Journal of
Biomedical Research and Clinical Case Reports

Review Open Access

Bioinformatics Tools to Make Possible to Use High-Throughput Data in Clinical Activities

Giuseppe Agapito[*]

[1]Department of Medical and Surgical Science, University Magna Græcia of Catanzaro, Catanzaro, Italy

[*]**Corresponding author:** Giuseppe Agapito, Department of Medical and Surgical Science, University Magna Græcia of Catanzaro, Catanzaro, Italy; E-mail: agapito@unicz.it

Received Date: January 20, 2018 **Accepted Date:** February 15, 2018 **Published Date:** February 17, 2018

Citation: Giuseppe Agapito (2018) Bioinformatics Tools to Make Possible to Use High-Throughput Data in Clinical Activitiess. J Biomed Res Clin Case Rep 1: 1-5.

Abstract

High-Throughput methodologies like Next Generation Sequencing (NGS), Genome Wide Association Study (GWAS), gene expression and Single Nucleotide Polymorphism - SNP microarrays, and Mass-Spectrometry are moving traditional science like medicine and clinical activities toward a data driven science, because these methodologies are able to produce lots amount of data for each single experiment. Therefore, to efficiently deal with these vast amounts of data produce daily using high-throughput methodologies, the development of efficient and scalable software tools able to analyze, store, visualize and manage sensible data (i.e., personal information, therapies and so on) arise. Thus, software tools could simplify the use of high-throughput data in clinical practice, since the models obtained by using software tools can be used easily in clinical practice.

Introduction

Nowadays new experimental methodologies like Next Generation Sequencing (NGS), Genome Wide Association Study (GWAS), gene expression and Single Nucleotide Polymorphism - SNP microarrays, and Mass-Spectrometry are moving traditional science like medicine and clinical activities toward a data driven science. These methodologies known as High-Throughput methodologies are able to produce lots amount of data for each single experiment.

NGS is a DNA sequencing methodology able to test in parallel multiple small fragments of DNA to determine a sequence, increasing the speed with which an individual's genome can be sequenced. Thus, the information produced by NGS can be used by clinicians to improve diagnostic and treatment decisions as stated in [23]. GWAS is a high-throughput genotyping methodology to study genetic variants in the form of SNPs. The genetic variants found can be related to the development of clinical and quantitative traits. For example, in [4] the authors describe the application of GWAS in the eyes' disease field.

Microarray can simultaneously analyze thousands of genes in parallel to identify new biomarkers, monitoring the drug adverse reaction of a drug, spurring light on the mechanisms involved in the progression of disease or responsible of adverse drug reaction, as explained in [7].

Mass Spectroscopy was employed in the measurement of the mass of several molecules within a sample i.e. the identification of large molecules like proteins. But the use for mass spectrometry today is extended to pharmacokinetic and drug metabolism studies, as well as it has also found application in medical practice, such as blood analysis, and drug trials and neonatal screening.

To make the data produced by using high-throughput methodologies more useful for researcher and clinic doctor is necessary to extract knowledge buried in the data. At this regard, to handle with high-throughput data is mandatory to develop efficient software able to preprocess the raw data. For example, after the hybridization process in microarray, it is necessary to read microarray by using a beam laser scanner obtaining optical images. Subsequently, the boundaries of each spot in the image must be identified, through software quantifies able to locate the signal intensity at each spot. Mathematical filters can be applied during quantification, to exclude pixels with very low or very high intensities, because these are usually due to background fluorescence or dust, introducing bias in the next steps of analysis.

These analyses are carried out by using proprietary software made available by the microarrays' vendor. Another example is the data produced by using NGS. NGS data present several sequence artifacts, including read errors (insertions/deletions), poor quality reads and primer/adaptor contamination, which can impose a significant negative impact on the downstream sequence processing/analysis. The quality of data is crucial for various downstream analyses, such as sequence assembly, single nucleotide polymorphisms identification. Thus, several tools for filtering, trimming, of NGS data have been developed [19]. These steps are known as data cleaning, data consolidation and belong to a general data analysis known as data pre-processing. Data pre-processing makes data in a format more suitable for the actionable knowledge extraction. For example, starting from a list of SNPs obtained by using GWAS, it is necessary to develop software methodologies able to discover SNPs related with the cancer growth by integrating GWAS data and Protein Protein Interaction (PPI) data and biological Pathways data. In the same time the problem of data integration has to be handled because, PPI data, Pathways data and SNP data are in different format, requiring a lot of effort to make data integration possible, thus the developing of specific algorithms arise. Moreover, to make more effective and useful the knowledge extracted from high-throughput data, data visualization is a key stone. In fact, presenting results in a graphical way can simplify and speed-up the understanding of the problem under investigation. Visual representation is more effective than textual representation. Finally, high-throughput data should be stored in order to be available for the downstream analysis. Data have to be stored in efficient and scalable databases, by developing customized data formats able to speed-up the data retrieving and indexing, allowing to more users at the same time to get access to the data. Thus, several public and private databases are available on the web, that makes data available for the scientific community.

