RESEARCH ARTICLE OPEN ACCESS

Computational Intelligence applied in Bioinformatics

C. Sunitha Ram [1], Swetha Gayathri Kuchimanchi [2]

CSE, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya - Kanchipuram

ABSTRACT
Computational intelligence poses Several possibilities in Bioinformatics, particularly by generating low-cost, low-precision, good solutions. Rough sets promise to open up an important dimension in this direction. The present article surveys the role of artificial neural networks, fuzzy sets and genetic algorithms, with particular emphasis on rough sets, in Bioinformatics. Since the work entails processing huge amounts of incomplete or ambiguous biological data, the knowledge reduction capability of rough sets, Learning ability of neural networks, uncertainty handling capacity of fuzzy sets and searching potential of genetic algorithms are synergistically utilized.
Keywords- Computational Intelligence; Bioinformatics; genetic algorithms

I. INTRODUCTION

Since the term 'bioinformatics' was coined in 1970, the field of bioinformatics has become relatively mature allowing high-throughput whole genome sequencing and making computer-aided drug design an essential part of drug discovery. With the needs of addressing ever more complex problems in a faster and more accurate manner, the bioinformatics community has exploited many different paradigms. Among them, 'Computational Intelligence' has proved particularly effective since nature-inspired computational approaches are able to extract patterns from large volumes of data, infer rules from sets of examples and adapt according to changing data and/or contexts. Many of those methods have been

applied to bioinformatics; they include: Artificial Immune Systems, Bayesian Networks, Evolutionary Algorithms, fuzzy Logic, Hidden Markov Models, Neural Networks, Rough Sets, Support Vector Machines (SVM) and Swarm Intelligence. In this special issue, we present six papers that illustrate the latest applications of Computational Intelligence in Bioinformatics.

Since more and more protein structure prediction tools are now available, it is crucial to be able to assess the quality of the generated models.

Using features extracted from sequence alignment between a target and its template a SVM-based method was developed to predict the quality score of a model. High correlation between predicted and actual values showed the effectiveness of their method.

When analysing data such as gene expression data where the number of features can be two orders of magnitude higher than the number of samples, advanced feature reduction is essential to produce robust classifiers. Xu et al. presented a new procedure based on the discriminative or predictive ability of variables via bootstrapped ROCAUCs (Area Under Receiver Operating Characteristic Curves). Simulations demonstrated the usefulness of the proposed methodology to build predictive models from bioinformatics data.

Since generally combining predictions of a set of classifiers produces more accurate predictions than the individual classifiers, Nguyen et al. adopted an assemble approach in order conduct experiments established the validity of their original approach, which would benefit further from integration of additional classifiers.

The aim of this study is to implement a comprehensive directory of all the methods, tools, and databases, which are available in different areas of contemporary hot topics in bioinformatics principally the machine learning and computational intelligence techniques [6]. Section two provides an overview of bioinformatics and a brief description of currently available papers on machine learning, big data, IoT and cloud computing based on bioinformatics. Section three is a literature review, which illustrates the computational intelligence methods in different areas of bioinformatics such as Neural Network, Gene Expression, Gene Selection, DNA Fragment Assembly (FA), Multiple Sequence Alignment (MSA), Protein Structure Prediction (PSP), Protein Sequence Classification, Human Genetics, and Microarray Classification. Section four lists and summarizes the free online available tools and databases in bioinformatics. Finally, section five provides a conclusion and a slight discussion of the future studies.

II. LITERATURE REVIEW

Although bioinformatics deals with immature data collected from researchers on a daily basis in order to create images, charts, and numbers, as well as sort the data collected with a wide range of tools,on the network database The significance and meaning of preliminary data collection tests, including experimental errors, principles, or to data collection for statistical correlation means that careful experimental design and multiplicity are required. Experiments in professional settings, reactants, Both the equipment and the time are expensive. Finally, biological information are never complete.Large amounts of data from recent biological tests This resulted in the creation of a massive database containing genes, proteins, and other information.[1]

Sprocket for genetic data and other data types Researchers are adamant about recovering data from some of the most important DB specifications, such as nuclide or amino acid chain, organism, marginal genes, or protein name. Thematic research categories are the most important tool in bioinformatics. A biologist used this tool to compare a new DNA or protein chain (blast) to all other DB chains in order to find more similarities between the chains.[2]

Computer simulations based on Computational Intelligence play a critical role in biological methods to increase the production of experimental data. Proteins are necessary molecules for cell manipulation. Protein is an enzyme, which catalyses a chemical reaction. Components of a stem cell act as a converter and respond to nature, or as powerful elements and regulatory elements that prevent certain processes in the cell such as metabolism, duplication, and transfer. Genes that are transcribed in mRNA must be precise and unequivocal.[3]

