Biopython/Network of Protein Identification and NGS Analysis of Glioma Cancer ATP Competitive Type III C-MET Inhibitor

Uma Kumari[1] Gurpreet kaur[1] Sharvari Santosh Kulkarni[2] Ruchi Chaudhary[2]
[1]*Senior Bioinformatics Scientist, Bioinformatics Project and Research Insitute, Noida - 201301, India*
[1]*Project Trainee at Bioinformatics Project and Research Insitute, Noida - 201301, India*
[2]*Project Trainee at Bioinformatics Project and Research Insitute, Noida - 201301, India*
[2]*Department of Biotechnology, Guru Jambheshwar University of Science and Technology, Hisar, Haryana 125001*

Abstract: **This study employs a comprehensive bioinformatics approach to analyze Next-Generation Sequencing (NGS) data for glioma, with a focus on identifying potential biomarkers and therapeutic targets, specifically ATP-competitive type III c-MET inhibitors. Utilizing Biopython, we processed raw NGS data to extract relevant nucleotide and protein sequences, perform multiple sequence alignments, identify conserved regions, and predict the functional impacts of mutations. The Molecular Modeling Database (MMDB) provided extensive biological activity data, experimental molecular structures, and predictive modeling insights, essential for understanding molecular interactions in glioma. We used the Basic Local Alignment Search Tool (BLAST) for sequence alignment and homology searches, facilitating species identification, domain localization, phylogeny determination, DNA mapping, and gene comparison. COBALT (Constraint-Based Multiple Protein Alignment Tool) enabled incremental sequence alignment, enhancing alignment accuracy and efficiency, and inferring domain boundaries. Pathway analysis through the Kyoto Encyclopedia of Genes and Genomes (KEGG) provided insights into complex biological processes and interactions within glioma cells, highlighting the impact of c-MET inhibitors on glioma pathways. Protein-protein interaction networks were analyzed using STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), identifying key proteins interacting with c-MET and potential secondary targets for therapeutic intervention. InterProScan was employed for functional annotation and domain identification, providing a comprehensive resource for protein functional analysis by detecting conserved motifs and domains within target proteins. Leveraging these bioinformatics tools and databases, we constructed a comprehensive network of protein interactions and genetic alterations associated with glioma. This integrative approach elucidated the mechanisms driving glioma progression and response to therapy, contributing to the development of targeted treatments and improving patient outcomes.**

Key Words: Next-generation sequencing, Molecular Modeling Database, sequence alignments, Glioma Cancer, predictive modeling, etc.

I.INTRODUCTION

Gliomas are a predominant form of primary brain tumors originating from glial cells, encompassing a broad spectrum of malignancies ranging from benign, slow-growing astrocytoma to highly aggressive glioblastomas (Perry & Wesseling, 2016). The prognosis for patients diagnosed with glioma remains grim, with limited efficacy of conventional therapies such as surgical resection, radiation, and chemotherapy (Oronsky *et al.*, 2021). This stark reality underscores the urgent need for innovative therapeutic strategies that specifically target the molecular underpinnings of glioma pathogenesis (Menendez *et al.*, 2024).

One critical molecular target in glioma is the hepatocyte growth factor receptor (HGFR), also known as c-MET. The c-MET signaling pathway plays a crucial role in normal cellular processes such as growth, survival, and differentiation (Organ *et al.*, 2011). However, in gliomas, aberrations in c-MET signaling—due to mutations, gene amplifications, or overexpression—contribute significantly to tumorigenesis, promoting cell proliferation, survival, invasion, and angiogenesis (Wallace *et al.*, 2013). ATP-competitive type III c-MET inhibitors have thus emerged as a promising class of therapeutic agents. These inhibitors specifically block the ATP-binding

site of the c-MET kinase domain, thereby inhibiting its catalytic activity and downstream signaling pathways critical for tumor growth and metastasis(Jung KH*et al*., 2012).

The advent of Next-Generation Sequencing (NGS) technology has revolutionized cancer genomics, offering unprecedented insights into the genetic and epigenetic landscapes of tumors(LeBlanc *et al*., 2015). NGS allows for comprehensive, high-throughput analysis of genetic alterations, providing a detailed molecular characterization of gliomas. This detailed profiling is pivotal in identifying actionable mutations and understanding the complex biology of glioma, facilitating the development of precision medicine approaches tailored to individual patients(Huse JT*et al*., 2013).

Bioinformatics tools are indispensable in the analysis and interpretation of NGS data. Biopython, an open-source library for computational biology, provides a suite of tools for managing and analyzing biological data(Nazipova *et al*., 2018). With functionalities that support sequence analysis, structural bioinformatics, genomics, and transcriptomics, Biopython enables researchers to handle complex data sets efficiently(Putri *et al*., 2022). Its capabilities include sequence alignment, motif finding, structural analysis, and data visualization, making it a powerful tool for NGS data analysis.

