculations and also provide an estimate of the electronic properties (electrostatic potential, polarizability, etc.) of the drug candidate that will influence binding affinity.

Molecular mechanics methods may also be used to provide semi-quantitative prediction of the binding affinity. Also, knowledge-based scoring functions may be used to provide binding affinity

estimates. These methods use linear regression, machine learning, neural nets or other statistical techniques to derive predictive binding affinity equations by fitting experimental affinities to computationally derived interaction energies between the small molecule and the target.[15][16]

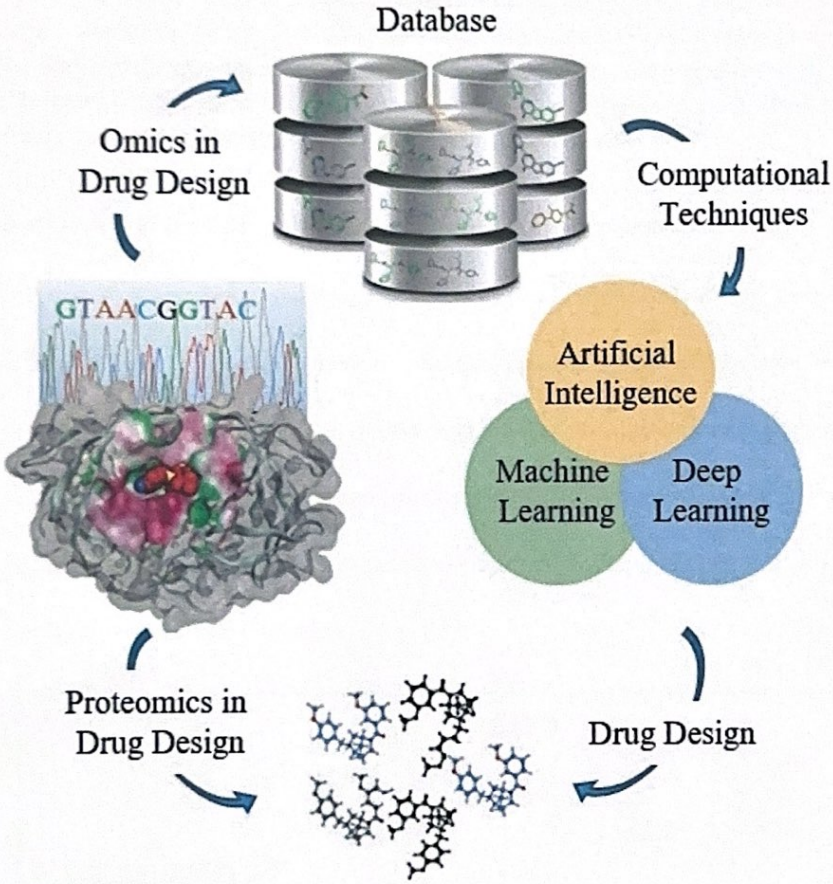
Ideally the computational method should be able to predict affinity before a compound is synthesised and hence in theory only one compound needs to be synthesised. The reality however is that present computational methods are imperfect and provide at best only qualitatively accurate estimates of affinity. Therefore in practice it still takes several iterations of design, synthesis, and testing before an optimal molecule is discovered. On the other hand, computational methods have accelerated discovery by reducing the number of iterations required and in addition have often provided more novel small molecule structures.

Drug design with the help of computers may be used at any of the following stages of drug discovery:

1. hit identification using virtual screening(structure- or ligand-based design)
2. hit-to-lead optimization of affinity and selectivity (structure-based design, QSAR, etc.)
3. lead optimization optimization of other pharmaceutical properties while maintaining affinity

Structure-based drug design:

Structure-based drug design is becoming an essential tool for faster and more cost-efficient lead discovery relative to the traditional method. Genomic, proteomic, and structural studies have provided hundreds of new targets and opportunities for future drug discovery. This situation poses a major problem: the necessity to handle the "big data" generated by combinatorial chemistry. Artificial intelligence and deep learning play a pivotal role in the analysis and systemization of larger data sets by statistical machine learning methods. Advanced AI-based sophisticated machine learning tools have a significant impact on the drug discovery process including medicinal chemistry. In this review, we focus on the currently available methods and algorithms for structure-based drug design including virtual screening and de novo drug design, with a special emphasis on AI- and deep-learning-based methods used for drug discovery.



