We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter 1

Bioinformatics: Basics, Development, and Future

Ibrokhim Y. Abdurakhmonov

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63817

Abstract

Bioinformatics is an interdisciplinary scientific field of life sciences. Bioinformatics research and application include the analysis of molecular sequence and genomics data; genome annotation, gene/protein prediction, and expression profiling; molecular folding, modeling, and design; building biological networks; development of databases and data management systems; development of software and analysis tools; bioinformatics services and workflow; mining of biomedical literature and text; and bioinformatics education and training. Astronomical accumulation of genomics, proteomics, and metabolomics data as well as a need for their storage, analysis, annotation, organization, systematization, and integration into biological networks and database systems were the main driving forces for the emergence and development of bioinformatics. Current critical needs for bioinformatics among others highlighted in this chapter, however, are to understand basics and specifics of bioinformatics as well as to prepare new generation scientists and specialists with integrated, interdisciplinary, and multilingual knowledge who can use modern bioinformatics resources powered with sophisticated operating systems, software, and database/networking technologies. In this introductory chapter, I aim to give an overall picture on basics and developments of the bioinformatics field for readers with some future perspectives, highlighting chapters published in this book.

Keywords: bioinformatics, databases, molecular sequence analysis, software and analysis tools, bioinformatics training

1. Introduction

Biological data can be described as molecular sequence information and "wet-bench" experimented content of genome and gene product analyses [1]. Being an interdisciplinary branch of the life sciences, bioinformatics targets to develop methodology and analysis tools to explore

INTECH
open science | open minds

© 2016 The Author(s). Licensee InTech. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

large volumes of biological data, helping to store, organize, systematize, annotate, visualize, query, mine, understand, and interpret complex data volumes. It uses conventional, modern computer science and cloud computing, statistics, and mathematics, as well as pattern recognition, reconstruction, machine learning, simulation and iterative approaches, and molecular modeling/folding algorithms [1, 2]. The emergence and advances of the bioinformatics field, however, are tightly associated with the computerized programming and software developments needed for the handling and structural and functional analysis of large volumes of molecular sequences of DNA, RNA, proteins, and metabolites.

Presently, although still core for genomics and genetics field, bioinformatics became an umbrella for wider range of biological studies analyzing variety types of biological data, structuring, systemizing, annotating, querying, mining, and visualizing available biological information and a variety of biomedical text records [1–3]. Although drawing a fine line between bioinformatics and some other related fields is difficult because of increased applications of computers, statistics, and mathematics to scientific problem solving and experiments of life sciences, there should not be a misperception about bioinformatics description and objectives. Bioinformatics should not be mixed with, for example, biometry and biostatistics, development of DNA computers, or computerized generation and filing of data from imaging.

Bioinformatics also should be differentiated from related scientific fields such as biological computation and computational biology [1, 2]. Biological computation aims to develop biological computers using advances of bioengineering, cybernetics, robotics, and molecular cell biology. In contrast, bioinformatics develops and utilizes computational algorithms to understand and interpret biological processes based on genome-derived molecular sequences and their interactions [2]. Therefore, in many aspects, bioinformatics seems similar to computational biology objectives. A computational biology is concentrated on building and/or developing theoretical models for biological analyses [1, 2], whereas bioinformatics focuses on providing practical tools to organize and analyze basic genomic, proteomic and other "omics" data, including sequence analysis and its visualization [1, 2]. Admittedly, computational biology and bioinformatics both target to use genome data, for example, multiple sequence alignments and/or genome assembly tools. This makes distinctive boundaries of these two fields less distinguishable if their theoretical and practical scales are forgotten [2]. Thus, as mentioned above, the common core aims of bioinformatics are to handle, analyze, and interpret the genome-derived molecular sequence data and its organizational principles in broad scales/spectra of comparative, simulative, and evolutionary/phylogenetics perspectives. These tools are applicable and widely used for studies related to genetics, genomics, biochemistry, physiology, biophysics, all agricultural, medical, and environmental sciences as well as evolution, system biology, and artificial intelligence [1–10].

For instance, bioinformatics tools such as the comparative analysis of genomic and genetic data and/or signal processing help to interpret and understand the molecular and evolutionary processes [9] and interactions from large volumes of raw data in the field of wet-bench experimental molecular biology [1, 2]. In the "omics" fields, it helps to sequence and annotate genomes, and identify distinct patterns, mutation profiles, genetic epistasis, gene/protein expression and regulation, and gene ontologies [1, 2, 4, 8–11] as well as be instrumental in

mining and querying the biological data and biomedical literature text [3, 4, 7]. When applied for system biology [2, 6], bioinformatics is a key instrument to analyze and catalogue the biochemical/genetic pathways and networks, which helps to integrate pieces of analyzed information to depict and model a full picture of the life processes. Application of reconstruction, pattern recognition, folding, simulation, and molecular modeling with bioinformatic tools can identify structural peculiarities and interactions of molecular sequences important for structural biology and medicinal drug design [12, 13]. All of these large scale, genome-derived, molecular sequence analyses of raw "Big Data" are impossible to be analyzed manually [1, 2]. This prompted the biology science research community to apply interdisciplinary methods and tools for "Big Data" analysis in combination with modern computing knowledge, which resulted in the emergence of novel interdisciplinary bioinformatics science. Let us, first, take a look the historic developments in the bioinformatics field.

1.1. History of emergence and development

Bioinformatics term was coined by Paulien Hogeweg and Ben Hesper in 1970 [2, 14]. Its meaning was very different from current description and referred to the study of information processes in biotic systems like biochemistry and biophysics [14–16]. However, the emergence of bioinformatics tracks back to the 1960s. It was appeared in concordance with the development of protein sequencing methods from a variety of organisms and with the availability of protein sequences after Frederick Sanger determined the sequence of insulin in the early 1950s [17, 18]. New computer methods to analyze and compare a large number of protein sequences of different organisms were needed because handling many amino acid sequences manually was impractical. This led in compiling the first "Protein Information Resources" (PIR) [1, 19, 20] by Margaret Oakley Dayhoff and her collaborators at the National Biomedical Research Foundation [1]. Dayhoff's team successfully organized the protein sequences into distinct groups and sub-groups based on sequence similarity and percent accepted mutation (PAM) matrices [1]. This was published as protein sequences atlas [21, 22] that has been widely used in performing protein sequence alignments and database similarity searches [1, 2, 23]. This was pioneered methods of protein sequence alignment and molecular evolution [22]. In the 1970s, Elvin A. Kabat further contributed to bioinformatics development by his extended protein sequence analysis of comprehensive volumes of antibody sequences, released in collaboration with Tai Te Wu between 1980 and 1991 [2, 24].

With the objective of providing the theoretical background to immunology experiments in 1974, George Bell and colleagues initiated the collection of DNA sequences into GenBank [1]. During 1982–1992, the first version of GenBank was prepared by Walter Goad's group [1] and the efforts resulted in the development of presently known and widely used DNA sequence databases of GenBank [25], "The European Molecular Biology Laboratory (EMBL) [26], and DNA DataBank of Japan (DDBJ) [27] in 1979, 1980, and 1984, respectively [1]. Most important development in DNA sequence databases, however, was incorporation of web-based searching algorithms allowing researchers to find and compare the target DNA sequences. Such first developments and resulting computer software called "GENEINFO" and its derivative version of "Entrez" were developed by David Benson and David Lipman and colleagues [1].

