APPLICATION OF DATA MINING AND INTERPRETATION IN BIOINFORMATICS

Saket Vinayak and Rakesh Ranjan
University Centre of Bioinformatics (Sub-DIC), T. M. Bhagalpur University, Bhagalpur- 812007.

Abstract

The emerging era of “new biology” accompanied by the birth of other branches of sciences, such as bioinformatics and computational biology, which have an integrated interface of molecular biology. Bioinformatics is the combination of biology and information technology which focuses on cellular and molecular levels for application in modern biotechnology. Recently, bioinformatics and genomics have evolved interdependently and promoted a historical impact on the available knowledge. Bioremediation is the recent technology which explores the microbial potentiality for biodegradation of xenobiotics compounds. Microorganisms display a remarkable range of contaminant degradation ability that can efficiently and effectively restore natural environmental conditions. Attempts have been made to interpret some areas of genomics and proteomics which have been employed in bioremediation studies. Bioinformatics requires the study of microbial genomics, proteomics, systems biology, computational biology, phylogenetic trees, data mining and application of major bioinformatics tools for determining the structures and biodegradative pathways of xenobiotic compounds. As discussed bioinformatics is an increasingly data rich industry and thus using data mining techniques helps to propose proactive research within specific fields of the biomedical industry. Additionally this allows for researchers to develop a better understanding of biological mechanisms in order to discover new treatments within healthcare and knowledge of life. In recent years the computational process of discovering predictions, patterns and defining hypothesis from bioinformatics research has vastly grown. This paper highlights the significance of data generation and interpretation in bioinformatics.

Key Words: Bioinformatics, Bioremediation, Computational biology, Databases, Biotechnology.

INTRODUCTION: Biological researches has generated an increasingly large amount of biological data. Drawing conclusions from this data requires sophisticated computational analysis in order to interpret the data. One of the most active areas of inferring structure and principles of biological datasets is the use of data mining to solve biological problems. Some typical examples of biological analysis performed by data mining involve protein structure prediction, gene classification, analysis of mutations in cancer and gene expressions. As biological data and research become ever more vast, it is important that the application of data mining progresses in order to continue the development of an active area of research within bioinformatics (Ritchie et al., 2015). In biology, the main challenge is to make sense of the enormous amount of structural data and sequences that have been generated at multiple levels of biological systems (Pevsner, 2015). Still, in bioinformatics, development of new tools is necessary capable of assisting in understanding the mechanisms underlying biological questions in the study. Bioinformatics has its origin a decade before DNA sequencing became feasible (Hagen, 2000). Historical moments can be highlighted for its development are the publication of the structure of DNA by Watson and Crick in 1953 and knowledge of biochemistry and protein structure with the studies of Pauling, Coren, and Ramachandran in the 1960s (Hunt, 1984,Verli, 2014).

TYPES OF DATABASES: Data mining is the method extracting information for the use of learning patterns and models from large extensive datasets. Data mining itself involves the uses of machine learning, statistics, artificial intelligence, database sets, pattern recognition and visualization. Often referred to as Knowledge Discovery in Databases (KDD) or Intelligent Data Analysis (IDA), the data mining process is not just limited to bioinformatics and is used in many differing industries to provide data intelligence. The application of data mining and machine learning models can involve varied systems. Due to the large volume of data that has been generated, its organization and storage become necessary. Therefore, databases were created, which constitute a large number of biological information stored and processed to allow the scientific community access (Luscombe et al., 2001; Prosdocimi, 2010). The increasing amount of data has been accompanied by an increase in the number of biological databases, whose compilation, updating and dissemination have been carried out. Public databases are classified into primary and secondary databases. The primary databases are composed of results of experimental data that are published without careful analysis related to previous publications. On the other hand, in the secondary databases, there is a compilation and interpretation of data, called content curation process (Prosdocimi, 2010). These databases are members of the International Nucleotide Sequence Database Collaboration (INSDC) and share among each other the deposited information daily (Prosdocimi et al., 2002). According to the latest update, the information sources used by bioinformatics can be divided into

i) Raw DNA sequences,
ii) Protein sequences,
iii) Macromolecular structures,
iv) Genome sequencing, among others.

