Current Impact of Bioinformatics Based Approaches in Providing Solutions against Rheumatoid Arthritis

[1,2]Priyangulta Beck, [1,3]Hewanti Kumari, [1,3]Sweety Guria Rani, [1,3]Purbasha Pati, [1,3]Sadaf Naaz, [1,2]Kumari Surekha Mahto, [1,2]Shekhar Marandi, [1,4]Anant Kumar Mehta, [4]Shristi Kumari, [4]Agatha Sylvia Khalko, [3]Soma Roy and [1*]Mukesh Nitin

[1]Department of Tech. Biosciences, Digianalix, South Samaj Street, Tharpakhna, Ranchi-834001, Jharkhand, INDIA.
[2]Department of Botany, Ranchi University, Ranchi, Jharkhand, INDIA.
[3]Department of Biotechnology, Ranchi Women's College, Circular Rd. Nagra Toli, Ranchi- 834001, Jharkhand, INDIA.
[4]Department of Biotechnology, Marwari College Ranchi, Jharkhand, INDIA.

***Abstract*: Rheumatoid Arthritis (RA) is a chronic autoimmune inflammatory disease characterized by symmetric synovial joint inflammation, with an incidence of 0.5-1 percent in prosperous nation. It occurs due to abnormal activity in immune response that encourages our immune system to attack its own cells and starts attacking the lining of joints. Occurrence of chronic inflammation is related to the significant involvement of vital inflammatory cytokines and TNF alpha (TNF- α). Maximum patients develop RA due to environmental, genetic and hormonal activities. 50% of RA patients develop RA due to genetic factors. It affects joints in palm, knees, wrists and several organs like skin, lungs, blood vessels, kidneys and heart. The exact cause of RA is still not found. Various types of drugs are implemented to treat rheumatoid arthritis-like non-steroidal anti-inflammatory drugs (NSAIDs), corticosteroids, Disease-Modifying Anti-Rheumatic Drugs (DMARDs). Past few decades, more than 60 molecular docking tools have been discovered to study the different properties of RA. Few of them are- Next-generation sequencing (NGS) and the omics approaches (transcriptomics, epigenomics and genomics), dispense novel views of genome-wide association studies (GWAS). Bioinformatics has also taken accurate steps together with genomics and docking studies to support the drug discovery in treating rheumatoid arthritis.**

***Keywords*: Rheumatoid arthritis, Docking tools, Omics, GEO, NGS**

1. INTRODUCTION

Rheumatoid arthritis is an autoimmune illness that affects the joints in which a person's immune system attacks its self-tissue, affecting many joints, involving hands and legs. RA is a chronic inflammatory rheumatic illness that affects both articular and extra-articular structures, causing discomfort, disability, and death. Constant inflammation leads to erosive joint damage and functional impairment in maximum patients. The beginning of the disease is not similar in all patients but varies regarding different types, number, and the pattern of joint involvement.

HLA gene is the most notable genetic risk factor for RA. Apart from Human Leukocyte Antigen (HLA), Protein Tyrosine Phosphatase Non-receptor Type 22 (PTPN22), Peptidyl Arginine Deiminase type 4 (PADI4), Signal Transducer and Activator of Transcription 4 (STAT4), Cytotoxic T-lymphocyte associated protein 4 (CTLA-4), Tumor Necrosis factor Receptor associated factor (TRAF1), Tumor Necrosis Factor (TNF) [1]. PTPN22 is the second strongest association with RA after HLA- DRB1 gene [2].

The potency of the inflammatory activity, and the existence or lack of different alters like genetical factors, regular swollen joints, autoantibody in serum, and the intensity of inflammatory activity, can action the development of the disorder. With the observation, we first determined differentially expressed genes (DEGs) of RA linked to the development of new blood vessels from pre-existing vessels by computational investigation. Collagen-induced arthritis model has numerous similar pathological and immunological characteristics to human RA, several animal experiments have used CIA mice as RA models.