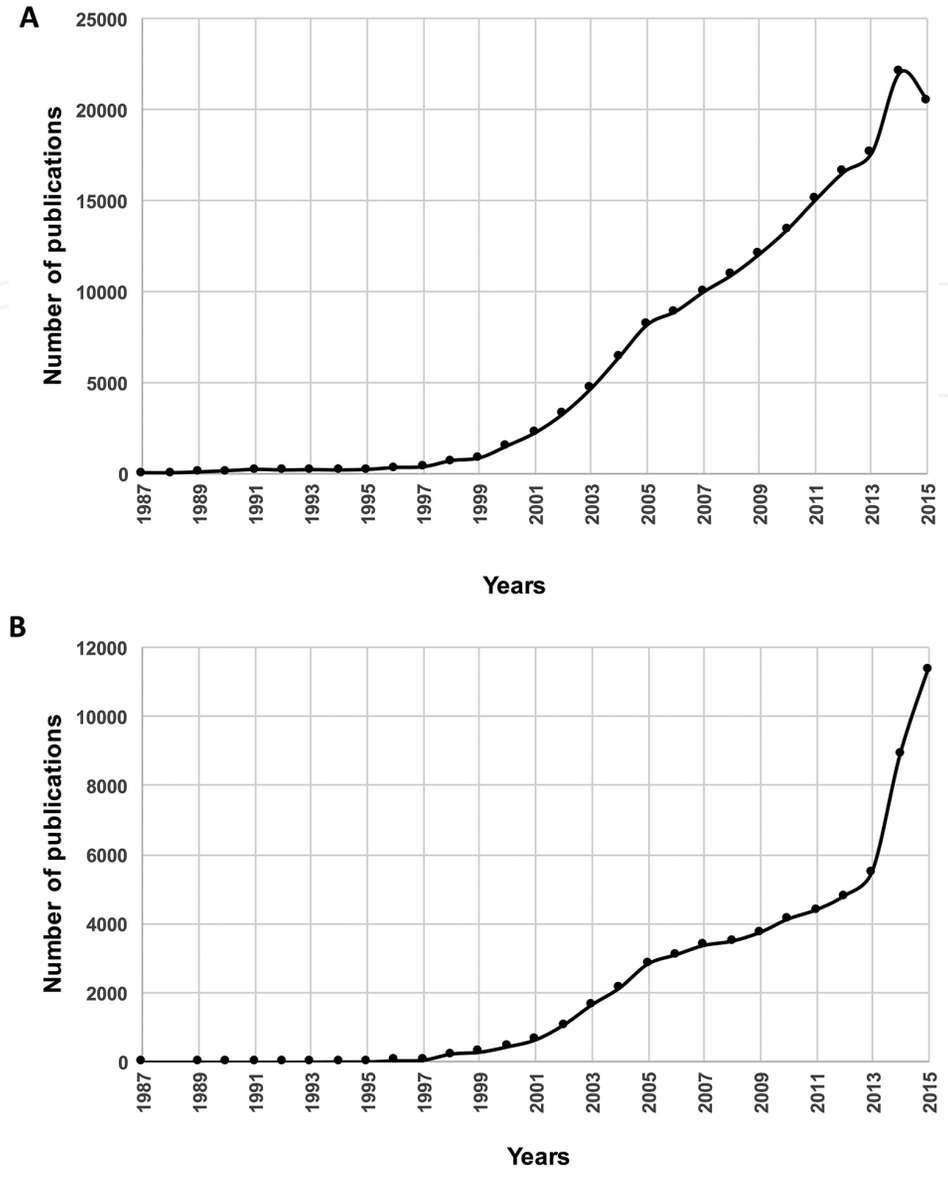


Figure 1. Dynamics of bioinformatics-related publications over the past four decades. (A) Unquoted and (B) quoted keyword retrieved scientific publications from PubMed [74].

This software allowed researchers to rapidly search database-indexed sequences and match them with queried sequence. Software became readily available through web-based interface of the National Center of Biotechnology Information (NCBI) database [28]. Molecular sequence analysis, comparison, and visualization methods have been improved, and many different methodologies have been contributed to bioinformatics advancements in this direction. Such advancements can be exemplified by the development of dot matrix and diagram methods [29], alignment of sequences by dynamic programming [30], finding of local alignments between sequences [31], multiple sequence alignment tools [32–35], predicting the secondary structures of RNAs [36, 37], determination of evolutionary relationships of sequences [38, 39], and assigning the gene function based on sequence similarity of known function from models [40]. Development of FASTA [41, 42], BLAST [43, 44], and their various modifications [45–47]

has further powered the bioinformatics field and greatly improved the biological data analysis. Development of tools for predicting the putative protein sequences, structure, and function of proteins/genes based on DNA sequences [48–58], completing full genome sequences, and building web-based genome databases for many prokaryotic and eukaryotic organisms [58] has provided a great advance in the bioinformatics field. In addition, rapid genome-wide gene expression profiling and analysis opportunities [59–62], biological pathway assignment and identification, data storing, and mining and querying for large volume of biological datasets [63–73] have further provided unprecedented popularity of bioinformatics in the scene of world science, which has been briefly reviewed below.

Since its emergence as an interdisciplinary scientific field in 1970, bioinformatics research has continuously increased over the past four-decade period. Unquoted search of keyword of bioinformatics in the PubMed database [74] has found nearly 181,000 scientific publications covering the period of 1958 to March of 2016. Repeating the search with the quoted keyword found 62,402 scientific publications over the four-decade period, demonstrating the starting point of increased publication efforts in the end of 1990s with its first raise in 2000/2001, following significant peaks in 2003/2004 and after 2013 (**Figure 1**). In this introductory chapter, I aim to give a brief highlight of these four-decade developments introducing the chapters presented in this book.

2. Bioinformatics help in handling and analysis of the genomics data, genome annotation, and expression profiling

Rapid and reliable determination of DNA molecules, because of the introduction of the sequencing technique of Sanger and Coulson [75] and Maxam and Gilbert [76], provided large-scale DNA sequence data that needed to be analyzed by computerized programming. This prompted the development of efficient bioinformatics methodologies. For example, a seminal effort of the Phage Φ-X174 [2, 77] and the *Haemophilus influenza* [2, 78] genome sequencing using shotgun sequencing techniques generated the sequences of many thousands of small DNA fragments, ranging from 35 to 900 nucleotides [2] and required the assembly of a complete bacterial genome. The ends of sequenced shotgun clones overlap and can be assembled using computerized similarity search algorithms into the complete genome although the assembly tasks are challenging due to the requirement for powerful computers with sufficient memory and issues of generating multiple gaps in assembled genome. Genome assembly algorithms are a critical area of bioinformatics research as fragmented genome sequencing methods have been the core approach for virtually all genomes sequenced today [1, 2].

Therefore, without bioinformatics tools, it is not possible to think about genome sequencing as present bioinformatics programs such as BLAST/sequence alignments not only provide rapid practical tools to handle, analyze, compare, relate, and visualize DNA sequences but also offer help with the sequencing process itself. The development of cost-effective, next generation sequencing (NGS) platforms [79, 80] has helped to completely decode nearly the entire genome of many different organisms including human and many other model and

specialty organisms, or crop genomes with complex polyploidy levels within a short period. For example, according to the listings in the Genomes OnLine Database (GOLD) as of March 8, 2016, there were 79,650 genome sequencing projects of which 8018 were completed projects, 33,489 were permanent drafts, 35,609 were incomplete projects, and 1553 were targeted projects [81]. There are 73,000 organism, including archaea (1201), bacteria (55,303), eukaryotes (11,990), and viruses (4473), listed for sequencing. These numbers should be increased if the sequencing of the 100,000 whole-human genomes [82] is added.

Bioinformatics tools are needed in annotation and prediction of genes from sequenced genomes that requires computerized approaches because genomes are large to be manually annotated as mentioned above. Bioinformatics-based gene finding and annotation including a search for protein-coding genes, RNA transcripts, and other functional sequences within a genome is possible because there are patterns to recognize the start, stop regions, introns, exons, motifs, repeats, and other regulatory and sensory as well as signaling regions with some variations between genes and among organisms. With the availability and need for analysis of *H. influenza* genome, the first genome annotation computer program system was designed in 1995 by Owen White [2, 78], which provided tools to find the genes and identify putative functions of annotated sequences. White's effort was basic for all currently available gene annotation and prediction software, which keep periodically improving [2].

