**A**

**Dissertation Project Report**

**On**

# Exploration of Sequence, Structure and Substrate Space for Development of Glycosyltransferase Biocatalyst

**Submitted To**

**Department of Biosciences**
**Integral University, Lucknow**
**For the partial fulfillment for the requirement of the degree**
**MASTER OF SCIENCE**
**IN**
**BIOTECHNOLOGY**
**(2020-2022)**
**BY**

**SHASHANK SINGH**
**M.Sc. Biotechnology (IV semester)**
**Department of Bioscience**
**Integral University, Lucknow**

**UNDER THE SUPERVISION OF**

**DR. KINSHUK RAJ SRIVASTAVA**
**(SENIOR SCIENTIST)**
**DIVISION OF MEDICINAL PROCESSES IN CHEMISTRY**
**CSIR-CENTRAL DRUG RESEARCH INSTITUTE, LUCKNOW**

# CERTIFICATE

सी.एस.आई.आर.-केन्द्रीय औषधि अनुसंधान संस्थान, लखनऊ
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
**CSIR-Central Drug Research Institute, Lucknow**
(Council of Scientific & Industrial Research)
पोस्ट बॉक्स नं. 173, सेक्टर 10, जानकीपुरम विस्तार, सीतापुर रोड, लखनऊ – 226031 (भारत)
Post Box No. 173, Sector 10, Jankipuram Extension, Sitapur Road, Lucknow-226031 (INDIA)
दूरभाष- Phone: +91-522-2772450, 2772550(PABX), फैक्स Fax: +91-0522-2771941,
टी-फैक्स T-Fax: 2771942, 2772793, ग्राम Gram: CENDRUG वेब Web:www.cdriindia.org

This is to certify that the Dissertation report / Thesis entitled **"Exploration of Sequence, Structure and Substrate Space for Development of Glycosyltransferase Biocatalyst"** being submitted to the Integral University, Lucknow in the partial fulfillment of the requirement for the award of Degree of **Master of Science** in **Biotechnology.**

It is an original piece of work carried out by **Mr. Shashank Singh** under the supervision of **Dr. Kinshuk Raj Srivastava at CSIR- Central Drug Research Institute, Lucknow.** The matter embodied in this dissertation report has not been previously submitted in part or full for the award of any degree at any other university or institution. It is further certified that the matter embodied in this report has been checked and passed through a standard plagiarism detection tool.

(Signature with seal)

Supervisor: Dr. Kinshuk Raj Srivastava

Date : 18/07/22

Place : CDRI, Lucknow, U.P., India

# ACKNOWLEDGMENT

The work presented in this project would not have been possible without my close association with many people. I take this opportunity to extend my sincere gratitude and appreciation to all those who made this project possible. I gratefully acknowledge to **Dr D. Srinivasa Reddy, Director, and Dr. Atul kumar, HOD Medicinal processes in Chemistry Division, CSIR-CDRI Lucknow** to allow me to carry my research in the eminent institution of CSIR.

I am profoundly indebted to my supervisor **Dr. Kinshuk Raj Srivastva, Senior Scientist, Department of Medicinal Processes in Chemistry, CSIR-CDRI, Lucknow** for accepting me as a project trainee and her continuous support, guidance and suggestion to put in the best of my efforts in my research work.

It is a matter of immense pleasure and pride for me to present my report entitled "**Exploration of Sequence, Structure and Substrate Space for Development of Glycosyltransferase Biocatalyst**" which is an outcome of my training in the Department of Chemistry and experimental medicine ,CDRI ,Lucknow.

I would like express my thanks to **Dr. Snober S. mir, Dean and Head Department of Bioscience and Biotechnology, Integral University** for her kind cooperation and support through out my Masters program.

I express my grateful thanks to **Mr. Shrawan Kumar Rathode** for their support and cooperation for my work.

I extended my warm thanks to my friends **Shilpa, Sargam, Vanshika, Naveen, Anjali and Neetu** for their cooperation during my work. I owe my special thanks and gratitude to my parents and family for their love, inspiration and constant support and encouragement that made me able to face all the hindrance in my path and to complete the project work. Above all, I am very thankful to "**Almighty**" for providing me strength for completion of my project work.

# INTEGRAL UNIVERSITY

Established Under U.P. Act no. 09 of 2004
By state Legisilation
Approved by University Grant Commisson
Phone no.: +91(0552)2890812,2890730,
Fax no.-0522-2890809


## TO WHOM IT MAY CONCERN

This is to certify that **Mr. Shashank Singh** a student of M. Sc. Biotechnology (II Year, IV semester), Integral University has completed his four months dissertation work entitled "**Exploration of Sequence, Structure and Substrate Space for Development of Glycosyltransferase Biocatalyst**" successfully. He has completed this work from 14th of February to 13th of June 2022 at CSIR-CDRI, under the guidance of **Dr. Kinshuk Raj Srivastava.** The dissertation was a compulsory part of his M. Sc. degree.

I wish him good luck and future endeavors.

**Dr. Snober S. Mir**

Head,

Department of Biosciences,

Integral University, Lucknow

# DECLERATION

I here by declare that this project work titled "**Exploration of Sequence, Structure and Substrate Space for Development of Glycosyltransferase Biocatalyst**" is a record of original work done by mender the supervision and guidance of **Dr. Kinshuk Raj Srivastava Senior Scientist at CSIR-CDRI, Lucknow** and this project work has not formed on the basis for the award of any Degree/Diploma Association/Fellowship similar title to any candidate of the university.

Signature with name

Place:Lucknow

Date:

# CONTENT

# LIST OF FIGURE

# ABBEREVEATIONS AND SYMBOLS

| | |
|---|---|
| **CD-HIT** | cluster database at high identify with tolerance |
| **EFI-EST** | enzyme function initiative enzyme similarity network tool |
| **MSA** | multiple sequence alignment |
| **MUSCLE** | multiple sequence comparison by long exception |
| **SSN** | sequence similarity network |

# 1-Introduction-

An enzyme is a biocatalyst, which enhances the rate of thermodynamically favourable biological reactions to several thousand to million folds.Enzymes have been employed for a wide variety of chemical processes for decades. Enzymes are highly specialised catalysts with extraordinary catalytic power and also with remarkable specificity, catalysing almost all cellular reactions.It can accelerate a reaction by several orders of magnitude. They accelerate both the forward and reverse reactions, so they have no effect on the equilibrium. Although proteins make up the majority of enzymes, some RNA molecules can also act as enzymes..

Biocatalysis has become an important aspect of modern organic synthesis, both in academia and across the chemical and pharmaceutical industries. The pharmaceutical industry's use of biocatalysis has been fueled by novel protein engineering tools that enable quick optimization of catalyst activity, including computational design and lab evolution.Its success has been largely due to a rapid expansion of the range of chemical reactions accessible, made possible by advanced tools for enzyme discovery coupled with high-throughput laboratory evolution techniques for biocatalyst optimization.Enzymes are examples of biological macromolecular catalysts. In chemical reactions, enzymes speed things up. Substrates are the molecules that proteins may interact with, and an enzyme converts them into new molecules known as products.. Most enzymes are proteins, although a few are catalytic RNA molecules.Proteins are increasingly being used as catalysts in the chemical synthesis of more complex molecules, such as pharmaceuticals. Because they combine the benefits of a catalyst and a directing group controlling selectivity in a single reagent, enzymes are particularly potent.
We go over the various stages of development that can result in a bioprocess, including reaction design, biocatalyst selection, and optimization. a variety of industrial goods.These phases must be carefully integrated because they are interdependent.For example,nitrile hydratases are used to make acrylamide on the thousands of tons scale These microorganisms were used to purify and characterise this enzyme. According to the metal involved, NHases can be divided roughly into two groups: Fe-type and Cotype. Because NHases can transform nitriles into the corresponding higher-value amides under benign conditions, they are anticipated to have enormous potential as catalysts in organic chemical processing. We have produced useful compounds using microbial enzymes: NHase has been utilised in the industrial production of acrylamide from acrylonitrile (capacity: 30,000 tons/year). This is the first instance of a biotransformation method that has produced a commercial chemical successfully· Proteins that act alone or as part of larger complexes are typically globular in shape. The structure of the enzyme is specified

by the amino acid sequence, which also controls the enzyme's catalytic activity. Although structure dictates function, it is still not possible to predict a novel enzymatic activity solely based on structure. When heated or exposed to chemical denaturants, the structures of enzymes unravel (denature), which typically results in a loss of activity.Typically, enzymes are much bigger than their substrates. The monomer of 4-oxalocrotonate tautomerase has just 62 amino acid residues, while the animal fatty acid synthase has over 2,500 residues. The catalytic site, which only makes up a small portion of their structure (about 2-4 amino acids), is directly responsible for catalysis. One or more binding sites where residues orient the substrates are close to this catalytic site. The enzyme's active site is made up of both the binding and catalytic sites. The majority of the enzyme structure that is still present is used to preserve the active site's precise orientation and dynamics.Some enzymes have binding and orientation sites for catalytic cofactors rather than amino acids being directly involved in catalysis. Allosteric sites, where the binding of a small molecule results in a conformational change that either increases or decreases activity, may also be present in the structures of enzymes.