2. MOLECULAR DOCKING APPROACHES

The objective of molecular docking research is to imitate the molecular identification procedure via computational operations. Molecular docking research attempts to achieve optimal justification for proteins as well as ligands, and considerable adaptation across proteins and ligands, in order to decrease the entire organization's free energy. Two methods are prevalent within the molecular docking community. The first method implies a matching strategy in which the protein and ligand are described as complementary surfaces. The second method calculates the ligand-protein pair-wise interaction energies, simulating the real docking process. These two methods of molecular docking can be performed by following any interrelated steps like small molecule preparation, protein preparation, generation of grid files and molecular docking. A brief pipeline for molecular docking studies is illustrated in Figure 1.

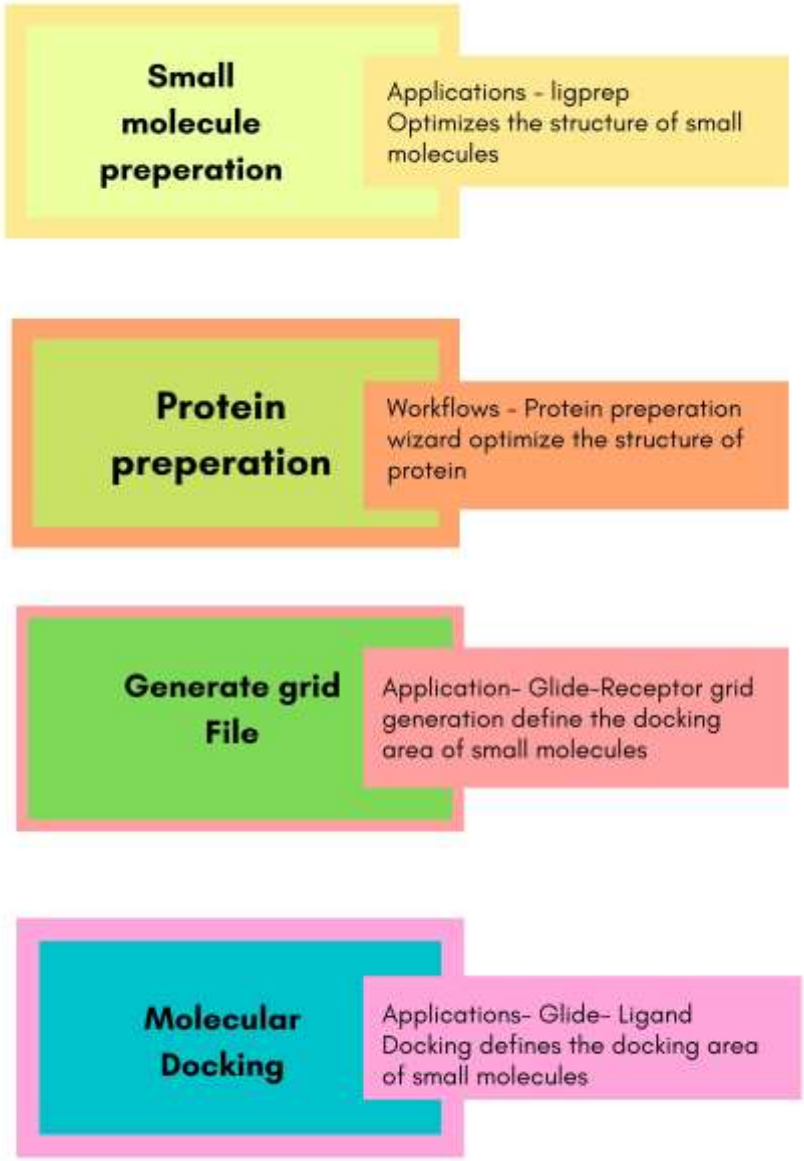

Figure 1- The pipeline for the studies of ligand receptor using molecular docking approaches.

