American Journal of Molecular Biology, 2013, 3, 115-130 **AJMB**
http://dx.doi.org/10.4236/ajmb.2013.32016 Published Online April 2013 (http://www.scirp.org/journal/ajmb/)

Genome sequencing and next-generation sequence data analysis: A comprehensive compilation of bioinformatics tools and databases

Jose C. Jimenez-Lopez[1], Emma W. Gachomo[2,3], Sweta Sharma[2,3], Simeon O. Kotchoni[2,3*]

[1]Department of Biochemistry, Cell and Molecular Biology of Plants, Estacion Experimental del Zaidin, High Council for Scientific Research (CSIC), Granada, Spain
[2]Department of Biology, Rutgers University, Camden, USA
[3]Center for Computational and Integrative Biology (CCIB), Rutgers University, Camden, USA
Email: simeon.kotchoni@rutgers.edu

Received 5 February 2013; revised 30 March 2013; accepted 25 April 2013

ABSTRACT

Genomics has become a ground-breaking field in all areas of the life sciences. The advanced genomics and the development of high-throughput techniques have lately provided insight into whole-genome characterization of a wide range of organisms. In the post-genomic era, new technologies have revealed an outbreak of prerequisite genomic sequences and supporting data to understand genome wide functional regulation of gene expression and metabolic pathways reconstruction. However, the availability of this plethora of genomic data presents a significant challenge for storage, analyses and data management. Analysis of this mega-data requires the development and application of novel bioinformatics tools that must include unified functional annotation, structural search, and comprehensive analysis and identification of new genes in a wide range of species with fully sequenced genomes. In addition, generation of systematically and syntactically unambiguous nomenclature systems for genomic data across species is a crucial task. Such systems are necessary for adequate handling genetic information in the context of comparative functional genomics. In this paper, we provide an overview of major advances in bioinformatics and computational biology in genome sequencing and next-generation sequence data analysis. We focus on their potential applications for efficient collection, storage, and analysis of genetic data/information from a wide range of gene banks. We also discuss the importance of establishing a unified nomenclature system through a functional and structural genomics approach.

*Corresponding author.

Keywords: Databases; Computational Biology; Genomics; Proteomics; Next-Generation Sequencing

1. INTRODUCTION

Information processing by bioinformatics tools and computational biology methods has become essential for solving complex biological problems in genomics, proteomics, and metabolomics. Such methods give new insights into areas such as genome evolution, systems biology, biotechnology, genome deciphering, and developments in medicine.

Bioinformatics is the application of computational tools to predict, manage and interpret biological data [1]. It is one of the essential tools for integrative and multidisciplinary understanding of metabolic network processes in systems biology [2]. For example, understanding -omics data requires both common statistical and computing-based methods due to the multi-dimensional and complexity level of the data.

Generally, there are three central biological principles around which bioinformatics tools must be developed: 1) DNA sequence determines protein sequence, 2) protein sequence determines protein structure, and 3) protein structure determines protein function.

Next Generation Sequencing (NGS) is revolutionizing the study of the genetics of many organisms, with immense biological implications. With the rapid advancement in NGS technologies and the subsequently fast-growing volume of biological data, diverse data sources (databases and web servers) have been developed to facilitate data management, accessibility, and analysis, which will be facilitated by following an adequate and sequential work-flow (**Figure 1**). Using deep sequencing,

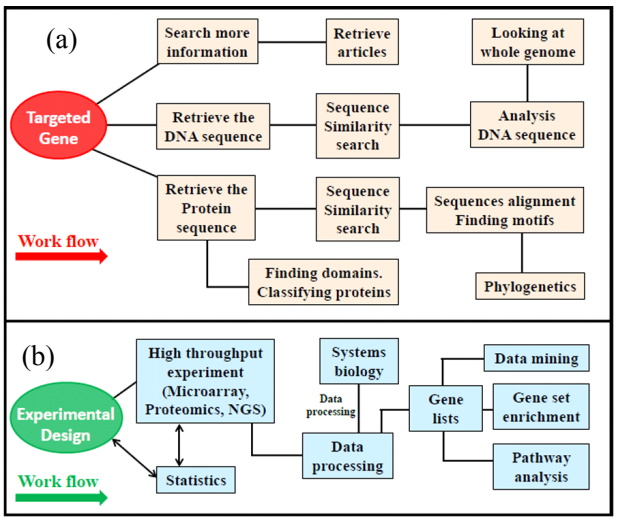

Figure 1. Hypothesis-generating bioinformatics (a) and experimental (b) workflow.

it is now possible to discover novel disease causing mutations [3] and detect traces of pathogenic microorganisms [4]. A single ultra high throughput sequencing run can produce millions of reads in various lengths per experiment [5]. Thus, integration of data from heterogeneous and voluminous data sources is significantly important in order to fully and efficiently exploit the huge and readily available biological data. The storage, processing, querying, parsing, analysis and interpretation of such an incredible amount of data is a significant task that also holds many obstacles [6]. As acquisition of genomic data becomes increasingly cost-efficient, genomic data sets are accumulating at an exponential rate and new types of genetic data are emerging. These come with the inherent challenges of new methods of statistical analysis and modeling. Indeed new technologies are producing data at a rate that outpaces our ability to analyze its biological meaning. Researchers are addressing this challenge by adopting mathematical and statistical software, computer modeling, and other computational and engineering methods. As a result, bioinformatics has become the latest engineering discipline. As computers provide the ability to process the complex models, high-performance computer languages have become a necessity for implementing state-of-the-art algorithms and methods.