Bioinformatics tools are very important to analyze gene and protein expression profiles. Large-scale sequencing of cDNA libraries has generated large volumes of serial analysis of gene expression (SAGE), expressed sequences tags (ESTs), massively parallel signature sequencing (MPSS), transcriptome profiling, or RNA-Seq, and various applications of multiplexed in-situ hybridization (microarray) profile data [83–95]. All of these gene expression techniques are extremely noise-prone and/or subject to bias in the biological measurement, which requires application of statistical tools to separate signal from noise in high-throughput gene expression studies. In this context, chapter by Zhao et al. in this book reviews and discusses the main tools and algorithms currently available for RNAseq data analyses, discussing rapidly evolving RNAseq technologies such as stranded RNAseq, targeted RNAseq, and single cell RNA-seq. Moreover, Sripathy et al. have comprehensively discussed transcriptome profiling, RNAseq, and micro-RNA expression studies in cotton (*Gossypium* species), whereas Younis et al. present a chapter on skin microbiome, transcriptome, and microarray data analyses. In this book, readers can find an interesting chapter on bioinformatics challenges and tools for Hepatitis B genome analysis written by Bell and Kramvis, which highlight features of this small genome virus for bioinformatics analysis.

Similarly, protein microarrays and high-throughput mass spectrometry require bioinformatics analysis to identify proteins through the complex sequence similarity searches using protein sequence databases [96–103]. Bioinformatics is a great help for analysis of gene regulation through searching and comparing the sequence motifs related to promoters and other regulatory elements. Using bioinformatics tools and sequence motifs/regulatory elements genes can be clustered by function, and the co-expression characteristics can be determined. Examples of such bioinformatics tools include k-means clustering, hierarchical clustering, and

consensus clustering methods such as the Bi-CoPaM, and self-organizing maps (SOMs) that can identify functionally active sequences from very complex microarray datasets [104–107]

Not only just these, bioinformatics plays a major role in data collection of the functional elements of sequenced genomes that use the next-generation DNA-sequencing technologies and genomic tiling arrays. This is best exemplified "Encyclopedia of DNA Elements (ENCODE)" [108] project developed by the National Human Genome Research Institute that describes the functional elements of the human genome. Thanks to bioinformatics and applications of its tools, genomes and genes, and protein sequences of different organisms can be rapidly compared, searched, and interpreted. In addition, mutations can be identified that help to judge and diagnose many complex human and plant diseases, crop traits, and interpret complex evolutionary process, such as genome duplications, polyploidization, adaptation, and speciation.

3. Structural bioinformatics: molecular folding, modeling, and design

One of the widely used applications of bioinformatics is identification of three-dimensional protein structures, molecular modeling, and folding to predict the possible function of proteins or other molecular structures, model behavior of molecules, fold the molecule to its native biologically functional three-dimensional structure, and design biomedical drugs for many complex human diseases. It helps *de novo* protein design, enzyme design, protein-ligand/drug docking, protein-peptide interaction, and structure prediction of biological macromolecules and macromolecular complexes [1, 2, 109].

From the coding DNA sequences, the primary structure of proteins can be easily determined that is vital in understanding the function of the protein(s). Further, based on homology patterns in primary structure of proteins and using homology modeling, important structural formations and interaction sites with other proteins can be determined. This helps to predict reliably the structure of a protein based on known structure of a homologous protein(s). Moreover, the identification of secondary, tertiary, and quaternary structures of proteins is very important to understand the function of proteins. The exact three-dimensional structure is essential for correct function, and a failure to fold into native structure generally produces inactive proteins or misfolded proteins that can be toxic [108]. Bioinformatics of protein folding includes (1) energy landscape of protein folding and (2) modeling of protein folding approaches [12, 13, 109].

One of the freely available and leading web server/stand-alone software tools for automated protein structure prediction and structure-based functional annotation can be exemplified by the "Iterative Threading ASSEmbly Refinement" (I-TASSER), which "first generates full-length atomic structural models from multiple threading alignments and iterative structural assembly simulations followed by atomic-level structure refinement" [110]. Using the I-TASSER, all above-mentioned functional and structural characteristics of proteins, including ligand-binding sites, enzyme commission number, and gene ontology terms can be explored in a comparative scale [110, 111].

Molecular modeling through molecular mechanistic and/or the quantum chemistry approaches is the key bioinformatics approaches to study the behavior of molecules. These are routinely used to investigate the structure, dynamics, surface properties, and thermodynamics of inorganic, biological and polymeric systems. It helps to explore conformational changes associated with biomolecular function, and molecular recognition of proteins, and membrane complexes. The protein folding, identification of catalysis sites of enzymes, and protein stability can be studied using molecular modeling. Vast different bioinformatics tools for modeling of biomolecules and designing are available [110–112]. In this book, the chapter by Leong et al. presents bioinformatics modeling and tools for biological membranes using molecular dynamic simulations, all-atom, united-atom, and coarse-grained membrane models of lipids and proteins. In addition, in this book, by Filntisi et al. a computational method for the generation of antibody-drug through site-specific cysteine conjugation using structural prediction methods based on PDB files of a drug, linker, and antibody. Moreover, Bórquez and González-Billault have presented an interesting chapter on computational algorithms of predicting kinase-substrate relationships in protein kinases; this chapter compares prediction tools and methods and discusses improving substrate prediction with contextual information.

4. Biological networks and system biology

Watts and Strogatz in 1998 [113, 114] and Barabási and Albert in 1999 [115–117] fueled the opinion that complex systems can be viewed as networks where components can be represented as nodes and they are linked through their interactions (i.e., edges). The properties of nodes and edges form the network topology. This approach has widely been applied to many scientific fields including bioinformatics that resulted in construction of large-scale biological networks denoted as "omes" like biome, interactome, microbiome [2, 6].

Above highlighted molecular sequence analysis, prediction and annotation, and molecular modeling-related bioinformatics approaches are also the core for building, organizing, and systematizing biological networks of molecules (e.g., metabolic, protein-protein interactions, etc.), and genetic and biochemical pathways of complex cellular processes. These include reception, signal transduction, and gene regulation and gene co-expression. Such molecular networks integrate many different data types including DNA sequences, regulatory RNA, proteins, secondary metabolites, gene expression data, and other small molecules, which may be all connected physically and functionally. The construction and organization of such physically and functionally connected molecular networks of cellular processes can be achieved only by applying the combination of simulative, iterative, and model-oriented bioinformatics approaches. Such biological networks are useful to analyze and visualize the complex connections of these cellular processes, helping understand other biological networks such as neuronal networks, food webs, between/within-species interaction networks, which are the central component of modern system biology [2, 6]. Examples of "omes"-related networks are the Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCyc database collection, BRaunschweig ENzyme DAtabase (BRENDA), Reactome, Comparative Toxicogenomics Database, and many other [118] biological networks. Some biological network

databases and their utilization in plant genomics/epigenomics have been discussed by the chapters of Sripathi et al. and Rahman et al. in this book.

5. Databases

An organized collection of data is referred to as database that aims to collect schemes, tables, queries, reports, images, and other objects. An access to information in the databases is provided by an integrated set of computer software, which is referred to as a "database management system" (DBMS) [119]. The DBMS allows users to access all of the data contained in the databases. It has general functions for data definition, entry, storage, update, administration, and retrieval of large quantities of information in an organized way that requires modeling (hierarchical and network models), clustering, query languages and query optimization, and visualization algorithms [1, 2, 119].

Development of databases, therefore, is significantly dependent on bioinformatics tools, advances, research, and applications. There is a large number of different types of databases available, which cover all aspects of biological data storage and organization. Some aforementioned databases such as GenBank, EMBL, DDJB belong to primary nucleotide sequence databases. There are meta-databases that incorporate data compiled from multiple other databases such as Entrez, mGen, Metascape, etc. Some others are specialized databases such as those specific to an organism, for example, TAIR, the p53 Knowledgebase (p53), the plant alternative splicing database (PASD); the plant secretome, and subcellular proteome knowledgebase (PlantSecKB) [119]. All databases vary in their data definition, usage, format, and access types. In this book, the chapter by Kadam et al. specifically describes databases and bioinformatics algorithms related to allergen informatics, discussing the concepts of allergen bioinformatics and the key areas for potential development in the allergology, whereas Bell and Kramvis highlight public sequence database for Hepatitis B virus. In this book, readers can find a comprehensive discussion for bioinformatics resources, including databases for plant "omics," written by Rahman et al.