2.1 Process of Molecular Docking

2.1.1 Docking Tools

For both academic and commercial purposes, more than 60 different docking tools and programs have been developed, such as DOCK [3], AutoDock [4], FlexX [5], Surflex [6], GOLD [7], ICM [8], Glide [9], Cdocker, LigandFit [10] MCDock, FRED [11] MOE-Dock [12], LeDock [13], AutoDock Vina [14], rDock [15], UCSF Dock [16] and many others, over the last two decades. Among these tools AutoDock Vina, GOLD and MOE-Dock are known as the best docking tools due to their top ranking poses with best scores.

Table 1. Docking tools and their applications

Docking tools	Application
Autodock	It is proposed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure
DOCK	It addresses firm body docking using a geometric matching algorithm to superimpose the ligand onto a negative image of the binding sac.
GOLD.	Protein–small molecule docking
FlexAID	Used as minute molecules as well as peptides as ligands and Proteins and nucleic acids as docking targets.
HomDock	It used to improve docking accuracy and efficiency in cases where a complex structure of a ligand with the target protein is known
ICM	Ligand-protein docking, peptide-protein docking, and protein-protein docking
Glide	Protein-small molecule docking
	X-ray crystallography, structure-based drug design, lead escalation, virtual screening, protein-protein docking
	Structure-based drug design, flexible protein-ligand molecular docking
FRED	Protein-ligand interaction
	Docking program based on an idealized active site ligand (a protomol), used as a target to generate putative poses of molecules or molecular fragments, which are resulted in using the Hammerhead scoring function
ParaDockS	Tool for molecular docking with population-based metaheuristics
MS-Dock	Free multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening
SwissDock	Predict the molecular interactions that may occur between a target protein and a small molecule
rDock.	Docking Ligands to Proteins and Nucleic Acids
HADDOCK	Protein-protein docking, protein-ligand docking
	Protein preparation (AMBER package and Reduce), molecular mechanics applications (AMBER package), and docking and scoring (AutoDock Vina and SLIDE) are used.
	It manages diverse ligands and arranges digital broadcasts, and arranges docking works with adjustable side-chains.

2.1.2 Docking studies highlighting target protein in RA

Molecular docking of RA related target proteins, i.e., BTK (Bruton's Tyrosine Kinase) mutant protein, PAD4(Protein Arginine Deiminase 4) and CCL28 (C-C motif chemokine ligand 28) have been organized to calculate the binding of different ligands with their active sites via Glide docking, AutoDock and MOE-Dock software respectively [17,18,19]. Cadherin-11 is a protein that takes part in the organization of synovium [20], C-reactive proteins may increase the risk of RA because of its sensitive marker of systematic inflammation in RA patients [21].

3. 'OMICS' APPROACHES BY NGS CONTRIBUTE IN CASE OF RA

Omics is a bioinformatics software system that makes it simple to get from reads to insights. Omics are used for the NGS data analysis of genomes, transcriptomes, metagenomes and many more. Modules are used to construct Omics. Chronic synovitis, systemic inflammation, and varying degrees of bone and cartilage degradation of designated RA, an inflammatory disease that primarily affects the lining of synovial joints. The systemic inflammation that characterizes RA is associated to a variety of extra-articular disorders, including cardiovascular disease, which leads to a greater mortality rate in RA patients. In modern countries, RA is the most frequent form of inflammatory arthritis. Rapid improvements taking place in the field of 'omics' equipments, over the last decade have resulted in significant improvements in our ability to interpret genetic and molecular reasons behind complicated disorders like autoimmune diseases. The etiology of RA is not fully understood, however, NGS(Next generation sequencing) Combining innovative multi-omics approaches, cell profiling technologies, and bioinformatics tools has allowed for a more comprehensive investigation and deeper insight into the pathogenesis and disease variants of RA, including the definition of RA-associated cell populations, specific gene expression profiles, susceptibility loci, gene-environment interactions, and genetic loci associated with subsets of patients and those linked with response to RA treatment.