Sequencing technologies are evolving rapidly, with an overwhelming increase in efficiency and throughput [5]. For example, the pyrosequencing method can sequence a microbial genome in one hour [7-9]. These improved technologies deploy random fragmentation of the nucleotide sequence of interest in order to increase throughput by simultaneously sequencing millions of fragments. Platforms such as Roche/454 [10], Illumina/Solexa [11], and Life/APG SOLiD [12,13] ligate these fragments with adapters and thereafter amplified using PCR primers. Alternatively, when a high amount of DNA is initially present, platforms such as Pacific Biosciences [14] use the fragments themselves as single molecule templates. The amount of introduced errors is correlated with the fidelity of the polymerase utilized in the reaction [15]. Read lengths vary with the technology, pyrosequencing generating long reads (~400 nts), while reverse termination and sequencing by ligation technologies produce shorter reads. Different technologies can thus result in significantly different output data and performance. The combination of more than one platform is potentially more cost effective and could yield higher fidelity and accuracy [16,17].

2. GENOME SEQUENCING

2.1. Pre-Analysis and Processing of Sequencing Data

To alleviate the above mentioned difficulties in NGS, platform developers should provide end-users with a sequencing quality scale for both automated and manual-based data filtration and refinement.

The most common sequence output format is a FASTA file accompanied by a numerical quality QUAL file, describing the per-base probability of incorrect sequencing based on the PHRED quality score [18]. Quality control of deep sequencing data refers to an overview of the base and quality distribution between lanes, tiles and cycles, and correlation of the initial sequence data with expected length, GC content, ambiguous bases, sequence complexity and alignment of ensuing location distributions which can hold information regarding possible sequencing bias, contamination or artifacts. Platform specific quality control tools and more general quality assessment software [19] can help circumvent such biases. A more common example is sequence duplication, usually an artifact of PCR amplification and other library preparation processes, that cause over-representation of certain sequences. We urge the user to consider sequencing data in the appropriate experimental context. The aforementioned quality control methods should be used prior to downstream analysis to increase the experimental validity and accuracy and, thus, ensure better, more reliable results [20].

2.2. Genomic Annotation

Annotation generates data that allows various types of research on model organisms. After sequencing the organism's genome, sequences must be mapped according to areas pertinent to the research objectives. Gene predictions can be made with computational techniques for recognizing gene sequences, including stop codons and the initial portions of nucleotide sequences. This is known as functional annotation, and can be done initially by com-

puter, using similarity in sequence alignment. However, no software is capable of generating a functional annotation without false positive results. Predicted genes need to be revised manually. When annotation is complete, the genome should subsequently be submitted a public-access site.

Gene Prediction Strategies

Gene prediction programs can be divided into two categories: empirical and *ab initio*. Empirical predictors search for sequence similarity in the genome; they predict genes based on homologies with known databases, such as genomic DNA, cDNA, dbEST and proteins. This approach facilitates the identification of well—conserved exons. Ab initio gene finders use sequence information of signal and content sensors. Usually, these programs are based on hidden Markov models. They can be classified as single, dual and multiple—genome predictors based on the number of genome sequences used in the analysis. Integrated approaches couple the extrinsic methodology of empirical gene—finders and intrinsic ab initio prediction. This technique significantly improves prediction protocols [21]. Gene prediction methodology for eukaryotes involves two distinct aspects. The first focuses on the information such as signal functions in the DNA strand for gene recognition. The second uses algorithms implemented by prediction programs for accurate prediction of gene structure and organization [22]. Unlike eukaryotes, the archaeal, bacterial and virus genomes are highly gene-dense. The protein coding regions usually represent more than 90% of the genome. The simplest approach in gene prediction is to look for Open Reading Frames (ORFs). An ORF is a DNA sequence that initiates at a start codon and ends at a stop codon, with no other intervening stop codon. One way to locate genes is to look for ORFs with the mean size of proteins [21]. Example of tools used for gene prediction are: 1) Glimmer, a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi), 2) FgenesB, a package developed by Softberry Inc. for automatic annotation of bacterial genomes (http://www.molquest.com/help/2.3/programs/FgenesB/about.html), 3) Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) is a microbial (bacterial and archaeal) gene finding program (http://prodigal.ornl.gov/), and 4) GeneMarkTM, a public access program for gene prediction in eukaryotes (http://exon.biology.gatech.edu/).

2.3. Solving the Problem: Biological Patterns

One of the key aspects in the analysis of biological sequences is the identification of interesting patterns [23]. We define an interesting pattern as one which shows an unusual behavior with respect to the sequence under analysis. The search for shared or over-represented patterns is motivated by a simple commonly accepted principle: if two or more sequences perform the same functions or have the same structure, then the common elements among the sequences might be responsible for the observed similarity. To identify biologically significant patterns, we must find those that are statistically significant.

A scoring function to evaluate the output also plays an important role in the identification of the searched patterns. However, traditional statistics are often unable to discriminate interesting motifs from motifs that are likely to occur by chance, necessitating development of different measures of statistical significance [24].

We define a statistical measure (SM) of a motif m, and ask the following three questions: 1) What is the value of this motif's statistical measure SM(m)? 2) How surprising is measure SM(m) with respect to the expected value according to some background distribution? and 3) How likely is it for the recorded values to occur by chance? These three questions can be answered by different computational means. The first one, for example, can be answered by exact counts or estimates. To answer the second, we need a score that measures over-representation, such as the z-score. The third one requires calculation of the p-value of a statistic.

