Computational tools to model and analyze biomolecular structures and interactions

L Ponoop Prasad Patro

BO14MTECH11002

A Dissertation Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Biotechnology

June, 2016

Declaration

I declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

L Ponoop prasad Patro

L Ponoop Prasad Patro
BO14MTECH11002

Approval Sheet

This thesis entitled "Computational tools to model and analyze the biomolecular structures and interactions" by L Ponoop Prasad Patro is approved for the degree of Master of Technology from IIT Hyderabad.

Dr. Anindya Roy
Associate Professor
Department of Biotechnology
IIT Hyderabad

Dr. D.S. Sharada
Assistant Professor
Department of chemistry
IIT Hyderabad

Dr. Thenmalarchelvi Rathinavelan
Assistant Professor
(Thesis Adviser)
Department of Biotechnology
IIT Hyderabad

Acknowledgements

In the first place I would like to like to extend my gratitude to my advisor, Dr. Thenmalarchelvi Rathinavelan, Assistant Professor, Department of Biotechnology, IIT Hyderabad. She has been supportive since the days I began working under her for my final year project. She had helped me come up with the thesis topic and guided me over almost a year of development. The joy and enthusiasm she has for her research was contagious and motivational for me. I am highly indebted to her for her contributions of time, ideas, valuable guidance and support in completing the project.

The members of my lab have contributed immensely to my personal and professional time at IIT Hyderabad. The group has been a source of friendships as well as good advice and collaboration. I would like to acknowledge all of them for their help and support for completion of this work. In particular, I would like to thank Narendar Kolimi for his great support and help in the project of 3DNuS.

Abstract

Development of the interdisciplinary branch, bioinformatics, is a blessing for life science research. It made the analysis, model, pattern recognition and visualization of biological data much easier, efficient and accessible. Now the research in life science is much stronger with efficient outcomes in a shorter period with minimized error, cost and labor. The bioinformatics tools have helped us in better understanding of a process or pathway with aiding us to analyze them with our convenience. In my thesis work I have worked on three fields, modelling of 3 dimensional nucleic acid structures, finding the steric hindrance in molecules and identifying the amino acids surrounding the sugar ligand in a catalytic pocket to identify catalysis sites in the protein as they are highly conserved invariant of the source organism. The nucleic acid models generated through our server can be utilized as starting model for different research purposes related to nucleic acids and interactions associated to them. Steric hindrance identifier will help in validating models and the identified conserved aminoacids will further be used in automated identification of catalytic sites in glycosidases and can also be useful in explaining the cause specificity and selectivity of the proteins towards a specific sugar moiety.

Keywords: bioinformatics tools, nucleic acids, glycosidases, server.

Nomenclature

EMBL: European Molecular Biology Laboratory

DDBJ: DNA Data Bank of Japan

INSDC: International Nucleotide Sequence Database Collaboration

PDB: Protein Data Bank.

HTML: Hyper Text Markup Language

CSS: Cascading Style Sheets

MSA: Multiple Sequence Alignment

TREDs: Trinucleotide repeat expansion disorders

NMR: Nuclear magnetic resonance

EC: Enzyme Commission

CAZy: Carbohydrate Active enzyme

GH: Glycosyl Hydrolase

Contents

Chapter 1

Introduction to bioinformatics

It took 15 years to completely sequence euchromatic human genome under human genome project [1], but now, only 26 hours is needed to completely sequence the whole human genome through next generation sequencing method. This is one of the example of the impact of the rapidly developing interdisciplinary field that develops methods and software tools for understanding biological data, called bioinformatics. Bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. In 1970, the term bioinformatics was coined by Paulien Hogeweg and Ben Hesper to refer to the study of information processes in biotic systems [2][3] and referred as a parallel field to biophysics (the study of physical processes in biological systems) or biochemistry (the study of chemical processes in biological systems) [2]. Now bioinformatics is an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques such as image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, and protein structures as well as molecular interactions.

1.1 Purpose of bioinformatics

The exponential growth of experimental and clinical data generated from systematic studies, the complexity in health and diseases, and the request for the establishment of systems models are bringing bioinformatics to the center stage of pharmacogenomics and systems biology. Bioinformatics plays an essential role in bridging the gap among different knowledge domains for the translation of the voluminous data into predictive, preventive, and personalized medicine [5]. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures [6]. The main objective of bioinformatics include:

- development and implementation of computer programs that enable efficient access to, use and management of, various types of information
- development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets.

Examples of activities of bioinformatics include: pattern recognition, data mining, machine learning algorithms, mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis [7].

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

1.2 Current bioinformatics tools for life science research

Here, an overview of bioinformatics tools are given that cover most aspects of protein structure prediction, including automated methods for primary, secondary and tertiary structure prediction from the amino acid sequence of the query protein.

There is a huge number of available databases covering almost everything from DNA and protein sequences, molecular structures, to phenotypes and biodiversity. The main repositories of biological sequences are the publicly available sequence databases. Data derived from sequencing projects are independently stored in the nucleotide databases. The primary public and comprehensive repositories of nucleotide sequence entries are: GenBank [7], the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) [8] and the DNA DataBank of Japan (DDBJ) [9]. These are members of the International Nucleotide Sequence Database Collaboration (INSDC) and they are cross-referenced against each other on a daily basis. Similarly, the protein databases contain amino acid sequences derived from translations of the sequences stored in the nucleotide databases or resolved protein structures. The major protein sequence databases are GenPept [7], RefSeq [10], the Protein Information Resource (PIR) [11], the UniProt Knowledgebase (UniProtKB) [12], which consists of the non-redundant, manually curated UniProtKB/Swiss-Prot and its computer-annotated supplement, UniProtKB/TrEMBL, which contains protein sequences translated from the EMBL nucleotide sequence database. The Protein Data Bank (PDB) [13] is the universal repository for the three-dimensional structural data of biological macromolecules (proteins and nucleic acids). Some special databases like ENSEMBL [18], Entrez Genome [19], EK3D, KEGG, LIPID MAPS, OMIM, EPITRANS are also available for a huge range of different information like structural details of *E. coli* K-antigens, diseases related to human genome, Epigenetic and transcriptomic data etc. [14, 15].
Annotation is a major aspect of bioinformatics in sequence analysis. This involves computational gene finding to search for protein-coding genes, RNA genes, and

other functional sequences within a genome. For this purpose, pairwise sequence similarity search methods are employed in order to search the sequence databases for template sequences similar to the target sequence; the target sequence is aligned with each of the template sequences in a database. The computational tools for local alignment include BLAST [16], and SSEARCH [17] and for global alignment include [20] and GGSEARCH [21]. To establish phylogenetic relation, accurate multiple sequence alignment (MSA) is essential and for this purpose, popular tools like CLUSTALW [22], T-Cofee [23] and MUSCLE [24] are available.

The basic information about the structure of a protein comes from its primary sequence. By applying methods like position-specific score matrix (PSSM) [25], Fragment Database Mining (FDM) [27] and Hidden Markov Models (HMMs) [26] are used to create tools for primary and secondary structure prediction. On the basis of these structure and sequence information homology modelling of unknown protein can be done through available tools like SWISS-MODEL [28] and these structures can be validated by 3-d structure analysis tools, LiveBench [29], CASP [30], CAFASP [31] and EVA [32].

Bioinformatics has also solves the problem of visualization and intuitive representation of biological data. Swiss PDB Viewer (also known as DeepView), developed by the Swiss institute of bioinformatics if free, well-known and powerful software package for protein visualization and modeling [33]. VMD (visual molecular dynamics) and PyMol are more powerful free software. Both support scripting on Python and have good quality of graphics. There are also more powerful commercial software packages. For instance Accelrys Discovery Studio [34] is a software package that can solve lots of tasks in molecular modeling. Being a complete software package, Discovery Studio can be integrated to Accelrys Pipeline Pilot to model, simulate and construct protein and their complexes, research their interactions dynamically, develop proteins and make QSAR (Quantitative Structure-Activity Relationship). Discovery Studio also allows dock sequences, research protein-binding site properties, run complex AB initio simulations etc. Discovery Studio backend grants access to NCBI (national center for biotechnology information) data banks and instruments, proteomics protocols, pharmacology, sequence analysis etc [35]. Other free software for docking Autodock and Autodock Vina.

1.3 Impact of development of bioinformatics tools

The first and foremost point to be noted that development of advanced efficient bioinformatics tools has reduced the cost, labor, time limit and errors in a research to many folds. Now the whole genome can be analyzed in only 26 hours. The search of a particular gene, genome, protein across a search sequences of more than 260 000 organisms, containing over 190 billion nucleotides by BLAST is possible in less than a minute. These sequence comparisons are now used to establish the evolutionary relationships between different organisms. Now shuttles or gene markers are available to identify different functional genes, mutations and their location across a genome. These information has opened many gates for personalized medicines for very dangerous diseases. The sequence annotation has unfolded many secrets related to diseases and different metabolic processes.

The development of bioinformatics tools and access to the large stored experimental data, enabled to unfold the mechanisms of interactions of atoms within a molecule and between molecules. Computational algorithms for molecular simulations has unfolded many mysteries like identifying new dual function of a protein (Wzi) in *E. coli* outer membrane [36]. The behavior of a macromolecule in a particular condition can also be studied through the bioinformatics tools. Now, *in-situ* study of ligand substrate interactions can be studied with a better approximation, thus proving much efficient information for drug design and development. All these points have made the life science research a stronger field in terms of research and production of efficient solutions.

1.4 Scope of the study

Compared with the traditional experimental studies, bioinformatics has provided many sophisticated, extremely valuable, easily-accessible and user-friendly tools for analysis in life science research and its updating every day. However, the integration of these bioinformatics tools and experimental data is the current challenge. With the rapid growth of both genomic and clinical phenotypic data, biomedical informatics would play even a more important role in every step of the development of bioscience and medical research, from design to analysis, from diagnosis to treatment, from prognosis to prevention. There is a need of developing many more efficient bioinformatics tools for aiding the analysis and discover many hidden secrets of biological phenomena. In this prospective, I have been working on three fields; modelling of 3 dimensional nucleic acid structures, finding the steric hindrance in molecules and identifying the amino acids surrounding the sugar ligand in a catalytic pocket to identify catalysis sites in the protein as they are highly conserved invariant of the source organism. The nucleic acid models generated through our server can be utilized as starting model for different research purposes related to nucleic acids and interactions associated to them. Steric hindrance identifier will help in validating models and the identified conserved aminoacids will further be used in automated identification of catalytic sites in glycosidases and can also be useful in explaining the cause specificity and selectivity of the proteins towards a specific sugar moiety.

1.5 References

[1] Chial, H. (2008) DNA sequencing technologies key to the Human Genome Project. Nature Education 1(1):219

[2] Hogeweg P (2011). Searls, David B., ed. "The Roots of Bioinformatics in Theoretical Biology". PLoS Computational Biology 7 (3): e1002021.

[3] Hesper B, Hogeweg P (1970). "Bioinformatica: een werkconcept" 1 (6). Kameleon: 28–29.

[4] Wong, KC (2016). Computational Biology and Bioinformatics: Gene Regulation. CRC Press (Taylor & Francis Group). ISBN 9781498724975.

[5] Yan Q (2010) Translational bioinformatics and systems biology approaches for personalized medicine. Methods Mol Biol 662:167–178.

[6] Attwood TK, Gisel A, Eriksson NE, Bongcam-Rudloff E (2011). "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". Bioinformatics – Trends and Methodologies. InTech. Retrieved 8 Jan 2012.

[7] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Sayers EW: GenBank. Nucleic Acids Res 37: D26-D31, 2009.

[8] Kulikova T, Akhtar R, Aldebert P, et al: EMBL nucleotide sequence database in 2006. Nucleic Acids Res 35: D16-D20, 2007.

[9] Sugawara H, Ogasawara O, Okubo K, Gojobori T and Tateno Y: DDBJ with new system and face. Nucleic Acids Res 36: D22-D24, 2008.

[10] (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-D65, 2007.

[11] Wu CH, Huang H, Arminski L, et al: The protein information resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res 30: 35-37, 2002.

[12] Bairoch A, Apweiler R, Wu CH, et al: The universal protein resource (UniProt). Nucleic Acids Res 33: D154-D159, 2005.

[13] Berman HM, Westbrook J, Feng Z, et al: The Protein Data Bank. Nucleic Acids Res 28: 235-242, 2000.

[14] Pavlopoulou A, Michalopoulos I. State-of-the-art bioinformatics protein structure prediction tools (Review). Int J Mol Med. 2011; 28:295–310.

