IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Advances in Multi-Omics Integration: Unveiling the Mechanisms of Metabolic Pathways, Proteomics, Genomics through Computational Tools

[1]Anesh SA, [1]Shriya Vikram, [1]Sloka Kumarswamy, [2]Dr. Shivandappa

[1]Student, [1]Student, [1]Student, [2]Assistant Professor
[1]Department of Biotechnology,
[1] R.V. College of Engineering
Bengaluru, India

Abstract : — Integration of proteomic and genomic data would be extremely useful for a more holistic understanding of metabolic pathways and biological functions. Herein, this review considers recent advances in computational methods that enhance the visualization and profiling of such data. It stresses on the vital role of bioinformatics in the detection of genomic selection signatures, cluster analyses for evolutionary biology, and functional understanding of metabolic pathways. Impactful analyses, on the other hand, can only be accomplished when effective computational tools are at hand, as next-generation sequencing and mass spectrometry are examples of high-throughput technologies generating voluminous data.

Among these techniques, PCA and KMeans are basic for the simplification and interpretation required in high-dimensional gene expression data. While PCA diminishes the dimensions to retain variance important for visualization, KMeans clustering does the work of grouping samples according to their similarities and helps in finding biological patterns and subtypes. This approach enables hypothesis generation, data summarization, visualization, and further research. Beyond that, visualization of protein interactions as a network using libraries like NetworkX supports the identification of key proteins and strengths of interactions, hence providing functional insights into possible pathways. The confusion matrix and classification reports will serve to further assess the performance of the models by pointing out aspects that have to be optimized for class balance performance.

3D visualization of protein structure enhances insights related to spatial configurations and interactions that lie central in structural biology. Stoichiometric matrices, represented as heatmaps, contribute to metabolic modelling by encoding the relationships between metabolites and reactions in the model, supporting their analysis and communication. Taken together, these approaches will not only provide a broad view of complex biological data but may also allow the forecasting of metabolic behaviour or the overcoming of some serious challenges standing in the way of complete understanding and modelling of metabolic pathways and diseases. In this review, future research directions in the integration of proteomics and genomics are discussed, with the aim of enhancing our power to resolve the mechanisms behind metabolic disorders.

IndexTerms **- Multi-Omics Integration, Proteomics, Genomics, Metabolic Pathways, Bioinformatics, Computational Tools, Gene Expression Data, High-Throughput Technologies, PCA (Principal Component Analysis), KMeans Clustering, Network Visualization, Protein-Protein Interactions**

INTRODUCTION

This rapid advancement in genomics and proteomics results in immense complexity and volume that have arisen in biological data; therefore, it calls for much advanced methods of computation to precisely represent the information and produce the right interpretation. Integration of genetic and proteomic data alone, especially in applications like metabolic pathways, calls for advanced methods of consideration against cellular functions and mechanisms of diseases. Recent breakthroughs in GWAS and protein profiling have identified genetic variation associations with metabolic functions, whereas integration of multi-omics data has resulted in complex systems toward more comprehensive representations of metabolic pathways. Modern machine learning frameworks have been invaluable in predicting protein-protein interactions and the elucidation of their roles within metabolic networks. Lumping these frameworks into multi-species modelling approaches enhances metabolic prediction accuracy and provides further insights into how genetic variation impacts metabolic pathways. Network-based methods, using graphical theoretical approaches, visually and mathematically represent metabolic reactions that enable the identification of critical regulatory nodes and assist in exploring the consequences of genetic variations. These are constraint-based models, like FBA, which integrate genomic and proteomic data for a holistic view of cellular metabolism by tracking the flux of metabolites. With such advances, new challenges arise in the form of high-dimensional datasets produced via high-throughput technologies such as mass spectrometry and next-generation sequencing. Development of effective computational tools and standard methods for storage and visualization will be important for sorting this complexity out to create meaningful insights from these data. This review, therefore, tried to

highlight recent advances in multi-omics integration from a computational perspective. Predictive modelling approaches and standardization necessary for data management and visualization were discussed. Synthesizing the current techniques with the emerging trends have formed the basis for this review with an objective of trying to outline the future directions for research in metabolic pathway modelling and resolution of mechanisms underlying metabolic diseases.