3.1 GENOMICS:

3.1.1 Genomics: Genomics is a wide association study and NGS have assisted in the finding of unique genomic variants (e.g., genetic polymorphisms) with the objective of better interpretation, complex disease, pathobiology and estimate the influence of nutrient input and genetic variants in humans.

3.1.2. Quality control and assessment: FastQC [22] and Trimmomatic [23] are the docking tools used to analyze the quality

control of samples, filter readings and separate low-quality bases.
3.1.3. **de-novo assembly:** the assembly features allow reconstructing whole-genome sequences without a reference genome or specific hardware requirements. Three different algorithms: AbySS [24], SPAdes[25], and Flye[26], are used to assemble sequencing data from short and long read technologies.
3.1.4. Repeat masking: mask repeats of genome assemblies with Repeat Masker 27] to improve downstream gene predictions.

3.2 TRANSCRIPTOMICS:
3.2.1. Transcriptomics: thanks to technologies like real-time PCR and powerful microarrays, transcriptomics investigations have become common. Furthermore, RNA sequencing has emerged as a significant option for transcriptome research since it covers a broader range of RNAs, resulting in more useful data.
3.2.2. Quality control: FastQC and Trimmomatic used to conduct the quality control of samples, filtering reads and separate low-quality bases.
3.2.3. de-novo assembly: assemble short reads with the trinity to obtain a de-novo transcriptome without a reference genome. Assess the completeness of the transcriptome with BUSCO[28], cluster similar sequence with CD-HIT[29], and predict coding regions with Trans decoder[30].
3.2.4 RNA-Seq alignment: Align RNA-seq to reference genome using of STAR(Spliced transcripts alignment to a reference)[31] OR BWA(burrow wheeler aligner)[32]. However, your hardware is stored in the cloud at Omics Box.
3.2.5. **Quantify expression:** at gene or transcript level through Htseq[33] or RSEM[34] and with or without a reference genome.
3.2.6. Differential expression analysis: detect differentially expressed genes between experimental conditions or over time with well-known and versatile statistical packages like NOISeq[35], edger[36] or maSigPro[37], affluent accommodations helps to interpret results.
3.2.7. Enrichment analysis: by mixing different expression that results in functional annotations, and is known as biological functions.

3.3 METAGENOMICS:
3.3.1. Quality control and assessment: This works similarly to transcriptomics and genomics.
Taxonomic classification - identify bacteria, archae, fungi and any microorganisms by Kraken2 [38].
3.3.2. Metagenomics assembly: select among MetaAPAdes [39] and MEGAHIT [40] to collect considerable dataset easy and fast in the cloud.
3.3.3. Gene prediction: use FragGeneScan [41] for plain reads Prodigal for assembled data to identify and extract possible genes and proteins.
3.3.4. Functional analysis: EggNOGMapper [42] and PfamScan [43] are used for the function analysis of metagenomics.

3.4 EPIGENOMICS: Epigenomics is the study of how nutrition and bioactive food ingredients affect global epigenetic systems that control gene activity and expression. DNA methylation analysis, ChIP seq[44], ATAC seq[45].

3.5 METABOLOMICS:
Metabolomics is the analysis of the entire set of metabolites or tiny molecules (metabolome) found in biological samples. In general, the main technologies used in metabolomics investigations have been Nuclear Magnetic Resonance (NMR), Proton Nuclear Magnetic Resonance (1HNMR) Spectroscopy, and Mass Spectrometry (MS).