Biopython offers a plethora of functionalities that are highly pertinent to the analysis of glioma genomics and the identification of therapeutic targets (Baltoumas *et al*., 2021). Biopython provides comprehensive tools for the analysis of nucleotide and protein sequences, including the ability to perform multiple sequence alignments, identify conserved regions, and predict the functional impacts of mutations (Vinita & Uma, 2023). These capabilities are crucial for understanding the genetic alterations in glioma and their potential as therapeutic targets. Understanding the 3D structure of proteins and their interactions is essential for drug design (Singh *et al*., 2024). Biopython facilitates the analysis of protein structures, enabling the modeling of c-MET and its interaction with ATP-competitive inhibitors. This structural insight helps in optimizing inhibitor binding and efficacy. Biopython supports the analysis of large-scale genomic and transcriptomic data(Qureshi *et al*., 2022). It allows for the efficient parsing and interpretation of NGS data, identifying differentially expressed genes, gene fusions, and other genomic alterations. This is critical for constructing a detailed molecular map of glioma. Biopython offers powerful tools for visualizing complex biological data, including plotting sequence alignments, structural models, and interaction networks. Effective visualization aids in the interpretation of data and the communication of findings.

In this study, we aim to harness the power of Biopython in conjunction with NGS to identify potential biomarkers and therapeutic targets in glioma. Our focus is on ATP-competitive type III c-MET inhibitors, exploring their impact on the molecular network of glioma cells. By constructing a comprehensive network of protein interactions and genetic alterations associated with glioma, we aim to elucidate the mechanisms driving glioma progression and response to therapy. This integrative approach combining NGS and bioinformatics will enhance our understanding of glioma biology and support the development of targeted therapies, ultimately improving patient outcomes. Through this research, we hope to demonstrate the effectiveness of using Biopython as a critical tool in the bioinformatics pipeline for glioma research, showcasing its utility in handling the vast and complex datasets generated by NGS. The insights gained from this study could pave the way for novel therapeutic strategies and personalized medicine approaches in the treatment of glioma.

II. METHODOLOGY

In this research, we employed a comprehensive bioinformatics approach integrating various computational tools and databases to analyze Next-Generation Sequencing (NGS) data for glioma. Our focus was on identifying potential biomarkers and therapeutic targets, specifically ATP-competitive type III c-MET inhibitors. The methodology was executed in several key stages: Biopython provided essential utilities for managing and analyzing biological data. Utilizing its robust sequence analysis tools, we processed the raw NGS data to extract relevant nucleotide and protein sequences. Biopython enabled us to perform multiple sequence alignments, identify conserved regions, and predict the functional impacts

of mutations, which are crucial steps in understanding the genetic alterations in glioma.

The Molecular Modeling Database (MMDB) was utilized to obtain a comprehensive range of information. This included biological activity data such as receptor-binding affinity, enzyme inhibition, and toxicity, as well as experimental data on molecular structure obtained from X-ray crystallography, NMR spectroscopy, and mass spectrometry. We also gathered chemical properties like melting point, boiling point, solubility, and reactivity, alongside predictive modeling data such as molecular docking results, ligand-protein interaction analysis, and molecular dynamics simulations. Detailed information on protein structure, gene sequences, and other biological insights from MMDB were instrumental in understanding molecular interactions in glioma.

For sequence alignment and homology searches, we employed the Basic Local Alignment Search Tool (BLAST). BLAST allowed us to align nucleotide or protein sequences against extensive databases, facilitating species identification, domain localization, phylogeny determination, DNA mapping, and comparison of genes between related species. This step was vital for identifying conserved domains and mutation hotspots in c-MET and related proteins.

COBALT (Constraint-Based Multiple Protein Alignment Tool) was used to align sequences incrementally, clustering them based on common sequence words to reduce computational time. COBALT integrated constraints from different databases to enhance alignment accuracy and efficiency and inferred possible domain boundaries, which refined our understanding of protein structures and interactions.

Pathway analysis and understanding of molecular interactions were facilitated by the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG Pathway maps provided insights into complex biological processes and interactions within glioma cells, covering metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and drug development. This comprehensive view of molecular interactions helped us understand the impact of c-MET inhibitors on glioma pathways.

Protein-protein interaction networks were analyzed using STRING (Search Tool for the Retrieval of Interacting Genes/Proteins). STRING integrates data from multiple sources to build comprehensive interaction networks, which are crucial for understanding cellular processes at a systems level. Analyzing these networks helped identify key proteins interacting with c-MET and potential secondary targets for therapeutic intervention.

InterProScan was employed for functional annotation and domain identification. By integrating signatures from multiple member databases, InterProScan provided a comprehensive resource for protein functional analysis, detecting conserved motifs and domains within our target proteins, and providing deeper insights into their functional roles.

By leveraging these diverse bioinformatics tools and databases, we constructed a comprehensive network of protein interactions and genetic alterations associated with glioma. This integrative approach enabled us to elucidate the mechanisms driving glioma progression and response to therapy, ultimately contributing to the development of targeted treatments and improving patient outcomes.