## 1.1-Various aspect-

The field of biocatalysis has grown tremendously, reaching the current state of the art and science. Problem-solving became possible to successfully and efficiently remove the associated bottlenecks due to cross-inspiration and connections between the art and know-how of various practical biocatalytic applications and the motivation for the creation of fundamental knowledge in the science of biocatalysis over time. Building bridges between various cultures, disciplines, and interests has historically been crucial to success; the current situation demonstrates that this is still true today.The need for system-relevant products, resource-efficient and sustainable processes, and resilient and strong supply chains has increased as a result of the current megatrends. Limiting factors like scarce resources, safety, health, environmental concerns, and sustainability issues have emerged as major forces driving innovation in science, business, and society.Although the focus is, on a different scale, addressing the molecular and engineering aspects of biocatalysis, the science of biocatalysis is crucial for developing new transformations and creating new sustainable value chains on a global scale. Enabling viable biocatalytic processes at scale requires research, development, and innovation of biocatalysts in the form of whole cells, cell-free extracts, or isolated enzymes, as well as their characterization, optimization,
and engineering for desired transformation.One-step and multistep biocatalytic reactions as well as total biocatalytic syntheses based on inexpensive and easily accessible

renewable resources like glucose and even CO2 are all included in the expanding research field of biocatalytic reaction design. Research on cascade reactions, synthetic metabolic pathways, and systems biocatalysis are all part of the very active and stimulating field known as the coupling of reactions and reactors, which is also bridging over to metabolic engineering, systems biology, and synthetic biology.

The development of novel chemoenzymatic synthetic methodologies at the intersection of biocatalysis and organic chemistry is a prime example of how well-suited synthetic biocatalysis methodologies are to bridging the scientific disciplines of biotechnology and chemistry. The molecular economy and eco efficiency perspectives of the toolboxes for biocatalytic transformation link engineering and social sciences to biocatalysis and the ideas of a circular economy and a bioeconomy.The substantial increase in global biocatalysis knowledge and the resulting facilitation of global communication present excellent opportunities for addressing bottlenecks and challenges both now and in the future.This Special Issue highlights some of the promising recent advancements in biocatalysis.

## 1.2-Application-

Our ability to discover and engineer new enzymes with expanded substrate scope and new reaction chemistries is now proceeding faster than ever before. Advances in enzyme discovery by, for example, bioinformatics,

metagenomics, enzyme promiscuity and de novo design coupled with enzyme engineering (by directed evolution and high-throughput screening) have led to a rapid increase in the number of biocatalytic tools that are available for synthetic chemistry.We are now clearly entering a golden age that offers unprecedented opportunities for the application of biocatalysis across a broad range of disciplines. This is especially true when combined with a much greater understanding of the target molecule synthesis steps for which biocatalysts can be used (biocatalytic retrosynthesis).A different enzyme will be evolved to more closely match the needs of the alternative synthetic scheme rather than using a biocatalyst that is optimised for the synthesis of one pharmaceutical intermediate.This paradigm, in our experience, cannot yet be accomplished using chemocatalysis and demonstrates the true power of biocatalysis. Additionally, these engineered biocatalysts can now be quickly screened against target substrates to identify lead hits for additional development and optimization in a 96-well format.Ketoreductases (KREDs) and transaminases are two examples of recently created biocatalytic enzyme families with broad substrate specificity.This strategy might be implemented using flow-format immobilised biocatalysts. We must be able to control all types of chemical matter

(both large molecules and small molecules) that are involved in biological processes if we are to effectively regulate biological activities and eradicate disease. As the primary method by which biological processes are controlled and signalling molecules are produced in living organisms By eliminating the need for protecting groups, biocatalysis can eliminate unnecessary steps from chemical synthesis. In fact, because enzymes typically catalyse a transformation with exquisite regioselectivity, they can be thought of as both a catalyst and a directing or protecting group.

Accelerating the biocatalyst screening process and spotting potential hits for quick scale-up and proof-of-concept studies are important issues in drug development. The availability of a much wider variety of new and evolved enzymes, along with advancements in modelling and bioinformatics, are all contributing to a noticeable increase in the speed of biocatalyst hit identification.

## 1.3-As a Drug Development-

In the pharmaceutical sector, a number of major factors influence the development of new medications and their cost- and safety-effective manufacturing.At this stage of the drug development process, it is imperative to find chemistries as soon as possible that allow for the preparation of a sufficient amount (0.1–1 kg) of a candidate drug for further research. In fact, some of these drug candidates belong to classes of molecules called modalities, which are not typical small-molecule active pharmaceutical ingredients and are stereochemically complex and have multiple functionalities (APIs).Biocatalysis has primarily been used to enable the scalable, economical, and controlled synthesis of complex single-stereoisomer drugs.This method of synthesis is evolving thanks to the true strength of biocatalysts: their exquisite selectivity in catalysing reactions, which not only eliminates the need for protecting group manipulation but also permits the running of multiple reactions simultaneously.Newly discovered biocatalytic enzymes, like reductive aminotransferases and a wider variety of aldolases, allow for the stereoselective formation of molecules, and biocatalytic cascades can synthesise entire pharmaceutical molecules from basic building blocks in a single step.Biocatalysis producing lead compounds with novel activities and metabolic profiles, biocatalysis helps to speed up the drug discovery process.

## 1.4-As a Drug Discovery-

In the last 10 to 15 years, there have been significant advancements in the field of biocatalysis, which has led to a rapid diversification of the enzymes that are available to synthetic chemists.One of the most significant developments, for instance, is the widespread accessibility of sequenced genomes, which has led to a massive increase in the number of gene and protein sequences that can serve as the foundation for the discovery of new enzymes.Synthesizing genes is now a quick and affordable method of creating novel biocatalysts for a variety of applications due to the decreasing cost of DNA sequencing and artificial gene synthesis. screening as a result, the difficulty in the field has the focus has shifted from finding new open reading frames (ORFs) to curate already-built genomic databases to find or anticipate the enzyme activity that an ORF encodes.

## 1.4.1-Bioinformetics-

Numerous databases and algorithms that are based on protein structure (such as 3DM from Bio-Prodict) or on both primary amino acid sequence and protein structure (such as Catalophor17 and Zymphore18) or both have been developed to relate the primary amino acid sequence and the structure of biocatalysts to their function. The CAVER algorithm has been developed to predict the presence of tunnels (that is, clefts, cavities, and pockets) in enzymes, which may be important for substrate access or product release19. Databases can also be used to identify target residues for mutagenesis. Because the primary amino acid sequence of an enzyme frequently cannot predict its activity, high-throughput screening techniques must be used to determine the activity needed.

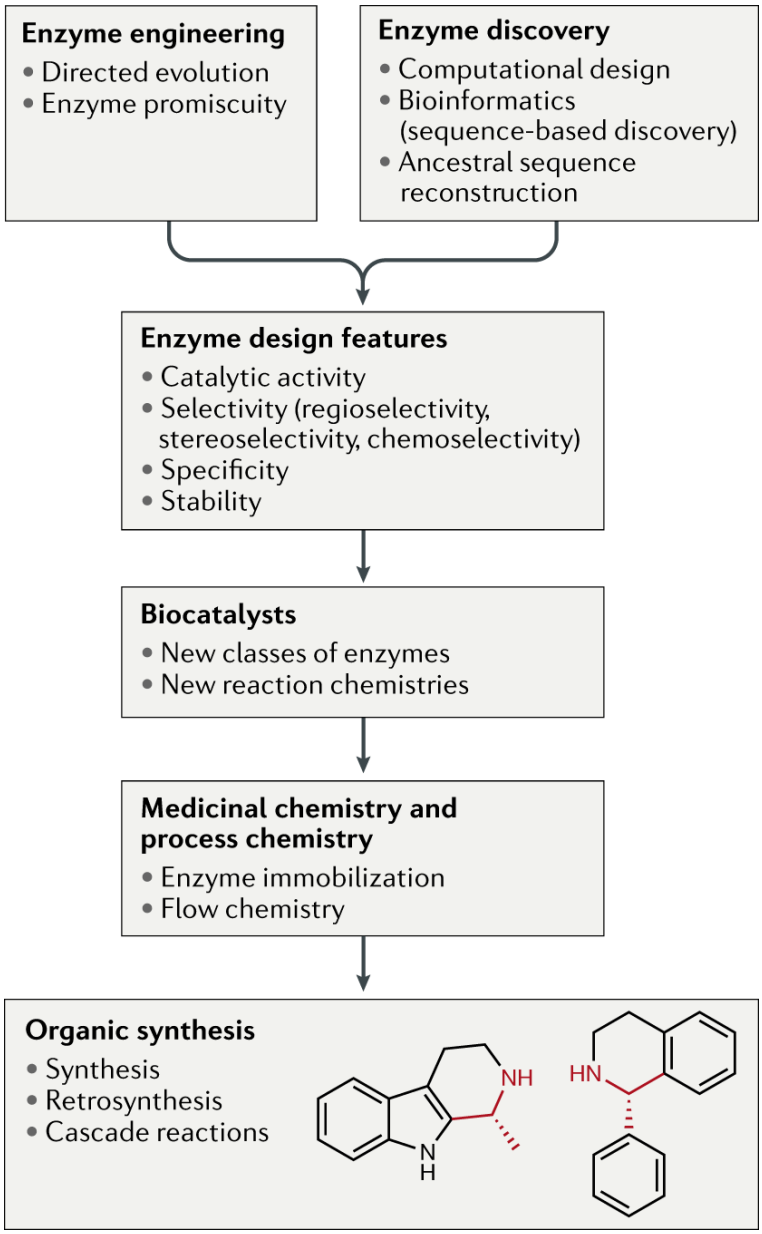
## 1.4.2-Metagenomics-

Large metagenomic libraries can be used to clone new enzymes; this method is particularly effective when there are distinct sequence motifs that can be used to locate homologues.

This method is now being used by a number of businesses, including Prozomix and c-LEcta, to produce large panels of biocatalysts (>500 enzymes) from particular enzyme families, like KREDs, transaminases, and imine reductases.

## 1.4.3-Enzyme promiscucity-

The search for new biocatalysts may be aided by enzyme promiscuity, which is the capacity of an enzyme to catalyse a secondary reaction in addition to its primary reaction. Atrazine chlorohydrolase (encoded by atzA) from Pseudomonas sp. is an example of an enzyme with promiscuous activities that are typically slow in comparison to the main activity and are subject to neutral selection. Melamine deaminase, which is encoded by triA and exhibits promiscuous activity toward the synthetic herbicide atrazine, gave rise to ADP.