[15] Qing Yan (ed.), Pharmacogenomics in Drug Discovery and Development, Methods in Molecular Biology, vol. 1175, DOI 10.1007/978-1-4939-0956-8_2, © Springer Science+Business Media New York 2014.

[16] Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 215: 403-410, 1990.

[17] Smith TF and Waterman MS: Identification of common molecular subsequences. J Mol Biol 147: 195-197, 1981.

[18] Hubbard TJ, Aken BL, Ayling S, et al: Ensembl 2009. Nucleic Acids Res 37: D690-D697, 2009.

[19] Wheeler DL, Barrett T, Benson DA, et al: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-D21, 2008.

[20] Pearson WR: Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183: 63-98, 1990.

[21] Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453, 1970.

[22] Larkin MA, Blackshields G, Brown NP, et al: Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948, 2007.

[23] Notredame C, Higgins DG and Heringa J: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217, 2000.

[24] Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797, 2004.

[25] Henikoff S and Henikoff JG: Embedding strategies for effective use of information from multiple sequence alignments. Protein Sci 6: 698-705, 1997.

[26] Gribskov M, McLachlan AD and Eisenberg D: Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84: 4355-4358, 1987.

[27] Cheng H, Sen TZ, Kloczkowski A, Margaritis D and Jernigan RL: Prediction of protein secondary structure by mining structural fragment database. Polymer (Guildf) 46: 4314-4321, 2005.

[28] Kiefer F, Arnold K, Kunzli M, Bordoli L and Schwede T: The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37: D387-D392, 2009.

[29] Rychlewski L and Fischer D: LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 14: 240-245, 2005.

[30] Moult J, Fidelis K, Zemla A and Hubbard T: Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins (Suppl 5): S2-S7, 2001.

[31] Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR and Elofsson A: CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 53 (Suppl 6): S503-S516, 2003.

[32] Eyrich VA, Marti-Renom MA, Przybylski D, et al: EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17: 1242-1243, 2001.

[33] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997; 18(15): 2714-23.

[34] Accelrys Software Inc. [Electronic resource]. URL: http://accelrys.com.

[35] Tikhvinskiy DA, Porozov YuB. Bioinformatics and tools for computer analysis and visualization of macromolecules. Russian Open Medical Journal 2013; 2: 0101.

[36] Shivangi Sachdeva, Narendar Kolimi, Sanjana Anilkumar Nair and Thenmalarchelvi Rathinavelan, Key diffusion mechanisms involved in regulating bidirectional water permeation across E. coli outer membrane lectins, ScientificRepoRts | 6:28157 | DOI: 10.1038/srep28157

Chapter 2

Extension of 3D-NuS: A Web Server for Automated Modelling and Visualization of 3-Dimensional Nucleic Acid Structures

(With emphasis on triplexes and G-quadruplexes)

(www.iith.ac.in/3dnus/)

2.1 Introduction

DNA, RNA and their complexes play essential roles in various cellular processes such as mismatch repair, replication, recombination, transcription, translation etc. by forming duplex, triplex, quadruplex, hairpin etc. structures. For instance, four stranded guanine quadruplex structures whose existence in mammalian cells is detected very recently [8] is expected to participate in the regulation of replication, recombination, transcription and chromosome stability [4, 14, 16, 21]. These structures are also considered to be a good anticancer drug targets due to their presence in telomeres [2] and promoter region of oncogenes [29]. Similarly, RNA-DNA hybrid structures are found to have major roles in transcription [20, 24], replication [13] and gene editing [28]. RNA-DNA hybrid duplex also plays an important role in the therapeutics of antisense strategy [11]. Yet another secondary structure formed by nucleic acids is triplexes (three stranded nucleic acid structures) that are assumed to have a role in gene regulation due to the predominant presence of long polypurine stretch in the genome of eukaryotes [3], a prerequisite for triplex formation. A very recent finding about the complex formation between miRNA & mRNA resulting in RNA.RNA*RNA (herein onwards, '*' indicates the interaction between purine strand of the duplex and the third strand) triplex and miRNA &

DNA duplex resulting in DNA.DNA*RNA triplex [18] further strengthens the role of triplexes in gene regulation. Further, presence of DNA.DNA*DNA triplex that leads to transcription inhibition of frataxin gene in Fridreich's ataixa patients is observed [7]. Triplex mediated antigene strategy of gene regulation is also well established since 1987 [10, 15].

To elucidate the functions of aforementioned nucleic acids in various biological processes, their 3 dimensional structural information is necessary at any sequence context and length. Nonetheless, determining the structures of these nucleic acids by experimental techniques are bounded by certain limitations. For instance, nucleic acids triplexes are not tractable to structure determination by X-ray crystallography or NMR techniques. Only limited RNA-DNA hybrid structures are deposited in protein databank. Further, exploring the structural distortions created by various base pair mismatches and the associated mechanisms behind DNA mismatch recognition by mismatch repair proteins require their 3dimensional structural information at different sequence contexts. Such structural information would be very helpful in understanding the ineffectiveness of mismatch repair proteins in certain circumstances such as in TREDs [22]. Base pair mismatches in RNA duplex also play a crucial role in the misregulation of alternative splicing [17] and thus, necessitate the structural details of base pair mismatch containing RNA duplexes. In order to addresses these issues, a reasonably good starting model is essential and thus, molecular modeling becomes as an alternate. In fact, molecular modeling studies have been exploited to derive structural information of triplexes [25, 26, 6] as well as to explore the influence of certain mismatches on DNA structures that are not readily accessible through experimental techniques [9]. Thus, to facilitate the establishment of biological functions of nucleic acid duplexes, triplexes and quadruplexes, our lab has developed a web server called 3D-NuS.

3D-NuS, 3D-NuS: A Web Server for Automated Modeling and Visualization of 3-Dimensional Nucleic Acid Structures, generates energy minimized models of nucleic acid structures for the user defined sequence and length. It can build the 3dimensional (3D) structures of i) RNA-DNA hybrid duplex, ii) DNA/RNA duplex with noncanonical base pairs in an automated way. The models built by 3D-NuS will be useful to explore the dynamics of the aforementioned molecules as well as their docking with proteins and ligand molecules to investigate their potential biological roles as well as their applications in pharmaceutical industry and synthetic biology. Although web-servers like w3DNA [12], model.it [27], and 3D-DART [26] are involved in nucleic acid structure modeling, our server is different from others in terms of providing energy minimized models of mismatch containing DNA&RNA duplexes. The server is freely accessible through http://iith.ac.in/3dnus/ without any login information.

2.2 Scope of the study

As mentioned previously in the introduction part, triplexes and quadruplexes also has crucial role in gene regulation, recombination and chromosome stability like biological processes, their structural study and analysis is very important. For instance, nucleic acids triplexes are not tractable to structure determination by X-ray crystallography or NMR techniques. Hence, to understand the structural details using computational methods, we need to have good starting models for triplexes and quadruplexes too. In this context the method duplex model generation from 3D-NuS server can be used to generate models of triplexes and quadruplexes. These structures may serve as good starting models for docking studies with proteins & small molecules, NMR structure determination, cryo-electron microscope modeling and molecular dynamics simulation studies.

2.3 Objectives

2.3.1 Automated generation of DNA/RNA triplexes and implementation of it in 3D-NuS server.

2.3.2 Automated generation of DNA/RNA G-quadruplexes and implementation of it in 3D-NuS server.

2.4 Materials and Methods

2.4.1 3D-NuS architecture

The user interface of 3D-NuS server is developed using PHP web server scripting language, HTML and Java script. 3D-NuS web server communicates with web browsers via TCP (Transmission Control Protocol) / IP (Internet Protocol). For the interactive visualization of the modeled 3D structures, 3D-NuS utilizes JSmol applet. When a query is submitted, 3D-NuS takes the job and runs it in the background and subsequently, the user is directed to a result page. Thus, the interlinked 3D visualization requires Java plug-in (https://java.com/en/download/) from the user end. The web interface is accessible through a web browser like Mozilla Firefox, Safari, Internet Explorer 6, Google Chrome etc. and provides the facility for submitting jobs & viewing results. Web interface visual enhancement is primarily supported by cascading style sheet (CSS) that enhances the style and display of HTML pages for better viewing by the end users. The algorithms used to generate the models are written in php, python and bash script. Runtime of the job depends on the size & complexity of the user input sequence. Further, 3D-NuS web server does not require any login information for access.

2.4.2 Model Optimization

Cartesian coordinates for Hoogsteen (G*GC, T*AT & C+*GC) & reverse Hoogsteen (G*GC, T*AT & A*AT) base triplets and G-quadrates interacting via Hoogsteen hydrogen bonding scheme, but, differs in their glycosyl conformations depending on the quadruplex group (parallel or antiparallel or mixed) are either taken from previous experimental data or modeled manually and stored in the web server along with the helical twist and rise information for each category. When a user requests for a model, these information have been used to generate the model.

As the generated models may contain steric hindrance and may not have proper connectivity between the adjacent nucleotides, they are subjected to energy minimization using Xplor-NIH by applying Powell minimization algorithm. During the energy minimization, base pair hydrogen bonding patterns, sugar-phosphate backbone & glycosyl torsion angle restraints are imposed along with a mild base pair planarity restraint.

2.5 Results and Discussion

2.5.1 Technical overview of 3D-NuS web server

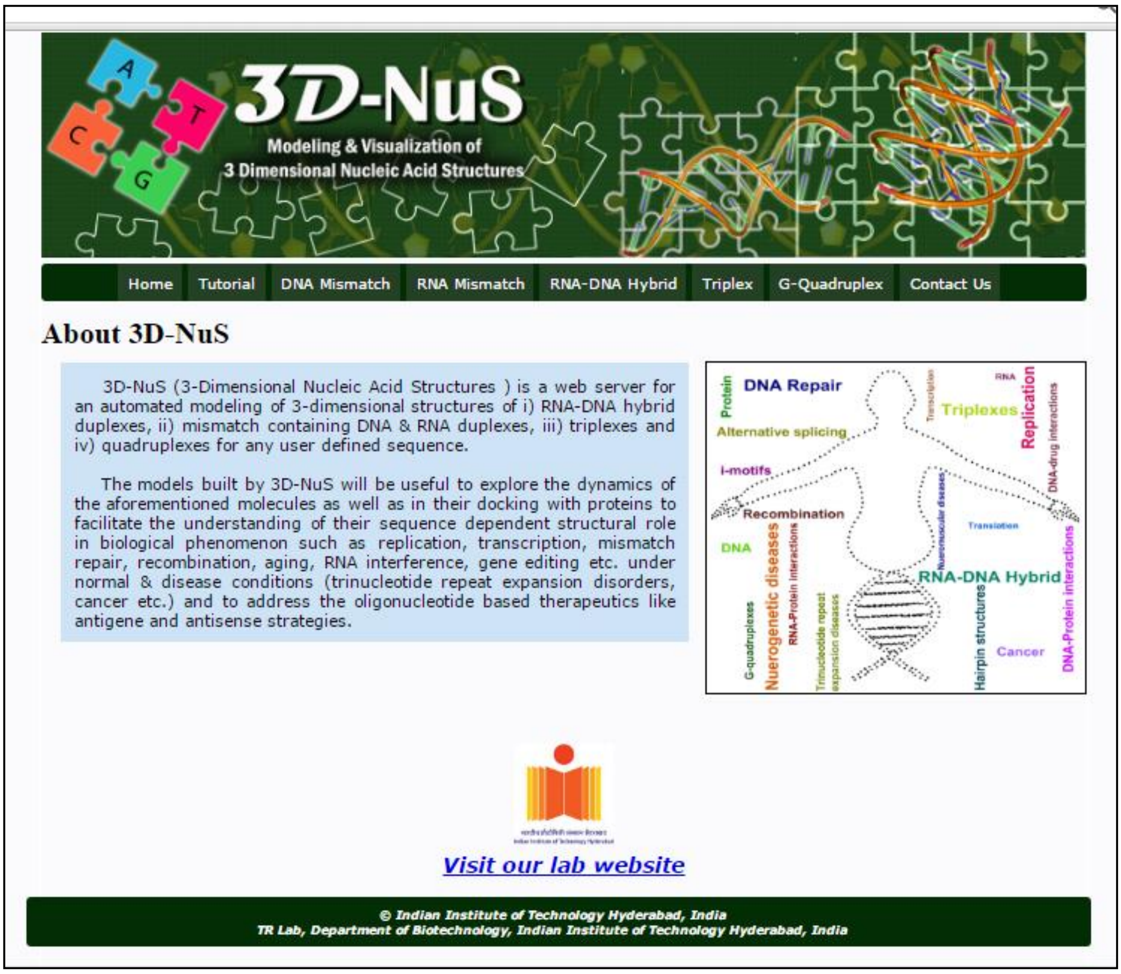
3D-NuS web server facilitates the automated modeling of 3dimensional structures of i) RNA-DNA hybrid duplexes, ii) mismatch containing DNA & RNA duplexes, iii) triplexes and iv) quadruplexes for any user defined sequence. The user can choose the nucleic acid structures of their requirement from the main menu that will consequently lead to the appropriate webpage. Triplex and Quadruplex modelling pages can be accessed through web address http://iith.ac.in/3dnus/triplex and http://iith.ac.in/3dnus/quadruplex respectively. The home page of 3D-NuS web server is shown in Figure 2.1

2.5.1.1 Input

Nucleic acid sequences of user desired length with sequences specified in case insensitive single letter code (e.g. A/a, T/t, U/u, C/c & G/g).