MOLECULAR DOCKING:

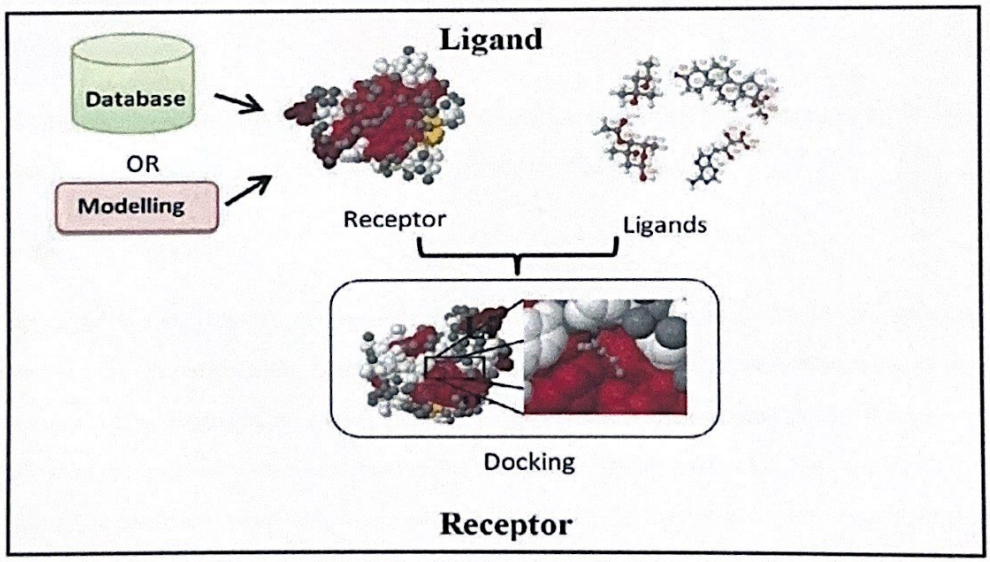
The docking process involves two basic steps: prediction of the ligand conformation as well as its position and orientation within these sites (usually referred to as pose) and assessment of the binding affinity.(65)

Ligand docking is used for docking simulations under the premise that the position of the binding site is already known, and meanwhile, it can also be used without prior knowledge of the binding site. However, most of the optimization search algorithms used in popular docking software are far from being ideal in the first case, and they can hardly be directly utilised for the latter case due to the relatively large search area. In order to design an algorithm that can flexibly adapt to different

sizes of the search area, we propose an effective swarm intelligence optimization algorithm in this paper, called diversity-controlled Lamarckian quantum particle swarm optimization. The highlights of the algorithm are a diversity-controlled strategy and a modified local search method. Integrated with the docking environment of Autodock, is compared with Autodock and other two Autodock-based search algorithms for flexible ligand docking.

Application of molecular modelling in modern drug development

It is used to evaluate potential harms produced by relationships with other proteins, such as proteases, cytochrome P450, and others. Docking can also be used to determine the specificity of a proposed medication against homologous proteins. Additionally, docking is a frequently utilised technique for identifying protein-protein interactions.Comprehension of cellular connections helps in the comprehension of a range of processes occurring in live organisms and the identification of potential pharmaceutical targets.



Types of Docking

Comprehensively utilised docking tools employ search algorithms such as genetic algorithms, fragment-based algorithms, Monte Carlo algorithms and molecular dynamics algorithms. Besides this, there are some tools such as DOCK,which are mainly used for high throughput docking simulations. There are various kinds of molecular docking procedures involving either ligand/target flexible or rigid based upon the objectives of docking simulations like flexible ligand docking (target as rigid molecule), rigid body docking (both the target and ligand as rigid molecules) and flexible docking (both interacting molecules as flexible).

Lead optimization

Molecular docking can predict an *optimised orientation* of ligands on its target. It can predict different binding modes of ligands in the groove of the target molecule. This can be used to develop more potent, selective and efficient drug candidates

Hit identifications

Docking in combination with scoring function can be used to *evaluate large databases* for finding potent drug candidates *in silico*, which can target the molecule of interest.

Drug-DNA interaction

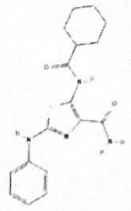
Molecular docking plays a prominent role in the initial prediction of a drug's binding properties to nucleic acid. This information establishes the correlation between a drug's molecular structure and its cytotoxicity. Keeping this in view, medicinal chemists are constantly putting their efforts to elucidate the underlying anticancer mechanism of drugs at molecular level by investigating the interaction mode between nucleic acid and drugs in presence of copper. Medicinal chemists are doing silico observations where their main finding is to predict whether the compound/drug is interacting with the protein/DNA. If the docking programme is predicting the said interaction, then the experimental procedures are made available to find out the real binding mode of the complex.

