Indian Journal of Science and Technology, *Vol 9(32), DOI: 110.17485/ijst/2016/v9i32/100745, August 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Computational Tools; Indispensable Armamentarium of Medical Biotechnology

Nutan, Swapnil Jain, Shilpa, Akanksha Tomar, Harish Changotra and Jitendraa Vashistt*

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan - 173234, Himachal Pradesh, India; nutanthakur11@gmail.com, jain.swapnil19@gmail.com, shilpa.anku1994@gmail.com, akankshatmr0@gmail.com, harish.changotra@juit.ac.in, jvashist@gmail.com

Abstract

Objective: Computational infrastructure of medical biotechnology provides insight into elements of genomics, proteomics for understanding diseases and biological systems in a more comprehensive and systematic way. The present study exclusively focuses on role of computational and bioinformatics tools employed in troubleshooting heterogeneous data of biology origin. **Method and Analysis:** Methodology adopted for study included analysis of online literature through Google Scholar, Pub-med, and Science Direct by searching with comprehensive input like computational tools, biotechnology, proteomics, genomics, drug design and discovery, metagenomics to include all relevant research/review articles within period from 1980 to 2016. Collection, inclusion and exclusion of published articles, review and reports were critically assessed and discussed within line of available databases. Furthermore, bibliography of relevant research articles was taken into consideration. **Findings:** Major conventional biological databases i.e., National Centre for Biotechnology Information (NCBI) and its subset PubMed, not only provide and exchange biological sequence information but also allow comprehensive analysis for sequence alignment and comparisons to find out disease biomarkers like Single Nucleotide Polymorphisms (SNPs). Likewise, other computational tools and servers are developed to be specific for organism, disease, biological pathway, microbial resistance and drug design. Use of currently available computational techniques provides rapid cross-reference search with higher accession of the sequences with statistical approach. However, a significant challenge in this field is to organize and provide user access, readability and skill sets to heterogeneous data repositories like databases and web servers as well as to keep data privacy. These temporal processes significantly enhanced the vision and understanding about cryptic biological processes. **Novelty/ Improvement:** Present study exclusively targets computational tools employed in deciphering complex biological systems through use of algorithms and software which made available large data repositories in public domain. Current study will enhance and compile amalgam of computational and bioinformatics pipelines engaged in vast perspectives of basic and applied biology.

Keywords: Biotechnology, Bioinformatics, Computational Tools, Genomics, Proteomics, Web-Servers

1. Introduction

Computational tools are invaluable for researchers and scientists indulged in various fields across the globe for data extraction, analysis and integration of complex heterogeneous data sets. Amalgam of classical biological techniques with high throughput computational technologies leads to beginning of a new stratum for precise understanding and answers of complex biological problems. In this regard, advances in computational biology and bioinformatics have made access to vast genome, transcriptome and proteome data which are being utilized for a variety of biomarkeridentification, nano-device engineering and drug design etc[1,2]. In the present paper we are discussing the usage of different computational approaches which render the information of biological processes systematic and more useful.