## 1.5-Promiscucity and selectivity of enzyme family-

Promiscuity plays a significant role in the evolution of novel functionalities, according to an analysis of enzyme families and superfamilies. Biochemists have been enthralled by these amazing catalysts' evolutionary roots for decades.

enzyme are good source of biocatalysis as they are promiscuous and substrate specific they alter the many of the metabolic pathways complex reaction to within few minutes which earlier without enzyme it takes hours to metabolize the reaction and these properties are well characteristics to emerge as biocatalysis as they are involve in biological synthesis of complex body metabolites.

Enzymes involved in the creation of specific metabolites or natural products are particularly helpful as biocatalysis starting points. Natural products contain a wide range of chemical structures, and research into the biosynthesis of such natural products has shown a comparably wide range of biosynthetic enzymes. The broad chemical and enzymatic variety seen in natural product biosynthesis is discussed in a recent study. The most crucial factors in choosing a potential biosynthetic enzyme from a biocatalytic perspective are its substrate specificity, cofactor dependence, turnover, stability, functional recombinant expression, and capacity to carry out a stand-alone function outside of its natural pathway inside a cell.

Oxidative Fe(II) and 2-oxoglutarate-dependent enzymes are one class of biosynthetic enzymes that are frequently found in natural product biosynthesis. These enzymes can catalyze difficult reactions like hydroxylation, halogenation, and oxidative cyclizations, typically for C (sp3)-H functionalization, due to the reactive intermediates generated in the catalytic cycle. These potent enzymes' development as useful biocatalysts for chemical synthesis. Strong biocatalysts that are frequently employed in industrial production are created through protein engineering. As an alternative to chemical catalysis, biocatalysis offers a wide range of uses. The most notable examples include the employment of enzymes in organic synthesis, particularly to create chiral chemicals for medicines as well as for the taste and fragrance business. Biocatalysts are also widely employed to produce speciality and even bulk chemicals. With a particular focus on scalable chemical manufacturing employing enzymes, this aims to provide exemplary examples in this area. Due to easier availability of enzymes and the capacity to modify those enzymes to match the needs of industrial processes, the usage of biocatalysis in the pharmaceutical sector is continuing to grow. The pharmaceutical industry is rapidly resorting to the utilization of engineered biocatalysts for both lead generation of active compounds and the sustainable production of active pharmaceutical components, spurred by a growing desire to offer novel and more effective treatments to patients. The question

of whether a greater use of biocatalysts in discovery chemistry to produce innovative lead compounds for further optimization and to find new methods of manufacturing target molecules is viable is raised by the availability of a wider range of biocatalysts with an enlarged substrate scope.

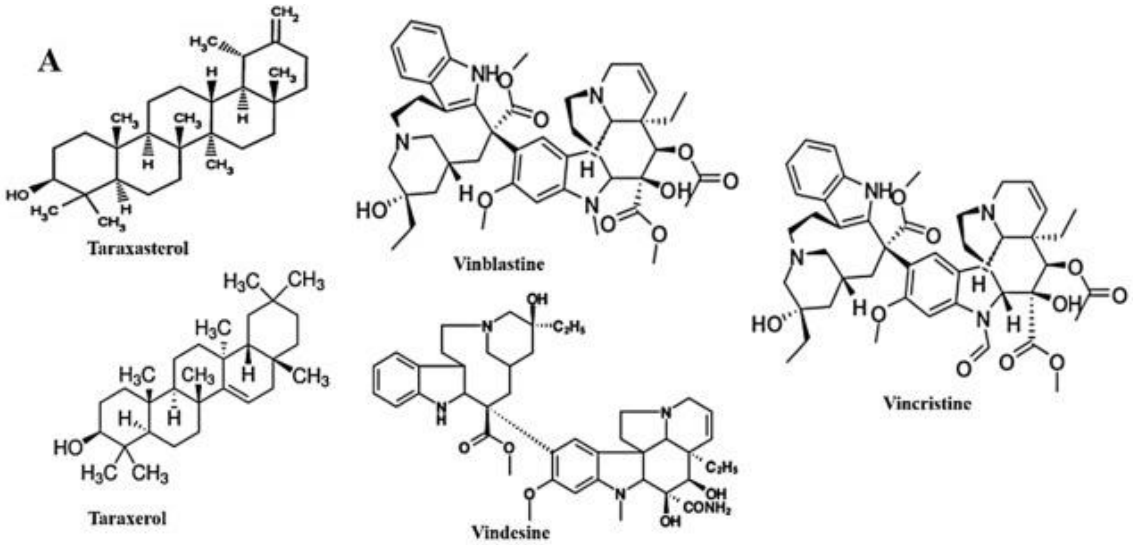
This strategy may be implemented using flow-format immobilized biocatalysts. We must be able to control all types of chemical matter (both large molecules and tiny molecules) that are involved in biological processes if we are to properly govern biological activity and eradicate sickness. We may employ the same strategy to more quickly reach the needed complicated chemical space since biocatalysis is the main method by which signaling molecules are produced and biological processes are controlled in live organisms.

This paradigm change opens the door to novel applications of biocatalysis in areas where conventional chemistry is undeveloped or inefficient, such as late-stage functionalization, C-H activation, reductive amination, halogenation, and oligonucleotide synthesis.Various important forces in the pharmaceutical business influence the discovery and development of novel drugs, as well as their cost-effective and safe production. One of these factors is the speed with which new medication candidates may progress through the preclinical and clinical development stages. Chemistries that enable the synthesis of a suitable quantity (0.1-1 kg) of a candidate drug for additional investigations must be discovered swiftly at these phases of the drug development process. Many of these medication possibilities are stereochemically complicated and have numerous functions; in fact, some candidates belong to modalities, which are not standard small-molecule active medicinal ingredients (APIs).

## 1.6-Lead optimization-

In drugs discovery and development processes lead optimization is critical process through which molecule are chemically modified and subsequently characterized in order to obtain compound with suitable properties to become a drugs.Iterative rounds of synthesis and characterization are used to establish a picture of the relationship between a putative drug's chemical structure and activity in terms of interactions with its targets and metabolism.It is effective to produce natural product derivatives for drug discovery and chemical biology through late-stage diversification of natural products.In the development of new drugs and in chemistry, natural products are crucial. Chemists have worked to create new techniques and plans for efficient chemical synthesis over the years[8].in order to identify their biosynthesis for synthetic biology and clarify their

biological activities for new medicinal medicines[9,10,11].These experiments were inspired by the historical use of natural compounds in drug development, including **penicillin (1), taxol (2), artemisinin (3), and vinblastine (4)**





*Taraxacum officinale* Weber

*Catharanthus roseus*

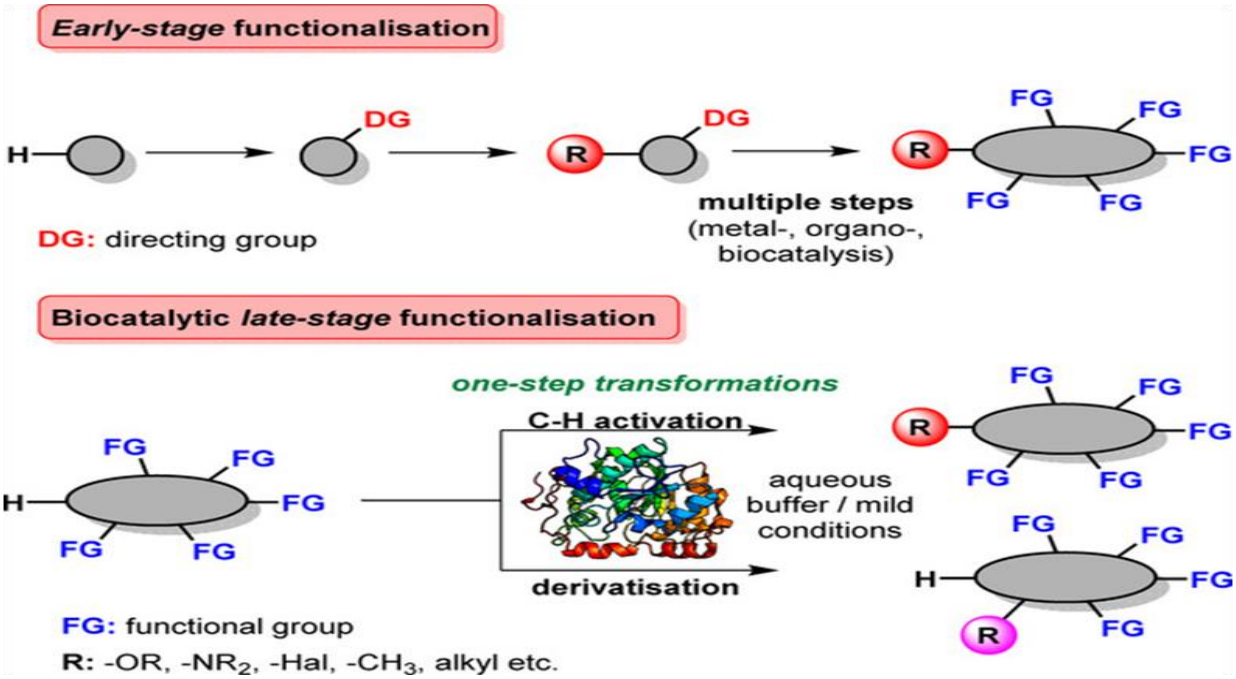
## 1.7-Late stage diversification and Drug discovery-

The research of structure-activity connections and the optimization of pharmacological characteristics are two areas of drug development where natural product diversity is very important (SARs).