2.5.1.2 Modeling nucleic acids triplexes

Triplex helical nucleic acid structures are categorized into 2 groups: when the orientation of the third strand is parallel (parallel triplex formed by Hoogsteen hydrogen bonds) and antiparallel (antiparallel triplex formed by reverse Hoogsteen hydrogen bonds) with respect to the purine strand of the duplex. This can be further subdivided into 8 classes depending on the molecular nature (DNA or RNA) of the strands involved in the formation of triplexes (Figure 2.1). 3D-NuS offers choice for the users to model both parallel and antiparallel triplexes for desired combinations of



About 3D-NuS

3D-NuS (3-Dimensional Nucleic Acid Structures) is a web server for an automated modeling of 3-dimensional structures of i) RNA-DNA hybrid duplexes, ii) mismatch containing DNA & RNA duplexes, iii) triplexes and iv) quadruplexes for any user defined sequence.

The models built by 3D-NuS will be useful to explore the dynamics of the aforementioned molecules as well as in their docking with proteins to facilitate the understanding of their sequence dependent structural role in biological phenomenon such as replication, transcription, mismatch repair, recombination, aging, RNA interference, gene editing etc. under normal & disease conditions (trinucleotide repeat expansion disorders, cancer etc.) and to address the oligonucleotide based therapeutics like antigene and antisense strategies.

© Indian Institute of Technology Hyderabad, India
TR Lab, Department of Biotechnology, Indian Institute of Technology Hyderabad, India

Figure 2.1. 3D-NuS Homepage.

RNA or/and DNA strands. Triplexes can be formed by both isomorphic (structurally similar) and nonisomorphic (structurally dissimilar) base triplets (Figure 2.2), depending on whether the third strand is pyrimindine (isomorphic) or purine (nonisomorphic) rich. Thus, 3D-NuS facilitate the modeling of both purine rich and pyrimidine rich triplexes, wherein, the third strand interacts with the purine strand of the duplex by a pair of hydrogen bonds. The web server allows triplex modeling for the following base triplet combinations: G*GC&T*AT (parallel), G*GC&T*AT (antiparallel), G*GC&A*AT (antiparallel) & C+*GC&T*AT (parallel) apart from homopolymeric T*AT (parallel), A*AT (antiparallel), T*AT (antiparallel), G*GC (antiparallel), G*GC (parallel) & C+*GC (parallel) triplexes. At this moment, 3D-NuS doesn't support base triplets with mismatches, in which, the purine strand of the duplex and the third strands are indulged through a single hydrogen bond. To model the triplex using 3D-NuS, the user has to select the appropriate group and class of the triplex from the dropdown menu along with the nucleic acid sequences of their choice for all the three strands. 3D-NuS assigns 1st, 2nd and 3rd sequences for the pyrimidine, purine and Hoosteen/reverse Hoogsteen (third) strands respectively. Here, the orientations of the first and second strands as mentioned before, whereas, the third strand orientation is depending on the user's choice. Hoogsteen and reverse Hoogsteen hydrogen bonding schemes used for various base triplets by 3D-NuS are shown in Figure 2.2.

2.5.1.3 Modeling nucleic acids quadruplexes

The G-quadruplex molecules generated by 3D-NuS are categorized into 3 groups (parallel, anti-parallel and mixed) according to their strand orientations and 6 classes according to the molecular nature (DNA or RNA) of the strands (Figure 2.3). The 4 guanines in the G-quadrates are interacting through Hoogsteen hydrogen bonding scheme (Figure 2.3) accompanied by syn or anti glycosyl conformation based on the orientations of the guanine strands. While the four neighboring guanine strands are parallel to each other in parallel quadruplexes (all anti conformations), they are antiparallel (alternate G-strands are in syn and anti conformations respectively) to each other in antiparallel quadruplexes. In the mixed quadruplexes, 2 guanine pairs are engaged through parallel mode with anti glycosyl conformations whereas, the remaining 2 are interacting through an antiparallel orientation with syn glycosyl conformations. Note that here we are considering only 3 groups of qudruplexes, as 3D-NuS generates only intermolecular (tetramolecular) quadruplexes.

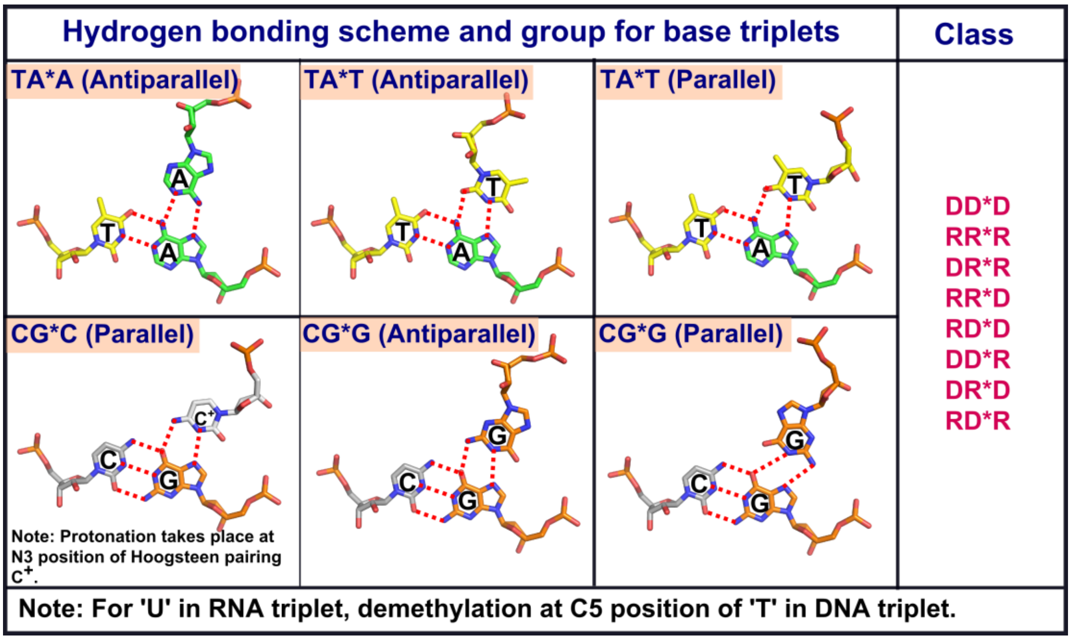

Hydrogen bonding scheme and group for base triplets
Class
TA*A (Antiparallel)
TA*T (Antiparallel)
TA*T (Parallel)
CG*C (Parallel)
CG*G (Antiparallel)
CG*G (Parallel)
A
T
A
T
T
A
T
T
A
C+
C
G
G
C
G
G
C
G
Note: Protonation takes place at N3 position of Hoogsteen pairing C+.
DD*D
RR*R
DR*R
RR*D
RD*D
DD*R
DR*D
RD*R
Note: For 'U' in RNA triplet, demethylation at C5 position of 'T' in DNA triplet.

Figure 2.2. Different combinations of DNA&RNA strands of the triplex, third strand orientation with respect to the purine strand of the duplex and base triplet hydrogen bonding schemes that are considered for triplex modeling in 3D-NuS. Note that D&R represent DNA&RNA respectively.

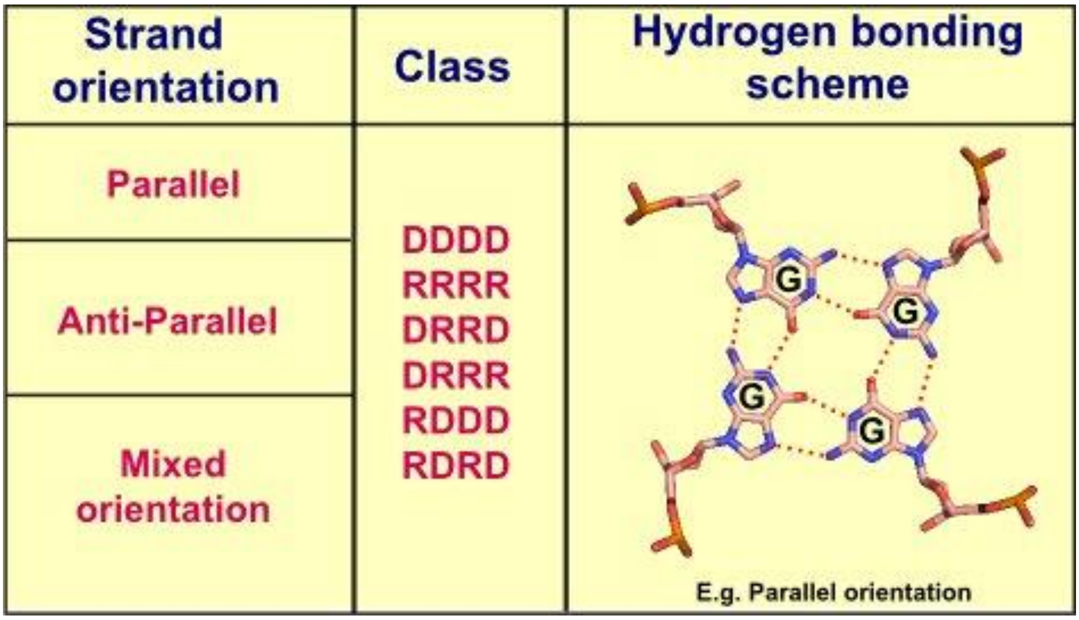
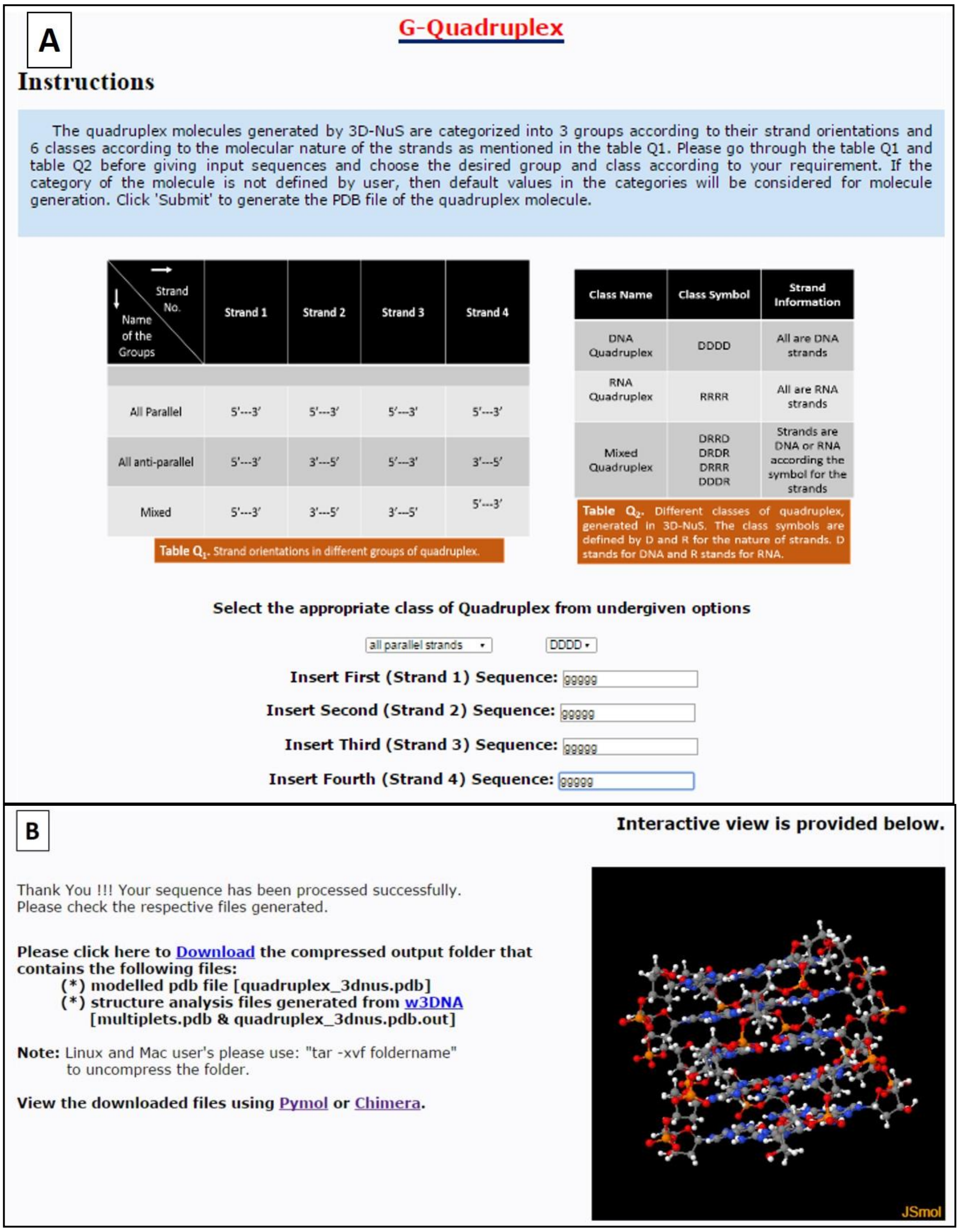


Figure 2.3. Different groups of G-quadruplex (Left) & combinations of DNA&RNA strands (Middle) used for quadruplex modeling in 3D-NuS and a representative G-quartet (Right). Note that D&R represent DNA&RNA respectively.