*Author for correspondence

2. Computational Platforms for Genomics and Proteomics

Complete set of genes and protein complements can be analyzed through genomics and proteomics, respectively. However, in both the techniques a statistical data analysis is needed to include relevant objects in the research study and to exclude various nonessential or impractical parameters for further experimental investigation. This objective refinement is well accomplished through use of various web servers and databaseswhich provide expression profile data, functional annotation, biomedical literatureand the sequence information search[3,4] e.g., screening of various potential drug resistant pathogen proteins through web-servers like MDRIpred[5], UniDrug-Target[6] provide outputs of specific and reproducible results. These output results also shorter the time of experimentations of the researchers. Another database; SNPedia (http://www.snpedia.com) provides DNA variations of human genome correlated with medical literature available. In 2006, FP6 STEP project launched with an initial vision of European funding under Framework 7 programmed to provide free access to model repositories of particularly clinical data by VPH Institute, a nonprofit umbrella organization[7]. Future perspective of such databases will facilitate unique drug target identification and thus to avoid antibiotic toxicity in humans, spread of resistant phenotype among microbial world Figure 1. Identification of biomolecules like protein through computational tools offers a great hope but also suffers with problem of false-positive identifications by techniques viz. Tandem mass spectrometry (MS/MS) data generated in shotgun proteomic experiments. Nesvizhskii and colleagues[8] addressed these hurdles one by one but of these most important is choice of protein sequence databases like Entrez (from NCBI), its subset databases RefSeq, and UniProt (Swiss-Prot and TREMBL) and International Protein Index (IPI) from EBI. However search against larger database such as Entrez Protein is suitable for sequence polymorphism studies but it lowers sensitivity of peptide identification by introducing more false identifications. Post-Translational Modifications (PTMs) in protein are directly indicated in peptide mass spectrometry by factors like neutral losses and modification-specific mass increments. However, database search engines become redundant on exhaustive search for multiple PTMs and flipped cleavage sites. An algorithm VEMS v3.0 (Virtual Expert Mass Spectrometrist) which utilizes novel elements such as automatic re-calibration of fragment ions during the search, unique scoring function taking into account recalibration of mass enhances Q-TOF MS/MS spectra analysis[9]. Biomarker identification in fields of proteomics and genomics is augmented by Support Vector Machines (SVM) which firstly learned to process set of elements belonging to various clusters afterwards gives output based on its own intelligence[10]. Current bioinformatics and computational methodologies are well utilized in recognition of cis-acting regulatory elements involved in transcription process and databases like AGRIS AtcisDB[11–13] (http://arabidopsis.med.ohio-state.edu/AtcisDB/) and CTCFBSDB and CTCFBSDB 2.0[14,15] can be used for data extraction and analysis of plant and vertebrate genome. Advent of bio-computational tools together with Genome-Wide Association Studies (GWAS) generated huge networks of metabolic pathways and accumulation of data sets of genes and their interactions for different phenotypes. Disease and Gene Annotation (DGA) database extract information from Disease Ontology (DO) and NCBI to construct disease molecular interaction networks[16]. Likewise, artificial intelligence and neural network approaches simulate human intelligence and learning processes and congregate scalability to locate random gene positions over large genome size[17,18].

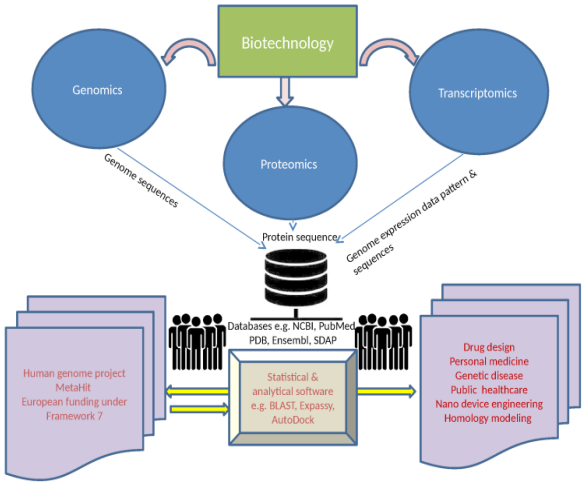

Figure 1. Role of computational biology is illustrated via various visionary elements like database/servers and software into public domain for the research purpose into biomedical science, healthcare service and bioengineering.

3. Web based Tools for Personalized Medicine and Drug Repositioning

The major research thrust into disease and drug came through elucidation of human genome project. It was initiated in 1990 to provide human genome sequence data to entire biomedical community, since then million of sequences were generated of which 30 millions entries correspond to human origin. Human genome project sequence information is deposited in public domain repositories which are national European Molecular Biology Laboratory (EMBL)/European Bioinformatics Institute[19] (EBI), the National Center for Biotechnology Information (NCBI, GenBank) database[20] and the DNA Database of Japan (DDBJ)[21]. User friendly access to these sequences is provided by different browsers like Entrez Gene browser[22], the UCSC genome browser[23] and the EBI/Ensemble browser. Other most important aspects of medical biology are drug toxicity to human host and alarming microbial resistance. These two issues are the most common and challenging aspects inpharmacy and therapeutic sectors. Two approaches i.e., drug based and disease based are proposed to eradicate problem. Idea of personalized medicine and drug repositioning could be possible on success of Connectivity Map (CMap) project and its extended project Library of Integrated Network-Based Cellular Signatures (LINCS)[24,25] and data extracted from functional genomics databases such as NCBI Gene Expression Omnibus (GEO). The project is aimed at large-scale gene expression profiles correlating with genome aberration, susceptibility to disease and drug targets. In disease based approach Drug versus Disease (DvD), the database for Annotation, Visualization and Integrated Discovery (DAVID)[26] and the Gene Set Enrichment Analysis (GSEA) provides huge amount of genome expression profiles to identify signature disease and drug targets. In another case, drug target interaction analysis is performed by integrated network system approach using DReSMin, an algorithm for mining semantically-rich networks for occurrences of a given semantic subgraph[27]. Personalized medicine and drug repositioning are very much important fields in anticancer treatment and therapy. Algorithms like basic local alignment search tool (http://www.ncbi.nlm.nih.gov/BLAST), ClustalW and FASTA are most widely used to compare sequences from genome[28–32]. Moreover NCBI dbSNP database, GENESCAN or AUGUSTUS are much exploited tools used for detection of single nucleotide polymorphism and gene prediction in areas of genetics and disease.