BACKGROUND

Different metrics—quantitative and qualitative—are followed in use with the various classification techniques for building, fine-tuning the predictive model.

K-Means Clustering: K-Means is an unsupervised machine learning algorithm that attempts to partition data into 'k' clusters that are non-overlapping. At the first step, all 'k' centres are initialised, and then every datum belongs to the closest centre regarding the distance. The cluster centre is then re-estimated as the mean of all points assigned to one cluster. This process is iterated until a stop condition is reached. This algorithm is important in the estimation of natural clusters present in the dataset and could be used to exhibit patterns or subgroups that may relate to the analyses.

Principal Component Analysis: is a way of cutting down dimensions in high-dimensional datasets. The technique reduces the dimensions of the original dataset in relation to the first few principal components in such a way that they account for the variance in the dataset. In that respect, this type of methodology enables one to project high-dimensional data into low-dimensional space and, thus, one can understand the basic pattern and correlations among variables.

The Protein-Protein Interaction Network: uses the principles of graph theory for the visualisation and analysis of interactions taking place amongst proteins of a biological system. Here, the proteins are represented by nodes and their interaction by the edges between the nodes. Such a representation would allow one to get highlighted key proteins known as hubs; bring out information on the strength of interactions; allow functional clustering or pathway mapping within a network. The understanding of structural and functional relationships among proteins and their contributions to cellular activities is extremely desirable.

Confusion Matrix: The confusion matrix is a quantitative measure that appraises or evaluates classification models. It provides a quick summary of predictions made by a model in comparison with the real labels. It consists of true positives, true negatives, false positives, and false negatives, which help to find out accuracy and the loophole in the classification of different classes. Consequently, it contains insight into useful knowledge regarding model performance and points out areas needing improvement.

Stoichiometric analysis: involves the process of developing and researching a stoichiometric matrix that will represent, in quantitative terms, relations among metabolites and reactions in metabolic pathways. Such a matrix provides important information with regard to metabolic fluxes and interactions in a biological system. Further support toward easy investigation of the metabolic pathways along with their regulatory mechanisms comes from additional analysis that incorporates heat maps of various visualisation methodologies, enabling the identification of the contained patterns, anomalies, and relationships.

3D Visualization of Atomic Coordinates: Conformation of three-dimensional atomic arrangements is represented through the method of 3D visualisation of atomic coordinates. It will do graphical representation of macromolecules with general information on the three-dimensional arrangement of the protein, which will be helpful for describing the overall spatial relationship, folding, and functional regions. The three-dimensional visualisation of atomic coordinates will aid in interpretation of the complex structure, identification of binding sites, and communication of results associated with a structure.

RESEARCH METHODOLOGY

3.1 Genomics

Data Collection and Pre-processing-

Data collection and pre-processing are foundational to any study, encompassing the volume, accuracy, speed, and diversity of the data. For this study, the focus is on genomic and proteomic data relevant to metabolic pathway mapping. The raw data, often retrieved from sources such as public repositories or experimental datasets, may come with various quality issues such as missing values, outliers, and inconsistencies. These issues are addressed through data cleaning methods, which include filling missing values, normalising data scales, and removing outliers to ensure the dataset is accurate and reliable for analysis.

Feature Analysis and Model Selection-

Feature analysis involves evaluating the dataset to identify and mitigate any biases or irrelevant features. This step is crucial as it ensures the model focuses on the most relevant features that demonstrate a strong relationship with the target variable. In this review, various analytical and predictive methods are applied:

Principal Component Analysis (PCA): PCA is employed for dimensionality reduction. It transforms high-dimensional data into a lower-dimensional space while preserving as much variance as possible. This helps in visualising and interpreting complex datasets by reducing them to two principal components.

K-Means Clustering: This algorithm is used for unsupervised classification. By partitioning the data into clusters based on feature similarity, K-Means identifies patterns within the data, helping to group similar samples and reveal underlying structures.