Figure 2.4 illustrate the functionality of the 3D-NuS web server in modeling G-quadruplex of class DDDD.



G-Quadruplex

Instructions

The quadruplex molecules generated by 3D-NuS are categorized into 3 groups according to their strand orientations and 6 classes according to the molecular nature of the strands as mentioned in the table Q1. Please go through the table Q1 and table Q2 before giving input sequences and choose the desired group and class according to your requirement. If the category of the molecule is not defined by user, then default values in the categories will be considered for molecule generation. Click 'Submit' to generate the PDB file of the quadruplex molecule.

Strand No. → Name of the Groups ↓	Strand 1	Strand 2	Strand 3	Strand 4
All Parallel	5′---3′	5′---3′	5′---3′	5′---3′
All anti-parallel	5′---3′	3′---5′	5′---3′	3′---5′
Mixed	5′---3′	3′---5′	3′---5′	5′---3′

Table Q_1. Strand orientations in different groups of quadruplex.

Class Name	Class Symbol	Strand Information
DNA Quadruplex	DDDD	All are DNA strands
RNA Quadruplex	RRRR	All are RNA strands
Mixed Quadruplex	DRRD DRDR DRRR DDDR	Strands are DNA or RNA according the symbol for the strands

Table Q_2. Different classes of quadruplex, generated in 3D-NuS. The class symbols are defined by D and R for the nature of strands. D stands for DNA and R stands for RNA.

Select the appropriate class of Quadruplex from undergiven options

all parallel strands ▾ DDDD ▾

Insert First (Strand 1) Sequence: ggggg

Insert Second (Strand 2) Sequence: ggggg

Insert Third (Strand 3) Sequence: ggggg

Insert Fourth (Strand 4) Sequence: ggggg

Interactive view is provided below.

Thank You !!! Your sequence has been processed successfully.
Please check the respective files generated.

Please click here to Download the compressed output folder that contains the following files:
(*) modelled pdb file [quadruplex_3dnus.pdb]
(*) structure analysis files generated from w3DNA [multiplets.pdb & quadruplex_3dnus.pdb.out]

Note: Linux and Mac user's please use: "tar -xvf foldername" to uncompress the folder.

View the downloaded files using Pymol or Chimera.

Figure 2.4. Functionality of 3D-NuS web server. Stepwise illustration of generating G-quadruplex using 3D-NuS from input page [A] to result page [B].

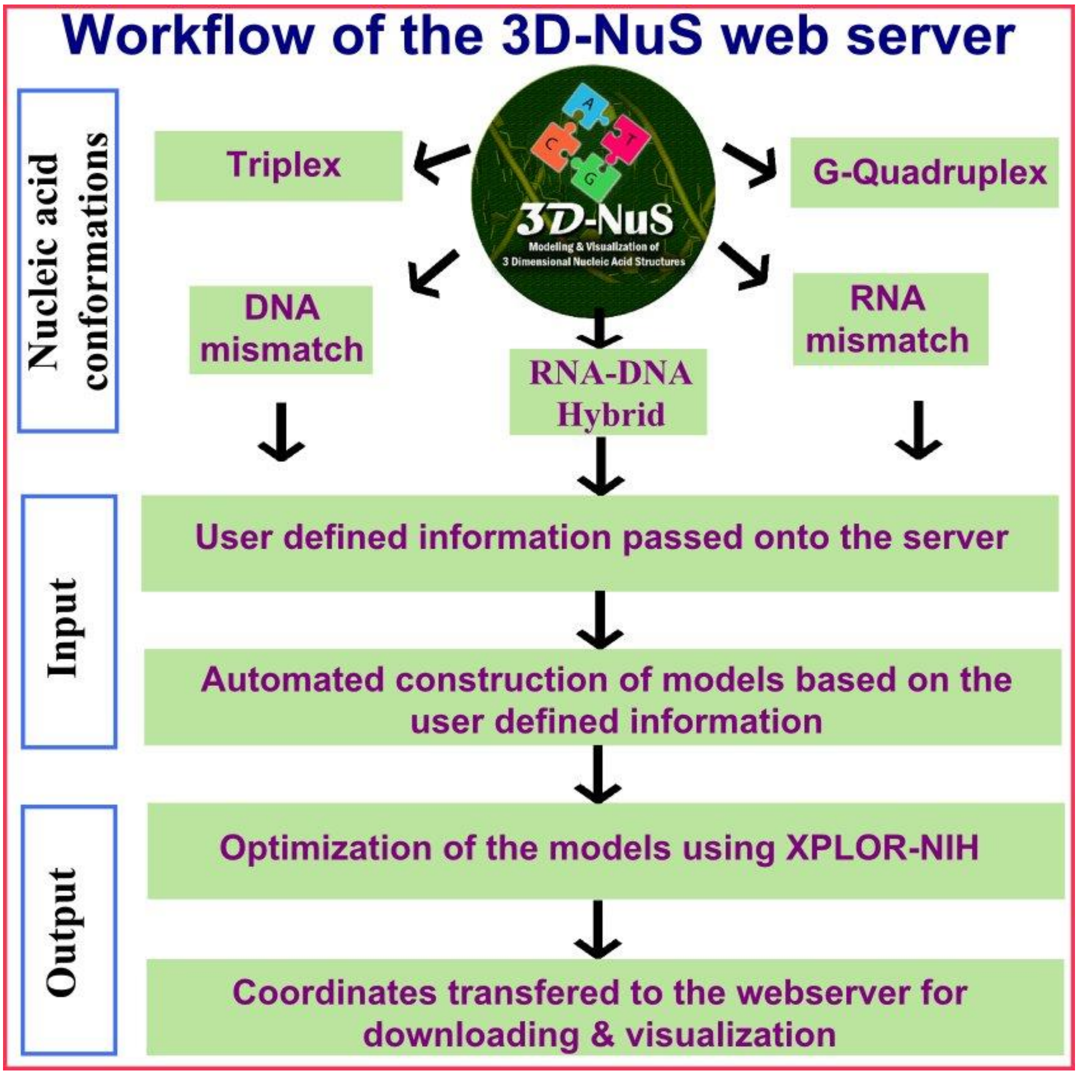

Workflow of the 3D-NuS web server
Nucleic acid conformations
Triplex
G-Quadruplex
3D-NuS
Modeling & Visualization of 3 Dimensional Nucleic Acid Structures
DNA mismatch
RNA-DNA Hybrid
RNA mismatch
Input
User defined information passed onto the server
Automated construction of models based on the user defined information
Output
Optimization of the models using XPLOR-NIH
Coordinates transfered to the webserver for downloading & visualization

Figure 2.5 Workflow in 3D-NuS. The user can define the molecule of their choice in the first place and subsequently the sequence information along with other information like type of mismatches (RNA & DNA duplexes), DNA & RNA composition (viz., class in triplexes & quadruplexes) and strand polarity (viz., group in triplexes & quadruplexes). Models generated as per the user's request is subjected to energy minimization using Xplor-NIH and the Cartesian coordinates are provided to the user in PDB format for the future use that can also be interactively visualized in the web page.

2.5.1.4 Output

Cartesian coordinates of the energy minimized 3D models can be downloaded in protein data bank (PDB) format for future use as well as visualized in an interactive manner using the JSmol (http://www.jmol.org/) Java applet in the webpage itself. The workflow of the web server is given in Figure 2.5.

2.6 Conclusion

The extended version of 3D-NuS can be used for automated modeling and visualization of 3dimensional nucleic acids structures like i) RNA-DNA hybrid duplex, ii) DNA/RNA duplex with noncanonical base pairs iii) triplexes and iv) quadruplexes. The web server is very flexible and user friendly such that, the user has to simply feed the nucleic acid sequences of their interest and length as described in the documentation and within couple of minutes the Cartesian coordinates of the energy minimized model can be downloaded in PDB format and can also be visualized interactively in the webpage itself. The models built by 3D-NuS can serve as good starting models for NMR structure refinement as well as for exploring the conformational dynamics of the aforementioned molecules. The models can also be useful for docking with proteins or nucleic acids or small molecules to facilitate the understanding of their sequence dependent structural role in biological phenomenon such as replication, transcription, mismatch repair, recombination, aging, RNA interference, gene editing etc. under normal & disease conditions (trinucleotide repeat expansion disorders, cancer etc.) and to address the oligonucleotide based therapeutics like antigene and antisense strategies.

2.7 References

[1] AGUILERA, A. & GARCIA-MUSE, T. 2012. R loops: from transcription byproducts to threats to genome stability. Mol Cell, 46, 115-24.

[2] BALASUBRAMANIAN, S., HURLEY, L. H. & NEIDLE, S. 2011. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? Nat Rev Drug Discov, 10, 261-75.

[3] BEHE, M. J. 1995. An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. Nucleic Acids Res, 23, 689-95.

[4] BOCHMAN, M. L., PAESCHKE, K. & ZAKIAN, V. A. 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet, 13, 770-80.

[5] BRUNGER, A. T. 1996. Recent developments for crystallographic refinement of macromolecules. Methods Mol Biol, 56, 245-66.

[6] GOLDSMITH, G., RATHINAVELAN, T. & YATHINDRA, N. 2016. Selective Preference of Parallel DNA Triplexes Is Due to the Disruption of Hoogsteen Hydrogen Bonds Caused by the Severe Nonisostericity between the G*GC and T*AT Triplets. PLoS One, 11, e0152102.

[7] GRABCZYK, E., MANCUSO, M. & SAMMARCO, M. C. 2007. A persistent RNA.DNA hybrid formed by transcription of the Friedreich ataxia triplet repeat in live bacteria, and by T7 RNAP in vitro. Nucleic Acids Res, 35, 5351-9.

[8] HENDERSON, A., WU, Y., HUANG, Y. C., CHAVEZ, E. A., PLATT, J., JOHNSON, F. B., BROSH, R. M., JR., SEN, D. & LANSDORP, P. M. 2014. Detection of G-quadruplex DNA in mammalian cells. Nucleic Acids Res, 42, 860-9.

[9] KHAN, N., KOLIMI, N. & RATHINAVELAN, T. 2015. Twisting right to left: A...A mismatch in a CAG trinucleotide repeat overexpansion provokes left-handed Z-DNA conformation. PLoS Comput Biol, 11, e1004162.

[10] LE DOAN, T., PERROUAULT, L., PRASEUTH, D., HABHOUB, N., DECOUT, J. L., THUONG, N. T., LHOMME, J. & HELENE, C. 1987. Sequence-specific recognition, photocrosslinking and cleavage of the DNA double helix by an oligo-[alpha]-thymidylate covalently linked to an azidoproflavine derivative. Nucleic Acids Res, 15, 7749-60.