These databases are curetted and present only information related to proteins, describing aspects of its structure, domains, function, and classification. To standardization, the INSDC adopted some identification systems of the sequences deposited that bring relevant information about the origin and nature of the data (Amaral et al., 2007). Some of these identifiers are the accession number (AN) represented by the combination of one to three letters and five or six digits, depending on the data type. The sequence identifier corresponds to a simple number assigned to every nucleotide or protein sequence (Prosdocimi, 2010). The GI is individual, non-transferable and non-modifiable (Amaral et al., 2007).

ANALYSIS OF BIOLOGICAL SEQUENCES: Widely used and essential for biological sequence comparison, alignment has been processed by the increase in availability of data generated by NGS technologies (Daugelaite et al., 2013). This process consists of comparing two or more nucleotide sequences (DNA or RNA) or amino acids by seeking a series of individual characters or patterns that are also arranged in the sequences (Manohar and Shailendra, 2012; Junqueira et al., 2014). However, why compare sequences? There are some applications for this procedure that allow information of the evolutionary relationship between organisms, individuals, genes, prediction functions and structures, among others (Junqueira et al., 2014). Furthermore, alignment techniques are necessary to whole genome analysis, in which the comparison between different genomes or from the same species allows us to identify variations in the sequences and associate them with specific phenotypes. Regarding proteins, the alignment of structures also stands out as an important bioinformatics tool. While the comparison of structures refers to the analysis of similarities and differences between two or more structures, alignment refers to the determination of what amino acids would be equivalent between such structures (Junqueira et al., 2014). Although apparently trivial, sequence similarity analysis is complex since the algorithm used calculates a "cost" to the alignment of such sequences to minimize the differences and obtain the "best possible result" (Manohar and Shailendra, 2012).

The sequence alignment is arranged in rows and the characters in columns. It is up to the algorithm used to search for the best match for the sequences, sometimes inserting gaps ("-") representing one or more nucleotide indel events (Prosdocimi, 2010). However, for the same sequences "*n*" alignments are possible. Therefore, to solve this question a scoring system, in which the matches are positively and the mismatches are negatively punctuated, was created. The most widely used punctuation/substitution matrices are those belonging to the PAM (Point Accepted Mutation) (Dayhoff et al., 1978; Pevsner, 2009; Sung, 2010) and BLOSUM (Blocks Substitution Matrix) families that relate the probability of substitution of one amino acid or nucleotide for another (Prosdocimi et al., 2002; Prosdocimi, 2010). Therefore, the best possible alignment will be one that maximizes the overall score (Junqueira et al., 2014). Alignment can be categorized by type according to the number of sequences that are compared, which can be:

i) simple and
ii) multiple.

By definition, the simple alignment specifically depicts the similarity relation between two sequences, while the multiple considers a value greater than three sequences. Concerning the extent of alignment, these can still be classified as global or local (Junqueira et al., 2014). About the algorithm used, it may be classified as optimal or heuristic (Prosdocimi, 2010). The optimum result is the best alignment possible, while the heuristic, although not presenting an optimal result, presents the best alignment for a given period of analysis.

Simple alignment: In this approach, the dynamic programming algorithms, dot matrix analysis, and *k*-tuple method are highlighted. The dynamic programming method is based on the Bellman's optimality principle that proposes that the solution to complex problems is solved by its various subproblems (Junqueira et al., 2014). This methodology can be applied to produce global and local alignments through Needleman-Wunsch and Smith-Waterman algorithms, respectively (Manohar and Shailendra, 2012). To alignment, a scoring scheme is required for matches and mismatches, for amino acids or nucleotides, and a penalty value for gaps. In this way, the algorithm will calculate the optimum alignment between the sequences. The dot matrix approach is conceptually simple and efficient in the detection of indels and repetitions (Manohar and Shailendra, 2012). Through of an identity matrix, it is possible to graphically visualize the regions of similarity (Junqueira et al., 2014). In this method, the sequences are arranged one vertically and the other horizontally, and regions with the same characters are signaled, representing the corresponding possible matches (Junqueira et al., 2014). A line on the diagonal will represent the regions of similarity, while the other points represent random correspondences (Junqueira et al., 2014). The *k*-tuple alignment method, or words, is a heuristic method that is significantly more efficient than dynamic programming (Manohar and Shailendra, 2012). This method is implemented in database search tools, such as FASTA and BLAST (Basic Local Alignment Search Tool). This approach identifies a series of subsequences of two to six characters. Likewise, the database sequences will also be subdivided, with the comparison being made. After the identity search, the algorithm will align the two complete sequences and extend the similarity analysis to neighboring regions. The highest score value will be determined for each alignment using a penalty matrix (Junqueira et al., 2014). A more detailed view of this approach will be presented when discussing the methodology adopted by BLAST.