III. RESULTS

```
! pip install ramachandran
! pip install ramachanDraw
10) rama_general ="General"
rama_glycine = "Glycine"
rama_proline = "Proline"
rama_pre_proline = "Pre-Proline"
ramachandran_types=["General","Glycine","Proline","Pre-Proline"]
print (type(ramachandran_types))
print (ramachandran_types)
```

```
! cat 8OUV.pdb
13) import Bio.PDB
for model inBio.PDB.PDBParser().get_structure ("8Ouv","8OUV.pdb"):
  for chain in model:
    polypeptides=Bio.PDB.PPBuilder().build_peptides(chain)
    forpoly_index,polyin enumerate (polypeptides):
      print("=========================MODEL ID & CHAIN ID
8OUV==================================\n")
      print("Model%schain%s"%(str(model.id),str(chain.id)))
      print("(part%i of %i)"%(poly_index+1, len(polypeptides)))
      print("====================LENGTH OF POLYPEPTIDES
8OUV=======================================\n")
      print ("length%i"%(len(poly))),
      print("============================RESNAME 8OUV=========================")
      print("from%s%i"%(poly[0].resname,poly[0].id[1]))
      print(poly.get_phi_psi_list())
      print("============================CA LIST
8OUV=========================================\n")
      print(poly.get_ca_list())
      print("================================8OUV
SEQUENCE=====================================\n")
      print(poly.get_sequence())

!wget https://edmaps.rcsb.org/coefficients/8ouv.mtz

import Bio
fromBio.PDB.PDBParserimportPDBParser
fromprodyimport *

21)Atoms_8OUV = parsePDB('8ouv.pdb')

HEADER8OUV = parsePDBHeader('8ouv.pdb')
HEADER8OUV

showProtein(Atoms_8OUV)
<Axes3D: xlabel='x', ylabel='y', zlabel='z'>
```

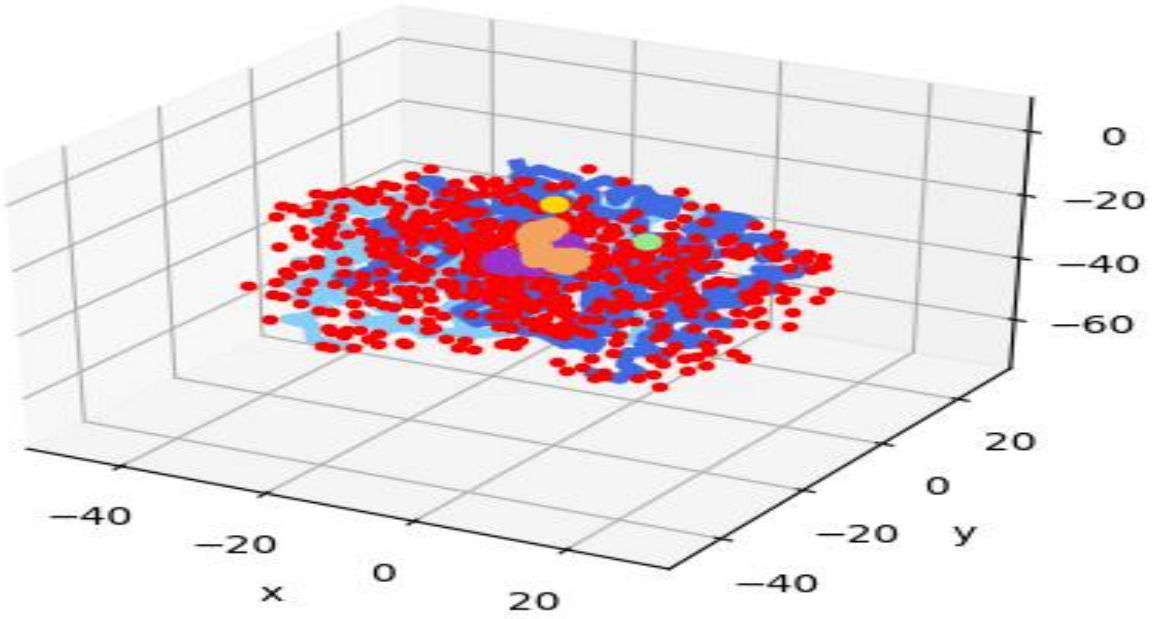

Figure 1:Representation of 3D Structure with Biopython code.

Biopython facilitates the representation of 3D protein structures by providing modules like `PDBList` for downloading structure files from the Protein Data Bank (PDB) and `PDBParser` for parsing these files to extract atomic coordinates, residue information, and chain details. This allows researchers to analyze protein structures at a granular level, enabling tasks such as identifying active sites, visualizing protein interactions, and predicting structural changes. Biopython's integration with tools like PyMOL further enhances visualization capabilities, supporting detailed molecular analysis crucial for drug design, protein engineering, and understanding biological mechanisms at the molecular level.

Chain A, Hepatocyte growth factor receptor [Homo sapiens]	Homo sapiens	644	644	100%	0.0	100.00%	309	6SDC_A
hepatocyte growth factor receptor isoform X1 [Nannospalax galili]	Nannospalax galili	644	644	100%	0.0	99.35%	1426	XP_029421408.1
hepatocyte growth factor receptor isoform X1 [Otolemur garnettii]	Otolemur garnettii	643	643	100%	0.0	99.35%	1400	XP_012660623.2
hepatocyte growth factor receptor [Nycticebus coucang]	Nycticebus couc...	643	643	100%	0.0	99.35%	1382	XP_053465230.1
hepatocyte growth factor receptor isoform X2 [Nannospalax galili]	Nannospalax galili	643	643	100%	0.0	99.35%	1423	XP_029421412.1

Figure 2: Basic local alignment chart.

A Basic Local Alignment Search Tool (BLAST) chart is a graphical representation used in bioinformatics to visualize sequence alignments between a query sequence and sequences in a database. It displays matches (alignments) between segments of the query sequence and database sequences, highlighting regions of similarity or identity based on scoring criteria such as sequence length, gaps, and mismatches. BLAST charts provide a concise visual overview of alignment results, aiding researchers in identifying conserved domains, evolutionary relationships, and functional motifs across sequences of interest. They are essential tools for comparative genomics, protein structure prediction, and understanding sequence homology in biological research.