4. SYSTEMICALLY APPROACHES CONTRIBUTE IN CASE OF RA

IRAMUTEQ is a text based study software that classified texts into four categories which is based on their textual form, into three main parts such as, in vivo models (class 1), clinical practice and traditional medicine (classes 2 and 3), and in vitro models (class 4). Beneficially it creates a similarity tree of the terms which found in the abstracts, as well as a word cloud with the most frequently mentioned terms. IRAMUTEQ software is used as a methodological tool which has been satisfying, as it allowed researchers to certify the main experimental models used, keywords, patho-physiological processes, and molecules involved in the etiology of rheumatoid arthritis without partial as well as being a tool for visual and impulsive results. The aim of this work was to use IRAMUTEQ that analyze text fragments qualitatively and quantitatively, to conduct overall literature review on experimental models in rheumatoid arthritis.
Step includes;

- The textual corpus is a study of the abstracts of publications that combined into a single file in text format (.txt).
- The following identifying directories have assigned to each abstract: primary author of the publication and experimental model used.

5. MD SIMULATION

MD simulation, which was first developed in the late 1970s [46], is a process of analyzing the physical movement of atoms and molecules by using numerical methods. This type of simulation captures the behavior of proteins and other biomolecules in full atomic detail and excellent temporal resolution. These simulations represents a large variation of biomolecular processes like constructional changes, ligand binding and protein folding.

The present generation of computers uses parallelism and accelerators to speed up the process. Most popular simulation codes are AMBER [47], CHARMM [48], GROMACS [49] or NAMD [50] have long been compatible with the Message Passing Interface (MPI).

In molecular docking studies of the cabability of *B.Sapida*, are found to be anti-rheumatoid arthritis. *B.Sapida* contains a lead chemical called quercetin, which has been shown to be useful in RA medication development studies. Here, quercetin has been used to check its drug possibility and molecular docking is executed by using AutoDock 4.2.1 between TNF-α and quercetin. The inhibitory effect of quercetin on RA plasma has been examined with the help of immunological assay ELISA, therefore, quercetin showed a non-carcinogenic reaction and may cross the membrane barrier easily. Ten different binding poses and best binding poses of TNF-α and quercetin showed -6.3 kcal/mol minimum binding energy and 23.94 μM inhibitory constant. In addition to this, ELISA indicated 2.2 down regulated expression of TNF-α in RA compared to control.

MD simulations have a history of more than forty years. Recently Molecuar Dynamics were efficient to gather time proportion that were suitable for biological processes. Constructional adjustment or ligand binding may now be efficiently modeled when routine simulations approach the microsecond scale. The advancement of computational equipment, particularly the use of GPUs (graphic processing units), as well as advances in MD algorithms, including coarse-grained algorithms, allow us to move away from the analysis of single compositions to the analysis of conformational ensembles, which is the foundation of molecular modeling.. Conformational ensembles provide much better depiction of fundamentals macromolecules since they account for flexibility and dynamic features (including all thermodynamic information) and make experimental data easier to match. Although the conceptual shift was understood and the technologies are improving, there is still a long way to go before biomolecular simulations, such as the development of conformational ensembles, become practice. Tools exist that make the setup of a macromolecular system much easier, and even allow the non-experts to enter the simulation world. However, shortage of representation required less improved searching device, and the difficulties in easily keeping and transferring vast number of data trajectory remain unsolved. In any case, MD is already a valuable tool in helping to understand biology.

6. BIOINFORMATICS APPROACH ON RA

For the diagnosis of RA there are many bioinformatics analytical tools were extensively used to identify target crucial biomarkers and potential etiology of RA. Three gene expression datasets were generalised using microarrays taken from the GEO (Gene Expression Omnibus) database for the investigation of the aetiology of RA. GSE55235 and GSE55457 datasets were aligned for the following study. The most significant sub network, which has 14 genes and 45 edges, was created with Cytoscape software and the MCODE plug-in[52]. For ROC curve analysis, eight genes with AUC >0.80 were designated as RA hub genes. Finally, the DEGs (differentially expressed genes) and their closely linked biological roles were defined, and they held chemokines and immune cell infiltration that accelerates the course of rheumatoid arthritis when compared to healthy controls.