2.4. Data Analysis Pathways and Tools

2.4.1. Alignment of Sequences

Bioinformatics and molecular evolutionary analyses most often start with comparing DNA or amino acid sequences by aligning them. Pairwise alignment measures the similarities between a query sequence and each of those in a database using BLAST search, the most used bioinformatics tools [25,26]. Multiple sequence alignment (MSA) is a useful tool in designing experiments for testing and modifying the function of specific proteins, predicting their functions and structures, and identifying new members of protein families. MSA of DNA, RNA, and protein is one of the most common and important tasks in bioinformatics. To process data efficiently, new software packages and algorithms are continuously being developed to improve protein identification, characterization and quantification in terms of high-throughput and statistical accuracy. In particular, for the analysis of plant proteins extensive data elaboration is necessary due to the lack of structural information in the proteomic and genomic public databases. The high dimensionality of data generated from these studies will require the development of improved bioinformatics tools and data-mining approaches.

When deep sequencing was initially introduced, established alignment tools, suited for the query of a limited

number of sequences, were inadequate for high throughput sequencing data which comprised millions of short fragment sequences. This spurred the design of novel alignment algorithms and tools which use heuristic techniques for alignment of millions of short sequences within an acceptable time requirement [27].

When choosing an alignment tool, one needs to consider some important features including the following: 1) Quality utilization and control—Most alignment software generate the alignment output in the Sequence Alignment Map (SAM) format, with a multitude of supporting downstream analysis tools. Alignment output contains a PHRED based quality score describing the probability of per-base false alignment. These quality scores can be re-assessed using currently available tools [28], 2) Gapped alignment. Alignment tools may or may not use a gap alignment algorithm. When specifically detecting for insertions and deletions (indels) [29] it is highly recommended to choose a tool that implements gapped alignment [30], 3) Mismatches and Gap penalties. Most alignment tools allow the user to set the number of allowed mismatches between the read and a reference location and the scoring scale for gap opening and extension, and 4) Multiple mapping. Usually, a portion of the reads will remain unmapped due to contaminant origin or sequencing errors. More commonly, they will ambiguously map to several different locations (multiple mapping) due to sequence homology and repetitiveness. Of the current approaches for allocation of these multiply mapped reads, one uses probabilistic models such as maximum likelihood to compute the most likely origin of each read. This greatly improves the results of quantitative deep sequencing experiments and differential expression [31].

1) *Assembly*. Assembly refers to the process of piecing together short DNA/RNA sequences into longer ones. These long sequences, called contigs, are then grouped to form scaffolds for computationally reconstructing a sample's genetic component. When the assembly process is performed with the assistance of a reference genome, it is referred to as mapping assembly; if no reference is available it is called de novo assembly.

2) *Variant calling*. This refers to the identification of single nucleotide polymorphisms (SNPs), insertions and deletions (indels), copy number variations (CNVs) and other types of structural variations, e.g. inversions, translocations etc, in a sequenced sample [32]. The process is complicated by areas of low coverage, sequencing errors, misalignment caused by either low complexity and repeat regions or adjacent variants and library preparation biases (e.g. PCR duplicates) [12].

2.4.2. Multiple Sequence Alignment

Evolutionary history among sequences is well reflected through MSA. When building a MSA, it is assumed that the sequences compared are derived from a common ancestral sequence. MSA infers homologous positions between the input sequences and place gaps in the sequences in order to align these positions. Gaps are caused by either insertions or deletions of nucleotides or amino acids on a particular lineage of sequences during the evolution. Some examples of bioinformatics methods that utilize information extracted from MSAs include: profile building in similarity search (PSIBLAST [33]), motif/profile recognition (PROSITE [34]), profile hidden Markov models for protein families/domains (Pfam [35]), and protein secondary-structure prediction [36].

Measuring the quality of MSAs requires a benchmark dataset and a scoring method. Benchmark datasets like OXBench [37], HOMSTRAD [38], PREFAB [39], BaliBASE [40], and SABmark [41] are built on real sequences by aligning structural elements and in some cases with hand-curation. Others like IRMBASE [42] are generated by simulating sequence evolution based on specific molecular evolutionary models.

Visual Inspection of MSAs

Currently, there are multiple MSA tools available, depending on the requirement and specific needs of the user as shown in **Table 1**.

2.5. Genomics

When reduced to its respective base letters (A, T, G, C), the genome sequence represents the unique identifier of biological species. This is a vital mechanism for computer scientists to store and retrieve data using a unique identifier (ID). A user can search and exactly pinpoint a particular gene in a database or flat file using the ID. Identification and classification of sequences led to the annotation of genes and retrieval of meaningful information about their history.

Two diverging paths appeared in the development of bioinformatics in terms of project concepts and organization, the -omics and the bio-. The latter focuses on molecular level resolution, while the focus of the -omics

Table 1. Examples of the most used MSA methods.

Method	Web
ClustalW2	http://www.clustal.org/
MUSCLE	http://www.drive5.com/muscle/
MAFFT	http://mafft.cbrc.jp/alignment/software/
PRALINE	http://www.ibi.vu.nl/programs/pralinewww/
PRANK	http://www.ebi.ac.uk/goldman-srv/webprank/
Probalign	http://probalign.njit.edu/probalign/login
PROMALS	http://prodata.swmed.edu/promals/

trend is on mapping information and objects such as genes, proteins, and ligands; finding interaction relationships among the objects; engineering networks and objects to understand and manipulate regulatory mechanisms; and integrating various omes and omics subfields. *Genomics* is the -omics science that deals with the discovery and noting of all the sequences in the entire genome of a particular organism. Genomic sequences are used to study the function of genes (functional genomics), compare the genes in one organism with those of another (comparative genomics), and generate the 3-D structure of one or more proteins from each protein family, thus offering clues to their function (structural genomics). The first eukaryotic organism to have its genome completely sequenced was *Saccharomyces cerevisiae*. Today, 131 eukaryotes' genomes have been sequenced. Among them 33 are protists, 16 are higher plants, 26 are fungi, 17 are mammals (including humans), 9 are non-mammalian animals ,10 are insects, and 4 nematodes.