Provided a description of how cloud computing and big data technologies such as Apache Hadoop project can be used for biology's big data sets and explained the current use of Hadoop in bioinformatics. Discussed the relationship between sequence database, IoT, and bioinformatics. Examined the Big Data and Bioinformatics to show that emerging the two sciences can provide richer representation. Considered the IoT criteria's in the health sector for sustainable development. Combined Human-Computer Interaction (HCI) and knowledge discovery from data theories, methods, and

approaches to obtain an interactive data set. Gathered all recent researches about Big Data tools in bioinformatics. Discussed the relationship between the physical world and IoT on bioinformatics.[4]

III. OVERVIEW OF METHODS

Computational intelligence (CI) is the area of developing generic intelligent information processing methods and systems with wider applications.CI methods, in its majority, are inspired by the human intelligence. They are characterised by learning, generalisation, adaptation, pattern recognition, rule extraction,knowledge representation, which are characteristics of the living systems too. There is a large variety of machine learning (ML) techniques used in the above methods. The methods of CI include:

•Probabilistic learning methods, e.g., Hidden MarkovModels;

•Statistical learning methods, e.g., Support VectorMachines (SVM), Bayesian classifiers;

•Case-based reasoning (e.g., k-NN; transductivereasoning);

•Decision trees;

•Rule-based systems (propositional logic dated back toAristotel) and fuzzy systems (introduced by L. Zadeh1);

•Neural networks;

•Evolutionary computation;

•Particle swarm intelligence;

•Artificial Life;

•Quantum computation;

•Hybrid systems (e.g., knowledge-based neural net-works; neuro-fuzzy systems; neuro-fuzzy-genetic systems;evolving connectionist systems).

CI adopts many principles from Biology, thus offering suitable methods and tools for BI. While CB aims at understanding the biology principles through their computational modeling, BI is aiming at the use and the development of new information methods and systems to enhance the storage, the analysis, modeling, and discovery from biological data. The synergism between the three disciplines, their methodologies, problems, and some current solutions are review in the paper. Some new methods and experimental results are introduced, such as feature and model optimization with genetic algorithms applied on gene expression data.

IV. PROBLEMS AND PERSPECTIVES

Biological systems are inherently non-linear and dynamic. Datasets resulting from the analysis of biological systems typically include additional noise (in light of the experimental conditions and/or methods used to generate the data). The proper interpretation of such systems demands interpretive methods that do not rely strictly on linearity and yet can provide reasonable solutions to the researcher in fast time. CI represents a burgeoning field in computer science that has broad, largely underappreciated, utility in biochemistry and medicine. However, in recent years there has been a dramatic increase in the use of these approaches, broadly, over many disciplines of biomedicine and biochemistry. Such methods can be complementary to previous approaches (such as local search or dynamic programming), and can be used to search very large solution spaces efficiently. The breadth of recent successful application makes it all the more apparent that the need for these methods will continue to grow in the near future.

Some of these approaches are considered 'black box' models, where it is difficult for the researcher to understand the underlying logic of the optimized pattern recognition model. For some problems, this is of little importance. However, there are problems where an understanding of the logic is of great importance. In these cases, more appropriate model representations can be developed that optimize the logical transform in a manner that is more easily understood by the end-user. Such approaches are very useful for automated feature reduction and pattern recognition.

Far too often, researchers fail to identify the right method or computational approach for the right problem and impose the use of one favored method over all problems. In reality, each problem requires its own model representation in light of input data types (e.g. categorical or numerical values, fuzzy terms or discrete variables) and end-user constraints and interpretation. The No Free Lunch theorem suggests that there is likely not to be one best representation or model optimization method for all problems. Thus the informed bioinformatician will do well to explore the potential of CI approaches in providing a unique and powerful versatility to overcome many hurdles, simultaneously, with model development and optimization in achieving the right method for the right problem.

V. CHALLENGES

The key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of the genome and its manifestation. In this chapter we first briefly review the source of this data. We then provide some informatics frameworks for organizing and thinking about challenges and opportunities in bioinformatics. We use then use one informatics framework to illustrate specific challenges from the informatics perspective. As a contrast we provide also an alternate perspective of the challenges and opportunities from the biological point of view. Both perspectives are then illustrated with case studies related to identifying and addressing challenges for bioinformatics in the real world.

In the broader context, the key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of the genome and its manifestation. The genome can be thought of as the machine code or raw instructions for creation and operation of biological organisms (its manifestation).