Microarray technology, which has been available for more than 20 years, allows researchers to examine the full transcriptional information of a variety of cell types and tissues. Studies based on gene expression analysis have yielded new insights, indicating how the transcriptome changes across different symptoms and stages of the illness. The Gene Expression Omnibus (GEO) is a user-friendly archive where users may search and download microarray, next-generation sequencing, and other genomics data. As a result, we used a combination of bioinformatics research tools, including R packages from Bioconductor, STRING database[53], CIBERSORT website, Cytoscape, and GSEA software, to deconstruct biomarkers and the inflammatory status of rheumatoid arthritis. It was identified as superior alternatives during study findings, which may add to innovative ideas for the diagnosis and therapy of RA.

STEPS:-

- **NGS Sample data and Preprocessing:-** Four steps were used to process the data**:** i) The Gene Expression Omnibus database's three probe expression matrix files (series matrix.txt) were regulated and log2 converted first. ii) The platform annotation files were compared with each probe expression matrix, and specifically annotated probes were used to process the given data. Basic expression value for many probes were corresponded to one gene were forwarded for analysis. iii) We combined the expression matrices of GSE55235 and GSE55457 and categorised the samples into datasets. iv) To remove heterogeneity generated by various experimental batches and platforms, the R-package sva was installed from Bio-conductor (https://bioconductor.org/).
- **Differential expression gene (DEG) analysis:-** Based on comparing expression levels between HC and RA samples, the limma tool was used to screen identify differentially expressed genes (DEGs). The adjusted value is 0.05, and the DEGs screening criteria are log2 fold change (FC) greater than 2 or less than -2. The study result of DEGs were described in the form of a heatmap and a volcanic map created in the R Studio software (version: 1.2.1335).

- **Functional Annotations:-** The bio-conductor package clusterProfiler was applied to carry out Gene Ontology (GO) and KEGG(Kyoto Encyclopedia of Genes and Genomes) pathway analysis for DEGs. GO keywords and signal pathways with high improvement were presented using a cutoff value of 0.05. The GSEA software was co-developed by the University of California San Diego and the Broad Institute, which decides that statistic signifies differences between two genetic phenotypes can be founded by using a predetermined gene list (e.g., HC and RA). With the help of given parameters, gene symbol and morphological information from the expression dataset were submitted to GSEA for enrichment analysis. Hallmark gene sets (h.all.v7.1.symbols.gmt) were retrieved from the Molecular Signatures Database for the current research (MSigDB). Statistical significance was defined as a nominal P-value of less than 0.01 and an FDR of less than

0.25.

- **Analysis of Immune Susceptibility:-** The approach for identifying the cell percentage of complex tissues based on gene expression patterns is large-scale analysis of RNA mixes using the online analytic tool CIBERSORT (https://cibersort.stanford.edu/), which outperforms other methods. In this study, we used CIBERSORT to distinguish the inflammatory state of RA from healthy joint tissue using the default signature gene file (22 types of immune cells). The outcome of immune susceptibility analysis was processed according to p value <0.05, and the resistant cell composition of each sample was described in bar-plot.
- **Network based Analysis:-** DEGs were made available on the STRING website (version:11.0) for next research to investigate the mutual interaction between proteins encoded by various genes. Because the lowest interaction score had to be larger than 0.4, the network's far-flung branches were removed. We then saved the analysis findings to a TSV file and utilised Cytoscape software (version:3.7.1) to analyse the modules and summarise the procedure. The MCODE is a Cytoscape App Store plug-in that uses topology to locate closely linked nodes in a complicated network. Additionally, using default values, we used this plug-in to discover key modules in the PPI network.
- **Validation of RA-related Hub Genes:** - The MCODE plug-in determined the most important module, from which candidate hub genes were chosen. The pROC programme in RStudio was used to perform receiver operating characteristic (ROC) curve analysis to identify the role of possible genes in RA diagnosis. RA hub genes were identified as those having an area under the curve (AUC) greater than 0.8 and a value less than 0.05.
  Chemokines and immune cell infiltration were identified as overly important variables in the onset of RA in the majority of the patients using a combination of bioinformatical techniques and methodologies. However, eight hub genes have been identified as possible treatment candidates for RA, and additional research is needed to back up our findings.