Genomics and biotechnology have become essential tools for understanding plant behavior at the various biological and environmental levels. The Arabidopsis Information Resource (TAIR) is a continuously updated database of genetic and molecular biology data of the model plant Arabidopsis thaliana. Available data include the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Genomics has also improved classical plant breeding techniques, well summarized in the Plants for the Future technology platform (http://www.epsoweb.eu/catalog/tp/tpcom_home.htm). New technologies now permit researchers to identify the genetic background necessary for crop improvement, explicitly the genes that contribute to the improved productivity and quality of modern crop varieties. Agronomically important genes are being identified and targeted to produce more nourishing and safe food. The genetical modification (GM) of plants is not the only technology in the toolbox of modern plant biotechnologies. Proteomics studies can provide information on the expression of transgenic proteins and their interactions within the cellular metabolism that affects the quality, health, and safety of food. Application of these technologies will substantially improve plant breeding, farming and food processing. In particular, the new technologies will make crops more traceable and enable different varieties to exist side by side, thereby expanding the consumer's freedom to choose between conventional, organic and GM foods. It will also expand the range of plant derived products, including novel forms of pharmaceuticals, biodegradable plastics, bio-energy, paper, and more. Plant genomics and biotechnology could potentially transform agriculture into a more knowledge-based business to address a number of socio-economic challenges. But the central challenge to identify genes underlying important traits and describe the fitness consequences of variation at these loci still remains [43].

Recently, various high-throughput technologies, including genomics, transcriptomics, proteomics and metabolomics have been employed to investigate medicinal plants for regulatory genes and metabolites that can modulate biological and metabolic processes, which in turn can confer specific physiological or pharmacological functions [44]. Bioinformatics and systems biology approaches are considered by many as needed to organize, manage, process, and understand the vast amounts of data obtained in various omics studies. Systems biology is aimed at understanding complex biology by integrating for network analysis experimental results from -omics studies which are most often not obtained or isolated as a single set of data points or events [45], thus evaluating the system as a whole. This approach will hopefully lead to new methods for classifying and authenticating potential medicinal plants, identifying new bioactive phytochemicals or compounds, and even improving medicinal plant species or cultivars that can tolerate stressful environmental challenges.

3. DATABASES

Data sources differ in data accessibility and dissemination. Different levels of provision are made by the data source managers for human-reading, computer-reading, or both. Certainly, data sources can also be classified by species of interest. Despite the challenges, the promise of data integration is high, because heterogeneous data sources provide biological data encompassing a wide range of research fields. Therefore, data integration has the potential to facilitate a better and more comprehensive scope of inference for biological studies. According to the 2010 update on the Bioinformatics Links Directory, there are almost 1500 unique publicly-available data sources. Based on their functions, data sources can be classified into diverse categories:

1) *Sequence databases*: GenBank (http://www.ncbi.nlm.nih.gov/genbank/), RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/), CMR (Comprehensive Microbial Resource) (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHome-Page.cgi/), PlantGDB (http://www.plantgdb.org/), Plant Genomic Resources (http://www.gramene.org/resources/), Plant Transcript Assemblies (http://plantta.jcvi.org/), Plant Cis-acting Regulatory DNA Elements Database (http://www.dna.affrc.go.jp/PLACE/), Plant Model Organism Databases (http://www.arabidopsis.org/por-tals/genAnnotation/othe

r_genomes/index.jsp/), 2) *Functional genomics databases*: ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), FFGED (Filamentous Fungal Gene Expression Database) (http://bioinfo.townsend.yale.edu/), GEO (Gene Expression Omnibus) (http://www.ncbi.nlm.nih.gov/geo/), 3) *Protein-protein interaction databases*: BIND (Biomolecular Interaction Network Database) (http://binddb.Org/), DIP (Database of Interacting Proteins) (http://dip.doe-mbi.ucla.edu/dip/Main.cgi/), IntAct (http://www.ebi.ac.uk/intact/), MINT (Molecular Interactions Database) (http://mint.bio.uniroma2.it/mint/). Protein databases, e.g., CluSTr, CSA, HPI, IntEnz, InterPro, IPI, PANDIT, Patentdata Resource, UniProt, UniSave (http://www.ebi.ac.uk/Databases/pro-tein.html), 4) *Pathway databases*: KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/), 5) *Structure databases*: CATH (http://www.cathdb.info/), PDB (Protein Data Bank) (http://pdb.org/), 6) *Annotation databases*: AmiGO (Gene Ontology) (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi/), NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/), and 7) Domain databases: Pfam v25.0 (pfam.sanger.ac.uk), Prosite (http://prosite.expasy.org/scanprosite), SMART v6.0 (http://smart.embl-heidelberg.de/), Conserved Domain. Database (CDD) v3.02, CDART (Conserved Domain Architecture Retrieval Tool) and CD-Search tools (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), InterPRO v35.0 (http://www.ebi.ac.uk/interpro/), ProDom release 2010.1 (http://prodom.prabi.fr/prodom/current/html/home.php), Superfamily v1.75 (supfam.cs.bris.ac.uk/SUPERFAMILY), PIRSF (pir.georgetown.edu), and functional search by PANTHER (www.pantherdb.org).

3.1. Bioinformatics Tools to Retrieve Biological Data

A large number of bioinformatics tools have been developed to address diverse biological questions. These include investigating relationship between protein structure and function, immune response, development of potential vaccine candidates, modeling pathways, discovery of drug targets and drugs.