[11] LIMA, W. F., ROSE, J. B., NICHOLS, J. G., WU, H., MIGAWA, M. T., WYRZYKIEWICZ, T. K., SIWKOWSKI, A. M. & CROOKE, S. T. 2007. Human RNase H1 discriminates between subtle variations in the structure of the heteroduplex substrate. Mol Pharmacol, 71, 83-91.

[12] LU, X. J. & OLSON, W. K. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res, 31, 5108-21.

[13] MASUKATA, H. & TOMIZAWA, J. 1990. A mechanism of formation of a persistent hybrid between elongating RNA and template DNA. Cell, 62, 331-8.

[14] MENDOZA, O., BOURDONCLE, A., BOULE, J. B., BROSH, R. M., JR. & MERGNY, J. L. 2016. G-quadruplexes and helicases. Nucleic Acids Res, 44, 1989-2006.

[15] MOSER, H. E. & DERVAN, P. B. 1987. Sequence-specific cleavage of double helical DNA by triple helix formation. Science, 238, 645-50.

[16] MURAT, P. & BALASUBRAMANIAN, S. 2014. Existence and consequences of G-quadruplex structures in DNA. Curr Opin Genet Dev, 25, 22-9.

[17] MYKOWSKA, A., SOBCZAK, K., WOJCIECHOWSKA, M., KOZLOWSKI, P. & KRZYZOSIAK, W. J. 2011. CAG repeats mimic CUG repeats in the misregulation of alternative splicing. Nucleic Acids Res, 39, 8938-51.

[18] PAUGH, S. W., COSS, D. R., BAO, J., LAUDERMILK, L. T., GRACE, C. R., FERREIRA, A. M., WADDELL, M. B., RIDOUT, G., NAEVE, D., LEUZE, M., LOCASCIO, P. F., PANETTA, J. C., WILKINSON, M. R., PUI, C. H., NAEVE, C. W., UBERBACHER, E. C., BONTEN, E. J. & EVANS, W. E. 2016. MicroRNAs Form Triplexes with Double Stranded

DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. PLoS Comput Biol, 12, e1004744.

[19] RATHINAVELAN, T. & YATHINDRA, N. 2006. Base triplet nonisomorphism strongly influences DNA triplex conformation: effect of nonisomorphic G* GC and A* AT triplets and bending of DNA triplexes. Biopolymers, 82, 443-61.

[20] REDDY, K., TAM, M., BOWATER, R. P., BARBER, M., TOMLINSON, M., NICHOL EDAMURA, K., WANG, Y. H. & PEARSON, C. E. 2011. Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. Nucleic Acids Res, 39, 1749-62.

[21] RHODES, D. & LIPPS, H. J. 2015. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res, 43, 8627-37.

[22] SCHMIDT, M. H. & PEARSON, C. E. 2016. Disease-associated repeat instability and mismatch repair. DNA Repair (Amst), 38, 117-26.

[23] SKOURTI-STATHAKI, K. & PROUDFOOT, N. J. 2014. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev, 28, 1384-96.

[24] TEMIAKOV, D., MENTESANA, P. E., MA, K., MUSTAEV, A., BORUKHOV, S. & MCALLISTER, W. T. 2000. The specificity loop of T7 RNA polymerase interacts first with the promoter and then with the elongating transcript, suggesting a mechanism for promoter clearance. Proc Natl Acad Sci U S A, 97, 14109-14.

[25] THENMALARCHELVI, R. & YATHINDRA, N. 2005. New insights into DNA triplexes: residual twist and radial difference as measures of base triplet non-isomorphism and their implication to sequence-dependent non-uniform DNA triplex. Nucleic Acids Res, 33, 43-55.

[26] VAN DIJK, M. & BONVIN, A. M. 2009. 3D-DART: a DNA structure modelling server. Nucleic Acids Res, 37, W235-9.

[27] VLAHOVICEK, K., KAJAN, L. & PONGOR, S. 2003. DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res, 31, 3686-7.

[28] YOSHIMI, K., KANEKO, T., VOIGT, B. & MASHIMO, T. 2014. Allele-specific genome editing and correction of disease-associated phenotypes in rats using the CRISPR-Cas platform. Nat Commun, 5, 4240.

[29] YUAN, L., TIAN, T., CHEN, Y., YAN, S., XING, X., ZHANG, Z., ZHAI, Q., XU, L., WANG, S., WENG, X., YUAN, B., FENG, Y. & ZHOU, X. 2013. Existence of G-quadruplex structures in promoter region of oncogenes confirmed by G-quadruplex DNA cross-linking strategy. Sci Rep, 3, 1811.

Chapter 3

STRIDER: Steric hindrance estimator

3.1 Introduction

As biomacromolecules like carbohydrates, nucleic acids, proteins and lipids are the pivotal building blocks of cell structure, they play a key role in organelle & cellular functions and inter- organelle & cellular communications. Establishment of their role in the formation & function of organelles and cells under normal and disease states necessitates the atomistic details about biomolecular structure & dynamics. As there are limitations in exploring these biomolecular structures through experimental techniques like X-ray crystallography, NMR and cryo-electron microscopy, molecular modeling becomes a promising strategy. For instance, molecular modeling acts as a supplement to experimental techniques to derive structural information of supramolecular nanomachines like bacterial secretion systems to establish the associated functional mechanisms in virulence exportation into the host. Further, nucleic acids triplexes are not tractable to structure determination by experimental techniques though it has been presumed to have a vital function in gene regulation [1] besides their promising role in antigene therapy [2, 3]. Thus, molecular modeling is crucial in establishing the biological significance of nucleic acids triplexes. Similarly, to understand (i) the association of various noncanonical base pair mismatches in mismatch repair pathway under various diseases/disorders, (ii) sequence dependent influence of various base step and base pair parameters in chromatin folding and (iii) the conformational dynamics of nucleic acids modeling molecular modeling comes into picture. Other important areas where bimolecular modeling has a key role is in homology modeling and de novo protein design. Further, one cannot avoid manual modeling during the docking of protein…nucleic acids, protein…protein, protein…lipid, protein…carbohydrate, nucleic acid…nucleic acid and protein…small ligands to facilitate the understanding of various biological processes like transcription, translation, recombination, gene editing, antimicrobial drug design, vaccine design and host…pathogen interaction. All the above require a good starting model that is lying close to the local minima in the energy landscape with accurate internal geometries and also free from steric hindrance (synonymously, short contact or bump). We have developed a web server namely, STRIDER (www.iith.ac.in/strider/) that calculates the intra- & inter-molecular interatomic distances and reports the user about the steric hindrance, if at all any.

3.2 Materials and Methods

3.2.1 Definition of steric hindrance and acceptable steric region

Each atom within a molecule occupies a certain amount of space. If atoms are brought too close together, there is an associated cost in energy due to overlapping electron clouds (Pauli or Born repulsion), thus steric effects arises and this may affect the molecule's preferred shape (conformation) and reactivity. We define a steric hindrance in a protein as any atomic overlap resulting in van der Waals repulsion. In our approach, identification of steric hindrance region is identified on the basis of distance between atoms. At first a boundary condition of 6Å distance from the target atom is defined. It helps in minimizing the calculation complexity and time consumption. Then, the hindrance cutoff is decided from the sum of individual van der Waals radius of each atoms in the 6Å distance and the target atom. The cutoff value for the hindrance is estimated with following formula.

$$\text{Cutoff} = VR1 + VR2 - 0.5$$

In this formula VR1 and VR2 are the van der Waals radius of target atom and the atom in 6Å distance of the target atom. The values will be in angstrom. 0.5 is subtracted to exclude the short contacts in the consideration. The van der Waals radius of 38 elements [4] are stored in database of server (Table 3.1). Automatically these values are accessed through the algorithm for calculation of cutoff value. Any atom in the 6Å distance of the target atom, other than those in covalent bonds and from neighbor residues, is at a distance less than the cutoff value from target atom will be considered for steric hindrance and will be shown in the output.

Element	Van der Waals radius (in Å)		Element	Van der Waals radius (in Å)
Hydrogen	1.20		Sulfur	1.80
Zinc	1.39		Lithium	1.82
Helium	1.40		Arsenic	1.85
Copper	1.40		Bromine	1.85
Fluorine	1.47		Uranium	1.86
Oxygen	1.52		Gallium	1.87
Neon	1.54		Argon	1.88
Nitrogen	1.55		Selenium	1.90
Mercury	1.55		Indium	1.93
Cadmium	1.58		Thallium	1.96
Nickel	1.63		Iodine	1.98
Palladium	1.63		Krypton	2.02
Gold	1.66		Lead	2.02
Carbon	1.70		Tellurium	2.06
Silver	1.72		Silicon	2.10
Magnesium	1.73		Xenon	2.16
Chlorine	1.75		Tin	2.17
Platinum	1.75		Sodium	2.27
Phosphorus	1.80		Potassium	2.75

Table 3.1 Van der Waals radius of elements.

3.2.2 STRIDER architecture

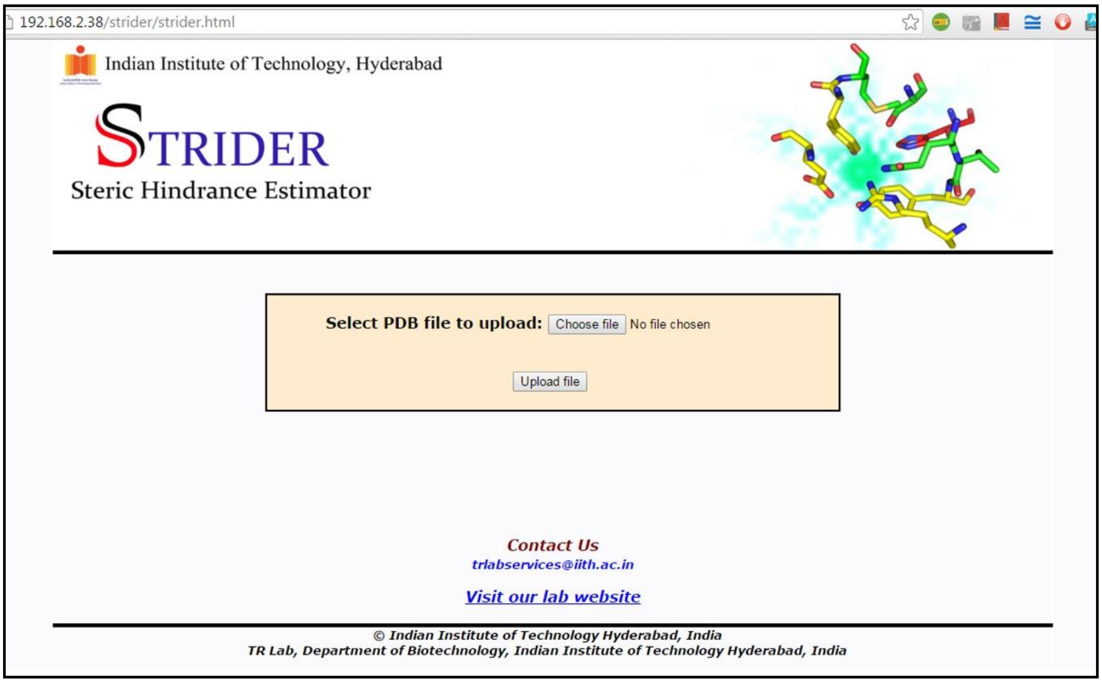
The user interface of STRIDER server is developed using PHP web server scripting language, HTML and Java script. Strider web server communicates with web browsers via TCP (Transmission Control Protocol) / IP (Internet Protocol). STRIDER takes the coordinates of the molecule of interest in the protein databank (PDB) format. When a query is submitted, STRIDER takes the job and runs it in the background and subsequently, the user is directed to a result page. In result page, Strider utilizes JSmol applet for the interactive visualization of the required output. Thus, the interlinked 3D visualization requires Java plug-in (https://java.com/en/download/) from the user end. The web interface is accessible through a web browser like Mozilla Firefox, Safari, Internet Explorer 6, Google Chrome etc. and provides the facility for submitting jobs & viewing results. Web interface visual enhancement is primarily supported by cascading style sheet (CSS) that enhances the style and display of HTML pages for better viewing by the end users. The algorithms used to generate the models are written in php, python and bash script. Runtime of the job depends on the size & complexity of the user input sequence. Further, STRIDER web server does not require any login information for access. The home page of STRIDER is shown in Figure 3.1.