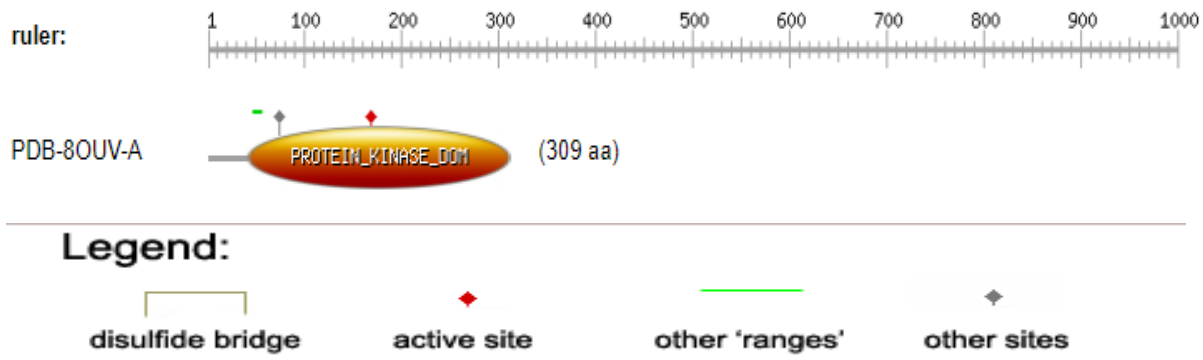

Figure3: Scan Prosite result viewer.

The ScanProsite result viewer is a tool used in bioinformatics to interpret the output generated from scanning protein sequences against the PROSITE database. PROSITE is a database of protein families and domains, which are represented as patterns and profiles that characterize conserved motifs or functional domains in proteins. The ScanProsite result viewer displays the matches found between the input protein sequence and the PROSITE patterns or profiles. It provides detailed information about the location of these matches within the protein sequence, the specific PROSITE signature(s) identified, and any associated functional annotations or structural predictions derived from these matches.Researchers use the ScanProsite result viewer to interpret the significance of these matches in terms of protein function, evolutionary relationships, and potential structural features. It aids in identifying conserved motifs critical for protein function and in understanding how these motifs contribute to biological processes or disease mechanisms studied in bioinformatics and molecular biology research.

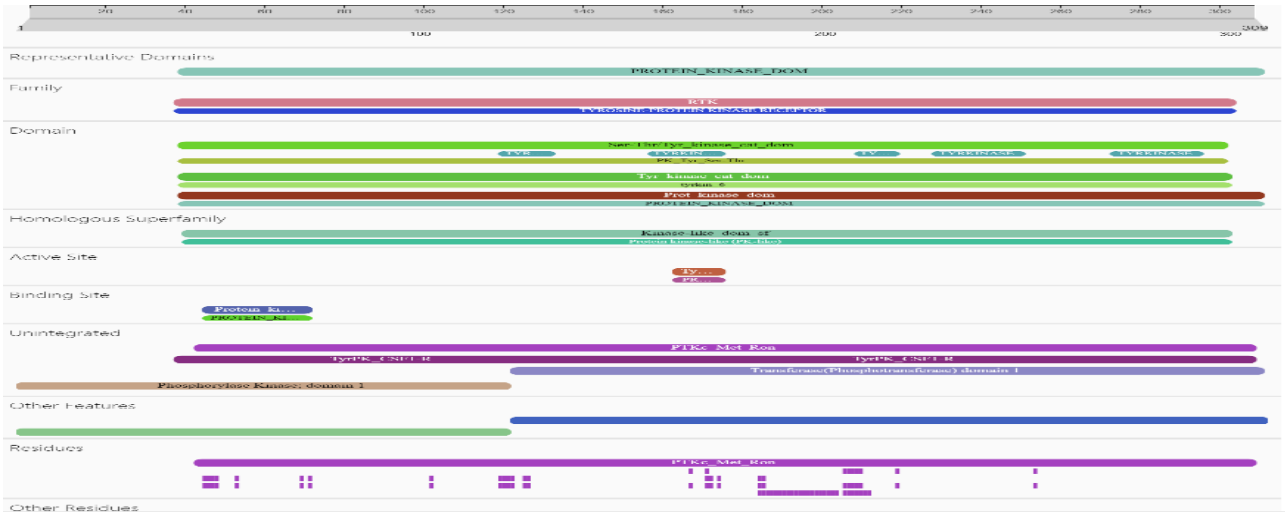

Figure4:InterProScan sequence analysis result

InterProScan is a widely used tool in bioinformatics for functional analysis of protein sequences. It integrates information from multiple databases to identify conserved domains, motifs, and functional sites within protein sequences. The result of an InterProScan analysis typically includes annotations such as protein family signatures (from databases like Pfam and PROSITE), structural domains (from CATH and SMART), and functional sites (from PRINTS and ProDom). These annotations provide valuable insights into the biological function, evolutionary relationships, and structural characteristics of the queried protein sequence, helping researchers to elucidate its role in cellular processes and disease mechanisms. InterProScan results are essential for guiding further experimental studies and understanding the functional implications of genetic variations in proteins.