3.2. Immunoinformatics Data

Immunoinformatics applies bioinformatics principles and tools to the molecular activities of the immune system. Immunoinformatics databases and predictive tools are used to fetch data on cells of involved in immune response. Immunological data can be broadly split into epitope and allergen categories. This data is used for vaccine discovery via computer aided vaccine design. An important aim here is identification of antigen epitopes. An epitope is a surface localized part of the antigen capable of eliciting an immune response. B-cell epitopes are regions of the antigen recognized by soluble or membrane bound antibodies. They are further classified as either linear or discontinuous epitopes. The former is a single continuous stretch of amino acids within a protein sequence, whereas the latter encompasses residues that are distantly placed in the sequence but are brought together by physico-chemical folding.

T-cell epitopes are short regions presented on the surface of an antigen-presenting cell, where they are bound to major histocompatibility complex (MHC) molecules. These epitopes are characterized based on their recognition by either MHC Class I molecule or Class II molecule. T-cell epitope prediction tools have been developed based on artificial neural networks and weight matrices such as NetMHC (http://www.cbs.dtu.dk/services/NetMHC/), predictive IC(50) values IEDB-ARB method (http://tools.immuneepitope.org/main/html/references.html/), predicted half-time of dissociation Bimas (http://www-bimas.cit.nih.gov/molbio/hla_bind/), and quantitative matrices ProPred (http://www.imtech.res.in/raghava/propred/). Reliable and accurate B-cell epitope prediction is still in development, though some tools are available such as ABCpred (http://www.imtech.res.in/raghava/), BepiPred (www.cbs.dtu.dk), BCPREDS (http://ailab.cs.iastate.edu/bcpreds/), Bcepred (http://www.imtech.res.in/raghava/), Ellipro (tools.immuneepitope.org), and COBEpro (http://scratch.proteomics.ics.uci.edu/) web servers. These tools help build epitope data from protein sequences.

Allergen identification holds major importance in vaccine discovery, as candidate vaccines should be non-allergenic. Allergens are substances like proteins, carbohydrates, particles, and pollen to which the body mounts a hypersensitive immune response typically of Type I. AlgPred (http://www.imtech.res.in/raghava/algpred/) allows prediction of peptide allergens through support vector machines, motif-based method, and database search of known IgE epitopes. Allermatch (http://www.allermatch.org/) performs BLAST search against allergen peptides using a sliding window approach. The results constitute allergen data from databases like the Structural Database of Allergenic Proteins (SDAP) (fermi.utmb.edu/SDAP), Allergome (http://www.allergome.org/), IUIS (www.allergen.org), AllergenIndex (www.expasy.ch/cgibin/lists?allergen.txt), BIFS (www.iit.edu/ sgendel/fa.htm), CSL (www.csl.gov.uk/allergen), FARRP (www.allergenonline.com), PROTALL

(www.ifr.bbsrc.ac.uk/Protall), and ALLALLERGY (www.allallergy. net).

3.2.1. Systems Biology Data

Systems biology deals with a system-level understanding of biological systems. It aims to integrate all the knowledge of networks that represent pathways. A network is mathematically modeled as a graph consisting of nodes and edges. The network can be shown diagrammatically by using *classical graph theory*. Many types of pathways, including gene regulatory networks, signal transduction and metabolic pathways can be modeled using qualitative (Data driven) and quantitative (Knowledge driven) modeling approaches. Data driven pathway modeling, as in for gene regulatory networks, requires DNA microarray data. Such models can be inferred by using logical networks like Boolean, probabilistic Boolean and dynamic Bayesian networks [46]. **Table 2** lists tools available for modeling systems based on given tasks.

3.2.2. Cheminformatics Data

Cheminformatics deals mostly with molecular modeling, chemical structure coding and searching, and data visualization. Cheminformatics deals mostly with molecular modeling, chemical structure coding and searching, and data visualization. Cheminformatics is especially useful in drug-like or lead identification and optimization steps of drug discovery. Databases for cheminformatics are listed in **Table 3**.

3.2.3. Text Mining

In the last few decades, there has been an enormous increase in available data from of scientific articles, abstracts and books, online databases and other resources. This text data may be structured or unstructured and may require hours of mining to extract useful information. Thus text mining has evolved interdisciplinary methods using computer sciences, linguistics and statistics. The database backend support also minimizes the memory demands to handle very large data sets in R. It accepts text data either from local database or directly from online database.

Table 2. Bioinformatics tools for systems modeling in different platforms.

Task	Tools	Web address
Model construction	CellDesigner	http://www.celldesigner.org/
	Jarnac	http://sys-bio.org/
	Jdesigner	http://sys-bio.org/
	Gepasi	http://www.gepasi.org/
Simulation	CellDesigner	http://www.celldesigner.org/
	COPASI	http://www.copasi.org/
	Gepasi	http://www.gepasi.org/
	SBaddon (MatLab tool)	http://www.mathworks.com/
Model Analysis	MatLab,	http://www.mathworks.com/
	R-environment	http://www.r-project.org/

3.2.4. Parallel Computing

When the data size is large (example in millions) and fast information calculation and retrieval is needed, a single modern computational processor fails to perform the task. Many processors are needed to work simultaneously, each carrying out same set of operations on different data objects. This is called *Parallelization on data level.* In this approach, the processing time for a single object is not being reduced but a number of data objects are being processed during the same time-interval by separate processors.

4. DATA INTEGRATION IN BIOINFORMATICS

One of the goals in bioinformatics is to establish automated and efficient ways to integrate large, biological datasets from multiple sources. This objective is challenging because data sources are heterogeneous in terms of their functions, structures, data access methods and dissemination formats. Several major approaches proposed for data integration can be classified into five groups [47,48]: 1) data warehousing, 2) federated databasing, 3) service-oriented integration, 4) semantic integration, and 5) wiki-based integration.