Diverted total synthesis is one method that has been used to make these compounds as a result of the identification of physiologically relevant natural product-like derivatives (DTS). Derivatizing natural products directly through selective reactions is an alternative

strategy to the labor-intensive and inefficient traditional synthesis of natural product derivatives from straightforward starting materials. This method may shorten synthetic routes and offer a more efficient way to produce these compounds. However, due to the synthetic difficulty of executing selective functionalizations in the presence of the numerous functional groups frequently found in natural goods, late-stage diversification of natural products has received little attention. Recent innovations in site-selective catalyst technology and other impressive advancements in synthetic organic methods.Modern C-H functionalization, photochemistry, electrochemistry, biocatalysis, photochemistry, and electrochemistry, as well as electrochemistry, have all been accomplished, which has sparked the development of several methods for directly functionalizing complex molecules.

The quick late-stage alteration of several natural compounds has been made possible by these selective transformations. The late-stage diversification of natural goods has grown dramatically during the past ten years. This method has made it efficient to access lead compounds and probes made of natural products. We discuss a few examples of late-stage diversification of complex natural products in this article, as well as the effects of this strategy on organic synthesis, chemical biology, and drug development. The essential method for compound diversification in synthesis is late-stage modification. In synthetic chemistry, enzyme catalysis is becoming increasingly significant. Particularly with regard to late-stage modification, biocatalysis offers outstanding possibilities that are frequently superior to traditional de novo synthesis. Enzymes have shown to be helpful for both the quick diversification of compound libraries as well as the direct insertion of functional groups into complicated scaffolds. Due to the great frequency of these motifs in medicines, enzyme-catalyzed oxy functionalizations, halogenations, methylations, reductions, and amide bond forms are particularly significant and very topical. Diversification is possible by targeted late-stage modification carried out as one of the last synthetic steps in a multistep process, allowing C-H and C-heteroatom bonds to be selectively addressed in the presence of other functional moieties.




Drug discovery and natural product derivatization are only two of the numerous applications that greatly benefit from advances in late-stage functionalization (LSF): In order to succeed in pharmaceutical research, compound libraries are often constructed and changed using readily available building blocks. Despite the clear benefits, compatibility and avoiding cross-reactivity are two important requirements that must be met during late-stage modification. Biocatalysis has thankfully become a new LSF approach in recent years. Enzymes are the perfect catalysts for LSF because they allow the synthesis of complex metabolites in aqueous systems without the requirement for protective or guiding groups, despite the enormous number of highly functionalized molecules in a cell.

The enormous potential for late-stage alterations is starting to be unlocked thanks to enormous research efforts and an infinite number of new enzymes. From the perspectives of drug discovery and high-throughput testing, where the availability of orthogonal and reliable methodologies is extremely desirable, this is of growing importance.

Due to their wide range of structural variations and biological activity, natural products are a significant source of therapeutic scaffolds. Endogenous and synthetic P450s are common catalysts for in vivo late-stage C-H oxyfunctionalization in a wide range of natural products, drug substrates, and metabolites.

## 1.8-Enzymatic Late stage functionalization-

When it comes to further modifying natural products Hydroxylation, Glycosylation, Methylation, prenylation, Halogenation, Acetylation frequently plays a crucial role. Late-stage diversification of bioactive compounds is a highly effective method for examining bioactivity and structure-activity connections (SAR).

### 1.8.1-Hydroxylation-

A chemical procedure known as hydroxylation entails adding a hydroxyl group (-OH) to an organic molecule.

For molecules with aromatic rings in their chemical structures, hydroxylation reactions are often quite prevalent. Because many medications and xenobiotics include aromatic rings, hydroxyl derivatives of pharmaceuticals and xenobiotics are frequently found in nature.Aliphatics exhibit a further hydroxylation reaction that occurs often. Since hydroxylation changes lipophilic substances into water-soluble (hydrophilic) molecules that are more easily removed by the kidneys or liver and expelled, it is crucial for detoxification. By hydroxylation, some medicines (like steroids) can be activated or deactivated.

### 1.8.2-Glycosylation-

The process of attaching a carbohydrate, or "glycan," or a glycosyl donor, to a hydroxyl or other functional group of another molecule, or a glycosyl acceptor, to create a glycoconjugate is known as glycosylation (also known as chemical glycosylation).And a kind of co-translational and post-translational modification is also glycosylation.Through the action of sugar-binding proteins called lectins, which are able to detect particular carbohydrate moieties, glycosylation also contributes to the method by which immune system cells adhere to one another.

### 1.8.3-Prenylation-

Prenylation is the process of adding hydrophobic molecules to a protein or other biomolecule (also known as isoprenylation or lipidation). A farnesyl or a geranylgeranyl moiety is transferred to the target protein's C-terminal cysteine(s) during protein prenylation. A cysteine residue is modified post-translationally by farnesylation, a form of prenylation, by the addition of an isoprenyl group. These characteristics are due to their higher affinity for their target proteins and improved permeability of biological

membranes. Prenylated molecules have a variety of biological functions as a result, and they are used in chemistry, agriculture, and medicine. Because of variations in prenylation location on the aromatic ring, prenylation of aromatic compounds adds to the variety of plant secondary metabolites. There are several prenylated molecules in plants as a result of different prenyl chain lengths and further alterations to the prenyl moiety, such as cyclization and hydroxylation. The secondary metabolites' prenyl moieties are crucial to the diversity of their chemical structures and biological activity. Additionally, it has been shown that prenylation boosts a compound's affinity for biological membranes and its interactions with cellular targets, leading to a higher degree of bioactivity compared to non-prenylated compounds.

## 1.9-Glycosyltransferases-

The glycosyltransferase, one of the most important and abundant enzymes in nature, is able to transfer a monosaccharide moiety from the corresponding sugar nucleotide donor to a specific hydroxyl group of a sugar acceptor, protein, or lipid. There are currently 9 Leloir sugar nucleotides, which are the most common sugar nucleotide donors in eukaryotic systems (UDP-Glc, UDP-Gal, UDP- GlcNAc, UDP-GalNAc, UDPGLcA, UDP-Xyl, GDP-Fuc, GDP-Man, and CMP-Neu5Ac). These serve as the high energy substrate, thereby allowing catalysis to occur. Therefore, a Leloir. glycosyltransferase will utilize one of these nine donors, transfer the corresponding sugar to the acceptor, and release the di/mono phosphate nucleotide. Furthermore, GTs are able to regulate the stereochemistry of the newly created glycosidic linkage between the donor and acceptor molecules, which is defined as either retention or inversion of the configuration at the anomeric carbon with respect to the sugar donor. Knowing that there are two pathways that the newly created glycosidic linkage can adopt, there are likely different mechanistic details, which result in either inversion or retention of the stereochemistry. It is proposed from the available crystal structures of the inverting type GTs, that these GTs utilize a direct displacement type mechanism which mimics an SN2-like reaction. Conversely, the mechanism of the retention-type GTs is a less characterized; however a prominent view in the field is that these GTs utilize a double displacement mechanism which is the result of a covalently bound glycosyl-enzyme intermediate. Due to genetic studies there are over 110 putative GTs that are housed in the CAZy database (www.cazy.org). Over the last ten years many of them have had their three-dimensional structures resolved, generally by X-ray crystallography. The available crystal structures revealed that there only two general folds that nucleotide-sugardependent GTs adopt, which are referenced as GT-A and GT-B type. Further threading analysis has revealed that many of the uncharacterized GTs adopt one of these two folds, suggesting that the majority of GTs

may have evolved from a small number of progenitor sequences. The two folds are analogous in that they both adopt similar β/α/β Rossmann domains; however, the orientation in which the domains are situated appears to differ.The GT-A fold consists of an open twisted β-sheet surrounded by α-helices on both sides, and is usually generalized as two abutting Rossmann-like folds. The two tightly associated β/α/β domains are usually similar in size, which leads to the formation of a continuous central β-sheet. This notion lead some to describe the GT-A fold as a single domain fold, however, there are distinct nucleotide and acceptor binding domains. Furthermore, most GT-A enzymes possess a signature DXD (Asp-X-Asp) motif, which is responsible for coordinating a divalent metal cation, thereby promoting catalysis. This motif is not a determining characteristic of the GT-A type GTs, as will be shown in this thesis, but it is a frequent characteristic of this class of transferases. Like the GT-A fold, GT-B enzymes consist of two Rossmann-like domains, however, the domains are much less tightly associated, and an active site is created within the cleft between the two domains. The two different domains are most likely associated with binding of either the acceptor or donor substrates. Regardless of the fold that a GT appears to adopt (GT-A vs. GT-B), it appears that the overall fold of the enzyme cannot dictate the stereochemical outcome of the reaction that is catalyzed, as there are numerous examples of both retaining and inverting GTs that adopt the differing folds. There is a third type of fold that has recently been discovered, notably the GT-C type fold. This type of GT is proposed to bind lipid phosphate-activated donor sugar substrates, as several lipid anchored GTs cannot adopt the traditional dual-Rossmann domains. There is little structural evidence proving the existence of the GT-C type fold, and this hampers the ability to definitively characterize a GT as adopting a GT-C like fold. However, further work is being done in order to determine if the putative GT-C type folds are evolutionarily related, and what their relationship is to other major classes of carbohydrate-active enzymes . Unlike the well characterized and studied GT-A and GTB folds, to date, all enzymes that adopt the predicted GT-C fold belong to inverting glycosyltransferase families, which may be a consequence of utilizing the lipid-linked activated donor. The GT-A, GT-B, and putative GT-C type folds appear to apply GTs from all domains of life, including prokaryotes and eukaryotes. While there are some similarities between the GTs from prokaryotes and eukaryotes, there are many distinct differences, which make the utilization of eukaryotic GTs more challenging to exploit than prokaryotic GTs for the synthesis of oligosaccharides. For example, looking at the putative GT sequences from various eukaryotic organisms, many if not most of the GT sequences contain at least one putative membrane domain, and thus require the membrane for complete enzymatic function. Membrane bound or anchored GTs are traditionally difficult to express in large amounts,

and have been known to display very tight substrate specificities. Conversely, bacterial GTs are usually only membrane ―associated‖, not membrane bound, and can easily be purified as soluble, active, proteins. Furthermore, it is relatively easy to obtain genomic DNAs from bacterial species, whereas cDNA libraries are usually required for the cloning of eukaryotic transferases. Lastly, there are numerous examples where bacterial GTs have exhibited promiscuous substrate specificities and thus are more advantageous than eukaryotic GTs for the in vitro synthesis of oligosaccharides. While enzymatic synthesis of oligosaccharides has been demonstrated to be useful, there are many drawbacks, and chemical synthesis is still advantageous in some cases. The yield for enzymatic glycosylation is usually high, however the purification and handling of certain GTs is not always easy, and many of the reagents required for GT-mediated reactions are very expensive, or not even commercially available (such as the sugar nucleotide donor substrates). Furthermore, there can be significant competitive inhibition of the glycosyltransferases by the by-products, such as the nucleotide phosphates. Thus, similar to chemical syntheses, cost effective methods are sought for large scale oligosaccharide synthesis, such as the regeneration of the donor molecules using the Super-bug and Super-bead technology, as well as alternatives to obtaining the required donors/acceptors