4. Screening of Pathogens and Diagnostic System

Genetic identification and characterization of circulating pathogens is necessary for suitable vaccine target. One such approach utilizes RotaC, web tool for genotyping Rota virus which is the leading cause of infant's morbidity and mortality due to diarrhoea[33]. Web tool interface is written in perl while classification script utilizes Java and is able to analyze 1000 nucleotide sequences at a time in FASTA format. Structural Database of Allergenic Proteins (SDAP) (http://fermi.utmb.edu/SDAP) is a web server written on CGI scripts and implemented with MySQL under Linux, provides information regarding IgE epitopes interaction with user supplied allergen peptide[34]. Based up on polarization anisotropy diagnostics technique a new smartphone application has been designed to identify causative agents of nosocomial infections. Firstly, polymerase chain reaction is used to amplify target bacterial RNA and amplified product is used as input for prototype PAD system consisting of optical tubes. MIT App Inventor 2 controlled softwareuses optical sensors to recognize signature bacterial sequences by comparing with 16s rRNA sequences[35]. Cancer bioinformatics grid CaGrid (http://www.cagrid.org/) and EGI (European Grid Infrastructure) (http://www.egi.eu/) grids serve to exchange data on cancer research in public domain. For avoiding big and time expensive problems GPUs clusters presents an effective way[36].

5. Metagenomics and Next Generation Sequencing

Field of metagenomics focuses on characterization of total microbial community, primarily uncultured microbes in a community[37]. The metagenomics RAST server provides user access to genomic data via in-built controlled manner to ensure data privacy. Taxonomic level classification of different lineages can be achieved by computer program like DOTUR[38]. DOTUR assigns diverse outer taxonomical units by rarefaction and collector's curve analysis. For, meta-transcriptomic shotgun sequencing rRNASelector (http://sw.ezbiocloud.net/rrnaselector) is

a computer program which uses Hidden Markov models for identifying prokaryotic 5S, 26S, and 23S rRNA genes on roche 454 platform[39].

6. Concluding Remarks

New approaches combining bioinformatics and computational tools demonstrate a great contribution and promise in functional genomics, genetics, disease and diagnostics. High-throughput, low-cost and high-performance computingis the necessity of current research because enormous data which is generated per day requires user friendly and systematic workflow in public domain.