4.1. Data Warehousing

The data warehouse approach offers a "one-stop shop" solution to ease access and management of a large variety of biological data from different data sources. Warehouses focus on data translation, fetching all accessible data from disparate sources. Currently the focus is on transforming the data and importing it into the data warehouse. Representative examples of data warehousing include the following list: 1) *Atlas* is a biological data warehouse that includes data from BIND, LocusLink, MINT, RefSeq, DIP, Entrez Gene, GO, GenBank, HomoloGene, HPRD (Human Protein Reference Database), IntAct, OMIM (Online Mendelian Inheritance in Man), Taxonomy, and UniProt [49], 2) *BioWarehouse* is an open source toolkit that incorporates data from ENZYME, GenBank, GO, BioCyc, CMR, KEGG, Taxonomy, and UniProt and integrates its component databases into a common representational framework within a single database management system [50], 3) *BIOZON* is a

Table 3. Softwares used in cheminformatics.

Function	Tools	Links
Databases for searching known inhibitors	Pubchem	http://pubchem.ncbi.nlm.nih.gov/search/search.cgi
	Pubmed	http://www.ncbi.nlm.nih.gov/pubmed
	Drug Bank	http://www.drugbank.ca/
Molecular visualization	Pymol	http://www.pymol.org
	Rasmol	http://rasmol.org
	SwissPDBviewer	http://spdbv.vital-it.ch/
Structures drawn and view	Marvin Sketch	http://www.chemaxon.com/products/marvin/marvinsketch
	Chemmine	http://bioweb.ucr.edu/ChemMineV2
File format translator	Smile Translator	http://cactus.nci.nih.gov/translate/
	OpenBabel	http://openbabel.org/wiki/Main_Page
Drug designing	AutoDock	http://autodock.scripps.edu/
	GOLD	http://www.ccdc.cam.ac.uk/products/life_sciences/gold/
Toxicity prediction	Toxtree	http://toxtree.sourceforge.net/
Molecular dynamics simulation	GROMACS	http://www.gromacs.org/

unified biological resource on DNA sequences, proteins, complexes and cellular pathways, including KEGG, PDB, RefSeq, Genbank, InterPro, Swiss-Prot, UniGene, BIND, DIP, and UniProt [51], 4) *COLUMBA* is an integrated database of information on proteins, structures and annotations. It integrates twelve different databases, including GO, CATH, PDB, SCOP, ENZYME, KEGG, and Swiss-Prot [52], and 5) *VINEdb* is a data warehouse for integration and interactive exploration of life science data. It manages diverse data from KEGG, OMIM, GO, IntAct, and UniProt and emphasizes the visualization of the integrated data in a comprehensible manner [53].

4.1.1. Federated Databasing

Federated databasing focuses on query translation. Representative examples for federated databasing include: 1) BioMart provides a user-friendly and unified way to retrieve data from one or multiple data sources located at diverse geographical locations, including Ensembl, HGNC, Uniprot, Reactome, Wormbase, and PRIDE [54], 2) DiscoveryLink is a system for integrated access to life sciences data from heterogeneous data sources, including GenBank, MedLine and Swiss-Prot [55], 3) K2/Kleisli is a federated database system, integrating data from EcoCyc, GenBank, GSDB, dbEST, GDB, KEGG and SRSindexed databases [56], 4) MRS allows for very rapid queries in a large number of flat-file data banks, including EMBL, UniProt, OMIM, dbEST, PDB, KEGG [57], 5) QIS (Query Integrator System) is based on a set of distributed network-based servers and includes CellPropDB [58], Brain Architecture Management System [59], Yale Microarray Database [60], a local Gene Annotation Database and GO, 6) SRS (Sequence Retrieval System) [61], and 7) TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources). The prototype version of TAMBIS contains five data sources, viz., BLAST, CATH, ENZYME, PROSITE [62], and Swiss-Prot.

4.1.2. Service-Oriented Integration

The service-oriented approach enables data integration from multiple heterogeneous data sources through computer interoperability. Examples for service-oriented integration include: 1) *BioMOBY* [63] is an open source ontology-based integration system for accessing distributed and heterogeneous data sources via WS, 2) *DAS* (Distributed Annotation System). It allows a single machine to collect all annotations from multiple distributed data sources and display them to the user in a single view. DAS is widely used in the genome annotation community (http://en.wikipedia.org/wiki/Distributed_Annotation_System) and adopted by several systems, including Ensembl, WormBase, and the Berkeley Drosophila Genome Project [64-66], and 3) *Taverna* [67] is a graphical workflow workbench application, aiming to integrate the growing number of molecular biology tools and databases.

4.1.3. Semantic Integration

The Semantic Web [68] aims to describe data in a way that computers can understand and build an interconnected network. Several studies have applied this tech-

nology in data integration and representative examples of semantic integration are described below: 1) Bio2RDF [69] Bio2RDF applies the Semantic Web technologies to multiple data sources, such as Entrez Gene, HGNC, KEGG, MGI, OMIM PDB, PubMed and UniProt, and converts data into RDF format based on RDFizer (a set of tools for converting various data formats into RDF; http://simile.mit.edu/wiki/RDFizers), Sesame (an open source framework for storage, inference and querying of RDF data; http://www.openrdf.org) and OWL ontology, 2) YeastHub [70] is an integrated database in RDF format for the yeast community. The ever-evolving next-generation Web (NGW), characterized as the Semantic Web, aims to provide information not only for human, but also for computers to semantically process large-scale data and automatically discover knowledge. The Semantic Web befits the exponential growth of biological data with promise to provide solutions for data integration and advancing translational research [71]. In order to manage large-scale data, it necessitates adopting advances in high performance computing [72]. In addition, a framework is also needed to set up Semantic WS workflows or pipelines [73].