VI. ALGORITHMS USED IN BIOINFORMATICS

There are many variants of algorithms in bioinformatics for different problems such as:

- Baum–Welch algorithm - used to find the unknown parameters of a hidden Markov model (HMM)

- BLAST (Basic Local Alignment Search Tool) - an algorithm for comparing primary biological sequence information

- Hirschberg's algorithm - a dynamic programming algorithm that finds the optimal sequence alignment between two strings

- Pairwise Algorithm - used for comparing a protein profile

- Ukkonen's algorithm - online algorithm for constructing suffix trees etc

Still, two most popular and used algorithms are:

- Needleman-Wunsch algorithm and

- Smith-Waterman algorithm.

Given two DNAs (or RNAs, or Proteins), high similarity means similar function or similar 3D structure. To compare the similarity of two biological sequences one can use global alignment: Needleman-Wunsch algorithm; or local alignment: Smith-Waterman algorithm.

VII. ONLINE SOFTWARE TOOLS

Everyday bioinformatics is done with sequence search programs line BLAST, sequence analysis programs like the EMBOSS and staden package, Structure prediction programs like THREADER or PHD or molecular imaging/modeling programs like RasMol and WHAT IF.

More:

- NetSurfP – Protein Surface Accessibility and Secondary Structure Predictions
- NetTurnP – Prediction of Beta-turn regions in protein sequences
- MODELLER – Used for homology or comparative modeling of protein three-dimensional structures
- AutoDock – Suite of Automated Docking Tools
- Gromacs – A molecular dynamics package primarily designed for biomolecular systems such as proteins and lipids
- OrfPredictor – The OrfPredictor (ORF-Predictor) server is designed for ORF prediction and translation of a batch of EST or Cdna sequences

VIII. TYPES OF ANALYSIS TOOLS

Next-generation-sequencing (NGS) and microarrays have vastly increased the ability to detect sequence variation and gene expression. These technologies require specific bioinformatics tools to extract information from the raw data. For analyzing human data, NGS pipelines work on three basic tiers: sequence generation, alignment to a reference genome, and interpretation of the results. Various bioinformatics tools are widely available to analyze NGS data sets specific to target/whole-genome sequencing or RNA- Sequencing. The so-called big data analytics, which have been made possible via high- throughput sequencing, have been vital to systems biology, especially gene regulatory networks (GRNs), which deal with gene-to-gene interactions. Different bioinformatics tools and pipelines exist for various GRN topics, including gene set analysis and identification of phenotype-differentiating pathways and subpathways with GRN information, topology, and regulatory mechanisms.

DNA sequencing

Before sequences can be analyzed they have to be obtained from the data storage bank example the Genbank. DNA sequencing is still a non- trivial problem as the raw data may be noisy or afflicted by weak signals. Algorithms have been developed for base calling for the various experimental approaches to DNA sequencing.

IX. TYPES OF DATABASES

There are basically 3 types of biological databases are as follows.

1. Primary databases :

- It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence,

macromolecular structure, etc. The data entered here remains uncurated(no modifications are performed over the data).

- It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change.
- The data are given accession numbers when they are entered into the database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely and it never changes.

Examples –

- Examples of Primary database- Nucleic Acid Databases are GenBank and DDBJ
- Protein Databases are PDB,SwissProt,PIR,TrEMBL,Metacyc, etc.

2. Secondary Database :

- The data stored in these types of databases are the analyzed result of the primary database. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database.
- The data here are highly curated(processing the data before it is presented in the database). A secondary database is better and contains more valuable knowledge compared to the primary database.

Examples –

Examples of Secondary databases are as follows.

- InterPro (protein families, motifs, and domains)
- UniProt Knowledgebase (sequence and functional information on proteins)

3. Composite Databases :

- The data entered in these types of databases are first compared and then filtered based on desired criteria.
- The initial data are taken from the primary database, and then they are merged together based on certain conditions.
- It helps in searching sequences rapidly. Composite Databases contain non-redundant data.

Examples –

Examples of Composite Databases are as follows.

- Composite Databases -OWL,NRD and Swissport +TREMBL

X. FREE AVAILABLE ONLINE DATABASES INBIOINFORMATICS

The obtained sequences are assigned to the researchers by online databases for this purpose. These databases are completely open to the public, and you can extract all similar sequences as well as a complete set of information about them by entering any desired sequence. These bioinformatics databases have stored massive amounts of genetic data and have quickly become the most valuable research tool in molecular biology.