Graph Theory for Protein-Protein Interaction Networks: Graph-based approaches are utilised to analyse protein-protein interactions. Nodes represent proteins, and edges indicate interactions, allowing for the visualisation and analysis of network structures and identifying key proteins and interaction patterns.

Confusion Matrix: This quantitative measure assesses the performance of classification models by comparing predicted labels with actual outcomes. It provides a detailed breakdown of prediction accuracy, helping to identify misclassifications and refine the model.
Stoichiometric Analysis: This involves analysing the stoichiometric matrix to understand the quantitative relationships between metabolites and reactions in metabolic pathways. It helps in exploring the fluxes of metabolites and integrating data from genomics and proteomics.
3D Visualization of Atomic Coordinates: This technique is used to visualise the three-dimensional structures of proteins. By representing atomic coordinates in 3D, researchers can gain insights into protein folding, interactions, and potential binding sites.

Implementation-
The implementation of these methods is conducted using Python in a Jupyter Notebook environment, supported by cloud-based platforms such as Google Colab for enhanced computational resources. The following tools and libraries are employed:
Pandas: For data manipulation and analysis.
Scikit-learn: To apply PCA, K-Means clustering, and other machine learning models.
Matplotlib and Seaborn: For data visualisation, including scatter plots and PCA results.
NetworkX: For graph-based analysis of protein-protein interactions.
SciPy: For advanced scientific computations including stoichiometric analysis.
PyMOL: For 3D visualisation of protein structures.

3.2 Proteomics
Data Collection and Pre-processing –
Visualising 3-D atomic coordinates and protein-protein interaction networks form the basis of this analysis. In the case of 3-D atomic coordinates, synthetic data are generated to simulate atomic positions in a protein structure. Random coordinates are generated between a predefined range of minimum and maximum values to simulate actual atomic coordinates. The protein-protein interaction network makes use of the clean dataset comprising columns of interacting and non-interacting protein pairs. Preloading checks the dataset for existence and correctness, while postloading includes cleaning and visualisation steps of data. This cleaning of data ensures that only valid and relevant information forms a part of the data, hence allowing for effective representation and analysis.

Feature Analysis and Model Selection-
In feature analysis, atomic coordinates are viewed for visualisation, and interaction data is viewed for network analysis. For the purpose of 3D visualisation, synthetic atomic coordinates are used in order to develop a scatter plot in three-dimensional space. The insight into the spatial arrangement of the atoms in a protein structure is provided by this visualisation. In network analysis, there are interacting pairs of proteins; such data can be viewed as a graph where those pairs serve as edges. It's an opportunity to see how proteins interact and define important proteins and their relationships.

Implementation-
The implementation has been done using Python, employing its libraries for visualisation and data processing, namely Matplotlib, NumPy, and NetworkX:
3D Visualisation: The code uses matplotlib and mpl_toolkits.mplot3d to visualise atomic coordinates in 3D. The plot of random data points in 3D space is done with specified labels for each axis. This, in turn, visualises the spatial distribution of atoms.
Protein-Protein Interaction Network: The following example uses a networkx library for the creation and visualisation of a protein-protein interaction network. First, it reads the dataset from a CSV file. Then it constructs a graph where nodes will represent proteins, and edges will show the interactions with weights associated. The spring_layout function from networkx is used here for node positioning, and the network will be drawn with labels of nodes and weights of edges so that interaction strength can be given.
Error Handling: Errors will be handled throughout the implementation. This will handle issues to do with file paths and data integrity. Specific exceptions will be caught such as FileNotFoundError and KeyError so that proper loading and processing of the dataset are ensured. This would involve creating a 3D scatter plot to visualise atomic coordinates and visualising protein interactions as a network. Figure sizes are increased to ensure clarity, while plt.show() is utilised in view to interactively show the visualisations.