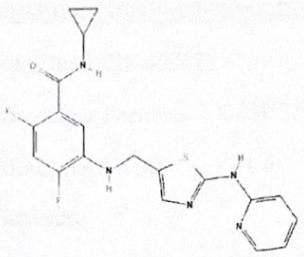
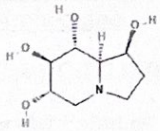
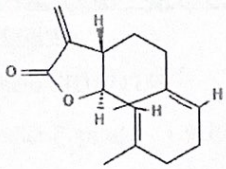
This leads to the development of new anticancer drugs. Furthermore, this knowledge would be instrumental in the detection of those structural modifications in a drug that could result in sequence/structure specific binding to their target

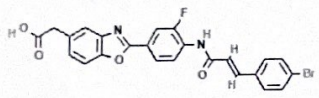
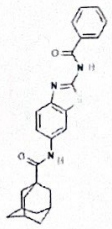
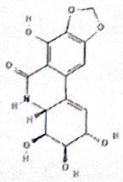
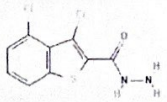
Experimental results revealed that the proposed algorithm has a performance comparable to those of Autodock for dockings within a certain area around the binding sites, and is a more effective solver than all the compared methods for dockings without prior knowledge of the binding sites.(65).

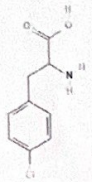
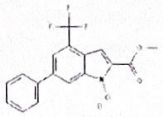
Docking of protein is done with Q7Z4J2(HUMAN) template(1G8O) and prediction of active site is PREDICTED by online website SCFBIO(**Supercomputing Facility for Bioinformatics & Computational Biology)** Docking of protein with another Q4R5T7(MACACA) template(2JCK) and prediction of active site is PREDICTED SCFBIO and with respective coordinates.

LIGANDS:

1)5-(Cyclohexanecarboxamido)-2-(phenylamino)thiazole-4-carboxamide (46355372) PubChem CID46355372 MF: C17H20N4O2S Molecular Weight 344.4 Structure 2D 	2)N-cyclopropyl-2,4-difluoro-5-((2-(pyridin-2-ylamino)thiazol-5-yl)methylamino)benzamide (11632737) PubChem CID11632737 MF: C19H17F2N5OS Molecular Weight 401.4 Structure 2D

	
3)Castanospermine PubChem CID 54445 Molecular Weight 189.21 MF: C8H15NO4 Structures 2D 	4)Costunolide PubChem CID 5281437 Molecular Weight 232.32 Molecular Formula C15H20O2 Structure 2D 

5)(E)-2-(2-(4-(3-(4-bromophenyl)acrylamido)-3-fluorophenyl)benzo[d]oxazol-5-yl)acetic acid PubChem CID44402523 Molecular Formula C24H16BrFN2O4 Molecular Weight 495.3 Structure 2D 	6)N-(2-benzamido-1,3-benzothiazol-6-yl)adamantane-1-carboxamide PubChem CID4096211 Molecular Formula C25H25N3O2S Molecular Weight 431.6 Structure 2D 
7)Narciclasine PubChem CID72376 Molecular Formula C14H13NO7 Molecular Weight 307.25 Structures 2D 	8)3,4-Dichloro-1-benzothiophene-2-carbohydrazide PubChem CID874733 Molecular Formula C9H6Cl2N2OS Molecular Weight 261.13 Structure 2D 

9)Fenclonine	10)Methyl 1-hydroxy-6-phenyl-4-(trifluoromethyl)-1H-indole-2-carboxylate
PubChem CID4652	PubChem CID51355147
Molecular Formula C9H10ClNO2	Molecular Formula C17H12F3NO3
Molecular Weight 199.63	Molecular Weight 335.28
Structure	Structures
2D	2D
	

About Ligands:

1)5-(Cyclohexanecarboxamido)-2-(phenylamino)thiazole-4-carboxamide (46355372)

Cyclohexane appears as a clear colourless liquid with a petroleum-like odour. Used to make nylon, as a solvent, paint remover, and to make other chemicals. Flash point -4°F. Density 6.5 lb / gal (less than water) and insoluble in water. Vapours heavier than air.