The rest of the paper is arranged as follows: in Section 2 the list of well-known databases is reported, Sections 3 presents the list of some very known and used software tools. Finally, Section 4 concludes the paper, highlighting the benefits of using bioinformatics tools in a clinical scenario.

Omics Databases

In this Section are presented the most known and used databases containing omics data. Databases are a decentralized data source gathering data submission from the community.

- dbGAP database of genotypes and phenotypes Data Browser [27,14,31] is available at the following we address: https://www.ncbi.nlm.nih.gov/gap/ddb/ To get access to the data it is mandatory create an account by following the instruction provide on the web site. dbGAP provide to researcher a fast access to a collection of genotype and phenotype data, together with a summary information simplifying the data exploration
- AtPID -Arabidopsis thaliana Protein Interactome Database - contains information about to PPI networks, domain architecture, orthologue information and GO annotation in the Arabidopsis thaliana proteome [13,6,12]. AtPIDis a database free for Academic or non-commercial purpose at the following web address http://www.megabionet.org/atpid/webfile AtPID, contains 5564 mutants where, 167 mutations are manually curated, and predicts 4457 high confidence gene-PO pairs with 1369 genes as the complement.
- GAD -Genetic Association Database- contains standardized genetic association study data [1]. GAD is a public repository without any restriction of use of genetic association studies, containing more than 5,000 human genetic association studies. GAD goal is to facilitate the studying of complex common human genetic disease. All datasets can be downloaded from the website https://geneticassociationdb.nih.gov

- SNPChip is a multi-species SNP-chip database [17, 18]. SNPChip provide the following function to simply the work of researcher: i) referencing the SNP mapping information from genome assembly, ii) retrieving of information from dbSNP and in commercially available bovine chips, and iii) identification of SNPs in common between more bovine chips. iv) link the information from the array vendor with data available in public databases. SNPChip is available at http://www.ncbi.nlm. nih.gov/snp without any restriction of use.
- TMC-SNPdb is the first open source, flexible, upgradable, and freely available for academic or non-commercial use single nucleotide polymorphism (SNP) database. TMC-SNPdb contains normal germline variants derived from Indian (non-European Caucasian population) [28]. TMC-SNPdb is available for download at the following web url:
http://www.actrec.gov.in/pi-webpages/AmitDutt/TMCSNP/TMCSNPdp.html
- dbSNP Single Nucleotide Polymorphism database [24] is a database of genetic plymorphisms. dbSNP includes single nucleotide polymorphisms - SNPs, deletion insertion polymorphisms or DIPs, and short tandem repeats or STRs. dbSNP is freely available at the follows web address:
https://www.ncbi.nlm.nih.gov/projects/SNP

- dbSAP Single amino-acid polymorphism database for protein variation detection [3]. dbSAP is freely available online at the following
http://www.megabionet.org/dbSAP/ dbSAP is a database containing human protein variations inferred from SNPs and genomic mutations. At the time of writing dbSAP contains a total of 16,854 SAP peptides involving in 439,537 spectra from various human tissues and cell lines.

- SILVA (Improved data processing and web-based tools) [20]. SILVA is a collection of databases of aligned ribosomal RNA (rRNA) gene sequences from the Bacteria, Archaea and Eukaryota species. SILVA contains more than 3, 000, 000 small subunits and more than 200,000 large subunits of rRNA gene sequences. SILVA provides a browser to explorer the database contents in a hierarchical view. SILVA is freely available online at the following address http://www.arb-silva.de.

- BioSD The BioSample Database at the European Bioinformatics Institute [8,10]. BioSD is a database at EBI contains information regarding biological samples used in molecular experiments, i.e., sequencing, gene expression or proteomics. BioSample provides storing and combining to store the data sample with information within EBI databases. BioSD minimizes data entry submitting in EBI database. Moreover, BioSD supports cross-database queries by sample features. BioSD includes a growing set of reference samples, i.e., cell lines, which are frequently used in experiments and easily accessed by their accession numbers. BioSD is freely available without restriction of use at the following address http://www.ebi.ac.uk/biosamples