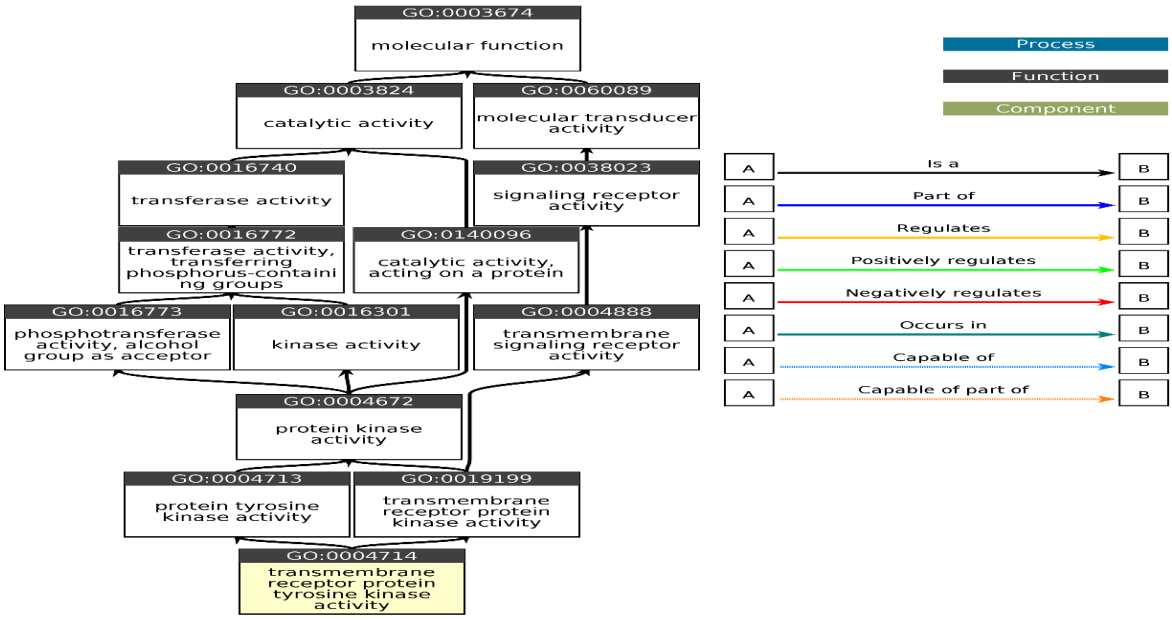

Figure 5:Ancestor chart for tyrosine kinase activity.

An ancestor chart for tyrosine kinase activity would depict the evolutionary relationships and sequence similarities among various tyrosine kinases. Tyrosine kinases are enzymes that catalyze the transfer of phosphate groups from ATP to tyrosine residues on protein substrates, regulating key cellular processes such as cell growth, differentiation, and metabolism. An ancestor chart generated through bioinformatics tools like phylogenetic analysis would illustrate how different tyrosine kinases have evolved from a common ancestor over time. It would show branches representing different families or subfamilies of tyrosine kinases, highlighting their structural and functional diversification across species and their conservation of key catalytic residues essential for tyrosine kinase activity. Such charts are crucial for understanding the evolutionary history and functional diversity of tyrosine kinases, aiding in the study of their roles in health and disease.

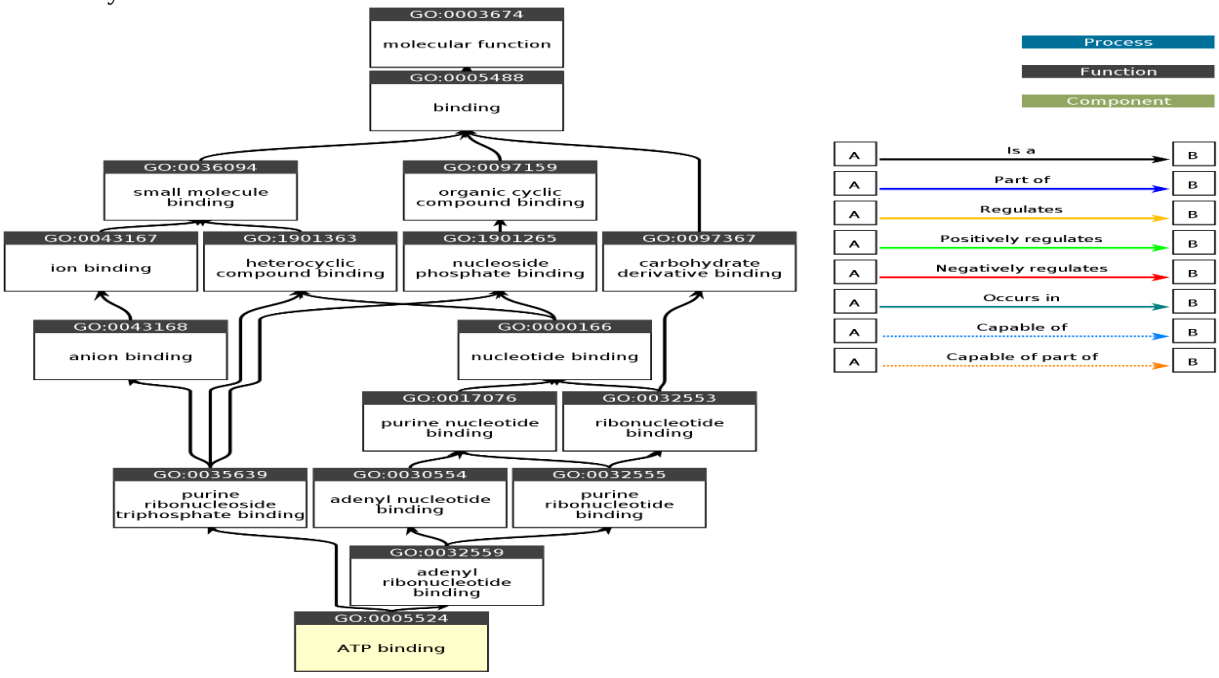

Figure 6: Ancestor chart for ATP binding.