Multiple alignments: Similar to simple alignments, the dynamic programming method is usually employed in global alignment. However, each possible pair formed is punctuated by a weighted sum of pairs, with the addition of similarity values (Junqueira et al., 2014). Besides that, alternative methods were developed to accelerate the calculations, among which we can highlight: progressive, iterative methods and hidden Markov models (Manohar and Shailendra, 2012).

BLAST: BLAST is a specific local alignment algorithm derived from the Smith-Waterman algorithm that presents a maximum alignment score of two sequences (Amaral et al., 2007). In addition to the dynamic programming arising from the algorithm mentioned above, BLAST employs a heuristic based on the *k*-tuple method to search the sequences in the database (Junqueira et al., 2014). The *k*-tuple method limits the search to those words that are more significant, being the size of 3 and 11 characters for amino acids and nucleotides, respectively (Amaral et al., 2007). The execution of BLAST is fast and reliable, whose search from the query sequence (Query) is compared to the database to be used. In a simplified way, the BLAST may be divided into four stages.
i) Compiling the word list (*k*-tuples);
ii) Searching for correspondence in the database;
iii) Extending alignments from the identified words, and
iv) Assembling the spaced alignments according to high-score segment pairs (HSP).

BLAST is a family of programs used for different purposes according to the type of sequence of interest and the database to be searched (Prosdocimi, 2010). Several applications available by BLAST. Although less common, there is megablast and PSI-BLAST. The E-value represents the statistical value that indicates the probability that the alignment did not occur at random, considering the alignment score and the database size

(Prosdocimi et al., 2002; Amaral et al., 2007). On the other hand, the score is attributed by the algorithm based on the matches and mismatches between the input sequences and database (Amaral et al., 2007).

Comparative Molecular Modeling: Homology modeling refers to the modeling of the 3-D structure of a protein from the structure of another homologous protein whose structure has already been previously determined (Capriles et al., 2014). This approach is based on the fact that evolutionarily related sequences share the same folding pattern of the tertiary structure (Calixto, 2013). The determination of the 3-D structure helps in the understanding of the function, in the dynamics and interaction of the proteins as well as in the functional prediction and identification of therapeutic targets (Madhusudhan et al., 2005). Although methodologies such as X-ray diffraction crystallography and nuclear magnetic resonance (NMR) may be applied in the determination of the structure, there are limitations to its use. Thus, experimental methods can be implemented, such as *ab initio* modeling or by homology (Madhusudhan et al., 2005). *Ab initio* protein modeling uses physical and chemical principles to calculate the most favorable conformation. On the other hand, homology modeling presents more accurate results (Wang, 2009). However, its accuracy is closely related to the degree of similarity between target and template structures (Capriles et al., 2014). Minimum identity values of 25 to 30% are acceptable, but the higher values present better predicted model quality (Calixto, 2013; Capriles et al., 2014). The prediction process consists of five main steps (Capriles et al., 2014):

1) reference identification; 2) selection of templates; 3) alignment;
4) construction, and 5) model validation.

The first step is identifying amino acid sequences of proteins whose structure has already been resolved and which have similarity to the target sequence (Capriles et al., 2014). This comparison can be performed using the BLAST, for which references with higher indexes of similarity and identity should be chosen (Calixto, 2013). The selection of templates is necessary to choose one or more structures, considering some criteria, such as if they belong to the same family or if they perform the same function (Capriles et al., 2014). Once the template structure is chosen, global alignment between the target and template sequences is carried out so that the identity is greater than 40%. However, it is worth noting that the final model is dependent on the quality of this alignment (Capriles et al., 2014). From this alignment the model will be constructed using one of the following methods: rigid body assembly, corresponding segment or spatial constraint (Madhusudhan et al., 2005), being the first and last most commonly used (Capriles et al., 2014). Softwares such as Modeller and SWISS-MODEL may be utilized for the construction of the models. Alignment functions as an input file for modeling that results in a set of atomic coordinates for *n* 3-D models for the target protein, containing the atoms of the major and side chains of the amino acid residues. For this, the software calculates several chemical and spatial constraints, which are parameters added to the force field to tend the calculations in a certain direction (Silva and Silva, 2007). The model validation consists in the verification of possible errors related to the methodology adopted. Therefore, evaluation of the model quality by factors, such as bonding lengths, the planarity of peptide bonds, ring planarity and torsion angles in the main and lateral chains, chirality, steric hindrance, and energy functional, is necessary (Capriles et al., 2014). The Ramachandran plot is a valuable tool for determining the quality of the protein structure since it points out the existence of stereochemical impediments in the main chain of amino acids (Calixto, 2013). Other software such as ProSA and Verify_3D also help validate the structure. If the analysis of the model was not satisfactory, it is possible to refine the model or start its prediction again (Capriles et al., 2014).