7. CONCLUSION

In this review, we studied different ideas or approaches of software for rheumatoid arthritis. Rheumatoid arthritis is generally weakening, continuous inflammatory disease and able to originate joint injury and persistent disablement. Now we have a greater knowledge of disease mechanism in molecular medicine, which can help us to design more efficient medicines. Old treatment modalities have been enhanced and new ones have been made, molecular docking tools are one of them at current date, and the above softwares are working at a wide range with advanced results.

8. ACKNOWLEDGEMENT

There is no funding for Research review.

9. CONFLICT OF INTEREST

The authors declare no conflict of interest.

10. AUTHOR'S CONTRIBUTION

PB, HK, SGR, PP, SN, KSM, SM, AKM, ASK, SK, SR did intensive research on various topics developed; PB, HK, KSM contributed in writing the manuscript; Scientist MN designed and supervised the present review article and assisted in writing the paper.

REFERENCES

[1] Korczowska Z. Rheumatoid arthritis susceptibility genes: An overview. World J Orthop. 2014; 5(4): 544-549.
[2] Prescott NJ, Fisher SA, Onnie C et.al. A general autoimmunity gene (*PTPN22*) is not associated with inflammatory bowel disease in a British population 2005;66(4):318-20.
[3] Venkatachalam CM, Jiang X, Oldfield T et.al. Ligand Fit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model. 2003;21(4):289–307.
[4] Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins. 2002; 46(1):34–40.
[5] Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol. 1996;261(3):470-89.
[6] Ajay N. Jain. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J. Med. Chem. 2003;46(4):499–511.
[7] Jones G , Willett P , Glen RC , Leach AR , Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267(3):727–748.
[8] Schapira M, Abagyan R, Totrov M. Nuclear hormone receptor targeted virtual screening. J Med Chem. 2003;46(14):3045–59.
[9] Richard A. Friesner, Jay L. Banks, Robert B. Murphy et.al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 2004;47(7):1739–1749.
[10] Venkatachalam CM, Jiang X, Oldfield T, Waldman M. Ligand Fit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model. 2003;21(4):289–307.
[11] McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Gaussian docking functions. Biopolymers. 2003;68(1):76–90.