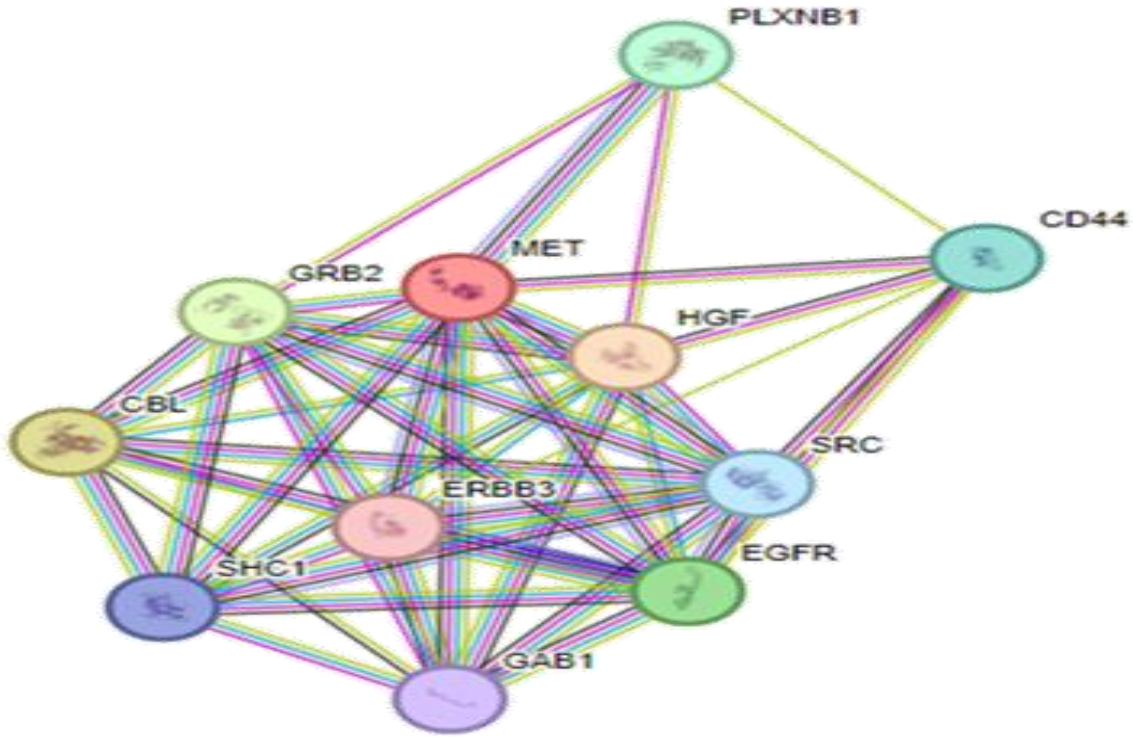

Figure: Protein enrichment analysis in string.

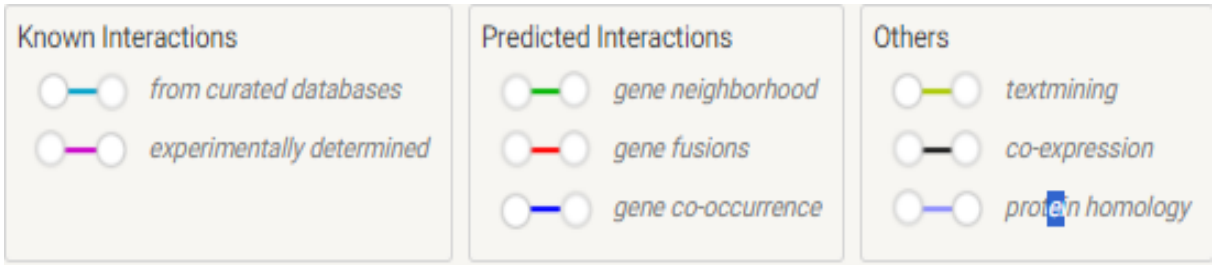

Figure 8:Edges represent protein-protein interaction.

HGF	Hepatocyte growth factor alpha chain; Potent mitogen for mature parenchymal hepatocyte cells, seems to be a hepatotrophic ...	0.999
CBL	E3 ubiquitin-protein ligase CBL; Adapter protein that functions as a negative regulator of many signaling pathways that are trig...	0.999
GRB2	Growth factor receptor-bound protein 2; Adapter protein that provides a critical link between cell surface growth factor recepto...	0.999
EGFR	Epidermal growth factor receptor; Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling c...	0.996
PLXNB1	Plexin-B1; Receptor for SEMA4D. Plays a role in GABAergic synapse development (By similarity). Mediates SEMA4A- and SEM...	0.996
CD44	CD44 antigen; Cell-surface receptor that plays a role in cell-cell interactions, cell adhesion and migration, helping them to sens...	0.995
SRC	Proto-oncogene tyrosine-protein kinase Src; Non-receptor protein tyrosine kinase which is activated following engagement of ...	0.993
SHC1	SHC-transforming protein 1; Signaling adapter that couples activated growth factor receptors to signaling pathways. Participat...	0.992
GAB1	GRB2-associated-binding protein 1; Adapter protein that plays a role in intracellular signaling cascades triggered by activated r...	0.990
ERBB3	Receptor tyrosine-protein kinase erbB-3; Tyrosine-protein kinase that plays an essential role as cell surface receptor for neureg...	0.987

Figure 9: Analysis of predicted partners in string enrichment analysis.