3.3 Results and Discussion

3.3.1 Technical overview of STRIDER web server

3.3.1.1 Implementation

Database of atomic van der Waals radii of 38 elements [4] are stored in a Linux based machine. Access to the data is enabled through Apache web-server management system and PHP web-server scripting language.



192.168.2.38/strider/strider.html

Indian Institute of Technology, Hyderabad

STRIDER

Steric Hindrance Estimator

Select PDB file to upload: Choose file No file chosen

Upload file

Contact Us

trlabservices@iith.ac.in

Visit our lab website

© Indian Institute of Technology Hyderabad, India

TR Lab, Department of Biotechnology, Indian Institute of Technology Hyderabad, India

Figure 3.1. STRIDER Home page

3.3.1.2 Input

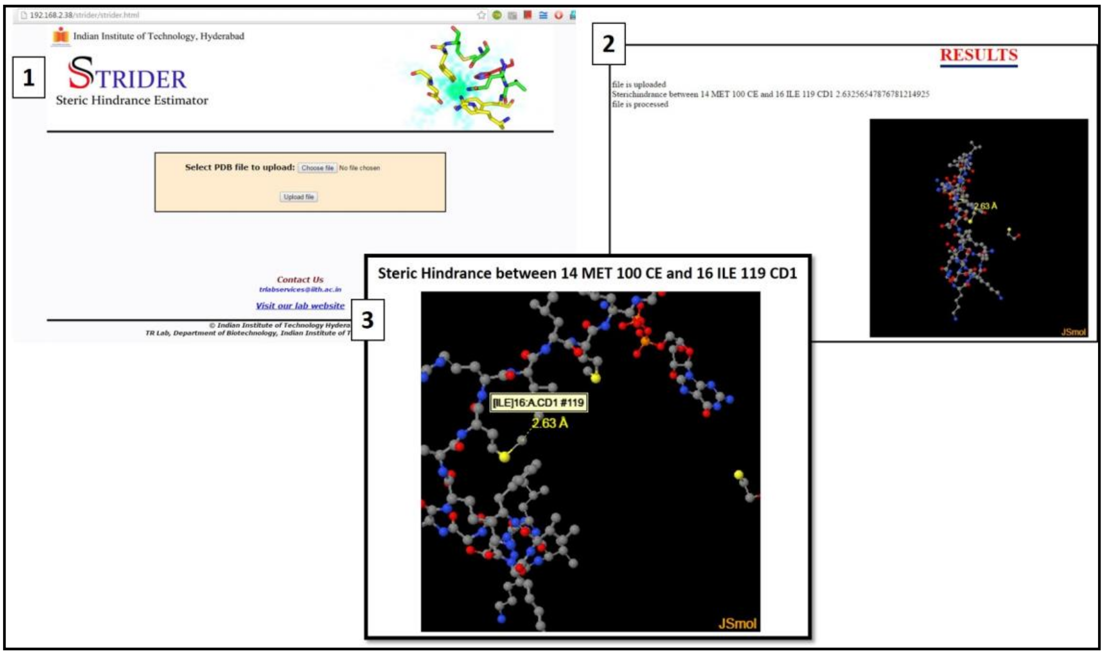
The user has to upload the coordinates of the molecule of interest in the protein databank (PDB) format.

3.3.1.3 Output

Subsequently, the web server identifies the atom given in the PDB and to save the computational time, it creates a list of inter- or intra- molecular atoms that are coming within 6Å distance. In the next step, the server calculates the interatomic distances between the atoms in the list and compares it with the sum of van der Waals radius in the database. Finally, it reports residue wise atoms that are coming within the sum of van der Waals radius along with the expected van der Waals radius. This report is also downloadable in text format. Alongside, it also displays the short contacts indicated by red dotted lines in the 3-dimensional interactive view of the corresponding molecule through Jsmol applet (JSmol: an open-source

HTML5 viewer for chemical structures in 3D). Indication of short contacts in Jsmol is enabled by a java script that gets the list of atoms whose interatomic distance is lower than the sum of van der Waals radius and draws a red dotted lines. Using the Jsmol applet, the user can bring changes to the representation of the molecule, background, etc.

Workflow of the STRIDER web server by considering a modified (incorporating steric region) PDB file, 1E0S, structure of a polypeptide as a case in point is explained in Figure 3.2.



192.168.2.38/strider/strider.html

Indian Institute of Technology, Hyderabad

STRIDER
Steric Hindrance Estimator

Select PDB file to upload: Choose file No file chosen

Upload file

Contact Us
trlabservices@iith.ac.in
Visit our lab website

© Indian Institute of Technology Hydera
TR Lab, Department of Biotechnology, Indian Institute of T

RESULTS

```
file is uploaded
Sterichindrance between 14 MET 100 CE and 16 ILE 119 CD1 2.63256547876781214925
file is processed
```

Steric Hindrance between 14 MET 100 CE and 16 ILE 119 CD1

Figure 3.2. Workflow of the STRIDER web server by considering modified pdb file, 1E0S. The numbering 1-3 shows the workflow of the server after submission of the PDB file.

3.4 Discussion

Molecular modeling is an essential tool in understanding host…pathogen interactions, biological pathways, drug and vaccine discovery, structure prediction, de novo protein design, nanotechnology etc. Examples include, modeling the supramolecular assemblies like bacterial secretion systems, capsular-, lipo- & expo-polysaccharides exportation machineries etc. to facilitate the design of antibacterial drug to hamper the bacterial function. Structure of such a supramolecular assembly cannot be determined just by a single experimental technique, but, in combination with several experimental techniques along with molecular modeling. Similarly, manual modeling comes to picture when experimental techniques fail to get the structure of the macromolecules such as triple helical nucleic acids. During modeling, one has to be cautious as the modeled structure may be good in terms of internal geometry, but, may have steric hindrance. Thus, the constructed models have to be verified for their steric free nature. In this context, STRIDER web server acts as an important tool not only in the modeling of biomolecular structures like carbohydrates, lipids, nucleic acids, proteins and their complexes with themselves and small molecules, but also, in the modeling of any chemical structure in protein databank format.

3.5 References

[1] BEHE, M. J. 1995. An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. Nucleic Acids Res, 23, 689-95.

[2] MOSER, H. E. & DERVAN, P. B. 1987. Sequence-specific cleavage of double helical DNA by triple helix formation. Science, 238, 645-50.

[3] LE DOAN, T., PERROUAULT, L., PRASEUTH, D., HABHOUB, N., DECOUT, J. L., THUONG, N. T., LHOMME, J. & HELENE, C. 1987. Sequence-specific recognition, photocrosslinking and cleavage of the DNA double helix by an oligo-[alpha]-thymidylate covalently linked to an azidoproflavine derivative. Nucleic Acids Res, 15, 7749-60.

[4] PERIODIC TABLE, Theodore Gray, Nick Mann, Max Whitby, RGB Research www.periodictable.com/Properties/A/VanDerWaalsRadius.v.html

Chapter 4

Identification of amino acids involved in the functional diversity of six glycosidase groups

4.1 Introduction

Glycosides or glycoside hydrolase is a sub class of glycosylases sub class of hydrolases enzyme class, as classified by the Enzyme Commission (E.C.) [1], which aid in hydrolysis of glycosidic bonds in complex sugars. Glycosides are found in all domains of life, from unicellular prokaryotes to complex organisms like mammals and higher plants with variety of functions. In prokaryotes, glycosides are involved in nutrient acquisition. One of the important occurrences of glycosidase in bacteria is beta-galactosidase (LacZ) enzyme, which is involved in regulation of expression of the lac operon in *E. coli*. Another example is enzyme neuraminidases, found on the surface of influenza viruses that enables the virus to be released from the host cell and infect the host cell. In higher organisms glycoside hydrolases are found within the Golgi apparatus and endoplasmic reticulum where they are involved in processing of N-linked glycoproteins, and in the lysosome as enzymes involved in the degradation of carbohydrate structures. Deficiency in specific lysosomal glycosidases can lead to a range of lysosomal storage disorders that result in developmental problems or death. Glycosidases are also found in the intestinal tract and in saliva where they degrade complex carbohydrates such as lactose, starch, sucrose and trehalose. The enzyme lactase is required for degradation of the milk sugar lactose and is present at high levels in infants, but in most populations will decrease after weaning or during infancy, potentially leading to lactose intolerance in adulthood. All these glycosidses are categorized into different groups in two major ways, EC classifications [1] and CAZy classifications [2][3].

4.1.1 Classification

Based on the context of classification, glycosidases can be classified in many different ways. They can be classified as endo- or exo-glycodiases, if the enzyme cleaves a glycosidic bond in the middle of a polysaccharide chain or the terminal glycosidic bond respectively [4]. On the basis of reaction mechanism of the catalytic enzymes, they can also be categorized into inverting glycosidases or retaining glycosidases [5]. But the classifications, that is going to be discussed here is based on Enzyme Commissioner number (EC) [1] and sequence similarity (CAZy) [2].

4.1.1.1 EC classification

EC numbers are codes representing the Enzyme Commission number. This is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. Every EC number is associated with a recommended name for the respective enzyme. EC numbers do not specify enzymes, but enzyme-catalyzed reactions. If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same EC number. Glycosidase enzymes has been given the EC number of 3.2.1. Under this sub class many different glycosidases are categorized into different groups with unique EC number 3.2.1.X, where X denotes the substrate specificity of that group. This number assignments are being updated on regular basis with respect to the new findings are their relationship with other groups.

4.1.1.2 CAZy classification

The Carbohydrate Active enzyme (CAZy) database contains glycosides in 135 different Glycosyl Hydrolase (GH) families, on june 2016, based on their sequence similarity [2]. Each family (GH family) contains proteins that are related by sequence, and by corollary, fold. This allows to predict the catalytic machinery and molecular mechanism is conserved for the vast majority of the glycosidase families [5] as well as the geometry around the glycosidic bond (irrespective of naming conventions) [6]. Usually within a GH family, the mechanism used for classification (i.e. retaining or inverting) is conserved.

These two major classifications, EC and CAZy has classified the glycosidases with respect to their unique sequence and catalytic characteristics, but the selectivity and specificity for a particular substrate by an enzyme is not well addressed and the classification is overlapping. It means one GH family contains many EC numbers and vice versa. Though the GH family addresses the catalytic action uniqueness, but does not explain the substrate specificity. So there is a clear need of structural analysis to understand this problem.

4.1.2 Catalysis mechanism

Hydrolysis of a glycoside is generally achieved via general acid and general base assistance from two amino acid side chains, normally glutamic or aspartic acids. Sometimes these amino acids act as acid/base and nucleophile. The catalytic mechanism is carried on following manner, one residue plays the role of a nucleophile, attacking the anomeric centre to displace the aglycon and form a glycosyl enzyme intermediate. At the same time the other residue functions as an acid catalyst and protonates the glycosidic oxygen as the bond cleaves. In the second step (known as the deglycosylation step), the glycosyl enzyme is hydrolyzed by water, with the other residue now acting as a base catalyst deprotonating the water molecule as it attacks [8]. The pKa value of the acid/base group cycles between high and low values during catalysis to optimize it for its role at each step of catalysis [9].

This mechanism was originally proposed by Dan Koshland, although at the time the identities of the residues was unclear [7]. The mechanism is explained in Figure 4.1, for enzymes acting on β-glycosides.

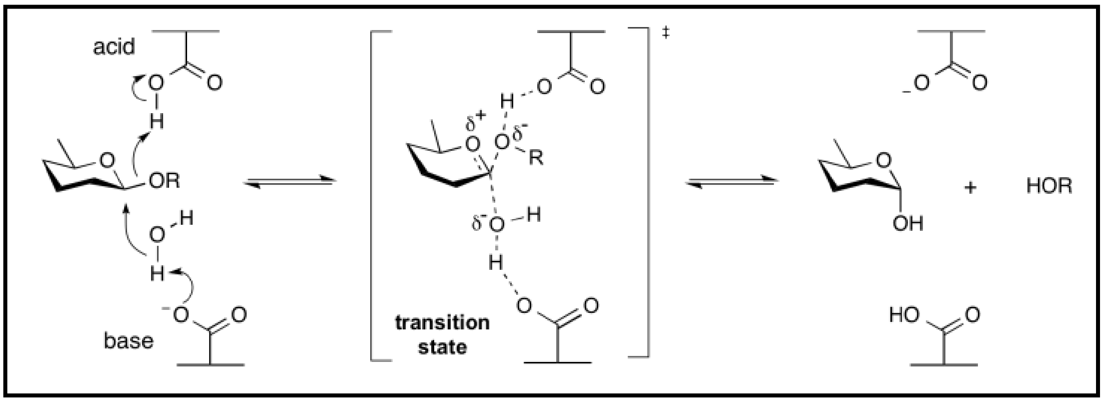


Figure 4.1. The catalysis mechanism and assistance of amino acids from glycosidases acting on β-glycosides [10].