## 2-Objective-

### 2.1-**The overall aim of my thesis is to decipher relationship between sequence structural and the substrate of Glycosyltransferase superfamilies, this will enable us to design/engineering enzyme for sterio selective glycosylation of targeted bioactive molecule to improve the pharmacological activity.**

## 2.2- Matrial and Method-

## 2.2.1-Gathering Glycosyltransferase sequence-

The gathering of the GT sequence is done from sequence which are available in interpro**(interpro is a database of protein families)**nd its contribute database of **SFLD** here every protein is given a unique IPR number as in case of glycosyltransferase IPR005076,IP002495,IPR006813,IPR002654,IPR008630,IPR004139,IPR003406,IPR00 2685,IPR007754,IPR002659,IPR007577,IPR005027,IPR006759,IPR021067 these are repreasented by the GT-A fold.

## 2.2.2-Getting a sequence dataset downloaded-

Sequence for generating SSN is downloaded from the web server of Cazyme and get the accession id and sequence for GT-6,GT-8,GT-13,GT-14,GT-15,GT-16,GT-17,GT-21,GT-25,GT-31,GT-32,GT-34,GT-43,GT-54,GT-60 is represented by 405,18482,207,2293,1793,126,604,1791,8198,2654,6423,937,395,228,466 protein for GT-A fold.

## Cazy. Web server link-

**http://www.cazy.org/**

## 2.2.3-Procedure-

**http://www.cazy.org/GlycosylTransferase-family** by clicking the following link choose the GT family and download the accession id.

Afterwards **https://www.ncbi.nlm.nih.gov/sites/batchentrez** clicking the following link and choose the protein database and upload your accession id file and retrieve it, then afterward get the result in NCBI**(https://www.ncbi.nlm.nih.gov/protein?cmd=HistorySearch&QueryKey=1)ad** and download the FASTA formet file.

## 2.2.4-Reducing the sequence dataset-

To enable the visualization and manipulation of SSNs on conventional computers, datasets containing a significant number of sequences (e.g., >10,000) must be condensed due to computational restrictions. An efficient way to choose representative sequences from a group of sequences that share a certain percent identity cutoff is to use the Cluster Database at High Identity with Tolerance (CD-HIT) web server (Li & Godzik, 2006). For example, one sequence can be chosen to represent a set of sequences sharing a specific sequence identity.

## CD-HIT web server-

**http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi**

## 3-Procedure-

The CD-HIT web server can be used to reduce the dataset above to one with Representative sequences that meet a particular percent identity criteria-

Upload the FASTA-formatted file to "h-CD-HIT." This choice does several CD-HIT runs: The non-redundant sequences chosen from this initial clustering procedure are then subsequently clustered at a lower identity after proteins are initially clustered at a high identity (i.e., 90%). (i.e., 70 percent ). For instance, after uploading the aforementioned FASTA file to the website, choose three CD-HIT runs at IDs of 90%, 70%, and 50%.

Download the entire CD-HIT calculation and save it in the "cdhit" folder. The reduced dataset, which consists of a representative sequence chosen for each 50 percent ID cluster, is stored in the "fas.3" file located in this folder.

## 3.1-Visualization of sequence similarity relationships among superfamily-

To visualize the relationships among the PT family SSN allows us to visualize the relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters.It is simple to create SSNs using the Enzyme Similarity Tool (EFI-EST).A user-provided dataset can be used to build networks using the EFI-EST webserver.To do this, enable the upload of a curated dataset by choosing the "User specified FASTA file" option.

## 3.2-Generate a SSN from provided sequences-

To determine the similarities between sequence pairings and calculate edge values to get the SSN, an all-by-all BLAST is used.
Enter a list of protein sequences in FASTA format, or upload a file with a sequence list in FASTA format.

There are many filtering options to curate the fasta and to get good cluster the options are enlisted below

## 3.3-Filter by Taxonomy-

It is possible to limit the input sequences (which come by default from the UniProt database) to those that correspond to particular taxonomic groupings (superkingdom, kingdom, phylum, class, order, family, genus, species). A union of several circumstances results from their combination.
Alternatively, the user can limit the returned sequences to those from species that may give genomic context (gene clusters/operons) suitable for inferring activities by selecting "Bacteria, Archaea, Fungi" from the "Preselected conditions.

## 3.4-Protein family addition option-

To enable the production of SSNs for extremely large Pfam and/or InterPro families, the EST gives access to the UniRef90 and UniRef50 databases.
The UniRef50 or UniRef90 databases will be used to produce SSNs for families with more than 50,000 sequences.
Sequences in UniRef90 are clustered and identified by a sequence known as the cluster ID when they share >90% sequence identity across 80% of the sequence length. Similar to UniRef50 but with a sequence identity of just about 50%.
If one of the UniRef databases is utilized, the output SSN is equal to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network, with each node corresponding to a UniRef cluster ID. In this instance, an extra node property is produced that specifies all the UniRef cluster IDs.

## 3.5-SSN edge calculation option-

This is one of the parameters which is used to generate SSN edge and that is E value .evalue is which is required for using BLAST to determine edge value similarities between sequences. Negative log of e value for all-by-all BLAST (≥1default 5)

To generate edges even between sequences with minimal similarity, default values are permissive. After setting all parameters for generating data or giving a job to a web server job name is required by which results can be easily identified on the webserver and to

restore the data or retain it E-mail ID should be provided so that after completing the work result is notified via Email.

- To see interaction of the protein with nearby residue and substrate structural following web server are used which are listed below-

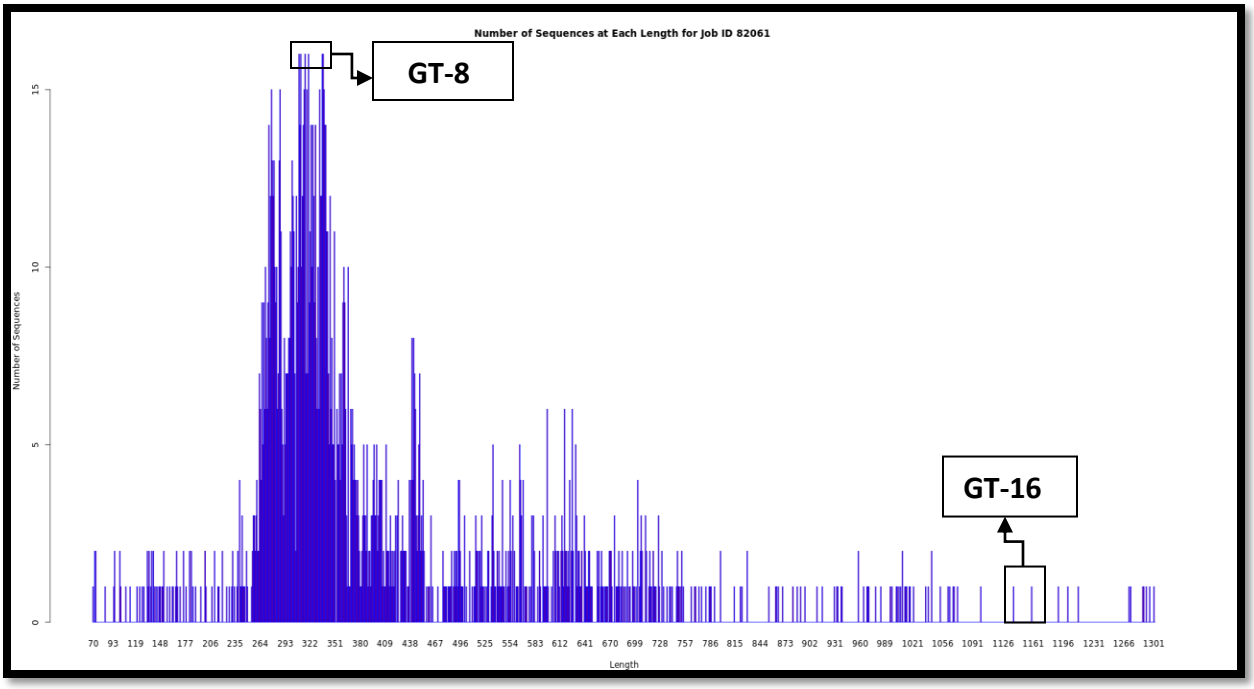
✓ **https://dyn.life.nthu.edu.tw/oENM/**
✓ **projects.biotec.tu-dresden.de/plip-web.**
✓ **http://www.rcsb.org/pdb/**

And to visualize the structure of the enzyme family, every protein has its unique uniprot ID (UniProt is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects).and every protein has its structural annotations and which is deposited in the protein data bank.and every protein has its unique PDB id And for structural study protein data bank structure are taken.