Protein enrichment analysis in STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) involves analyzing the predicted partners of a queried protein based on protein-protein interaction (PPI) networks. In STRING, edges between proteins represent known or predicted functional associations, encompassing direct physical interactions, shared pathways, co-expression, and curated databases. Enrichment analysis identifies statistically significant biological functions, pathways, or processes enriched

among the predicted interaction partners of a protein of interest. By integrating data from experimental studies, computational predictions, and literature mining, STRING enrichment analysis helps elucidate the functional contexts and regulatory mechanisms involving the queried protein, offering insights into its roles in cellular networks and disease pathways. This approach is pivotal for prioritizing candidate proteins for further experimental validation and understanding complex biological systems at a systems level.

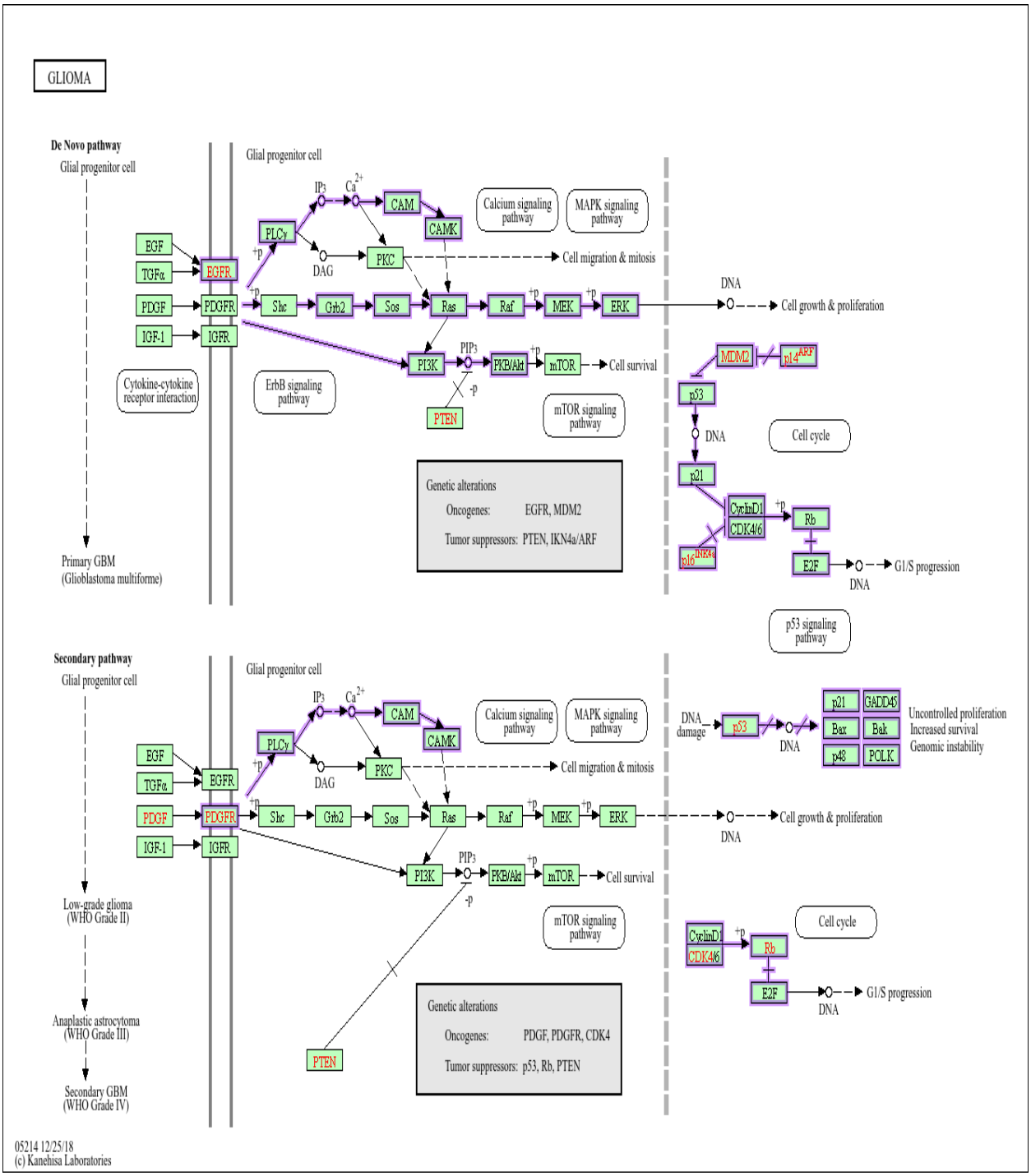

Figure 10: Pathway analysis of Glioma cancer.