[12] Corbeil CR, Williams CI, Labute P. Variability in docking success rates due to dataset preparation. J Comput Aided Mol Des. 2012; 26(6):775–86.
[13] Zhao H, Caflisch A. Discovery of ZAP70 inhibitors by high throughput docking into a conformation of its kinase domain generated by molecular dynamics. Bioorg Med Chem Lett. 2013;23(20):5721–6.
[14] Arthur OT, Olson J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–461.
[15] Carmona SR, Garcia DA et.al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. Plos Comput Biol. 2014.
[16] Allen WJ, Balius TE, Mukherjee S, Brozell SR, et.al. DOCK 6: impact of new features and current docking performance. J Comput Chem. 2015;36(15):1132–1156.
[17] Friesner RA, Banks JL, Murphy RB et.al. A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 2004;47(7): 1739–1749.
[18] Rohini K et al. Journal of Pharmacy Research 2011;4(8):2712-2714.
[19] Esther MYJ, Subramaniyan V et.al. Molecular docking, ADMET analysis and dynamics approach to potent natural inhibitors against sex hormone binding globulin in male infertility. Pharmacogn J. 2017;9(6s): 35–43.
[20] Wein H. Protein Implicated in Rheumatoid Arthritis. NIH Research Matter 2007. https://www.nih.gov/news-events/nih-research-matters/protein-implicated-rheumatoid-arthritis.
[21] Shadick NA, MD, MPH; Cook NR, ScD; Karlson EW, MD, et al Paul C-Reactive Protein in the Predication of Rheumatoid Arthritis in Women. Arch Intern Med. 2006;166(22):2490-2494.
[22] Leggett RM, Ramiez RH, Gonzalez et.al. Sequencing quality assessment tools to enable data-driven informatics for high throughput. Genomics Front Genet. 2013;4:288.
[23] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15): 2114–2120.
[24] Simpson JT, Wong K, Jackman SD et.al. AbySS: A parallel assembler for short read sequnece data. Genome Res. 2009;19(6):1117-1123.
[25] Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Current Protocols. 2020.
[26] Kolmogorov M, Bickhart DM, Behsaz B et.al. metaFlye: scalable long read metagenome assembly using repeat graphs. Nature Methods. 2020;17:1103-1110.
[27] Tarailo GM, Chen N. Using RepeatMasker to identify reptitive elements in genomics sequences. Current Protocols in Bioinformatics. 2009;4:4.
[28] Felipe A. Simao, Robert M. Waterhouse et.al, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19): 3210-3212.
[29] Weizhong Li, Adam G. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-1659.
[30] Shiyuyun T, Alexandre L, Mark B. Identification of protein coding regions in RNA transcripts. Nucleic Acids Research. 2015;43(12): e78.
[31] Alexander D, Carrie A. Davis et.al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1): 15-21.
[32] Heng Li and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760.
[33] Simon A, Paul TP, and Wolfgang H. HtSeq a Pyhton framework to work with high throughput sequnecing data. Bioinformatics. 2015;31(2): 166-169.
[34] Bo Li and Coiln ND (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).
[35] Sonia T, Pedro FT, David T, Antonio Di P, et.al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43(21):e140.
[36] Mark DR, McCarthy DJ, and Gordon KS. edgeR: a Bioconductor package for differntial expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.
[37] Ana C, Matia JN, Alberto F, Manuel T. maSigPro: a method to identify significantly differntial expression profiles in time course microarray experiments. Bioinformatics. 2006;22(9):1096-1102.
[38] Derrick EW, Jennifer L and Ben L. Improved metagenomic analysis with Kraken2. Genome Biology. 2019; 20(257).
[39] Sergey N, Dmitry M, Anton K and Pavel AP. metaSPAdes: a new versatile metagenomic assembler. Genome Research. 2017;27: 824-834.
[40] Dinghua L, Chi-Man L, Ruibang L, et.al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijin* graph. Bioinformatics. 2015;31(10): 1674-1676.
[41] Mina R, Haixu T, Yuzhen Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20): e191.
[42] Carlos PC, Ana HP, et.al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution. 2021;38(12): 5825-5829.
[43] Robert DF, Jaina M, John T, Penny C, et.al. The Pfam protein families database Nucleic Acids Res. 2010;38(database issue): D211-D222.
[44] Ryuichiro N, Toyonori S. Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods. 2021;187:44-53.

[45] Jason B, Beijing W, Howard C, and William G. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015,109:21.29.1-21.29.9.
[46] McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977;267(5612):585–590.
[47] Case DA, Darden TA, Cheatham TEI, et al. AMBER 12. San Francisco, CA: University of California; 2012.
[48] Brooks BR, MacKerell AD, et.al. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545–1614.
[49] Berk H, Kutzner C, David VS and Lindahl E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory. Comput. 2008;4(3):435–447.
[50] Mark TN, William H, Attila G et.al. NAMD: a parallel, object oriented molecular dynamics program. Int J Supercomput Appl High Perform Comput. 1996;10(4):251–268.
[51] Paul S, Andrew M, Owen O, Nitin SB et.al. Cytoscape: A Software Enivronment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13(11): 2498-2504.
[52] Gary DB and Christopher WV Houge. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4(2).
[53] Damian S, Annika LG et.al, The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research. 2021;49(D1): D605-D612.