6. Software, analysis tools, services, and workflow

As mentioned above, astronomical accumulation of genomic and proteomic as well as metabolomic data, and their expression profiles and annotation, storage, organization, systematization, and integration into biological networks as well as database systems and their wide utilization by the science research community *a priori* required computer programming algorithms, analysis tools, services, and workflow systems. Therefore, software and analysis tools, and bioinformatics services and workflow have been the main fields and core targets of bioinformatics since its emergence. Because of the contributions of various bioinformatics companies or public institutions, bioinformatics software, and tools started to exist as simple command-line tools, but later improved to more complex graphical programs standalone

packages, and web services. Since development of the first bioinformatics software and analysis tools for molecular sequence evaluations in the early 1980s, many free and open-source software tools have been developed and continue to grow and improve with the advancement made in genomics sciences [2, 120].

The main driving forces for the current and future development of bioinformatics software and tools have been made on the past-decade advances of genome decoding technologies, accumulation of large volume biological data, consequent need for their analyses, as well as advancements of computer technologies, graphics, visualization, and molecular modeling and networking techniques. Moreover, the availability of various open-source codes, shared object models, and community-supported plug-ins has facilitated gathering innovative ideas from the community and performing innovative *in silico* experiments on existing "Big Data." These all-created golden opportunities for all research groups and bioinformatics companies to work, experiment, and develop more new generation of bioinformatics software and tools that are user friendly, capable of performing extended and integrated analysis with better visualization and graphical outputs. The range of open-source software packages includes titles such as UGENE, EMBOSS, GenGIS, GENtle, MOTHUR, BioPerl, PathVisio, BioJava, GenoCAD, Biopython, GeWorkbench, GenomeSpace, Bioclipse, .NET Bio, Apache Taverna, BioJS, Bioconductor, and BioRuby [121, 122].

Development of sharing models and web access tools is also an important bioinformatics objective that allows users to utilize and access bioinformatics tools over the internet and from their computer systems to the main computing resources via servers in other parts of the world. Simple Object Access Protocol (SOAP) [123] and Representational State Transfer (REST) [124–126] are two bioinformatics tools to provide web services. SOAP is a standard-based web service access protocol, originally developed by Microsoft. REST, providing very simple web service access, has been developed to fix the problems with SOAP [127]. Both tools share similarities over the HTTP protocol and have its own issues and challenges, differ in messaging patterns, rules, architecture style, and flexibility. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads [127].

There are several basic bioinformatics services, for example, "Sequence Search Services" (SSS), "Multiple Sequence Alignment" (MSA), and "Biological Sequence Analysis" (BSA) [2, 128]. These web service-based bioinformatics analysis resources represent a collection of standalone or web-based interface data analysis tools as well as integrative, distributed, and extensible bioinformatics workflow management systems (BWMS). The BWMSs are designed specifically to compose and execute a series of interactive computational or data manipulation steps (i.e., a workflow) in a bioinformatics analyses. Such systems provide interactive analysis of biological data, build the specific workflows for the analysis, enable the visualization of the analysis outputs in real time, and simplify the process of sharing and reusing workflows between scientists. Some of the platforms giving this service: Galaxy, UGENE, Taverna, etc. [2, 121]. Several chapters of this book cover bioinformatics software, web-based analysis tools, and bioinformatics services for membrane analysis (see Leong et al.), in plant science and crop genomics (see chapters by Rahman et al. and Sripathi et al.), medicine, viral genome analysis and drug design (see chapters by Younis et al., Bell and Kramvis, and Filntisi et al.).

7. Text mining

Part of objectives in bioinformatics research and application is the utilization of computational algorithms and bioinformatics tools to collect, organize, and structure the growing body of biomedical literature allowing scientists to query, mine, read, and synthesize the specific literature and published articles of their research interest [2–4, 7, 129, 130]. Biomedical literature and text mining, therefore, are very important for scientific development, innovations, and integration and application of discoveries to society through extracting information (EI) and assessing the relationships of publications [3, 4]. Analysis of world literature demonstrates that more than 80% of text data remain unstructured that what makes it challenging to read every paper, resulting in disjointed sub-fields of research [3]. Biomedical literature text mining uses a variety of "text mining & data mining" tools, applying techniques such as data clustering, visualization and navigation, information retrieval, and extraction, and text categorization and summarization [3]. The use of IE and "Natural Language Generation and Understanding" (NLG and NLU) that have tokenizing, morphological or lexical, and syntactic analysis components helps to build structured text, and extract, collect, organize structured information [129, 130]. Pattern recognition and matching such as the recognition of biological abbreviations, terms, and interactions are important methods in text mining [2–4].

8. Education

Advances of life sciences and high-throughput biology fields in particular "omics" disciplines, the scale, and complexity of "Big Data," and growing demand for specialists with multilingual and cross-field expertise to understand and solve multidisciplinary scientific concepts and tasks underlie a great need for training and education in the field of bioinformatics. Bioinformatics training and education aim to create, collect, deliver, and share educational and training materials and techniques as well as develop university degree-program curricula on bioinformatics. This is to prepare scientists and specialists, who can utilize modern bioinformatics tools with the sophisticated operating systems, software and algorithms, and database/networking technologies to handle, analyze, interpret, and publish high-throughput complex biological data. This is a great bottleneck and critical need of current life sciences and bioinformatics field, especially in all developing countries, for example, analyzed by some recent reports for African [82] and Central American [131] countries.

To address this, bioinformatics research community has put specific efforts to develop local and global platforms for bioinformatics training and education. Such examples include "Bioinformatics Training Network" (BTN) [132] and "The Global Organization for Bioinformatics Learning, Education, and Training" (GOBLET) [133] that provide a community educational and training resource for bioinformatics trainers and trainees. As an outcome of European 7th Framework grant, BTN targeted to develop and share educational materials, short courses, and training delivery methods as well as discuss the challenge, issues, and needed requirements for bioinformatics training [132]. Furthermore, GOBLET continues

similar efforts beyond Europe, aiming to coordinate efforts at the global scales with concentrated strategy and within the frame of single, dedicated foundation although it requires much time, focused strategic efforts, and modern innovative approaches [133].

"The Swiss Institute of Bioinformatics" training portal [134] also provides online courses for software platforms designed to teach bioinformatics concepts and methods including Rosalind [135]. There are open-access website videos and slides from the "Canadian Bioinformatics Workshops" [136]. Similarly, many different, large bioinformatics conferences, and seminars contribute for training and education on bioinformatics such as Intelligent Systems for Molecular Biology (ISMB), European Conference on Computational Biology (ECCB), Research in Computational Molecular Biology (RECOMB), and the annual Bioinformatics Open Source Conference (BOSC) of the non-profit Open Bioinformatics Foundation [2, 128]. As public bioinformatics databases, the MediaWiki engine with the WikiOpener extension, extensively referenced in this chapter, also contributes for training and education of bioinformatics through gathering research materials and descriptions of tools that can be accessed and updated by all experts in the field [128].

With the specific objectives to develop bioinformatics research and application, its integration to genomics research, and training and education as well as to prepare well-qualified new generation scientists to life sciences, we established a dedicated organization—Center of Genomics and Bioinformatics in the developing country Uzbekistan [137]. As in other developing countries, there are many challenges and limitations in funding and in accessing to sophisticated bioinformatics tools and computer operating systems as well as lack of sufficient experience to carry bioinformatics research and resource development. However, our first step goal is to integrate genomics and bioinformatics curricula to the higher education system of Uzbekistan, develop training and educational materials, provide basic training and research practices to the university students and biology field specialists, and establish international collaborations on this direction. The long-term objective is to efficiently and broadly apply genomics and bioinformatics approaches to all areas of life sciences in national and regional levels that would contribute the development of biological sciences in Central Asia. Some efforts are ongoing regarding the establishment of international collaborations [138] and providing training and education in both national and regional levels.