7. References

1. Aksimentiev A, Brunner R, Cohen J, Comer J, Cruz-Chu E, Hardy D, Rajan A, Shih A, Sigalov G, Yin Y, Schulten K. Computer modeling in biotechnology: A partner in development. Methods Molecular Biology. 2008; 474:181–234.
2. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. Bio Medical Centre Bioinformatics. 2008 Sep; 9:386.
3. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: A new gateway to drug repositioning? Drug Discovery Today. 2013 Apr; 18(7):350–7.
4. Pacini C, Iorio F, Goncalves E, Iskar M, Klabunde T, Bork P, Saez-Rodriguez J. DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. Bioinformatics. 2013 Jan; 29(1):132–4.
5. Singla D, Tewari R, Kumar A, Raghava GP. Designing of inhibitors against drug tolerant Mycobacterium tuberculosis (H37Rv).Chemistry Central Journal. 2013 Mar; 7(1):49.
6. Chanumolu SK, Rout C, Chauhan RS. UniDrug-target: A computational tool to identify unique drug targets in pathogenic bacteria. PloS one. 2012 Mar; 7(3):e32833.
7. Hunter P, Chapman T, Coveney PV, De Bono B, Diaz V, Fenner J, Frangi AF, Harris P, Hose R, Kohl P, Lawford P. A vision and strategy for the virtual physiological human: 2012 update. Interface Focus. 2013 Apr; 3(2):1–10.
8. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of proteomics. 2010 Oct; 73(11):2092–123.
9. Matthiesen R, Trelle MB, Hojrup P, Bunkenborg J, Jensen ON. VEMS 3.0: Algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. Journal of proteome research. 2005 Dec; 4(6):2338–47.
10. Pang S, Havukkala I, Kasabov N. Two-class SVM trees (2-SVMT) for biomarker data analysis. InInternational Symposium on Neural Networks; 2006 May. p. 629–34.
11. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E. AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics. 2003 Jun; 4(1):25.
12. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. AGRIS and AtRegNet, a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant physiology. 2006 Mar; 140(3):818–29.
13. Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E. AGRIS: The Arabidopsis Gene Regulatory Information Server, an update. Nucleic acids research. 2011 Jan; 39(S1):D1118–22.
14. Bao L, Zhou M, Cui Y. CTCFBSDB: A CTCF-binding site database for characterization of vertebrate genomic insulators. Nucleic Acids Research. 2008 Jan; 36(S1):D83–7.
15. Ziebarth JD, Bhattacharya A, Cui Y. CTCFBSDB 2.0: A database for CTCF-binding sites and genome organization. Nucleic Acids Research. 2013 Jan; 41(Database issue):D188–94.
16. Peng K, Xu W, Zheng J, Huang K, Wang H, Tong J, Lin Z, Liu J, Cheng W, Fu D, Du P. The Disease and Gene Annotations (DGA): An annotation resource for human disease. Nucleic Acids Research. 2012 Nov; 41(Database issue):D556–60.
17. Syarifahadilah MY, Abdullah R, Venkat I. ABC algorithm as feature selection for biomarker discovery in mass spectrometry analysis. 2012 4th Conference on Data Mining and Optimization (DMO); 2012 Sep. p. 67–72.
18. Azuaje F. A computational neural approach to support the discovery of gene function and classes of cancer. IEEE Transactions on Biomedical Engineering. 2001 Mar; 48(3):332–9.
19. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V. Ensembl 2005. Nucleic Acids Research. 2005 Jan; 33(Database issue):D447–53.
20. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY. Database resources of the national center for biotechnology information. Nucleic Acids Research. 2007 Jan; 35(S1):D5–12.
21. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Research. 2002 Jan; 30(1):27–30.

22. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: Gene-centered information at NCBI. Nucleic Acids Research. 2005 Jan; 33(S1):D54–8.
23. Kent WJ, Sugnet CW, Furey T, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Research. 2002 Jun; 12(6):996–1006.
24. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006 Sep; 313(5795):1929–35.
25. Vidovic D, Koleti A, Schurer SC. Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. Comprehensive Systems Biomedicine. 2014 Sep; 5:342.
26. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology. 2003 Aug; 4(5):P3.
27. Mullen J, Cockell SJ, Tipney H, Woollard PM, Wipat A. Mining integrated semantic networks for drug repositioning opportunities. Peer J. 2016 Jan; 4:e1558.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990 Oct; 215(3):403–10.
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997 Sep; 25(17):3389–402.
30. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences. 1988 Apr; 85(8):2444–8.
31. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Bioinformatics Methods and Protocols. 1999; 185–219.
32. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994 Nov; 22(22):4673–80.
33. Maes P, Matthijnssens J, Rahman M, Van Ranst M. RotaC: A web-based tool for the complete genome classification of group a rotaviruses. Bio Medical Centre Microbiology. 2009 Nov; 9(1):238.
34. Ivanciuc O, Schein CH, Braun W. SDAP: Database and computational tools for allergenic proteins. Nucleic acids research. 2003 Jan; 31(1):359–62.
35. Park KS, Huang CH, Lee K, Yoo YE, Castro CM, Weissleder R, Lee H. Rapid identification of health care–associated infections with an integrated fluorescence anisotropy system. Science Advances. 2016 May; 2(5):1–10.
36. Johnson D, Shafer B, Lee JJ, Chen JY. Multi-biomarker panel selection on a GPU. IEEE International Conference on Electro/Information Technology (EIT); Beijing. 2012 May. p. 1–6.
37. Handelsman J. Metagenomics: Application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews. 2004 Dec; 68(4):669–85.
38. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology. 2005 Mar; 71(3):1501–6.
39. Lee JH, Yi H, Chun J. rRNA Selector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. The Journal of Microbiology. 2011 Aug; 49(4):689–91.