DATA INTEGRATION AND USE OF NETWORKS: Differential expression studies have been widely adopted as a method to investigate the functions of genes on a global scale. In this approach, the genes are treated individually, without considering the interactions between them (Hong et al., 2013). However, biological functions exhibit a complex behavior, resulting from a set of genes interacting with each other (Zhao et al., 2010). In this context, the integrated biological systems approach using gene networks of coexpression has been widely used to understand the genetic architecture of complex phenotypes (Xu et al., 2014). Different levels of information can be integrated with networks. For example, the body is made of

JETIR

multiple networks (genes, molecular, cellular, and organ networks) that are incorporated and communicate at multiple scales (Institute for Systems Biology, 2016). Among the multiple approaches that can be used in the identification of biological networks, the use of genetic co expression networks as multistage analysis method will be highlighted. In this analysis, it is assumed that all genes (nodes) are connected, and their connection strength (connectivity) is quantified by the correlation of the expression between them (Zhao et al., 2010). Thus, it is possible to detect groups of highly co expressed genes (modules) that share a common function for which they are believed to act cooperatively (guilt by association) in a metabolic pathway (Kogelman et al., 2014). The connectivity of the gene (*ki*) describes the relative importance of the gene in the network. Genes with high *ki* are biologically relevant and reflect heavily regulated processes (Kogelman et al., 2014). Since the modules may correspond to biological pathways, it is possible to investigate whether the modules identified are associated with certain phenotypes as well as the significance of the gene on the traits under analysis (Zhao et al., 2010). This analysis is an assumption of the WGCNA software (Weighted Gene Co-expression Network Analysis). The detailed description of other methods of data analysis can be obtained in Ritchie et al(2015).

CONCLUSIONS: This intelligence or knowledge discovery gained from data mining has a vast amount of aims, including the likes of forecasting, validation, diagnosis and simulations. Typically the process for knowledge discovery through databases includes the storing and processing of data, application of algorithms, visualisation/interpretation of results. It’s important to state that the process of data mining or KDD encompasses a multitude of techniques, such as machine learning. As a result the process of data mining includes many steps needed to be repeated and refined in order to provide accuracy and solutions within data analysis, meaning there is currently no standard framework of carrying out data mining. The extensively vast science of data mining within the domain of bioinformatics is a seemly ideal fit due to the ever growing and developing scope of biological data. As this area of research is so extensive it is apparent that attributes of biological databases propose a large amount of challenges. Improving the quality and the accuracy of conclusions drawn from data mining is ever more key due to these challenges. As a result it is important for the future directions of research to adapt for the integration of new bioinformatics databases in order to provide more methods of effective research. Advances in the capabilities of data generation and analysis, as well as in the interpretation of results, have pointed to a promising future. On the other hand, data integration is not the end. It is the beginning of new discoveries and hypotheses, generating a feedback system. This prospect points out the need for scientists with mastery in multiple areas of knowledge, as well as the performance of multidisciplinary research groups, in which the complementarities of the different abilities will allow remarkable advances in science.

REFERENCES

Altelaar AFM, Munoz J and Heck AJR (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 14: 35-48

Altmann A, Weber P, Bader D, Preuss M, et al. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* 131: 1541-1554

Amaral AM, Reis MS and Silva FR (2007). O programa BLAST: guia prático de utilização. 1st edn. Embrapa Recursos Genéticos e Biotecnologia. EMBRAPA, Brasília.

Calixto PHM (2013). Aspectos gerais sobre a modelagem comparativa de proteínas. *Cienc. Equat.* 3: 10-16.