9. Conclusions and future perspectives

Bioinformatics has become an essential interdisciplinary scientific field to the life science helping to "omics" field and technologies and mainly handling and analyzing "omes" data. Accumulation of high-throughput biological data due to the technological advances in "omics" fields required and prioritized the use of bioinformatics resources, and research and application for the analysis of complex and even further enlarging "Big Data" volumes, which would be impractical and useless without bioinformatics. Therefore, as highlighted herein, there is a critical need for the preparation of well-qualified, new generation scientists with integrated knowledge, multilingual ability, and cross-field experience who are capable of using sophis-

ticated operating systems, software and algorithms, and database/networking technologies to handle, analyze, and interpret high-throughput and increasing volume of complex biological data.

Community resources and a globally coordinated foundation of bioinformatics training and education platforms as well as research conferences, workshops, short online training, and web-based educational courses and materials are available to accomplish toward this goal. However, there is an urgent need for the development of bioinformatics education and training, in particular in developing countries, which requires innovative platforms, training techniques, better funding, web and network access, and high-performance computing systems.

In the research side, bioinformatics tools need to be improved for analysis of the growing body of high-throughput pangenomics, metagenomics, proteomics, and metabolomics data. There are needs for "effective tools" to perform better genome assembly and annotation with high accuracy; however, it requires the improvement of quality of sequenced genomes without gaps, and sequencing of more genome representatives, sub-genomes, polyploidy species, genomes of single cells, and specific tissues that would generate information to work, modify, and correct bioinformatics algorithms and programming approaches.

The use of third generation sequencing approaches and platforms as well as efforts on whole genome sequencing of, for example, 1000 or 100,000 human genome representatives [82] or transcriptome/exon sequencing of 1000 distinct plant species (e.g., 1KP) [139] will ultimately improve and advance the bioinformatics analysis tools. These efforts also help to improve orthologous gene identification tools that currently need attention [120]. There is a great need for sampling and handling diverse strains in pangenomic analysis, integration of prokaryotic genome-organization frameworks (GOFs) as well as integration of non-coding RNAs, pseudogenes, and epigenetics elements into the bioinformatics annotation and ontology tools and software [120]. There is a need to make sequenced genome data more functional and integrated through the construction of more organized, user friendly, cell-wide biological networks, and metabolic pathways [140] with better visualization effects, graphics outputs [120], and knowledge base construction (KB) [141]. This, however, requires the development of real-time imaging systems and high throughput phenotyping (referred to as "phenomics") tools that would help for efficiently determining biologically meaningful associations between genomic and phenotypic data, advancing the translational sciences, personal genomics, and personalized medicine [7] and/or agriculture [142].

Acknowledgements

I thank Academy of Sciences of Uzbekistan and Committee for Coordination Science and Technology Development of Uzbekistan, the Office of International Research Programs (OIRP) of the United States Department of Agriculture (USDA)—Agricultural Research Service (ARS) and U.S. Civilian Research & Development Foundation (CRDF) for research Grants FA-F5-T030, FA-A6-T081, FA-A6-T085, I-2015-6-15/2, I5-FQ-0-89-870, P120, P120A, P121, P121B, UZB-

TA-31016, UZB-TA-31017, and UZB-TA-2992, which have been the key factors for development of plant genomics and bioinformatics in Uzbekistan. I greatly acknowledge the Uzbekistan government support and investments/guide from Academy of Sciences of Uzbekistan, Ministry of Agriculture and Water Resources of Uzbekistan, Cotton Industry Joint Stock Company of Uzbekistan, Ministry Foreign Economic Relations, Investments and Trade of Uzbekistan, USDA-ARS, and Texas A&M University for establishment of Center of Genomics and Bioinformatics in Uzbekistan. I also thank Prof. Gilbert S. Omenn, Center for Computational Medicine & Bioinformatics, University of Michigan, USA for critical reading of this introductory chapter, and Mr. Mirzakamol Ayubov and Mr. Muhammad Mirzahmedov, Center of Genomics and Bioinformatics, Uzbekistan, for their technical assistance while preparing this chapter material.

Author details

Ibrokhim Y. Abdurakhmonov

Address all correspondence to: ibrokhim.abdurakhmonov@genomics.uz and genomics@uzsci.net

Center of Genomics and Bioinformatics, Academy of Science of the Republic of Uzbekistan, Tashkent, Uzbekistan

References

[1] Mount WD. Bioinformatics: sequence and genome analysis. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 2004. 692 p. doi:10.1086/431054

[2] Bioinformatics [Internet]. 2016. https://en.wikipedia.org/wiki/Bioinformatics. Accessed: 2016-03-10

[3] Vijaya S, Radha R. Text mining in biosciences-a review. International Journal of Scientific & Engineering Research. 2015;6:769–776.

[4] Raza K. Application of data mining in bioinformatics. Indian Journal of Computer Science and Engineering.2010;1:114–118.

[5] Shah VA, Rathod DN, Basuri T, Modi VS, Parmar IJ. Applications of bioinformatics in pharmaceutical product designing: a review. World Journal of Pharmacy and Pharmaceutical Sciences. 2015;4:477–493.

[6] Ma'ayan A. Introduction to network analysis in systems biology. Science Signaling. 2011;4:tr5. doi:10.1126/scisignal.2001965

[7] Soualmia LF, Lecroq T. Bioinformatics methods and tools to advance clinical care. Yearbook of medical informatics. 2015;10:170–173. doi:10.15265/IY-2015-026

[8] Altman RB, Miller KS. 2010 Translational bioinformatics year in review. Journal of the American Medical Informatics Association. 2011;18:358–366. doi:10.1136/amiajnl-2011-000328

[9] Saxena A., Soni B.P., Gupta V. A chronological review and comparison of four evolutionary based algorithms. European Journal of Advances in Engineering and Technology, 2015;2:35–41.

[10] Ezziane Z. Applications of artificial intelligence in bioinformatics: A review. Expert Systems with Applications. 2006;30:2–10. doi:10.1016/j.eswa.2005.09.042

[11] Upton A, Trelles O, Cornejo-GarcíaJA, Perkins JR. Review: High-performance computing to detect epistasis in genome scale data sets. Briefings in Bioinformatics. 2015;pii: 1–12. DOI: 10.1093/bib/bbv058

[12] Molecular Modelling [Internet]. 2016. https://en.wikipedia.org/wiki/Molecular_modelling. Accessed: 2016-03-10

[13] Protein Folding [Internet]. 2016. https://en.wikipedia.org/wiki/Protein_folding. Accessed: 2016-03-10

[14] Hesper B, Hogeweg P. Bioinformatica: een werkconcept. Kameleon. 1970;1:28–29 (in Dutch).

[15] Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Computational Biology.2011;7:e1002021. doi:10.1371/journal.pcbi.1002021

[16] Hogeweg P. Simulating the growth of cellular forms. Simulation. 1978;31:90–96. doi: 10.1177/003754977803100305

[17] Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The identification of lower peptides from partial hydrolysates. Biochemical Journal. 1951;49:463–481. doi:10.1042/bj0490463

[18] Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochemical Journal. 1951;49:481–490. doi:10.1042/bj0490481

[19] Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, KourtesisP, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC. The protein information resource. Nucleic Acids Research. 2003;31:345–347. doi:10.1093/nar/gkg040

[20] Protein Information Resources (PIR) [Internet]. 2016. http://pir.georgetown.edu. Accessed: 2016-03-10

[21] Dayhoff M, Eck R. Atlas of protein sequence and structure 1967–1968. Maryland (Silver Spring): National Biomedical Research Foundation; 1968. 356 p.

[22] Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science. 1966;152:363–366. doi:10.1126/science.152.3720.363

[23] Moody, G. Digital code of life: how bioinformatics is revolutionizing science, medicine, and business. Chichester: Wiley; 2004. 400 p.