4.1.4. Wiki-Based Integration

Wikipedia features collaborative integration that is continuous and frequently updated. It has a broad coverage and low maintenance costs. Content can be freely and anonymously changed in the wiki, Wikipedia outperforms the traditional Encyclopedia in accuracy (http://www.wikipedia.org). Representative examples include: 1) WikiGenes (a wiki system that combines gene annotation with explicit authorship (http://www.wikigenes.org/), 2) Wikiproteins (a wiki-based system for protein annotation (http://www.omegawiki.org/Portal:Wikiproteins), 3) BO-Wiki (a ontology-based wiki for data annotation and knowledge integration (https://onto.eva.mpg.de/trac/BoWiki), 4) Gene Wiki (a wiki for human gene annotation (http://en.wikipedia.org/wiki/Gene), and 5) PDBWiki (a scientific wiki for the community annotation of protein structures (http://pdbwiki.org/wiki/Main_Page). However, the wiki-based integration has its own shortcomings, including the unstructured data generated, the lack of a standard format for data exchange, the lack of credit for authorship and vulnerability to malicious editing [74,75].

5. DATABASES: ORGANIZATIONS AND INFORMATICS

The enormous quantity of information produced by NGS is handled via computers that systematically analyze and store the accumulating sequence and structure data. The idea that molecular information can be collected and distributed from electronic repositories is still in its infancy and faces significant challenges.

5.1. The Protein Data Bank (PDB)

PDB is one of the earliest scientific databases established in 1965 at the Cambridge Crystallographic Data Centre (CCDC) (http://beta-www.ccdc.cam.ac.uk/pages/Home.aspx) as a repository of small-molecule crystal structures. In February 2011, the archive housed 71,415 structures (http://www.rcsb.org/pdb/home/home.do).

5.2. The EMBL Nucleotide Sequence Data Library

At the end of 2010, the database contained 199,720,869 entries (http://www.ebi.ac.uk/embl/).

5.3. GenBank

In April 2011, GenBank contained 135,440,924 sequence records. It became the responsibility of the NCBI to maintain the database, where it remains today (http://www.ncbi.nlm.nih.gov/genbank/).

5.4. The PIR-PSD

The Protein Information Resource (http://pir.georgetown.edu/). In 2003, with 283,000 sequences, the PSD was the most comprehensive protein sequence database in the world.

5.5. UniProtKB/Swiss-Prot

It is a comprehensive, annotated and non-redundant high-quality and freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects, computed features, and research literature (http://web.expasy.org/docs/swiss-prot_guideline.html).

5.6. The European Molecular Biology Network (EMBnet)

EMBnet has promoted the development of distributed computing services to share workload among international servers. It has contributed to the development and maintenance of advanced database systems and has been an advocate of the deployment of Grid technologies for the life sciences through its contributions to major European Grid projects. EMBnet developed, and continues to promote the use of, an e-learning system both to support distance learning in bioinformatics and to complement face-to-face bioinformatics teaching and training. It is also committed to bringing the latest software and algorithms to users, free of charge.

The combined expertise of its Nodes has allowed EMBnet to provide services to its local European life science communities. Currently, the network connects 31 member Nodes extending over 27 countries. Together, the Nodes work to disseminate data, share computing resources and provide training support thousands of users (http://journal.embnet.org/index.php/embnetjournal/article/view/185).

5.7. Prosite

It consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them (http://prosite.expasy.org/).

5.8. The European Bioinformatics Institute (EBI)

EBI is a centre for research and services in bioinformatics. It maintains and distributes the EMBL Nucleotide Sequence database, Europe's primary nucleotide sequence data resource (http://www.ebi.ac.uk/).

5.9. TrEMBL

At the beginning of 2011, with millions of entries, TrEMBL was almost 26 times larger than Swiss-Prot, illustrating the vast disparity between manual and computer assisted annotation strategies (http://www.uniprot.org/news/2004/03/02/full).

5.10. InterPro

With 21,185 entries in February 2011 (release 31.0), InterPro is the most comprehensive integrated protein family database in the world (http://www.ebi.ac.uk/interpro/). This family database is integrated by GENE3D, HAMAP, PANTHER, PIRSF, PRINTS, PROSITE patterns, PROSITE profiles, Pfam, ProDom, SMART, SUPERFAMILY, and TIGRFAMs.

5.11. UniProt

By 2011 (http://www.uniprot.org/), UniProt also included a Metagenomic and Environmental Sequence component, termed UniMES (The UniProt Consortium, 2011); by this time, UniProtKB: Swiss-Prot contained 525,207 entries, accompanied by UniProtKB: TrEMBL, with a staggering 13,499,622 entries.

5.12. The Swiss Institute of Bioinformatics (SIB)

Today, the SIB (http://www.isb-sib.ch/) leads and coordinates the field of bioinformatics in Switzerland. Its vision is to help shape the future of the life sciences through excellence in bioinformatics services, research and education. SIB's mission is to provide world-class core bioinformatics resources to both national and international research communities in fields spanning genomics, proteomics and systems biology. Many of its core activities, including maintenance of databases such as UniProt and InterPro, are carried out in close collaboration with the EBI.

5.13. The European Nucleotide Archive (ENA)

Today, ENA (http://www.ebi.ac.uk/ena/) holds more than 20 terabases of nucleotide sequence data, which, combined with its annotation information, and so on, occupies more than 230 terabytes of disk space.

5.14. ELIXIR

Europe's databases (estimated to number around 500), especially those hosted by the EBI, will become the foundation of the new ELIXIR infrastructure, the aim of which is to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine, the environment, bio-industries and society (http://www.bbsrc.ac.uk/science/international/elixir.aspx).