## 4-Result-

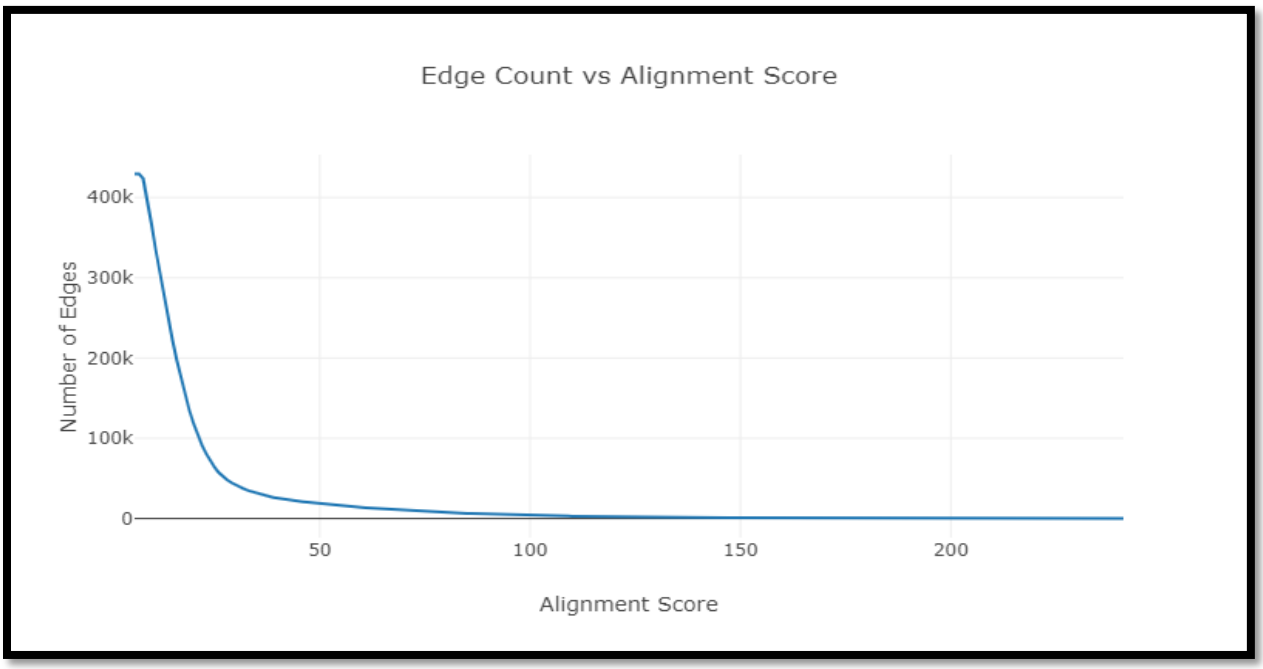
The application then gathers the sequences and runs an all-by-all BLAST on the EFI's cluster with 24 processors to get alignment scores (edges). Only node pairs with an internode alignment score of greater than 5 are kept, meaning that the edges of the SSNs will have alignment scores.

The software creates four graphs following the BLAST to help with the selection of the alignment score minimum required to produce the initial SSNs: After the above job submission on the web server the following results and figures are obtained and which are interpreted as listed and conclusions drawn from results are as follows.




**Figre-: 1**

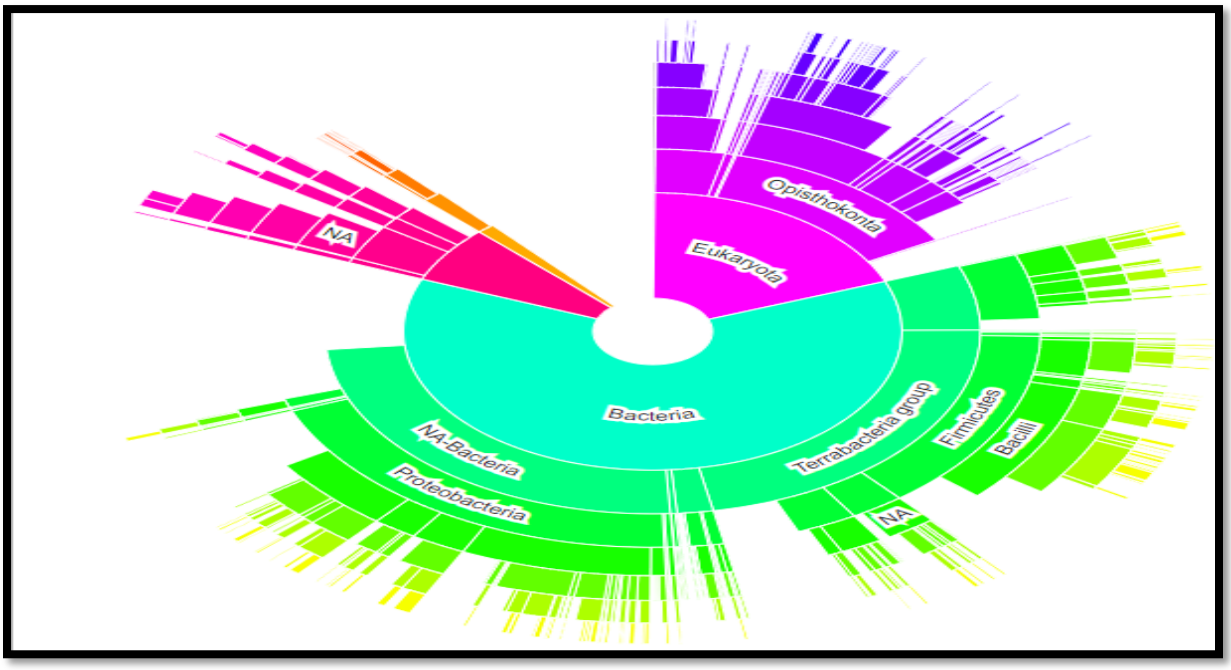
Inspection of the histogram permits identification of fragments, single domain proteins, and multidomain fusion protein of GT-A fold we got many region in histogram form SSN. The dataset can be length-filtered using the Minimum(GT-16) and Maximum(GT-8) "Sequence Length Restrictions" in the "SSN Finalization" tab to remove fragments.




**Figre-: 2**

Next plot which we got from the SSN is edge count vs aligament score of GT-A fold superfamily which represents the nodes(protein/sequence) of all superfamiy and how much of nodes are align are represented in the plot from the plot we figure out the most of the protein are having aligment score of 50 about 300k of the sequence are align at this score.

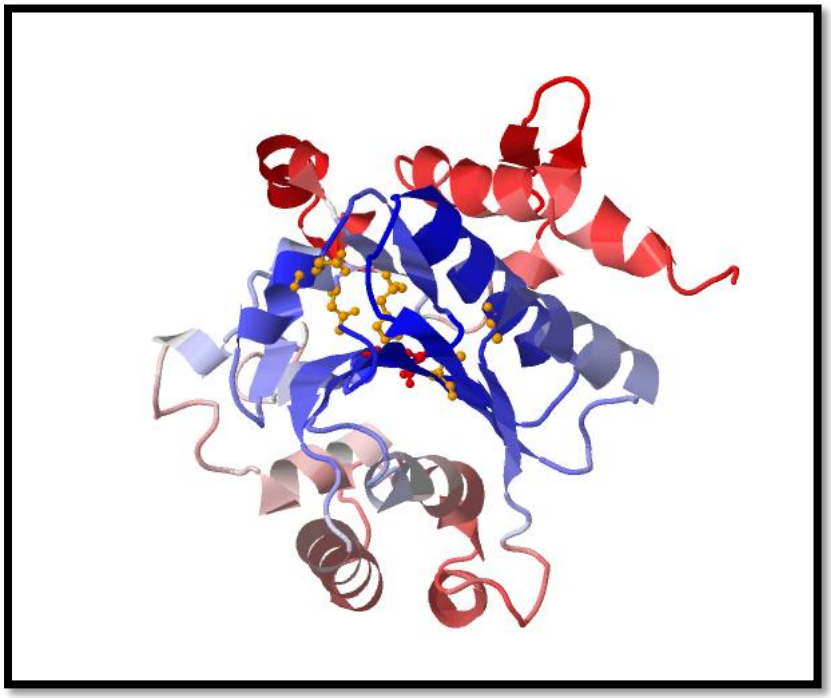
From this we get that most of the superfamily protein are having same identity due to common origin of ancestor links only few of the family members shows the divergence sequence from there ancestor origin.




**Figre-:3**

This is the third out put of EFI-EST sequence similarity tool and this represents taxonomical Distribution of the the GT-A fold superfamily from the above of the sunburst we got to know that most of the GT-A are bacterial, fungal and viral in origin , from and these are associated with plants and together they synthesize the secondary metabolites . metabolites produced from these are highly bioactive molecules.and this sunburst is represented 50 Uniref identity .