[24] Johnson G, Wu TT. Kabat Database and its applications: 30 years after the first variability plot. Nucleic Acids Research. 2000;28:214–218. doi:10.1093/nar/28.1.214

[25] GenBank [Internet]. 2016. http://www.ncbi.nlm.nih.gov/genbank. Accessed: 2016-03-10

[26] The European Molecular Biology Laboratory [Internet]. 2016. http://www.embl.org. Accessed: 2016-03-10

[27] DataBank of Japan [Internet]. 2016. http://www.ddbj.nig.ac.jp. Accessed: 2016-03-10

[28] The National Center of Biotechnology Information (NCBI) [Internet]. 2016. http://www.ncbi.nlm.nih.gov. Accessed: 2016-03-10

[29] Gibbs AJ, McIntyre GA. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. European Journal of Biochemistry. 1970;16:1–11. doi:10.1111/j.1432-1033.1970.tb01046.x

[30] Pearson WR, Miller W. Dynamic programming algorithms for biological sequence comparison. Methods in Enzymology. 1992;210:575–601. doi:10.1016/0076-6879(92)10029-D

[31] Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981;147:195–197. doi:10.1016/0022-2836(81)90087-5

[32] Johnson MS, Doolittle RF. A method for the simultaneous alignment of three or more amino acid sequences. Journal of Molecular Evolution. 1986;23:267–278 doi:10.1007/BF02115583

[33] 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;22:4673–4680. doi:10.1093/nar/22.22.4673

[34] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology. 2000;302:205–217. doi:10.1006/jmbi.2000.4042

[35] Rius J, Cores F, Solsona F, van Hemert JI, Koetsier J, Notredame C. A user-friendly web portal for T-Coffee on supercomputers. BMC Bioinformatics. 2011;12:150. doi:10.1186/1471-2105-12-150

[36] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research. 1981;9:133–148. doi: 10.1093/nar/9.1.133

[37] Jacobson AB, Good L, Simonetti J, Zuker M. Some simple computational methods to improve the folding of large RNAs. Nucleic Acids Research. 1984;12:45–52. doi:10.1093/nar/12.1Part1.45

[38] Felsenstein J. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics. 1988;22:521–565. doi:10.1146/annurev.ge.22.120188.002513

[39] Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proceedings of the National Academy of Sciences of the USA. 1996;93:9188–9193. doi:10.1073/pnas.93.17.9188

[40] Barker WC, Dayhoff MO. Viralsrc gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase. Proceedings of the National Academy of Sciences of the USA.1982;79:2836–2839. doi:10.1073/pnas.79.9.2836

[41] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the USA. 1988;85:2444–2448. doi:10.1073/pnas.85.8.2444

[42] Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods in Molecular Biology. 2000;132:185–219. doi: 10.1385/1-59259-192-2:185

[43] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403–410. doi:10.1016/S0022-2836(05)80360-2

[44] Pearson WR. BLAST and FASTA similarity searching for multiple sequence alignment. Methods in Molecular Biology. 2014;1079:75–101. doi:10.1007/978-1-62703-646-7_5

[45] Altschul SF, Madden TL, Schäffer 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;25:3389–3402. doi:10.1093/nar/25.17.3389

[46] Ladunga I, Wiese BA, Smith RF. FASTA-SWAP and FASTA-PAT: pattern database searches using combinations of aligned amino acids, and a novel scoring theory. Journal of Molecular Biology. 1996;259:840–854. doi:10.1006/jmbi.1996.0362

[47] Worley KC, Wiese BA, Smith RF. BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Research. 1995;5:173–184. doi:10.1101/gr.5.2.173

[48] Tsunoda T, Takagi T. Estimating transcription factor bindability on DNA. Bioinformatics. 1999;15:622–630. doi:10.1093/bioinformatics/15.7.622

[49] Loh SK, Low ST, Mohamad MS, Deris S, Kasim S, Wen CY, Wardani AK. A review of software for predicting gene function. International Journal of Bio-Science and Bio-Technology. 2015;7:57–70. doi:10.14257/ijbsbt.2015.7.2.06

[50] Höglund A, Kohlbacher O. From sequence to structure and back again: approaches for predicting protein-DNA binding. Proteome Sciences. 2004;2:3. doi: 10.1186/1477-5956-2-3

[51] Liu Y, Wei L, Batzoglou S, Brutlag DL, Liu JS, Liu XS. A suite of web-based programs to search for transcriptional regulatory motifs. Nucleic Acids Research. 2004;32:204–207. doi:10.1093/nar/gkh461

[52] Nagarajan V, Elasri MO. Structure and function predictions of the Msa protein in Staphylococcus aureus. BMC Bioinformatics. 2007;8:S5. doi:10.1186/1471-2105-8-S7-S5

[53] Pavlopoulou A, Michalopoulos I. State-of-the-art bioinformatics protein structure prediction tools. International Journal of Molecular Medicine. 2011;28:295–310. doi: 10.3892/ijmm.2011.705

[54] Ma X, Guo J, Liu HD, Xie JM, Sun X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012;9:1766–1775. doi: 10.1109/TCBB.2012.106

[55] Lv H, Han J, Liu J, Zheng J, Zhong D, Liu R. ISDTool: a computational model for predicting immunosuppressive domain of HERVs. Computational Biology and Chemistry. 2014;49:45–50. doi:10.1016/j.compbiolchem.2014.02.001

[56] Tuvshinjargal N, Lee W, Park B, Han K. Predicting protein-binding RNA nucleotides with consideration of binding partners. Computer Methods and Programs in Biomedicine. 2015;120:3–15. doi:10.1016/j.cmpb.2015.03.010

[57] Fujimoto MS, Suvorov A, Jensen NO, Clement MJ, Bybee SM. Detecting false positive sequence homology: a machine learning approach. BMC Bioinformatics. 2016;17:101. doi:10.1186/s12859-016-0955-3

[58] Reddy TBK, Thomas A, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos E and Kyrpides N. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta) genome project classification. Nucleic Acids Research. 2014;43:D1099–106. doi:10.1093/nar/gku950

[59] Lin E, Tsai SJ. Genome-wide microarray analysis of gene expression profiling in major depression and antidepressant therapy. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2016;64:334–40. doi:10.1016/j.pnpbp.2015.02.008

[60] Lim do H, Kim WS, Kim SJ, Yoo HY, Ko YH. Microarray Gene-expression profiling analysis comparing PCNSL and non-CNS diffuse large B-cell lymphoma. Anticancer Research. 2015;35:3333–3340.

[61] Kim HS, Lee NK. Gene expression profiling in osteoclast precursors by insulin using microarray analysis. Molecules and Cells. 2014;37:827–832. doi:10.14348/molcells.2014.0223

[62] Xing C, Zhang R, Cui J, Li Y, Li G, Yang Y, Pang L, Ruan X, Li J. Pathway crosstalk analysis of non-small cell lung cancer based on microarray gene expression profiling. Tumori. 2015;101:111–116. doi:10.5301/tj.5000225

[63] Steiger S, Perez-Fons L, Cutting SM, Fraser PD, Sandmann G. Annotation and functional assignment of the genes for the C30 carotenoid pathways from the genomes of two bacteria: *Bacillus indicus* and *Bacillus firmus*. Microbiology. 2015;161:194–202. doi:10.1099/mic.0.083519-0

[64] Chitale M, Palakodety S, Kihara D. Quantification of protein group coherence and pathway assignment using functional association. BMC Bioinformatics. 2011;12:373. doi:10.1186/1471-2105-12-373

[65] Haw R, Hermjakob H, D'Eustachio P, Stein L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. Proteomics. 2011;11:3598–3613. doi:10.1002/pmic.201100066

[66] Seoane JA, Day IN, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014;30:838–45. doi:10.1093/bioinformatics/btt610

[67] Veldhoven A, de Lange D, Smid M, de Jager V, Kors JA, Jenster G. Storing, linking, and mining microarray databases using SRS. BMC Bioinformatics. 2005;6:192. doi:10.1186/1471-2105-6-192

[68] Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Mueller W, Goble C. SEEK: a systems biology data and model management platform. BMC Systems Biology. 2015;9:33. doi:10.1186/s12918-015-0174-y

[69] Cremaschi P, Carriero R, Astrologo S, Colì C, Lisa A, Parolo S, Bione S. An association rule mining approach to discover lncRNAs expression patterns in cancer datasets. BioMed Research International. 2015;2015:146250. doi:10.1155/2015/146250

[70] Cheng CP, DeBoever C, Frazer KA, Liu YC, Tseng VS. MiningABs: mining associated biomarkers across multi-connected gene expression datasets. BMC Bioinformatics. 2014;15:173. doi:10.1186/1471-2105-15-173

[71] Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics. 2014;15:S11. doi:10.1186/1471-2105-15-S11-S11

[72] Wang M, Shang X, Li X, Liu W, Li Z. Efficient mining differential co-expression biclusters in microarray datasets. Gene. 2013;518:59–69. doi:10.1016/j.gene.2012.11.085.