3.3 Metabolic Pathways
Data Collection and Preprocessing-
Data used in this research can be defined by the 4Vs: volume, veracity, velocity, and variety, attributes that determine the size, accuracy, speed of collection, and diversity of data, respectively. In this work, we will be using datasets retrieved from protein structures and metabolic pathways that are bound to inherit quality issues such as a lot of missing values and inconsistencies, which usually require a considerable amount of preprocessing. Especially, there may be some anomalies in the raw data about protein structure or pathway classification. The data cleaning techniques are applied to the input data including missing value handling and feature normalisation such that the data becomes appropriate and accurate for further processing and analysis.

Feature Analysis and Model Selection-
Feature analysis can be done by analysing the dataset for those features that would lead to bias. Focus on relevant features which relate consistently to the target variable in order to give good performance of the model. For the models employed, we have
Principal Component Analysis: PCA helps reduce dimensionality and determines important features by changing the data into principal components that capture most of the variance.
K-Means Clustering: Data has been segregated into a number of k clusters based on the proximity of centroids, which helps in identifying the pattern or the relationship among data.
Stoichiometric Analysis: Stoichiometric matrices are drawn out for the analysis of metabolic pathways by determining quantitative relationships among metabolites and reactions.

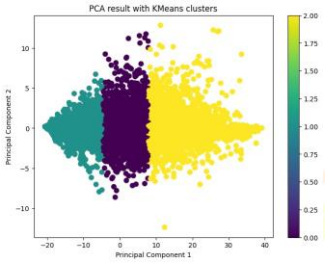
3D Visualisation: 3-D scatter plots represent the spatial distribution of atoms in protein structures by visualisation of atomic coordinates.
Protein-Protein Interaction Network: Graph theory is applied to visualise interaction networks that provide insight into relationships between proteins through interaction data.

Implementation-
This work has been implemented in Python and done in the environment of an IPython notebook for which Google Colab is preferred because of their free cloud computation services. The different tools and libraries which are used in this work are mentioned below:
Pandas: This is a software library which is used for data manipulation and cleaning. Pandas has been used to load the dataset. Handling of missing values was done.
NumPy: for numerical operations and to generate random data for visualisation. Matplotlib: to create plots/visualisations, including 3D scatter plots and heatmaps. Seaborn: enhancement of visualisations, especially plotting heatmaps for stoichiometric matrices and confusion matrices. Scikit-learn: this will be used to actually implement the machine learning algorithms of PCA and K-Means clustering, along with the evaluation of classification performance using confusion matrices and classification reports.

RESULTS

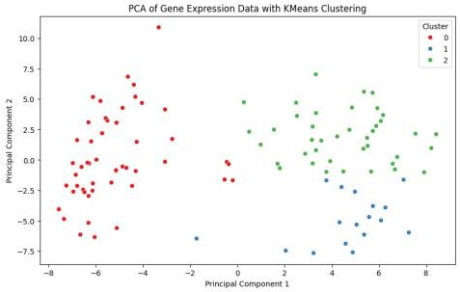
A. Analysis of Data Clustering using PCA and KMeans

The PCA output with KMeans clustering also demonstrates three clear clusters in the dataset, which essentially means that, based on their principal components, the data could be segregated effectively into three different groups. The first principal component has a wide spread from -20 to 40, and the second component varies between -10 and 10; this evidence shows that each component captures different levels of variance within the data. The separation between the teal, purple, and yellow clusters is well-contrasted, which indeed says these are very representative of the underlying structure. On the other hand, partial overlaps between clusters hint that there may be ambiguities or shared characteristics between data points, which could be further analysed or to which different clustering techniques can be applied.