Capriles PVSZ, Trevizani R, Rocha GK and Dardenne LE (2014). Modelos tridimensionais. In: Bioinformática da biologia à flexibilidade molecular (Verli H,ed.)SBBq, São Paulo, 147-171.

Chaisson MJP, Wilson RK and Eichler EE (2015). Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* 16: 627-640

Daugelaite J, O' Driscoll A and Sleator RD (2013). An overview of multiple sequence alignments and cloud computing in bioinformatics. *Int. Sch. Res. Not.* e615630.

Dayhoff MO, Schwartz R and Orcutt BC (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure (vol.5, supl 3 ed.). Nat. Biomed. Res. Found., Washington, D.C.

Goff LA, Trapnell C and Kelley D (2012). CummeRbund : visualization and exploration of Cufflinks high-throughput sequencing data. R package version 2.16.0.

Hagen JB (2000). The origins of bioinformatics. *Nat. Rev. Genet.* 1: 231-236

Hawkins RD, Hon GC and Ren B (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11: 476- 486 10.1038/nrg2795.

Hong S, Chen X, Jin L and Xiong M (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.* 41: e95.

Hunt LT (1984). Margaret Oakley Dayhoff 1925-1983. *Bull. Math. Biol.* 46: 467-472.

Institute for Systems Biology (2016). What is a systems biology. institute for systems biology. Available at https://www. systemsbiology.org/about/what-is-systems-biology/.

Jensen ON (2006). Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* 7: 391-403.

Junqueira DM, Braun RL and Verli H (2014). Alinhamentos. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 38-61.

Kitano H (2002). Systems biology: a brief overview. *Science* 295: 1662-1664.

Kogelman LJA, Cirera S, Zhernakova DV, Fredholm M, (2014). Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med. Genomics* 7: 57.

Luscombe NM, Greenbaum D and Gerstein M (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* 40: 346-358

Madhusudhan MS, Marti-Renom MA and Eswar N (2005). Comparative protein structure modeling. In: The proteomics protocols handbook (Walker, J.M., ed.). Human Press, New Jersey, 831-860.

Malone JH and Oliver B (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9: 34.

Manohar P and Shailendra S (2012). Protein sequence alignment: A review. *World Appl. Program.* 2: 141-145.

Marioni JC, Mason CE, Mane SM, Stephens M, (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*18:1509-1517

Miller JR, Koren S and Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315-327.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464: 773-777

Pevsner J (2009). Pairwise sequence alignment. Bioinformatics and Functional Genomics (2nd edn). Wiley-Blackwell.

Pevsner J (2015). Bioinformatics and functional genomics, 3rd ed. John Wiley & Sons Inc, Chichester.

Prosdocimi F (2010). Introdução à bioinformática. Curso Online.

Prosdocimi F, Cerqueira GC, Binneck E and Silva AF (2002). Bioinformática: Manual do usuário. *Biotec. Cienc. Des.* 12-25.

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16: 85-97.

Schmidt A, Forne I and Imhof A (2014). Bioinformatic analysis of proteomics data. *BMC Syst. Biol.* 8: 2, S3.

Silva VB and Silva CHT (2007). Modelagem molecular de proteínas-alvo por homologia estrutural. *Rev. Elet. Farm.* 4: 15-26.

Sims D, Sudbery I, Ilott NE, Heger A, (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15: 121-132.

Sung WK (2010). Algorithms in Bioinformatics: a practical introduction. CRC Press.

Staats CC, Morais GL de and Margis R (2014). Projetos genoma. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 62-79.

Trapnell C, Pachter L and Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7: 562-578.

Verli H (2014). O que é Bioinformática? In: Bioinformática da biologia à flexibilidade molecular (Verli H ed.). SBBq, São Paulo, 1-12.

Wang J (2009). Protein structure prediction by comparative modeling: an analysis of methodology. *Comp. Gen. Pharmacol.* 218: 1-13.

Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.

Xu L, Zhao F, Li L, (2014).Co-expression analysis of fetal weight-related genes in ovine skeletal muscle during mid and late fetal development stages. *Int.J.Biol.Sci.*10:1039-1050.

Zhao W, Langfelder P, Fuller T, Dong J, (2010). Weighted gene coexpression network analysis: state of the art. *J. Biopharm. Stat.* 20: 281-300.

Zhou X, Ren L, Meng Q, Li Y, et al. (2010). The next-generation sequencing technology and application. *Protein Cell* 1: 520-536.