- PRIDE the proteomics identifications database [15,29]. PRIDE is a public database of proteomics data, including protein and peptide identifications, post-translational modifications and spectral evidence. PRIDE belongs to the ProteomeXchange (PX) consortium, providing an entry point to submit mass spectrometry data in public-domain repositories. Submitted data are handled by expert bio-curators. PRIDE is available for free at http://www.ebi.ac.uk/pride
- GPMDB is a database of proteomics experimental information [9,5]. GPMDB is based on a combination of data analysis servers, a user interface, and a relational database, and all the system is made avail-able as open source development projects at the following web site: http://www.thegpm.org/gpmdb/index.html

- LMSD (LIPID MAPS structure database) [26] is a collection of structures and annotations of biologically relevant lipids. LMSD provides to the users several features: i) hierarchical classification and consistent nomenclature based on a classification scheme defined by LIPID MAPS, ii) a LIPID MAPS identifier to all structures without no duplicate structures and finally, iv) the capability to search entry by using structure. LMSD is freely available at the following address: www.lipidmaps.org/data/structure/

A more complete list of omics databases can be found in the web portal OMICX tool at the following web address: https://omictools.com

Omics Software Tools

In this Section are listed some the most known and used software to deal with omics data. Software tools allow researcher to speed-up the analysis and the understanding of complex diseases such as extract new biomarker, or highlight unknown gene interactions and so on.

- Trimmomatic [2] is a Java, multithreaded command line tool to trim and crop Illumina FASTQ data format, as well as to remove adapters, that represent a problem on the downstream application. Trimmomatic is available for Unix/Linux, Mac OS, and Windows operating systems under GNU General Public License version 3.0, it requires Java 1.5 or above, and it is available at http://www.usadellab.org/cms/index.php?page=trimmomatic

Trimmomatic provides a collection of useful trimming tools for illumina paired-end and single ended data. The set-up of trimming tasks is obtained suppling values by the command line.

- SNPTools [25] is a framework of software tools that enables integrative SNP analysis coming from next generation sequencing from large number of samples. SNPTools is wrote in C++ and command are supplied through command line interface. SNPTools is available only for Unix/Linux operating system under BSD 2-clause ”Simplified” License, MIT, at http://www.birc.au.dk/snptools along with example data sets and tutorials.

- SNPinfo [32] is software web-suite designed to predicted functional SNPs that have differential affect between reference allele and alternative allele, experimental and epidemiological information linked with GWAS data, and linkage disequilibrium (LD) information to prioritize SNPs for further analysis. It is compatible with all operating system, and it is freely accessible through a standard web-browser at http://www.niehs.nih.gov/snpinfo
- SNPchip [22] is an R package that provides a command line interface for storing, visualizing and analyzing high-density SNP data. SNPchip extends the open source R tools available at Bioconductor. The extended version has added methods useful for producing visual and descriptive summaries. In specific, the plotting methods are helpful to identify regions of probable chromosomal anomalies and, views of copy number and genotype data. SNPchip is available for Unix/Linux, Mac OS and, Windows operating system, under the GNU General Public License version 2.0 at the Bioconductor web-page at www.bioconductor.org

- GWASpi GWAS-pipeline is a desktop application for genome-wide SNP analysis and management [16]. GWASpi is freely available on the web at http://www.gwaspi.org GWASpi is implemented in Java, Apache-Derby and NetCDF-3, and it is available for all the major operating systems. GWASpi provide a command line interface with which to deal and analyze GWAS data.
- Postgwas is a comprehensive toolkit for post-processing, visualization and advanced analysis of GWAS results, through command line interface [11]. Postgwas supports virtually all model organisms and represents the first cohesive implementation of such tools for the popular language R. Postgwas package can be found on CRAN for download or direct installation using the ”install.packages()” function in R. Postg was is compatible with Unix/Linux, Mac OS and, Windows operating systems.

- OpenMS is a geometric approach for the alignment of liquid chromatography-mass spectrometry data [21]. OpenMS is wrote in C++ and Python, it is available for Unix/Linux, Mac OS and, Windows operating systems, under BSD 3-clause ”New” or ”Revised” License, GNU Lesser General Public License version 3.0 at the following address http://www.openms.de OpenMS is a library integrated into KNIME, Galaxy and WS-PGRADE, providing a ready-to-use tools for the analysis of both proteomics and non-targeted metabolomics data.

- MSPLIT-DIA Sensitive peptide identification for data-independent acquisition [30]. MSPLIT-DIA is available for Unix/Linux, Mac OS and, Windows operating systems at the following address http://proteomics.ucsd.edu MSPLIT-DIA allows user to easily match spectra to each DIA spectrum. Allowing to evaluate the similarity of the matched peaks between library spectra and multiplexed spectra across multiple consecutive DIA spectra.