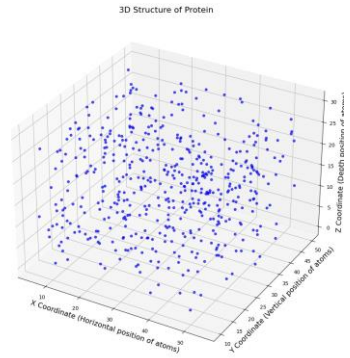
B. Clustering Analysis of Gene Expression Data Using PCA and KMeans

The following biplot of the PCA from gene expression data yields three obvious clusters from KMeans clustering that allow us to infer segregation of the samples into three groups according to their expression profiles. Well-separated and dense cluster 0 may indicate high within-cluster similarities in red, whereas Clusters 1 and 2, coloured in blue and green, respectively, had more spread, which could mean more variability within those clusters. These clear separations between the clusters may indicate meaningful biological differences among the groups, but the overlap and dispersion in Cluster 1 and Cluster 2 could mean that there is some heterogeneity that may be further investigated or maybe a way in which the clustering has been done in a more refined manner.



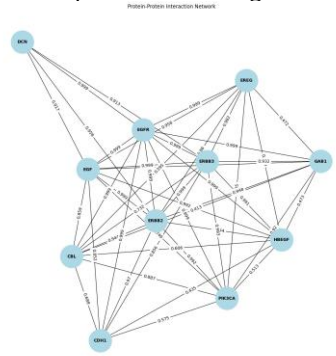
C. Three-Dimensional Atomic Distribution and Structural Conformation of a Protein

The three-dimensional scatter plot gives, in a spatial manner, the distribution of atoms in this protein. The X, Y, and Z coordinates for each atom are given by the axes. This representation will visualise how atoms are positioned in three-dimensional space to illustrate the overall conformation of this protein. These scattered points reflect the dense and sparse regions within the protein structure: the high atomic concentration areas, which could be the core or folded regions, and less-dense areas probably comprising loops or extended regions. The following distribution pattern can be imperative in understanding stability, function, and interactions of a protein with other molecules. This three-dimensional view forms the basis for further studies, such as docking analysis, protein folding, or dynamics simulations, which enables complete insight into the biological role of the protein.



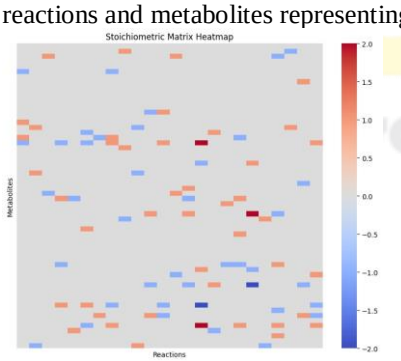
D. Protein-Protein Interaction Network of EGFR Family in Cellular Signalling Pathways

The following PPI network illustrates the interactions of several key proteins playing their role in cellular signalling pathways, with a particular emphasis on those implicated by the EGFR family and associated signalling components. Notice that EGFR is right in the centre, interacting with such proteins as ERBB2, ERBB3, and PIK3CA-players already well-known in the development and progression of cancer. This high connectivity points toward strong interactions and, possibly, cooperative involvement in signal transmission. Interacting proteins like the pairs EGFR-ERBB2 and EGFR-ERBB3 exhibit high interaction scores, reflecting their critical involvement in signalling by growth factors. The overall topological structure in this PPI network underlines the complex and interconnected nature of the signalling pathways mediated by these proteins, potentially implicating their involvement in cellular proliferation, migration, and oncogenesis.



E. Heatmap Visualization of Metabolite-Reaction Interactions in a Metabolic Network

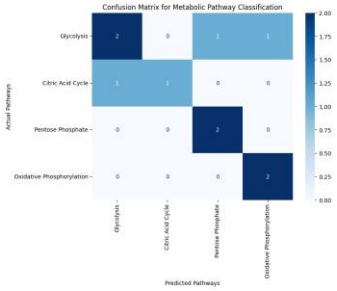
Heatmap of the stoichiometric matrix showing interactions between metabolites and reactions of the metabolic network. The range of values is set between -2.0 to 2.0. Colours used are red for positive values, indicating metabolite production in a reaction, and blue for negative values, indicating metabolite consumption in respective reactions. Red and blue colours vary in saturation to show the magnitude of metabolite involvement with reactions. Notice that some reactions have high positive values, such as the deep red cells at about (row 7, column 12) and (row 18, column 22), indicating high production of metabolites. Darker blue cells, such as those near (row 15, column 20) and (row 25, column 30), indicate reactions with high consumption of metabolites. The sparsity of the matrix indicates that each reaction typically involves only a few metabolites, reflecting the highly selective nature of metabolic processes. Such a visualisation will give insight into the distribution and intensity of metabolic activity and thus highlights key reactions and metabolites representing putative targets for further study.