[73] Wu WS, Wang CC, Jhou MJ, Wang YC. YAGM: a web tool for mining associated genes in yeast based on diverse biological associations. BMC Systems Biology. 2015;9:S1. doi: 10.1186/1752-0509-9-S6-S1

[74] PubMed database [Internet]. 2015. http://www.ncbi.nlm.nih.gov/pubmed. Accessed from 2016-03-10

[75] Sanger F, Coulson AR. A rapid method for determining sequences in DNA byprimed synthesis with DNA polymerase. Journal of Molecular Biology. 1975;94:441–448. doi: 10.1258/jrsm.99.2.62

[76] Maxam AM, Gilbert W. A new method for sequencing DNA. Proceeding of National Academy Sciences of USA. 1977;74:560–564. doi:10.1073/pnas.74.2.560

[77] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265:687–695. doi:10.1038/265687a0

[78] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269:496–512. doi:10.1126/science. 7542800

[79] Oliver GR, Hart SN, Klee EW. Bioinformatics for clinical next generation sequencing. Clinical Chemistry. 2015;61:124–135. doi:10.1373/clinchem.2014.224360

[80] Krampis K, Wultsch C. A review of cloud computing bioinformatics solutions for next-gen sequencing data analysis and research. Methods in Next Generation Sequencing. 2015;2:2084–7173, DOI: 10.1515/mngs-2015-0003

[81] The Genomes OnLine Database (GOLD) [Internet]. 2016. https://gold.jgi.doe.gov/ index. Accessed: 2016-03-08

[82] Karikari TK. Bioinformatics in Africa: The Rise of Ghana? PLoS Computational Biology. 2015;11:e1004308. doi:10.1371/journal.pcbi.1004308.

[83] Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, Turuspekov Y, Walia H, Akhunov ED, Kianian P, Otto C, Simons K, Deal KR, Echenique V, Stamova B, Ross K, Butler GE, Strader L, Verhey SD, Johnson R, Altenbach S, Kothari K, Tanaka C, Shah MM, Laudencia-Chingcuanco D, Han P, Miller RE, Crossman CC, Chao S, Lazo GR, Klueva N, Gustafson JP, Kianian SF, Dubcovsky J, Walker-Simmons MK, Gill KS, Dvorák J, Anderson OD, Sorrells ME, McGuire PE, Qualset CO, Nguyen HT, Close TJ. Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum* L.). Genetics. 2004;168:595–608. doi:10.1534/ genetics.104.034785

[84] Henry RJ, Edwards M, Waters DL, Gopala Krishnan S, Bundock P, Sexton TR, Masouleh AK, Nock CJ, Pattemore J. Application of large-scale sequencing to marker discovery in plants. Journal of Biosciences. 2012;37:829–41. doi:10.1007/s12038-012-9253-z

[85] Uenishi H, Morozumi T, Toki D, Eguchi-Ogawa T, Rund LA, Schook LB. Large-scale sequencing based on full-length-enriched cDNA libraries in pigs: contribution to annotation of the pig genome draft sequence. BMC Genomics. 2012;13:581. doi: 10.1186/1471-2164-13-581

[86] Rai KM, Singh SK, Bhardwaj A, Kumar V, Lakhwani D, Srivastava A, Jena SN, Yadav HK, Bag SK, Sawant SV. Large-scale resource development in *Gossypium hirsutum* L. by 454 sequencing of genic-enriched libraries from six diverse genotypes. Plant Biotechnology Journal. 2013;11:953–963. doi:10.1111/pbi.12088

[87] Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008;92:255–64. DOI:10.1016/j.ygeno.2008.07.001

[88] Archer SK, Shirokikh NE, Preiss T. Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage. BMC Genomics. 2014;15:401. doi: 10.1186/1471-2164-15-401

[89] Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. Large-scale genomic sequencing of extra intestinal pathogenic *Escherichia coli* strains. Genome Research. 2015;25:119–128. doi:10.1101/gr.180190.114

[90] Hou Z, Jiang P, Swanson SA, Elwell AL, Nguyen BK, Bolin JM, Stewart R, ThomsonJA. A cost-effective RNA sequencing protocol for large-scale gene expression studies. Scientific Reports. 2015;5:9570. doi:10.1038/srep09570

[91] Hough CD, Sherman-Baust CA, Pizer ES, Montz FJ, Im DD, Rosenshein NB, Cho KR, Riggins GJ, Morin PJ. Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. Cancer Research. 2000;60:6281–6287

[92] El-Meanawy MA, Schelling JR, Pozuelo F, Churpek MM, Ficker EK, Iyengar S, Sedor JR. Use of serial analysis of gene expression to generate kidney expression libraries. American Journal of Physiology Renal Physiology. 2000;279:383–392.

[93] Zhang XL, Gao F, Li W, Tang WZ, Zhang S. Serial analysis of gene expression in adenocarcinoma samples and normal colonic mucosa in a Chinese population. Genetics and Molecular Research. 2015;14:12903–12911. doi:10.4238/2015.October.21.11

[94] Han X, Wei YB, Tian G, Tang Z, Gao JY, Xu XG. Screening of crucial long non-coding RNAs in oral epithelial dysplasia by serial analysis of gene expression. Genetics and Molecular Research. 2015;14:11729–11738. doi:10.4238/2015.October.2.6

[95] Mackintosh CG, Griffin JF, Scott IC, O'Brien R, Stanton JL, MacLean P, Brauning R. SOLiD SAGE sequencing shows differential gene expression in jejunal lymph node samples of resistant and susceptible red deer (*Cervus elaphus*) challenged with Mycobacterium avium subsp. paratuberculosis. Veterinary Immunology and Immunopathology. 2016;169:102–110. doi:10.1016/j.vetimm.2015.10.009

[96] Loo JA, Brown J, Critchley G, Mitchell C, Andrews PC, Ogorzalek Loo RR. High sensitivity mass spectrometric methods for obtaining intact molecular weights from gel-separated proteins. Electrophoresis. 1999;20:743–748.

[97] Figeys D, Pinto D. Proteomics on a chip: promising developments. Electrophoresis. 2001;22:208–216.

[98] Wark AW, Lee HJ, Corn RM. Multiplexed detection methods for profiling microRNA expression in biological samples. Angewandte Chemie International Edition in English. 2008;47:644–652. doi:10.1002/anie.200702450

[99] Tom I, Lewin-Koh N, Ramani SR, Gonzalez LC. Protein microarrays for identification of novel extracellular protein-protein interactions. Current Protocol in Protein Sciences. 2013;Chapter 27:Unit 27.3. doi:10.1002/0471140864.ps2703s72.