A more detailed list of omics software tools can be found in the web portal OMICX tool at the following web address: https://omictools.com/

Conclusions

To efficiently deal with the vast amount of data produce daily using high-throughput methodologies, the development of efficient and scalable software tools able to analyze, store, visualize and manage sensible data (i.e., personal information, therapies and so on) arise. The primary challenge for bioinformaticians and computer scientists is to develop software tools able to handle these vast and heterogeneous amounts of data efficiently, providing to the researchers in life science the instruments with which to simplify their tasks. Efficient and scalable software tools are essential to spur light in complex diseases like cancer, diabetes, and Alzheimer. In fact, complex disorders are characterized by an intricate network of interacting element, thus only using efficient tools able to integrate biological pathways data together with clinical and gene expression data, can provide a better knowledge of the disease under investigation. Therefore, data integration should provide a better view of the diseases impossible to obtain analyzing singularly the same data, as well as the data analysis, should get benefits from useful data visualization. In fact, representing the biological pathways data together with clinical and gene expression data using a network representation, allow to researcher to carry out visual data analysis. Finally, software tools can simplify the use of high-throughput data in clinical practice, since the models obtained by using software tools simplify the data making possible to use the inferred knowledge in clinical practice. The inferred knowledge could avoid to dispense a particular drug in patients affected by a specific SNP in a specific chromosome, knowing that it will not produce benefits. In this way, it is possible to save time to treat the patient only with the right drug avoiding to proceed by attempts, in a field where time is precious, as well as save money avoiding to test several drugs without getting benefits.

References

1) Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nature Genetics 36:431.

2) Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30: 2114-2120.

3) Cao R, Shi Y, Chen S, Ma Y, Chen J, Yang J, Chen G, Shi T (2017) dbsap: single amino-acid polymorphism database for protein variation detection. Nucleic Acids Research 45(Database issue): D827-D832.

4) Chandra A, Mitry D, Wright A, Campbell H, Charteris DG (2014) Genome-wide association studies: applications and insights gained in ophthalmology. Eye 28:1066-1079.

5) Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. Journal of Proteome Research 3:1234-1242.

6) Cui J, Li P, Li G, Xu F, Zhao C, Li Y, Yang Z, Wang G, Yu Q, Li Y, Shi T(2008) Atpid: Arabidopsis thaliana protein interactome database-an integrative platform for plant systems biology. Nucleic Acids Research 36(Database issue): D999-D1008.

7) Deyholos MK, Galbraith DW (2001) High-density microarrays for gene expression analysis. Cytometry 43:229-238

8) Faulconbridge A, Burdett T, Brandizi M, Gostev M, Pereira R, Vasant D, Sarkans U, Brazma A, Parkinson H (2014) Updates to biosamples database at european bioinformatics institute. Nucleic Acids Research 42(Database issue): D50-D52 .

9) Feny¨o D , Beavis RC (2015) The gpmdb rest interface. Bioinformatics 31: 2056-2058.

10) Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, Parkinson H (2012) The biosample database (biosd) at the european bioinformatics institute. Nucleic Acids Re- search 40(Database issue): D64-D70.

11) Hiersche M, Ru¨hle F, Stoll M (2013) Postgwas: Advanced gwas interpretation in r. PLOS ONE 8: e71775.

12) Li P, Zang W, Li Y, Xu F, Wang J, Shi T (2011) Atpid: the overall hierarchical functional protein interaction network interface and analytic platform for arabidopsis. Nucleic Acids Research 39(suppl 1): D1130- D1133.

13) Lv Q, Lan Y, Shi Y, Wang H, Pan X, Li P, Shi T (2017) Atpid: a genome-scale resource for genotype-phenotype associations in ara-bidopsis. Nucleic Acids Research 45(D1):D1060-D1063.

14) Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, et al. (2007) The ncbi dbgap database of genotypes and phenotypes. Nature genetics 39: 1181-1186.

15) Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R (2005) Pride: The pro- teomics identifications database. PROTEOMICS 5: 3537-3545.

16) Mun˜iz-Fernandez F, Carren˜o-Torres A, Morcillo-Suarez C, Navarro A (2011) Genome-wide association studies pipeline (gwaspi): a desk-top application for genome-wide snp analysis and management. Bioinformatics 27: 1871-1872.

17) Nicolazzi EL, Caprera A, Nazzicari N, Cozzi P, Strozzi F, et al. (2015) v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. BMC Genomics 16: 283.

18) Nicolazzi EL, Picciolini M, F.Strozzi F, Schnabel RD, Lawley C, Pirani A, Brew F, Stella A (2014) Snpchimp: a database to disentangle the snpchip jungle in bovine livestock. BMC Genomics 15:123.

19) Patel RK, Jain M (2012) Ngs qc toolkit: A toolkit for quality control of next generation sequencing data. PLOS ONE 7: e30619

20) Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glo¨ckner FO (2013) The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(Database issue): D590-D596.