F. Confusion Matrix Analysis of Metabolic Pathway Classification Model

Below is the metabolic pathway classification confusion matrix performance of the model in four pathways: Glycolysis, Citric Acid Cycle, Pentose Phosphate, and Oxidative Phosphorylation. Here, the model correctly predicted 2 true positives each in the class of Glycolysis, 1 true positive in the class of Citric Acid Cycle, 2 true positives in the class of Pentose Phosphate, and 2 in the class of Oxidative Phosphorylation. It mislabeled 1 instance of Glycolysis as Citric Acid Cycle and another instance of Glycolysis as Oxidative Phosphorylation. Also, 1 instance of Citric Acid Cycle was misclassified as Glycolysis. However, all instances of Pentose

Phosphate and Oxidative Phosphorylation were classified correctly. This matrix indeed brings into light that, though performing reasonably, the model lacks clear distinction in certain pathways between Glycolysis and the Citric Acid Cycle.



CONCLUSION

In conclusion, this review provides a critical analysis of genomics and proteomics data with respect to metabolic pathway mapping. The integration of different computational approaches in the current work-PCA, K-Means clustering, and graph theory-has enabled distinct patterns of the data to be found, including the three clusters of gene expression data whose principal components range from -20 to 40 and from -10 to 10, respectively. Methods listed herein reflect meaningful biological differences, although the degree of overlap suggests avenues for improvement.

These atomic coordinates for proteins are then visualized in 3D, giving a very important insight into the structure of proteins. Denser regions are the places that may be of interest for further research. The protein-protein interaction network points to important relationships, such as those very strong interactions within the EGFR signaling pathway, to show just how complex and important these networks are to cellular processes. The stoichiometric matrix heatmap, ranging in value from -2.0 to 2.0, depicts the selectivity of metabolic reactions through the generation and consumption of metabolites. The confusion matrix in pathway classification for metabolism is showing true positives with some misclassifications across four pathways, demonstrating the fact that reasonable performance is achieved but there is room for improvement, especially with respect to differentiating between Glycolysis and the Citric Acid Cycle. The study, in summary, underlines the potentiality presented by integrated genomics and proteomics data, with strong support from robust computational tools for gathering higher knowledge of metabolic pathways. Logical outcomes would, therefore, be that future research efforts need to be directed toward refinement of methodology, minimizing inconsistencies in data, and further refinement of clustering and classification methods to achieve higher accuracy of biological interpretation.

REFERENCES

[1] Murray Cadzow, James Boocock, Hoang T. Nguyen, Phillip Wilcox, Tony R. Merriman, Michael A. Black “A bioinformatics workflow for detecting signatures of selection in genomic data”, Volume 5 - 2014 | https://doi.org/10.3389/fgene.2014.00293

[2] Ming Chen, Ralf Hofestädt “A medical bioinformatics approach for metabolic disorders: Biomedical data prediction, modelling, and systematic analysis” Journal of Biomedical Informatics Volume 39, Issue 2, April 2006

[3] Carlotta De Filippo, Matteo Ramazzotti, Paolo Fontana, Duccio Cavalieri “Bioinformatic approaches for functional annotation and pathway inference in metagenomics data” *Briefings in Bioinformatics*, Volume 13, Issue 6, November 2012, Pages 696–710

[4] Julia Handl, Joshua Knowles, Douglas B. Kell “Computational cluster validation in post-genomic data analysis” *Bioinformatics*, Volume 21, Issue 15, August 2005, Pages 3201–3212

[5] Claudia Perea, Juan Fernando De La Hoz, Daniel Felipe Cruz, Juan David Lobaton, Paulo Izquierdo, Juan Camilo Quintero, Bodo Raatz & Jorge Duitama “Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP” Volume 17, article number 498, (2016)

[6] Stuart Maudsley, Wayne Chadwick, Liyun Wang, Yu Zhou, Bronwen Martin, and Sung-Soo Park “Bioinformatic Approaches to Metabolic Pathways Analysis” Methods Mol Biol. Author manuscript; available in PMC 2016 Jan 4. *Published in final edited form as:* Methods Mol Biol. 2011; 756: 99–130.