Cyclohexane is an alicyclic hydrocarbon comprising a ring of six carbon atoms; the cyclic form of hexane, used as a raw material in the manufacture of nylon. It has a role as a non-polar solvent. It is a cycloalkane and a volatile organic compound.Cyclohexane is a natural product found in Helichrysum odoratissimum, Terminalia chebula, and other organisms

2)N-cyclopropyl-2,4-difluoro-5-((2-(pyridin-2-ylamino)thiazol-5-yl)methylamino)benzamide (11632737)

Cyclopropyl ring in drug development to transition drug candidates from the preclinical to clinical

stage. Important features of the cyclopropane ring are, the (1) coplanarity of the three carbon atoms, (2) relatively shorter (1.51 Å) C–C bonds, (3) enhanced π-character of C–C bonds, and (4) C–H bonds are shorter and stronger than those in alkanes. The present review will focus on the contributions that a cyclopropyl ring makes to the properties of drugs containing it. Consequently, the cyclopropyl ring addresses multiple roadblocks that can occur during drug discovery such as (a) enhancing potency, (b) reducing off-target effects.

3)Castanospermine

Castanospermine inhibits all forms of α- and β-glucosidases, especially glucosidase I (required for glycoprotein processing by transfer of mannose and glucose from asparagine-linked lipids),target α- and β-glucosidases. IC 50: 1.2 uM Castanospermine is a potent and specific inhibitor of mammalian and plant α-and β-D-glucosidases with castanospermine, an inhibitor of the glucosidases that convert protein N-linked high mannose carbohydrates to complex oligosaccharide

4)Costunolide

Costunolide is a germacranolide with anthelmintic, antiparasitic and antiviral activities. It has a role as an anthelmintic drug, an antiinfective agent, an antineoplastic agent, an antiparasitic agent, an antiviral drug and a metabolite. It is a germacranolide and a heterocyclic compound.

5)(E)-2-(2-(4-(3-(4-bromophenyl)acrylamido)-3-fluorophenyl)benzo[d]oxazol-5-yl) acetic acid

Bromophenyl is an organophosphorus compound with the formula $(C_6H_4Br)P(C_6H_5)_2$. It is a white crystalline solid that is soluble in nonpolar organic solvents. The compound is used as a precursor to the 2-lithiated derivative of triphenylphosphine,[1] which in turn is a precursor to other phosphine ligands.

6)N-(2-benzamido-1,3-benzothiazol-6-yl)adamantane-1-carboxamide

Benzamide is a white solid with the chemical formula of $C_6H_5C(O)NH_2$. It is the simplest amide derivative of benzoic acid. It is slightly soluble in water, and soluble in many organic solvents. A number of substituted benzamides are commercial drugs: sulpiride, remoxipride, amisulpride, tiapride, sultopride, veralipride, aminohippuric acid, cisapride, imatinib, and procainamide.

7)Narciclasine

Narciclasine is a member of phenanthridines. It has a role as a metabolite.Narciclasine is a natural product found in Lycoris sanguinea, Lycoris squamigera, and other organisms.Narciclasine (205) and its glucoside 207 showed similar toxic activity against *Artemia salina* with LD50 values of 0.29 and 0.88 μg/ml, respectively.

8)3,4-Dichloro-1-benzothiophene-2-carbohydrazide

Benzothiophene is an aromatic organic compound with a molecular formula C_8H_6S and an odour similar to naphthalene (mothballs). It occurs naturally as a constituent of petroleum-related deposits such as lignite tar. Benzothiophene has no household use. In addition to benzo[b]thiophene, a second isomer is known: benzo[c]thiophene.

9)Fenclonine

Fenclonine, also known as p-chlorophenylalanine, is an inhibitor of tryptophan hydroxylase, the enzyme that plays a rate-limiting role in the biosynthesis of serotonin. Fenclonine was studied for the treatment of carcinoid syndrome, but the psychological side effects prevented further development.

10)Methyl 1-hydroxy-6-phenyl-4-(trifluoromethyl)-1H-indole-2-carboxylate

The trifluoromethyl group is a functional group that has the formula $-CF_3$. The naming of is group is derived from the methyl group (which has the formula $-CH_3$), by replacing each hydrogen atom by a fluorine atom

Interaction of glycotransferase(Q7Z4J2) with LIGANDS

1) Interaction of protein (Q7Z4J2) with 5-(Cyclohexanecarboxamido)-2-(phenylamino)thiazole-4-carboxamide (46355372)