## 4.1-Identification of Functional site-

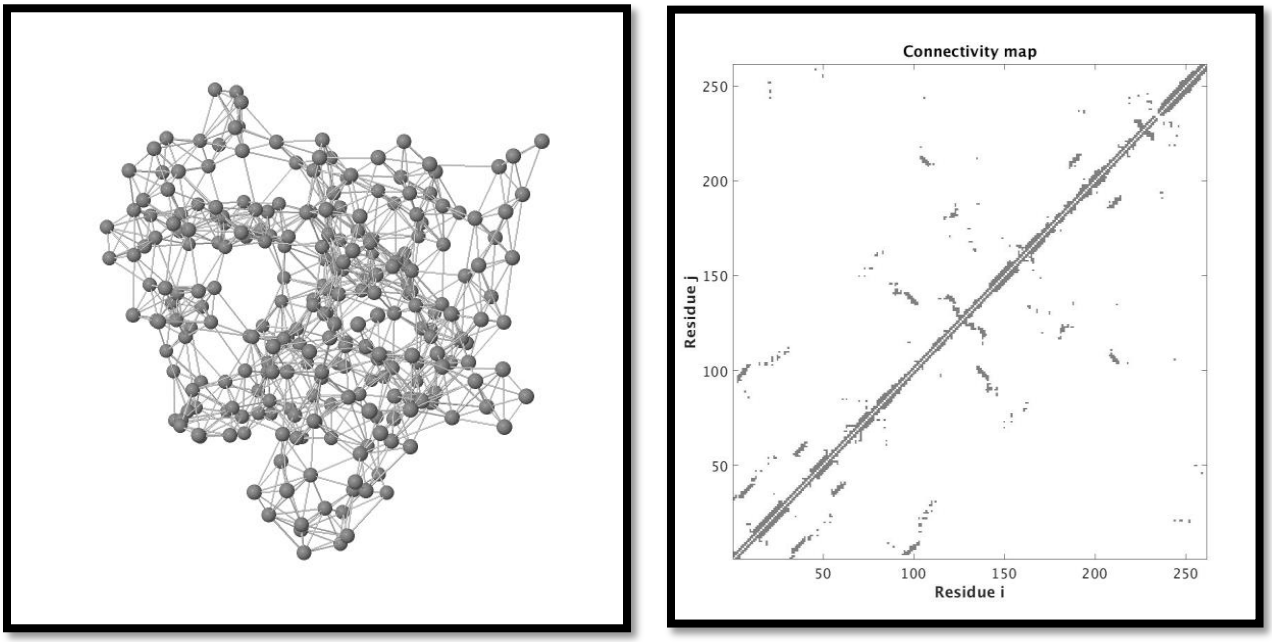


**Figre-:4(GT-8)**

Mobility ( → increase)

**Predicted potential Functional/Active Sites based on dynamics features:** A:ASP101, A:GLY134, A:THR7, A:HIS211, A:ASN187, A:ALA18, A:ASP111, A:ASN 109, A:TYR14, A:PHE114 Top 2 and top 3-7 predictions (sites) are colored red and orange respectively and shown in ball and stick. The rest of predictions, if any, are listed in green text.

## 4.2-Interlining network residue-



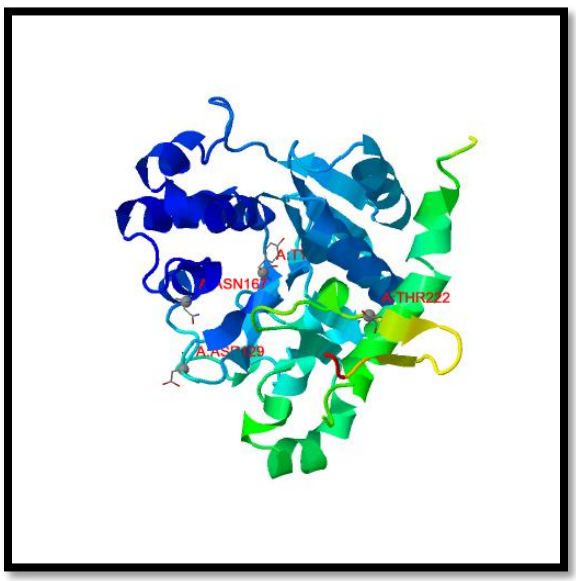

**Figre-:5 (GT-8) PDB-ILL3**

**Figre-:(5A)** **Figre-:(5B)**

**Fig-:(5A)-**The above fig decipher that how the residue are interconnected in the crystal structure of GT-8 family and interaction among the residue are necessary for the dynamic of the protein and also to maintain the functionality. And to we also got the connectivity map of each residue and maped are show in the fig NO as we see in the map we observed that residue which have highly interlinking connectivity are mapped at central. And mapping is started from residue 1-250 and on each side different residue are showing connectivity to each other.

**Fig-:(5B)-** structural representation of **affinity (B)** histogram showing the affinity of atomic As in **Figre(5B)** atomic affinity is represented in crystal structural and representing few of the residue which are having higher atomic affinity and in **Fig(B)** affinity is mapped between pair of the residue. Higher affinity represent a higher

contact/interaction between the pair of residue ,as we move above the affinity among the residue are also increasing as observed and vertical line is representing affinity score for all the residue site of **(1LL3)** which increase from white to black as leabelled .

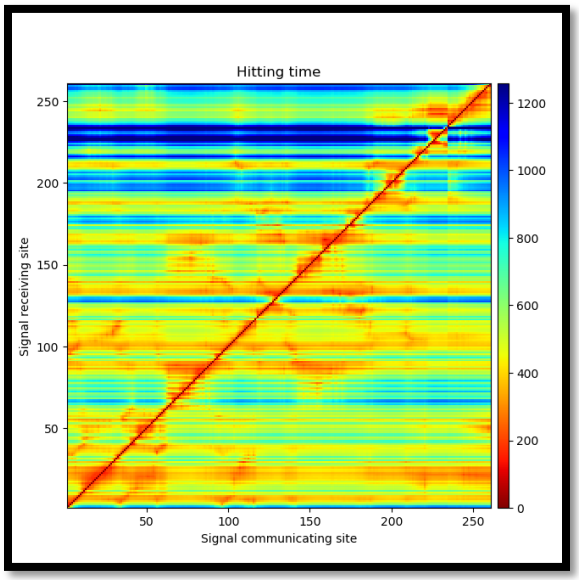


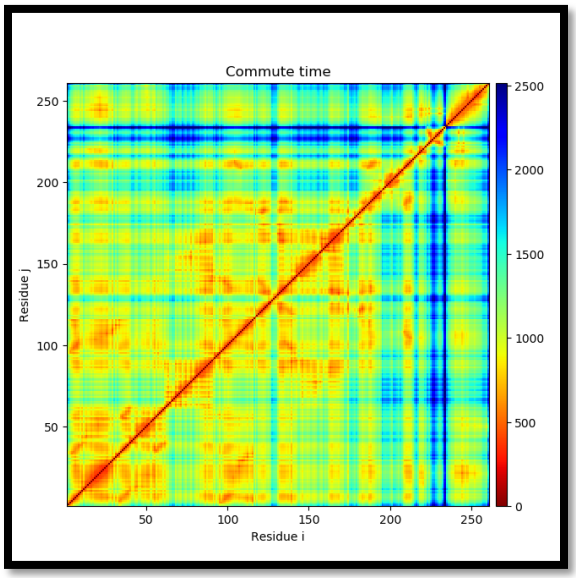


**Figre-:6**

**Figre-:(6A)** **Figre-:(6B)**

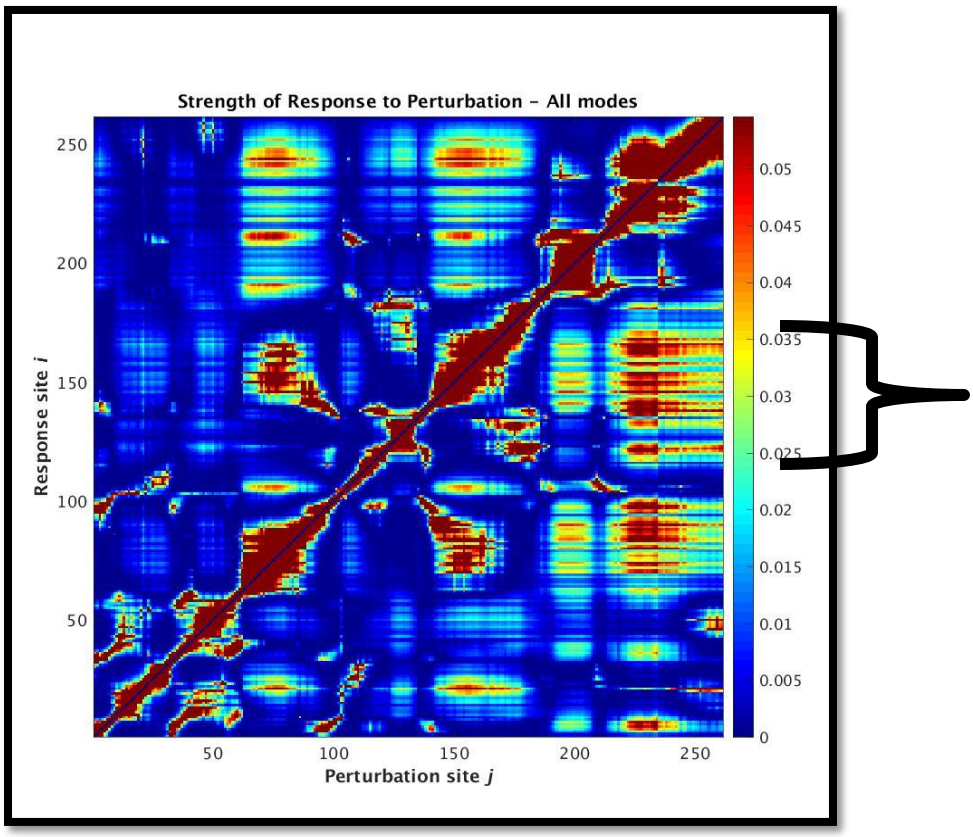
Increasing propensity to act as broadcaster/receiver →



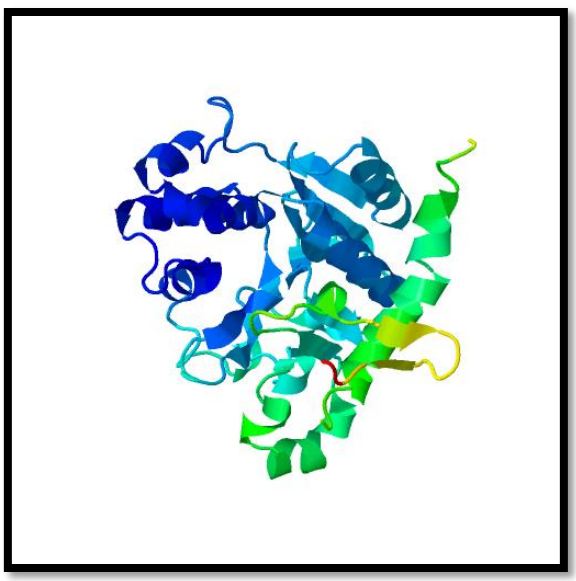

**Figre-:(6C)** **Figre-:(6d)**

**Fig-6 (A)** signal communication efficiency **(B)** signal reciving efficiency **(C)** properties for communication and signalling. **(D)** The hitting time matrix H for **GT-8 PDB 1LL3** is shown on the left as a function of signal communicating sites (abscissa) and signal receiving sites (ordinate).Red regions indicate efficiently communicating pairs. The average responses of all residue are shown on the (vertical) curve.