4.1.3 Structural analysis of catalytic pocket

A lot of studies and experimental data is available for structural analysis of catalytic pockets of different glycosidases from various source organism. Crystal structures of various glycosidases with different substrate ligands are also available and their analysis has given a plenty of information about the catalytic pockets. These information include the shape of the pockets and crucial aminoacids stabilizes the catalytic pockets and it has also revealed the different catalytic mechanisms. These information has helped in establishing relationships between different classes of glycosides but a more clear analysis is needed to address the specificity of glycosidases for different substrates.

Due to the development of structural analyzing instruments and bioinformatics tools, a lot of information has accumulated about the binding of different sugar to catalytic pockets. It is witnessed that the catalytic pockets have multiple specificities and are not highly selective. However, subtle variation in type of variation in the binding pocket can affect binding affinity. Such difference may accommodate functional groups at hydroxyl groups of sugar ligand in the pocket and enhances binding affinity or may affect stacking interaction in case of change of aromatic amino acids and thus affect the binding affinity. Some pockets can accommodate more than one monomer of sugar ligand, but it is observed that the pocket has more affinity for a particular monomer among all the ligands in that pocket [11]. It is also evident that the conformation of ligand in the pocket also play an important role in catalytic mechanism. These information are very much essential for drug designing and drug development.

The tremendous amount of carbohydrate-protein complex structural data generated over the past few years is available for the public use through different databases. The increasing number of bioinformatics tools and the availability of databases to store, retrieve, and analyze these data in an efficient way has boosted a lot to the progress in glycobiology [12, 13, 14, 15]. A wide variety of databases are available for glycobiology research [16-21, 23] and they can be grouped into databases that build the glycans, that contain information on the proteins themselves, that store information on the enzymes and pathways and carbohydrate structure databases. But very few of them have data on non-covalent carbohydrate-protein interaction. GLYCOSCIENCES.de [19], Glyco3D.com [16], BRENDA [22] are few of the databases along with protein database, PDB, contain information about the carbohydrate-protein complex having non-covalent interactions. The data available with the available databases is insufficient for a precise structural analysis of these carbohydrate binding proteins. So this is still an open problem due to the lack of broadly accepted matrices on carbohydrate binding pocket structures. The analysis of amino acids in the catalytic pocket may reveal a relationship between different classes of glycosidases and may give a clear explanation for substrate specificity and selectivity.

4.2 Scope of the study

Although the information about catalytic mechanism and amino acids crucial for the action is reported for most of the structures of these groups, but the high specificity of these pockets towards sugar ligands and their characterization is not available. As these enzymes are involved in crucial metabolism processes, the 3 dimensional geometry of their catalytic pockets are essential for drug designing and drug development. There is no specified study on the relationship between different classes of glycosides acting on closely related ligands on the basis of structure and function. This study may help in addressing the high specificity of the catalytic pocket towards a specific ligands over their epimers and anomers. Further analysis of the pocket is needed to study the position of conserved amino acids in the pocket and these amino acid positions can be used to develop a particular geometric pattern for a conserved binding pocket. This information can be stored in a database which can be used to identify catalytic sites in an unknown protein complex.

Our work is about finding the conserved amino acids in the catalytic pockets and to use this information identification of catalytic pockets in query protein structures and to address the specificity and selectivity of the catalytic pockets.

4.3 Objectives

4.3.1 Survey and characterization of catalytic pockets of glycosidase families that act on glucose and its epimers.

Structural analysis of the catalytic pockets of six glycosidase families, acting on epimers of glucose: α-galactosidase, β-galactosidase, α-glucosidase, β-glucosidase, α-mannosidase and β-mannosidase, to understand the type of interactions in the catalytic pockets, specificity of the pocket for sugar ligands and identification of conserved amino acids if any, beyond the source organism.

4.4 Materials and Methods

4.4.1 Short listing structures for analysis

Six glycosidase families, acting on epimers of glucose: α-galactosidase, β-galactosidase, α-glucosidase, β-glucosidase, α-mannosidase and β-mannosidase, are selected for the analysis purpose in order to observe the specificity and selectivity towards sugar ligands having very high structural similarity. The PDB (Protein Data Bank) structures are selected from database. In order to avoid redundancy of PDB structures, for each class of glycosidase, one PDB file is selected for the protein complex with ligands from each group of source organisms considering all the source organisms available for that particular class, in PDB database. Toatal 53 structures are considered from all six glycosidase families varying from bacteria, fungus to mammals and plants.

For structure analysis, in particular for analysis of catalytic pocket, in each class of enzymes, PDB structures, having same unique ligand in more than two structures are considered. This different selection process gives different data sets for the structural analysis as shown in Table 4.1.

4.4.2 Sequence analysis

To check the relevance between same class enzymes from all the source organisms, multiple sequence alignment (MSA) is done through EMBL-EBI tool, MUSCLE (multiple sequence comparison by log-expectation). The phylogenetic tree from MSA is derived and used for further analysis.

4.4.3 Structure analysis

All the available protein structures of these six groups of enzymes were downloaded from PDB database, considering the shortlisting criteria and analysis is done by pymol, a molecule visualizer. The analysis of catalytic pockets is done in two steps. First analysis is done for the amino acids surrounding 25Å around the sugar ligand in the pocket and those amino acids are referred “25Å sugar binding pocket” in this study and the second analysis is done to study the specificity and selectivity of the sugar binding proteins for the ligands in the active site. In this study, the amino acids surrounding 5Å around the sugar ligand in the active site of the protein are considered for the analysis. These amino acids are referred “5Å sugar binding pocket”. The structures of both 25Å and 5Å sugar binding pockets are stored for further analysis and the distance table for amino acids in 5Å sugar binding pockets are created for each group of enzymes.

For analysis of 5Å region around sugar ligand, structures considered should have the same ligand or at least more than two structures, having same ligand, should be available for one class of enzymes to avoid the randomness of selection of conserved amino acids. Aminoacids having a direct interaction with the ligand, either polar, stacking or hydrophobic interactions or are in very close proximity to the ligands are considered as conserved aminoacid selection. And for these six classes of glycosidases carbohydrate monomers are considered for target ligands.

The number of structures selected for each class along with the EC number is given in Table 4.1.

4.5 Results and Discussion

4.5.1 Sequence analysis

MSA is done for all 10 β-galactosidase enzymes from 10 different organisms including *Penicillium sp.* (PDB ID-1XC6), *Escherichia coli* (PDB ID-1JZ7), *Trichoderma reesei* (PDB ID-1T0O) and *Homo sapiens* (PDB ID-3THC). The MSA results depicted that there is very less sequence conservation in all these enzymes. Though they belong to the same class of enzyme having the same catalytic action in all these organisms, the highest sequence similarity between any two groups is 45%. The phylogenetic tree derived from the MSA by MUSCLE tool show the distant relationship between the sequences of these enzymes as shown in Figure 4.2.

EC Number	Name of the Enzyme	Total no of PDBs considered (25Å Pocket)	Total no of PDBs considered (5Å Pocket)
3.2.1.22	α-galactosidase	9	9
3.2.1.23	β-galactosidase	10	10
3.2.1.20	α-glucosidase	11	2
3.2.1.21	β-glucosidase	16	8
3.2.1.24	α-mannosidase	4	
3.2.1.25	β-mannosidase	3	2

Table 4.1. List of PDB structures in different analysis consideration for each class of glycosidses.

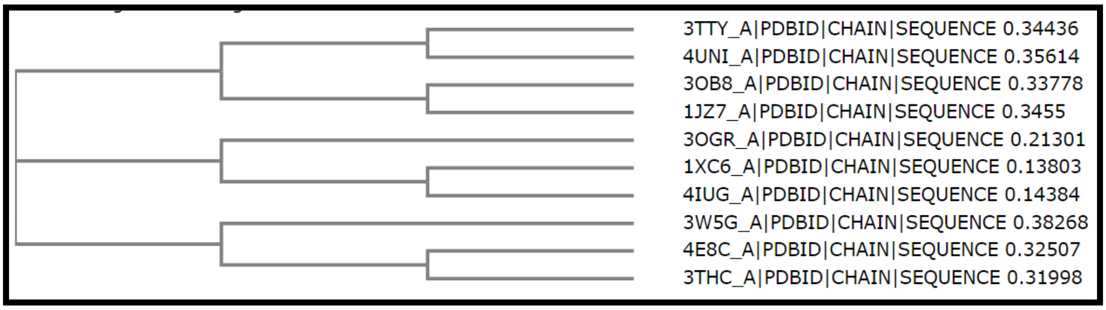

3TTY_A|PDBID|CHAIN|SEQUENCE 0.34436
4UNI_A|PDBID|CHAIN|SEQUENCE 0.35614
3OB8_A|PDBID|CHAIN|SEQUENCE 0.33778
1JZ7_A|PDBID|CHAIN|SEQUENCE 0.3455
3OGR_A|PDBID|CHAIN|SEQUENCE 0.21301
1XC6_A|PDBID|CHAIN|SEQUENCE 0.13803
4IUG_A|PDBID|CHAIN|SEQUENCE 0.14384
3W5G_A|PDBID|CHAIN|SEQUENCE 0.38268
4E8C_A|PDBID|CHAIN|SEQUENCE 0.32507
3THC_A|PDBID|CHAIN|SEQUENCE 0.31998

Figure 4.2 The cladogram representation of all 10 β-galactosidases showing their distant relationship derived from MSA

4.5.2 Structure analysis

The results obtained from MSA of β-galactosidases clarified that enzyme sequences will not be helpful in analysing the catalytic pockets as studies also have reported the same [4]. As mentioned before the structural analysis is done in two parts, 25Å sugar binding pocket and 5Å sugar binding pocket.

4.5.2.1 25Å sugar binding pocket

Analysis of 25Å sugar binding pocket of all these four groups depicted that TIM barrel motif [4] is highly conserved in all considered classes except α-mannosidase as shown in Figure 4.3. When the structural analysis is done in same class of enzymes, it is observed that the helix and sheets arrangement in TIM barrel was different for different GH families [2, 3], as it can clearly be seen for GH1 and GH3 families of β-glucosidase (Figure 4.4). Though all the 25Å sugar binding pockets are of TIM barrel motif, their structural conformation is not conserved. Rather the conservation of the conformation of TIM barrel is seen for one GH family under one class of enzyme i.e. one EC number.

4.5.2.2 5Å sugar binding pocket

The conclusion inferred from analysis of 25Å sugar binding pocket depicted that structural motif, that contains the catalytic site cannot explain the specificity and selectivity of an enzyme towards a substrate. So the 5Å sugar binding pocket or the aminoacids surrounding the 5Å region around sugar ligand are analyzed.

At first two structures, each from GH1 (PDB ID-2E9L) and GH3 (PDB ID-2X41) of β-glucosidase class are selected for analysis and it is seen that the orientation of ligand, β-glucose, and shape of the catalytic pocket is different for both structures (Figure 4.4 B).

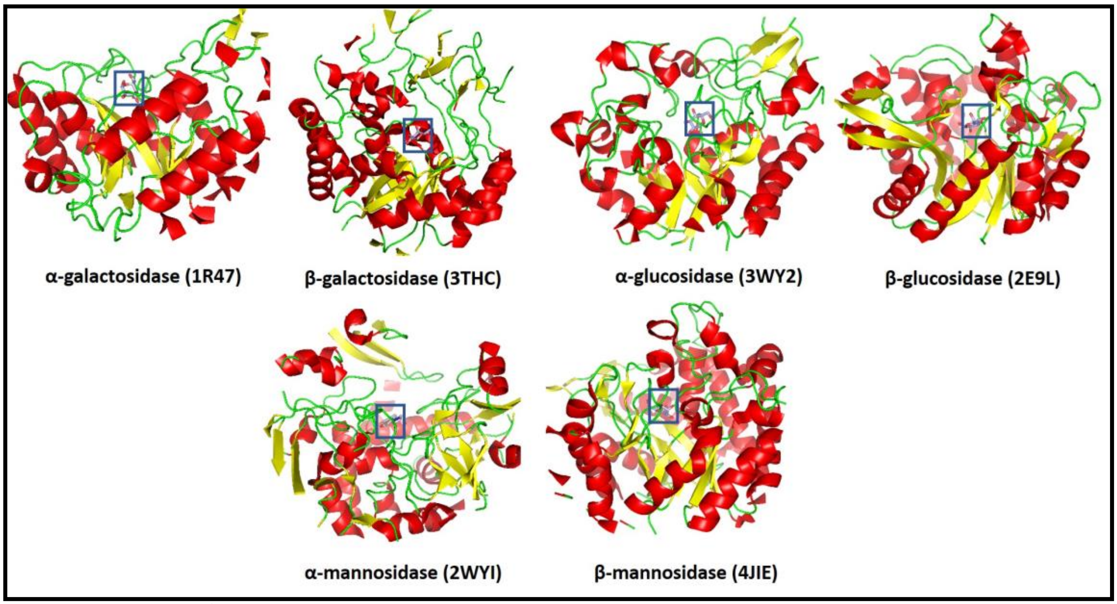


Figure 4.3. 25Å sugar binding pockets of all six glycosidases, represented by one PDB structure with PDB IDs from each group. Except α-mannosidase all other classes have TIM barrel motif that carries the catalytic pocket and ligand. The blue square box shows the ligand.