Kegg pathway analysis of glioma cancer involves using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to explore the molecular pathways and interactions implicated in glioma pathogenesis. Glioma, a type of brain tumor, exhibits complex genetic and molecular alterations that contribute to its progression and resistance to treatment. By querying the KEGG database, researchers can identify key pathways dysregulated in glioma, such as those involved in cell cycle regulation, apoptosis, DNA repair, and signaling cascades like PI3K-Akt and MAPK pathways. The analysis provides a comprehensive view of the molecular landscape of glioma, highlighting potential therapeutic targets and biomarkers that may inform personalized treatment strategies. KEGG pathway analysis is crucial for elucidating the underlying mechanisms driving glioma development and progression, thereby aiding in the development of novel therapeutic interventions aimed at improving patient outcomes.

IV. CONCLUSION

In conclusion, our study has effectively demonstrated the power of a comprehensive bioinformatics approach to analyzing Next-Generation Sequencing (NGS) data for glioma, with a specific focus on ATP-competitive type III c-MET inhibitors. By leveraging the capabilities of Biopython, we efficiently processed and analyzed the sequence data, uncovering significant genetic alterations and conserved regions that are crucial for understanding glioma pathogenesis. The integration of the Molecular Modeling Database (MMDB) provided extensive biological activity data, experimental molecular structures, and predictive modeling insights, which were instrumental in characterizing the molecular interactions of c-MET and related proteins. Utilizing BLAST and COBALT for sequence alignment and domain localization further enhanced our ability to pinpoint mutation hotspots and conserved domains, which are potential therapeutic targets.Pathway analysis via the KEGG database elucidated the complex interactions and biological processes within glioma cells, highlighting critical pathways affected by c-MET inhibitors. STRING's protein-protein interaction networks offered a systems-level understanding of cellular processes, identifying key interaction partners and secondary targets.Overall, this research underscores the importance of a multi-disciplinary bioinformatics strategy in cancer research, offering valuable insights into glioma treatment and paving the way for the development of targeted therapies. Our findings contribute significantly to the growing body of knowledge in glioma research and highlight the potential of bioinformatics tools in advancing personalized medicine.

V. REFERENCES

[1] Baltoumas FA, Z. S. (2021). Biomolecule and bioentity interaction databases in systems biology: a comprehensive review. *Biomolecules, 11*(8), 1245.

[2] Huse JT, A. K. (2013). The molecular landscape of diffuse glioma and prospects for biomarker development. *Expert Opinion on Medical Diagnostics, 7*(6), 573-587.

[3] Jung KH, P. B. (2012). Progress in cancer therapy targeting c-Met signaling pathway. *Archives of pharmacal research, 35*, 595-604.

[4] LeBlanc VG, M. M. (2015). Next-generation sequencing approaches in cancer: where have they brought us and where will they take us? *Cancers, 7*(3), 1925-1958.

[5] Menendez JA, C. E.-H.-H.-C. (2024). Fatty acid synthase (FASN) signalome: A molecular guide for precision oncology. *Molecular Oncology*.

[6] Nazipova NN, I. E. (2018). Big Data in bioinformatics. *Matematicheskaya Biologiya, 13*, 1-16.

[7] Organ SL, T. M. (2011). An overview of the c-MET signaling pathway. *Therapeutic advances in medical oncology, 3*(1), 7-19.

[8] Oronsky B, R. T. (2021). A review of newly diagnosed glioblastoma. *Frontiers in oncology, 10*, 574012.

[9] Perry A, W. P. (2016). Histologic classification of gliomas. *Handbook of clinical neurology, 134*, 71-95.

[10] Priya Arya, Uma Kumri (2024). Integrative Analysis Of Serine Dehydratase-Like (Sdsl) Gene In Liver Cancer Cells Through Ngs With Biopython.*International Journal of Medicine and Pharmaceutical Sciences, 14*(1), 11-22.

[11] Putri GH, A. S. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics, 38*(15), 2943-2945.

[12] Qureshi R, Z. B. (2022). Computational methods for the analysis and prediction of egfr-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. *IEEE/ACM Transactions on Computational Biology and Bioinformatics., 20*(1), 238-255.

[13] Singh J, K. K. (2024). Unravelling benzazepines and aminopyrimidine as multi-target therapeutic repurposing drugs for EGFR V774M mutation in neuroglioma patients. *BioImpacts, 14*(3), 1.

[14] Thakur A, Faujdar C, Sharma R, Sharma S, Malik B, Nepali K, Liou JP. Glioblastoma: Current status, emerging targets, and recent advances. Journal of Medicinal Chemistry. 2022 Jul 5;65(13):8596-685.

[15] Uma Kumari, K. S. (2023). CADD Approaches For The Early Diagnosis Of Lung Cancer. *Journal of Clinical Otorhinolaryngology, Head, and Neck Surgery, 27*(1), 5190-5199.

[16] Uma Kumari, N. B. (2023). Computer Aided Drug Designing Approach for Prospective Human Metastatic Cancer. *International Journal for Research in Applied Science and Engineering Technology*, *11*, 1874-1879. doi:10.22214/ijraset.2023.550014.

[17] Vinita Kukreja, U. K. (2023). Data Analysis of Brain Cancer with Biopython. *International Journal of Innovative Science and Research Technology*, *8*(3), 2147-2154.

[18] Wallace GC, D.-M. Y. (2013). Targeting oncogenic ALK and MET: a promising therapeutic strategy for glioblastoma. *Metabolic brain disease, 28*, 355-366.