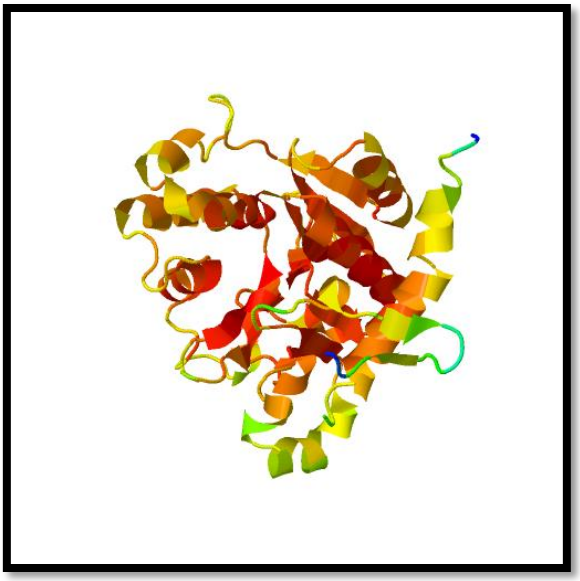



**Figre-:7(7A)**

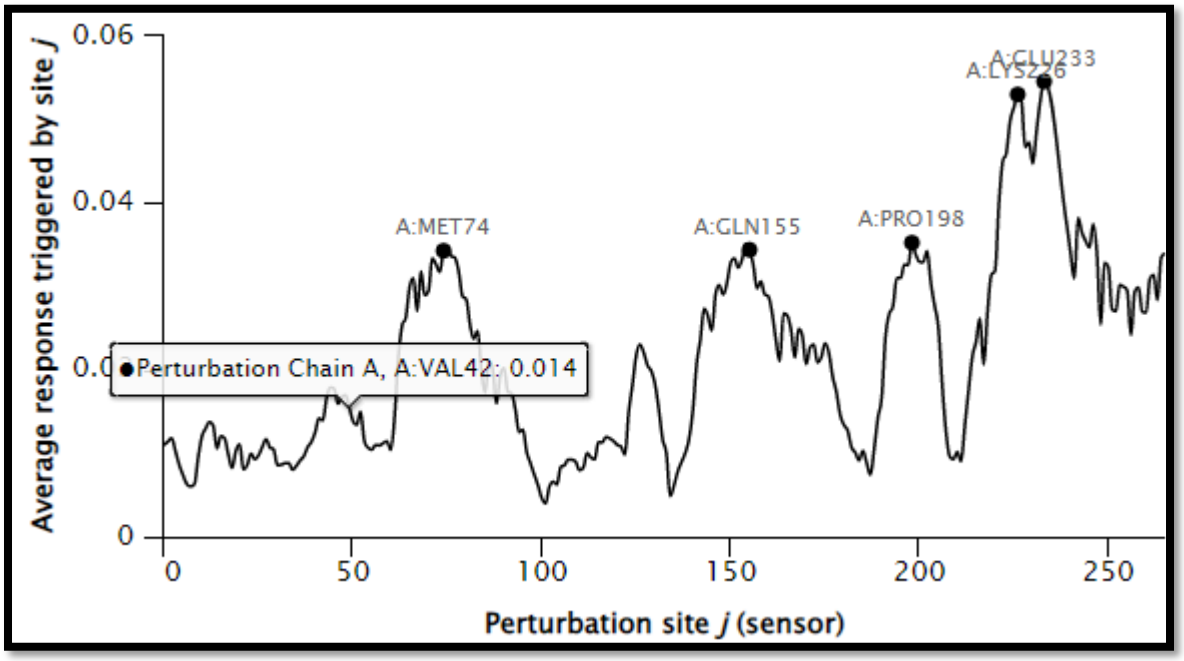
**Fig-:7** Peturbation effect on GT-8 residues The ribbon diagrams display the residues with highest propensities (colored red) to serve as **(A)** sensors and **(B)** senser residue **(C)** effector of perturbations. **(D)** PRS map, In **(7A)** where areas in dark red denote strongly positive responses. The column and row averages are shown as curves (on the left and at the bottom).



**Figre-:(7B)**



**Figre-:(7C)**




**Figre-:(7D)-** higly perturbation sites are **highlyted** for **GT-8 (1LL3)**

As there is ligand binding residue changes the dynamics in context of celluar interaction after the binding of ligand in GTes perturbation changes are seen which are highlited in circular form and they are also highlighted in structure of GTes these regions are highly responsed to the ligand binding perturbation.

## 5-Discussion-

The important finding of the research work carried out for the study of evolutionary key feature of enzyme Glycosyltransferase.

- Taxonomic signature
- GT-A fold is consisting of 15 subfamily and there are listed in Material and Method.

The comprehensive revelation of the relationships between protein sequence, structure, and function is one of biochemistry's ultimate objectives. Determining a protein fold's full functional potential is crucial for this reason. In order to achieve this goal with practical experimental efforts (especially taking into account the ongoing and rapid expansion of sequence data),
Detailed bioinformatic analyses offer a very potent tool for classifying sequence diversity and locating unexplored sequence space. A vast
Amount of untapped potential for cutting-edge protein and enzyme discoveries Many large superfamilies still have hidden functions.

To see the dynamic of structure as intermolecular interaction facilitate the catalysis of cognate substrate, trigger structure changes that enable biological activities or stimulate allostearic response that selectively modulate different cellular pathways. We plotted perturbation plots and got result that A:GLU233,A:LYS226 are highly perturbation sites and these are highlighted in **Fig (7D)** Residues that respond to structural changes brought on by legends binding, complications, or any deformation brought on by externally applied force fields are known as effectors or sensors. The robust reaction that sensors have to perturbations sets them apart as in **Fig (7A)** sensor residues are low in beta sheet sides and they act low to perturbation while in alpha helix its moderate type perturbation is seen in fig. Effectors, which are frequently found close to sensors, are distinguished by their capacity to effectively transfer disturbances or associated "information" to other sites.

A measure of the effectiveness of allosteric communication is provided by hitting times. In the Markovian process, the first passage time is the average time (number of steps) for residue/node I to convey the "message" to node j for the first time, also known as the

hitting time. As showing in Fig (7B) signal communicating residues are presenting both low and high as showing in the figure the beta sheet residues shows the modrate signaling communication sites while alpha helix are consisting of higher signal communicating residues. In Fig (7B) reciving residues are showing the alpha helix are represting more signal reciving sites to perturbation and in Fig (7A) atomic contact affinity between pair of nodes (i-j) define the conditional probability of signal transmission low value of signaling rate indicate an efficient signaling broadcasting receiving.

# CONCLUSION

Using these techniques, we can draw the following conclusion from the discussion above. We can use the secondary metabolites enzyme to produce structurally diverse and biologically active unnatural novel molecule scaffold for drug discovery by revealing the family specificity and difference in the sequence as well as how they are diverse and evolved in nature to maintain the reaction specificity.

# REFERENCES


**1-** https://www.shivajicollege.ac.in/sPanel/uploads/econtent/ed8ad70c5da6e71f70db998d cc27987e.pdf

**2-** Elizabeth L. Bell , William Finnigan , Scott P. France , Anthony P. Green,Martin A. Hayes , Lorna J. Hepworth , Sarah L. Lovelock , Haruka Niikura,Sílvia Osuna, Elvira Romero , Katherine S. Ryan , Nicholas J. Turner and Sabine L. Flitsch

**3-** Roland Wohlgemuth,Bruno Bühler

**4-** Paul N. Devine, Roger M. Howard, Rajesh Kumar, Matthew P. Thompson, Matthew D. Truppo & Nicholas J. Turner

**5-** Scott, T. A. & Piel, J. The hidden enzymology of bacterial natural product biosynthesis. Nat. Rev. Chem. 3, 404–425 (2019)

**6-** Truppo, M. D. Biocatalysis in the pharmaceutical industry: the need for speed. ACS Med. Chem. Lett. 8, 476–480 (2017).

**7-** Jackson RC, Handschumacher RE. 1970. Escherichia coli L-asparaginase. Catalytic activity and subunit nature. Biochemistry 9:3585–90

**8-** Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. 30:409–25

**9-** Scott, T. A. & Piel, J. The hidden enzymology of bacterial natural product biosynthesis. Nat. Rev. Chem. 3, 404–425 (2019)

**10-** Liao, C. & Seebeck, F. P. S-Adenosylhomocysteine as a methyl transfer catalyst in biocatalytic methylation reactions. Nat. Catal. 2, 696–701 (2019)

**11-** Wilson, R. M.; Danishefsky, S. J. Small Molecule Natural Products in the Discovery of Therapeutic Agents: The Synthesis Connection. *J. Org. Chem.* 2006

**12-** Danishefsky, S. J. On the potential of natural products in the discovery of pharma leads: A case for reassessment. *Nat. Prod. Rep.* 201

**13-** K. R. Campos, P. J. Coleman, J. C. Alvarez, S. D. Dreher, R. M. Garbaccio, N. K. Terrett, R. D. Tillyer, M. D. Truppo, E. R. Parmee, Science 2019, 363, eaat0805

**14-** E. K. Davison, M. A. Brimble, Curr. Opin. Chem. Biol. 2019, 52, 1–8.

**15-** Nicholas Roman Pettit Graduate Program in Chemistry The Ohio State University 2011 Dissertation Committee: Dr. Peng George Wang, Advisor Dr. Zucai Suo Dr. Karin Musier-Forsyth