The rate of growth of the information contained in databases is increasing exponentially, so that the volume of data in the "Gen Bank" database, for example, doubles every 14 months. Similarly, for proteins, proteome projects are defined, and the workload is significantly higher. However, useful methods, such as those used in genome projects, are available. Furthermore, many proteins and enzymes are unknown. The goal of these projects is to sequence proteins and determine their three-dimensional structure. The last one, in particular, is important for protein function. If the Genome Project took 12 years to complete, Because of Proteome Project's workload, it seems that nearly a century

requires finishing this project which is being done in collaboration with research centers all over the world. Some of the online databases are listed below:
1000 Genomes, ACSN, Array Express, ArrayMap, ASCAT, ASG, BIND, BioCarta Pathways, BioGPS, BioGRID, BioLINCC, BioMuta v2, BioProject, BioSample, Cancer Genome Anatomy Project, Cancer Genomics Browser etc.

XI. APPLICATIONS

Bioinformatics is an integrative field in life sciences that combines biology and information technology. Its application includes the study of molecular sequences and genomics data.

Being a combination of different branches of life sciences, the objective of bioinformatics is to develop methodologies and tools to study large volumes of biological data in order to organize, store, systematize, visualize, annotate, query, understand and interpret those data.

Bioinformatics utilizes modern computer science that includes cloud computing, statistics, mathematics and even pattern recognition, reconstruction, machine learning, simulation and iterative approaches, and molecular modelling
/algorithms.

In simpler terms, bioinformatics involves the application of computer technology to manage large volumes of biological information.

Applications of bioinformatics in medicine
Bioinformatics has proven quite useful in medicine as the complete sequencing of the human genome has helped to unlock the genetic contribution for many diseases. Its applications include drug discovery, personalized medicine, preventative medicine and gene therapy.

1. Drug discovery

Infectious diseases are currently the world's major killer of children and young adults. According to WHO, infectious diseases account for over 13 million deaths yearly.

Developing countries record the most number of deaths from infectious diseases and this was contributed to the non-availability of drugs and high cost associated with the drugs if available.

2. Personalized medicine

Personalized medicine is a model of healthcare that is tailor-made to each person's unique genetic make-up.

A patient's genetic profile can assist the doctor to predict susceptibility to certain diseases, provide proper medication and with the proper dose to reduce side-effects. It is applied in the treatment of personalized cancer medicine, diabetes-related disease and HIV.

Bioinformatics is used in personalized medicine to analyse data from genome sequencing or microarray gene expression analysis in search of mutations or gene variants that could affect a patient's response to a particular drug or modify the disease prognosis.

3. Preventive medicine

Preventive medicine focuses on the health of individuals, communities and defined populations. It uses various research methods, including biostatistics, bioinformatics and epidemiology, to understand the patterns and the causes of health and disease, and to transform such information into programs designed to prevent disease, disability and death.

An example of preventive medicine is the screening of newborns immediately after birth for health disorders, such genetic diseases or metabolic disorders, that are treatable but not clinically evident in the newborn period.

4. Gene therapy

Gene therapy is the method of replacing defective genes with a functional one in the cells of the patient. Gene therapy has not been widely used because developing a generic gene therapy method is quite complicated, as each person's genetic profile is different.

Bioinformatics could help to identify the best gene target site for each individual by taking their genetic profile into consideration. This can reduce the risk of unintended side effects.

The application of bioinformatics is not limited to the field of medicine. It is wide-ranging and constantly evolving as more areas in life sciences are transformed by it.

XII. CONCLUSION

Bioinformatics in Atomic innovation age embraced numerous concepts of organic advancement such as autonomous organizations with the heading from foot to best, crisis, the capacity to adjust the arrangements, straightforward rules, codes and naturally being at the atomic level. We will utilize the common organic advancement designs to anticipate the period of atomic advancement innovation. Unused apparatuses such as hereditary calculations, organize planning and atomic can be cases of this period. We will conclude that the expectation impact of meeting and atomic data is in a way that puts premonition and arranging before exceptional changes.

Big data can be utilized to advise the chiefs of the modern generation and unused openings. Besides, shed the disappointments as before long as conceivable. Computational Insights speedy behavior can open a few unused entryways and hence empower the management to embrace unused strategies by the plausibility of processes' enhancement. These witnesses recommend that computational insights and bioinformatics would be valuable in changes of imaginative handle as coordinated evolution technology can genuinely accomplish an totally modern concepts.

REFERENCES

[1] © 2008 Computational Intelligence in Bioinformatics

Editors: Kelemen,Arpad, Abraham**,** Ajith, Chen, Yuehui (Eds.)

[2] Computational intelligence techniques in bioinformatics

Aboul Ella Hassanien [1], Eiman Tamah Al-Shammari, Neveen I Ghali

[3] Computational Intelligence in Bioinformatics by Jean-Christophe Nebel

[4] Computational Intelligence in Solving Bioinformatics Problems: Reviews, Perspectives, and Challenges by Aboul-Ella Hassanien1,2, Mariofanna G. Milanova3, Tomasz G. Smolinski4, and Ajith Abraham5