[7] Alessio Fallani, Leonardo Medrano Sandonas, and Alexandre Tkatchenko “Inverse mapping of quantum properties to structures for chemical space of small organic molecules” Nat Commun. 2024; 15: 6061.Published online 2024 Jul 18. doi: 10.1038/s41467-024-50401-1

[8] Zak Costello & Hector Garcia Martin “A machine learning approach to predict metabolic pathway dynamics from time-series multi omics data” *npj Systems Biology and Applications* Article number: 19 (2018)

[9] Yue Zhang, Mengqi Luo, Peng Wu,5 Song Wu, Tzong-Yi Lee and Chen Bai “Application of Computational Biology and Artificial Intelligence in Drug Design” Int J Mol Sci. 2022 Nov; 23(21): 13568.

Published online 2022 Nov 5. doi: 10.3390/ijms232113568

[10] Joseph M Dale, Liviu Popescu & Peter D Karp “Machine learning methods for metabolic pathway prediction” *BMC Bioinformatics,* Article number: 15 (2010)

[11] Todd J. Dolinsky, Paul Czodrowski, Hui Li, Jens E. Nielsen, Jan H. Jensen, Gerhard Klebe and Nathan A. Baker “PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations” Nucleic Acids Res. 2007 Jul; 35(Web Server issue): W522–W525.

[12] Alice Cambiaghi, Manuela Ferrario, Marco Masseroli “Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration” *Briefings in Bioinformatics,* Volume 18, Issue 3, May 2017, Pages 498–510

[13] Xiaogang Wu, Mohammad Al Hasan, and Jake Yue Chen “Pathway and Network Analysis in Proteomics” J Theor Biol. Author manuscript; available in PMC 2015 Dec 7.

[14] Lixia Yao, James A. Evans, and Andrey Rzhetsky “Novel opportunities for computational biology and sociology in drug discovery” Trends Biotechnol. Author manuscript; available in PMC 2013 May 15.

[15] Andreas Schmidt, Ignasi Forne and Axel Imhof “Bioinformatic analysis of proteomics data” BMC Syst Biol. 2014; 8(Suppl 2): S3

[16] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences" Brief Bioinform. 2018 Mar; 19(2): 286–302.

[17] Debasis Mitra, Debanjan Mitra, Mohamed Sabri Bensaad, Somya Sinha, Kumud Pant, Manu Pant, Ankita Priyadarshini, Pallavi Singh, Saliha Dassamiour, Leila Hambaba, Periyasamy Panneerselvam, Pradeep K. das Mohapatra "Evolution of bioinformatics and its impact on modern bio-science in the twenty-first century: Special attention to pharmacology, plant science and drug discovery" Computational Toxicology Volume 24, November 2022, 100248

[18] Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein "*MolLM*: a unified language model for integrating biomedical text with 2D and 3D molecular representations" Bioinformatics. 2024 Jul; 40(Suppl 1): i357–i368.

[19] Erol E. Gulcicek, Christopher M. Colangelo, Walter McMurray, Kathryn Stone, Kenneth Williams, Terence Wu, and Hongyu Zhao "Proteomics and the Analysis of Proteomic Data: An Overview of Current Protein-Profiling Technologies" Curr Protoc Bioinformatics. Author manuscript; available in PMC 2013 Dec 15.

[20] Kobe De Becker, Niccolò Totis, Kristel Bernaerts, Steffen Waldherr "Using resource constraints derived from genomic and proteomic data in metabolic network models" Current Opinion in Systems Biology Volume 29, March 2022, 100400