[100] McKee CJ, Hines HB, Ulrich RG. Analysis of protein tyrosine phosphatase interactions with micro arrayed phosphopeptide substrates using imaging mass spectrometry. Analytical Biochemistry. 2013;442:62–67. doi:10.1016/j.ab.2013.07.031

[101] Choi HM, Beck VA, Pierce NA. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. ACS Nano. 2014;8:4284–4294. doi:10.1021/nn405717p

[102] Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW. Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. J. Proteome Res. 2015;14:3452–3460. doi:10.1021/acs.jproteome.5b00499

[103] Strack R. Highly multiplexed transcriptome imaging. Nature Methods. 2015;12:486–487

[104] Abu-Jamous B, Fa R, Roberts DJ, Nandi AK. Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery. PLoS One. 2013;8:e56432. doi:10.1371/journal.pone.0056432

[105] Lord E, Diallo AB, Makarenkov V. Classification of bioinformatics workflows using weighted versions of partitioning and hierarchical clustering algorithms. BMC Bioinformatics. 2015;16:68. doi:10.1186/s12859-015-0508-1

[106] Chen GK, Chi EC, Ranola JM, Lange K. Convex clustering: an attractive alternative to hierarchical clustering. PLoS Computational Biology. 2015;11:e1004228. doi:10.1371/journal.pcbi.1004228

[107] Bouvier G, Desdouits N, Ferber M, Blondel A, Nilges M. An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. Bioinformatics. 2015;31:1490–1492. doi:10.1093/bioinformatics/btu849

[108] Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe G, Podduturi NR, Tanaka

F, Hong EL, Cherry JM. ENCODE data at the ENCODE portal. Nucleic Acids Research. 2016;44:726–732. doi:10.1093/nar/gkv1160

[109] Selkoe DJ. Folding proteins in fatal ways. Nature. 2013;426:900–904. doi:10.1038/nature02264

[110] Yang J, Zhang Y. Protein structure and function prediction using I-TASSER. Current Protocols in Bioinformatics. 2015; 52:5.8.1–5.8.15. doi:10.1002/0471250953.bi0508s52

[111] I-TASSER [Internet]. 2016. http://zhanglab.ccmb.med.umich.edu/I-TASSER. Accessed: 2016-04-02

[112] List of software for molecular mechanics modeling [Internet]. 2016. https://en.wikipedia.org/wiki/List_of_software_for_molecular_mechanics_modeling. Accessed: 2016-02-11

[113] de Moura AP. Thin Watts-Strogatz networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics. 2006;73:016110. doi:10.1103/PhysRevE.73.016110

[114] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393:440–442. doi:10.1038/30918

[115] Campbell C, Shea K, Albert R. Network models. Comment on control profiles of complex networks. Science. 2014;346:561. doi:10.1126/science.1256492

[116] Albert R, Barabasi AL. Topology of evolving networks: local events and universality. Physical Review Letters. 2000;85:5234–5237. doi:10.1103/PhysRevLett.85.5234

[117] Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286:509–12. doi:10.1126/science.286.5439.509

[118] Biological pathway [internet]. 2016. https://en.wikipedia.org/wiki/Biological_pathway. Accessed: 2016-03-11

[119] Database and list of biological databases [Internet]. 2016. https://en.wikipedia.org/wiki/Database; https://en.wikipedia.org/wiki/List_of_biological_databases. Accessed: 2016-03-12

[120] Xiao J, Zhang Z, Wu J, Yu J. A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics. 2015;13:73–76. doi:10.1016/j.gpb.2015.01.007

[121] List of open-source bioinformatics software [Internet]. 2016. https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software. Accessed: 2016–03-11

[122] Stajich JE, Lapp H. Open source tools and toolkits for bioinformatics: significance, and where are we? Briefings in Bioinformatics. 2006;7:287–296. doi:10.1093/bib/bbl026

[123] Sugawara H, Miyazaki S. Biological SOAP servers and web services provided by the public sequence data bank. Nucleic Acids Research. 2003;31:3836–3839. doi:10.1093/nar/gkg558

[124] Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GR, Ruffier M, Taylor K, Vullo A, Flicek P. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics. 2015;31:143–145. doi:10.1093/bioinformatics/btu613

[125] Genereaux BW, Dennison DK. REST enabling the report template library. Journal of Digital Imaging. 2014;27:331–336. doi:10.1007/s10278-013-9668-6

[126] Sundvall E, Nyström M, Karlsson D, Eneling M, Chen R, Örman H. Applying representational state transfer (REST) architecture to archetype-based electronic health record systems. BMC Medical Informatics and Decision Making. 2013;13:57. doi: 10.1186/1472-6947-13-57

[127] Understanding SOAP and REST Basics and Differences [Internet]. 2016. http://blog.smartbear.com/apis/understanding-soap-and-rest-basics. Accessed; 2016–03-11

[128] Open Bioinformatics Foundation [Internet]. 2016. https://en.wikipedia.org/wiki/Open Bioinformatics Foundation. Accessed: 2016-03-11

[129] Coulet A, Garten Y, Dumontier M, Altman RB, Musen MA, Shah NH. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. J Biomed Semantics. 2011;2 Suppl 2:S10. doi:10.1186/2041-1480-2-S2-S10

[130] Ahmed A, Xing EP, Cohen WW, Murphy RF. Structured correspondence topic models for mining captioned figures in biological literature. KDD. 2009;2009:39–48. doi: 10.1145/1557019.1557031

[131] Orozco A, Morera J, Jiménez S, Boza R. A review of bioinformatics training applied to research in molecular medicine, agriculture and biodiversity in Costa Rica and Central America. Briefings in Bioinformatics. 2013;14:661–670. doi:10.1093/bib/bbt033

[132] Schneider MV, Walter P, Blatter MC, Watson J, Brazas MD, Rother K, Budd A, Via A, van Gelder CW, Jacob J, Fernandes P, Nyrönen TH, De Las Rivas J, Blicher T, Jimenez RC, Loveland J, McDowall J, Jones P, Vaughan BW, Lopez R, Attwood TK, Brooksbank C. Bioinformatics Training Network (BTN): a community resource for bioinformatics trainers. Briefings in Bioinformatics. 2012;13:383–389. doi:10.1093/bib/bbr064

[133] Attwood TK, Bongcam-Rudloff E, Brazas ME, Corpas M, Gaudet P, Lewitter F, Mulder N, Palagi PM, Schneider MV, van Gelder CW; GOBLET Consortium. GOBLET: the Global Organisation for Bioinformatics Learning, Education and Training. PLoS Computational Biology. 2015;11:e1004143. doi:10.1371/journal.pcbi.1004143

[134] The Swiss Institute of Bioinformatics training portal [internet]. 2016. http://www.isb-sib.ch/training. Accessed: 2016-03-11

[135] Nunes R, Barbosa de Almeida Júnior E, Pessoa Pinto de Menezes I, Malafaia G. Learning nucleic acids solving by bioinformatics problems. Biochemistry and Molecular Biology Education. 2015;43:377–383. doi:10.1002/bmb.20886

[136] Canadian Bioinformatics Workshops [Internet]. 2016. http://bioinformatics.ca. Accessed: 2016-03-11

[137] The Center of Genomics and Bioinformatics, Academy of Sciences of Uzbekistan [Internet]. 2016. http://en.genomics.uz. Accessed: 2016-03-12

[138] Uzbekistan-US Life Sciences Report [Internet]. 2016. http://www.aaas.org/sites/default/files/migrate/uploads/Uzbekistan-US-Life-Sciences-Report_2013.pdf. Accessed: 2016-03-12

[139] Leebens-Mack J, OneKp Capstone Wiki [Internet]. 2015. https://pods.iplantcollaborative.org/wiki/display/iptol/OneKP+Capstone+Wiki. Accessed: 2015-11-10

[140] Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. BMC Bioinformatics. 2006;7:168

[141] Kim JD, Kim JJ, Han X, Rebholz-Schuhmann D. Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task. BMC Bioinformatics. 2015;16 Suppl 10:S3. doi:10.1186/1471-2105-16-S10-S3.

[142] Stokes ME, McCourt P. Towards personalized agriculture: what chemical genomics can bring to plant biotechnology. Frontiers in Plant Science. 2014;5:344. doi:10.3389/fpls.2014.00344

IntechOpen

IntechOpen

IntechOpen