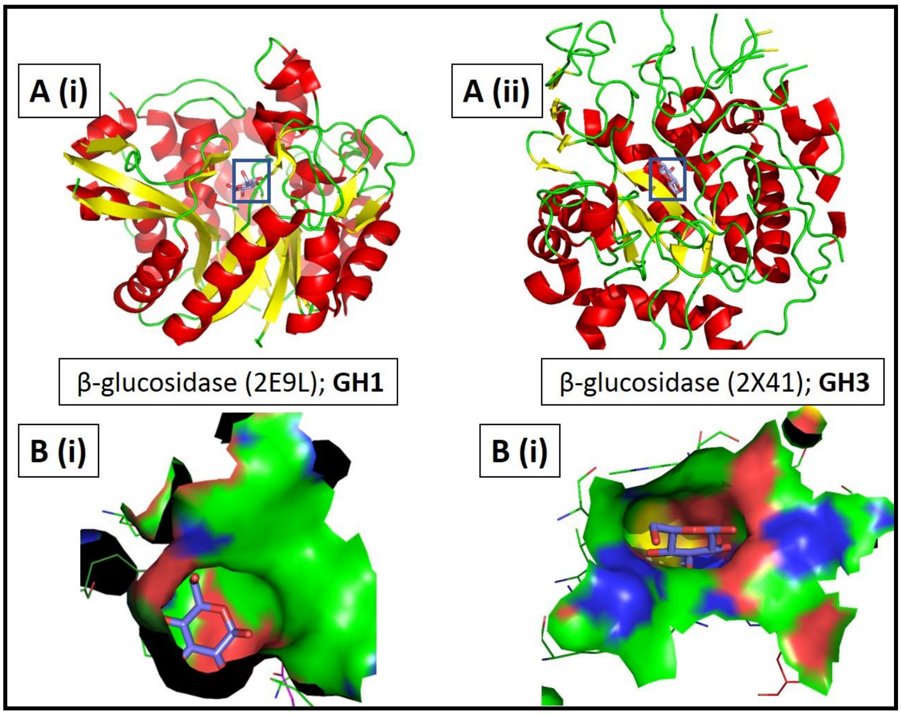


Figure 4.4. Structural comparison between two GH families (GH1 and GH3) of β-glucosidase. [A] Cartoon representation of 25Å sugar binding pocket and [B] surface representation of 5Å sugar binding pocket.

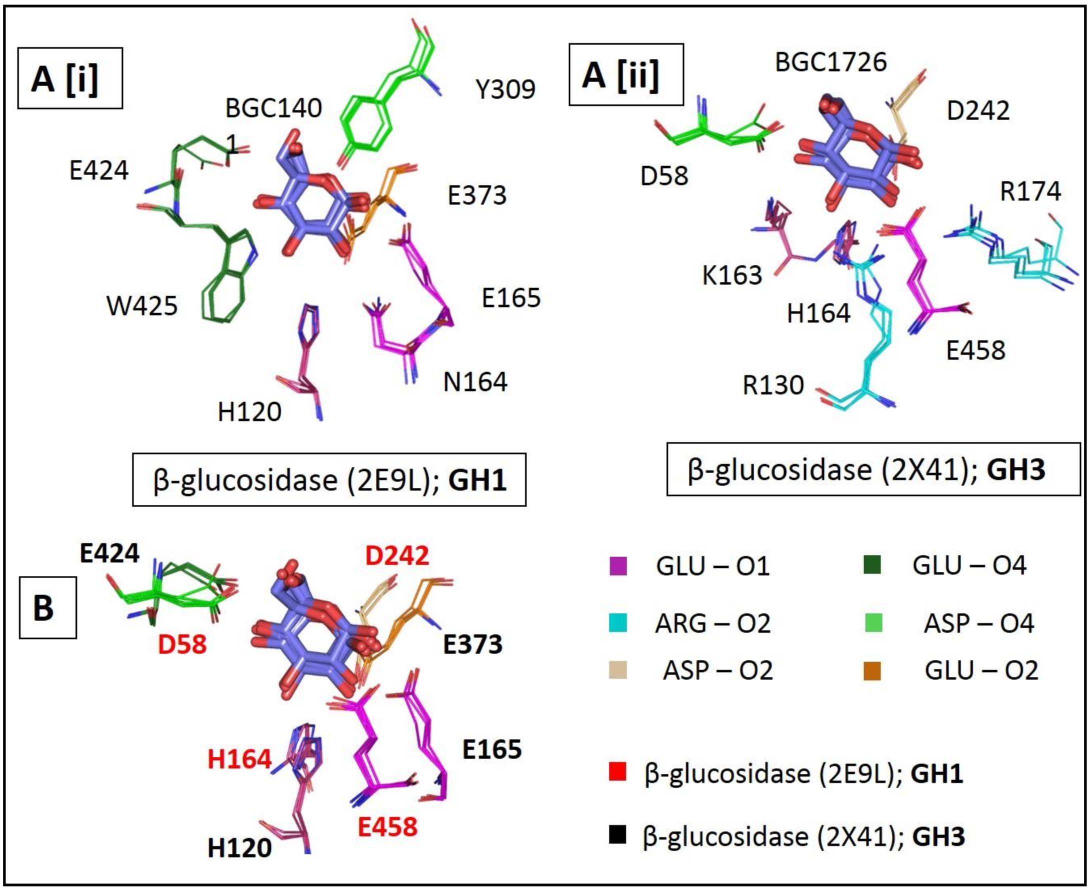


Figure 4.5. Structural comparison of conserved amino acids in 5Å sugar binding pockets for GH families of β-glucosidase. [A] Structural superimposition of all β-glucosidases from GH1 (A[i]) and GH3 (A[ii]). [B] Structural superimposition of all β-glucosidases invariant of the GH families. Amino acids are represented through single letters and sequence number and colors used to represent the aminoacids are described at right-down corner. The description of colors for amino acids are described by the interactions of amino acids represented by three letter code with oxygen atom from ligand, β-glucose (BGC), represented with single letter and position number.

Then to identify the conserved amino acids in the pocket, structural superimposition of all β-glucosidase structures are done in pymol, for two GH families separately and it is observed that conserved amino acids in 5Å sugar binding pockets of GH1 and GH3 family are different (Figure 4.5 A). It is seen that the conformation and position of amino acids in 5Å sugar binding pocket for one class of glycosides are conserved. It is also observed that certain amino acids are conserved in the same class of enzyme beyond the different GH families (Figure 4.5 B). The same observation is found with other glycosidases too. So the catalytic pockets of different classes of glycosidase having different EC number can be differentiated on the basis of the conserved aminoacids in the catalytic pocket, thus concludes objective of the research.

4.6 Conclusion

The analysis of these six glycosidases concludes that particular amino acids are conserved among all the structures of an enzyme class, invariant of the source organism and some amino acids are conserved between different groups of glycosidases too. It is also observed that the conserved amino acids of different GH families in one enzyme class is also different. So, with respect to the conserved amino acids in the catalytic pocket of the glycosidases, the GH families of one enzyme class can be differentiated and can be categorised into groups. This analysis can also be extended to all available glycosidases to define the catalytic pockets and can also be used for automatic prediction of protein-carbohydrate interaction sites and functionalities.

4. 7 References

[1] NC-IUBMB (1992) In the nomenclature committee of the International Union of Biochemistry and Molecular Biology (ed.), Enzyme Nomenclature, 1992. Academic Press, New York.

[2] Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The Carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–D495.

[3] Henrissat B (1991) A classification of glycosyl hydrolases based on amino-acid sequence similarities. Biochem. J. 280:309-316.

[4] Davies G and Henrissat B. Structures and mechanisms of glycosyl hydrolases. Structure. 1995 Sep 15;3(9):853-9. DOI:10.1016/S0969-2126(01)00220-9.

[5] Gebler J, Gilkes NR, Claeyssens M, Wilson DB, Béguin P, Wakarchuk WW, Kilburn DG, Miller RC Jr, Warren RA, and Withers SG. Stereoselective hydrolysis catalyzed by related beta-1,4-glucanases and beta-1,4-xylanases. J Biol Chem. 1992 Jun 25;267(18):12559-61.

[6] Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, and Davies G. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc Natl Acad Sci U S A. 1995 Jul 18;92(15):7090-4.

[7] Koshland, D. (1953) Biol. Rev. 28, 416.

[8] McCarter JD and Withers SG. Mechanisms of enzymatic glycoside hydrolysis. Curr Opin Struct Biol. 1994 Dec;4(6):885-92.

[9] McIntosh LP, Hand G, Johnson PE, Joshi MD, Körner M, Plesniak LA, Ziser L, Wakarchuk WW, and Withers SG. The pKa of the general acid/base carboxyl group of a glycosidase cycles during catalysis: a 13C-NMR study of bacillus circulans xylanase. Biochemistry. 1996 Aug 6;35(31):9958-66.

[10] http://www.cazypedia.org/index.php/Glycoside_hydrolases

[11] Rao V.S.R., Qasba P.K., Balaji P.V., Chandrasekaran R.(1998) Conformation of Carbohydrates, Harwood academic publishers, chapter 10

[12]Haslam SM, Julien S, Burchell JM, Monk CR, Ceroni A, Garden OA, Dell A (2008) Characterizing the glycome of the mammalian immune system. Immunol Cell Biol 86:564–573

[13]Ruhaak LR, Deelder AM, Wuhrer M (2009) Oligosaccharide analysis by graphitized carbon liquid chromatography-mass spectrometry. Anal Bioanal Chem 394:163–174

[14]Royle L, Campbell MP, Radcliffe CM, White DM, Harvey DJ, Abrahams JL, Kim YG, Henry GW, Shadick NA, Weinblatt ME, Lee DM, Rudd PM, Dwek RA (2008) HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Anal Biochem 376:1–12

[15]Packer NH, von der Lieth C-W, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS (2008) Frontiers in glycomics: Bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda, MD (September 11–13, 2006). Proteomics 8:8–20

[16]S. Perez; A. Sarkar, Ch. Breton, S. Drouillard, A. Rivet & A. Imberty (2013), Glyco3D: A Portal for Structural Glycoscience, http://glyco3d.cermav.cnrs.fr, Methods Mol Biol. 2015;1273:241-58. doi: 10.1007/978-1-4939-2343-4_18.

[17]Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH (2014), UniCarbkb: building a knowledge platform for glycoproteomics Nucleic Acids Res.;42(1):D215-21

[18]Maeda M, Fujita N, Suzuki Y, Sawaki H, Shikanai T, Narimatsu H. (2015s), JCGGDB: Japan Consortium for Glycobiology and Glycotechnology Database, Methods Mol Biol.1273:161-79. doi: 10.1007/978-1-4939-2343-4_12.

[19]Lutteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth C-W (2006): GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. Glycobiology, 16:71R-81R.

[20]Lutteke T (2008) Web Resources for the Glycoscientist. Chembiochem 9:2155–2160

[21] Ranzinger R, Herget S, Lutteke T, Frank M (2009) Carbohydrate Structure Databases. In: Cummings RD, Pierce JM (eds) Handbook of glycomics. Elsevier, Amsterdam, pp 211–233

[22] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004), BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res.;32(Database issue):D431-3.

[23] Kunduru, B.R., Nair, S.A. and Rathinavelan, T. (2016), “EK3D: an E. coli K antigen 3-Dimensional Structure Database”, Nucleic Acids Res. 44 (D1), D675-D681