6. NOMENCLATURES AND NAMING SYSTEMS

In biological research, there are thousands of specialized data repositories which offer sets of richly annotated records. To ensure data of the highest quality, manual data entry and curation (annotation) processes are generally performed on these databases. This process makes the information searchable through a variety of automated techniques, given that the curators use standardized terminologies or ontologies.

The task of gene annotation by means of a controlled vocabulary becomes laborious when an expert is required to inspect carefully all of the literature associated with each gene, to identify the appropriate terms. To reduce the cost of obtaining annotations, several initiatives for collaborative curation like the pseudomomas project (http://www.pseudomonas.com/), and wiki-based prototypes (e.g., http://www.wikiprofessional.org/) have been prompted. As of now, PubMed remains the richest and most updated source of information about biological data despite its unstructured nature. Text mining technology can contribute to this field by operating together with curators to minimize their involvement and speed up the pace of research; however it should not completely supplant their role.

6.1. Automated Functional Annotation

Automated functional annotation of genomes can bequite efficient because it takes advantage of knowledge con-

cerning alignment of ORFs of homologous organisms [76], saving considerable time in manual curation [77]. However, care must be taken with fully automated functional annotation, since similarity of sequences can easily incur false positives [78]. The following are some tools for automatic annotation of entire genomes.

6.1.1. GenDB

GeneDataBank is included among a selected set of tools for automatic annotation of genomes because it was developed as a web platform [79]. Geographically dispersed research groups can benefit from web interfaces using standard tools and a centralized database. The GeneDataBase program performs sequence alignments using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and allows incorporation of predictions of conserved domains of protein families based on InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/), as well as transmembrane domains based on TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), and indications of export to the extracellular medium through SignalP (http://www.cbs.dtu.dk/services/SignalP/).

6.1.2. BLAST2GO (B2G)

This tool was designed as an interface for Gene Ontology (GO); additional features have transformed it into a more comprehensive annotation platform (http://www.blast2go.com/b2ghome). The program menus include various steps initiating annotation, with an automatic alignment of genome sequences against a protein-based non-redundant (NR) NCBI database, through prediction of conserved domains (InterPro-Scan), GO annotation ratings against the enzymatic English Enzymatic Code (EC) and subsequent visualization of molecular interactions in a genome by means of maps in the format of the Kyoto Encyclopedia of Genes and Genomes (KEEGO).

6.1.3. CpDB

The *Corynebacterium pseudotuberculosis* DataBase (CpDB) is a relational database schema and tools for bacterial genomes annotation and pos-genome research. Its tutorial has approximately 100 steps, including software installation and configuration, edition of files by Linux commands or through interfaces with biological sequence manipulation programs. All of the steps in this manual can be automated in order to develop an automatic pipeline for annotation, allowing CpDB to become another web-based automatic annotation environment.

6.2. Manual Curation

Genome annotation is a process that consists of adding analyses and biological interpretations to DNA sequence information. This process can be divided into three main categories: 1) annotation of nucleotides, 2) proteins, and 3) processes. Annotation of nucleotides can be done when there is information about the complete genome (or DNA segments) of an organism. It involves looking for the physical location (position on the chromosome) of each part of the sequence and discovering the location of the genes, RNAs, repeat elements, etc. Annotation of proteins involves searching for gene function. Besides general predictions about gene and protein function, other information can be found in an annotation, such as biochemical and structural properties of a protein, prediction of operons, gene ontology, evolutionary relationships and metabolic cycles [80]. Consequently, manual curation is a fundamental part of the process of assembling and annotating a genome, in which the curator is responsible for validating all of the predicted genes [81]. A more detailed description of the gene or gene family product is obtained through similarity analyses using protein data banks that contain well-characterized and conserved proteins [82].

6.2.1. Steps for Manual Curation

Manual curation is a very complex task and is susceptible to errors. One of these is a lack of padronization in the interpretation of BLAST results. Another is propagation of errors, which involves prediction of protein function based on proteins that were also predicted but could have imprecise or even incorrect annotation [83]. The fundamental step for avoiding this error is mining data obtained from similarity analyses of BLASTp data banks. It is also important to observe whether there is any consensus among the first ten hits. In cases where there is no consensus or when the E-value of the best hit is significantly larger than that of the following sequences, it is preferable to transfer the annotation of the best hit [84]. In cases of non-significant alignments, run a similarity search at the nucleotide level (BLASTn). Other measures such as percentage identity between the sequence being analyzed and the sequence in the data bank, score value and E-value, as well as pair-by-pair alignment evaluation to check the texture of the alignment (evaluating the number of gaps, size of the gaps, and the number of conserved substitutions of amino acids) are also informative.

6.2.2. Pseudogenes

Comparisons between non-coding regions of genomes from various prokaryotic species has aided in the identification and characterization of genome segments with regulatory roles [85] such as pseudogenes. These are DNA sequences that are highly similar to functional genes but do not express a functional protein. Loss of function is probably due to deleterious mutations, such as a nonsense mutation that introduces a premature stop co-

don, resulting in an incomplete protein and a later change in the open reading frame [86]. Whenever possible, addition or removal of erroneous bases can restore the reading frame. If there is no data that justifies addition or removal of bases, the genes should be classified as pseudogenes.

6.2.3. Sequence Similarity Searches

1) Blast

BLAST (Basic Local Alignment Search Tool) is a tool that is widely used for the characterization of products coded by genes that are identified by gene prediction. This program is available on the NCBI—National Center for Biotechnology Information site (http://www.ncbi.nlm.nih.gov/), the central databank for genome information. BLAST has programs for alignment of protein and nucleotide sequences, according to the needs of the researcher. Through this type of algorithm, we can compare any DNA sequence or protein (query) with all of the genome sequences in the public domain (subject) [33]. BLAST parameters such as the number of points obtained (score), gap opening/extension penalties, number of expected alignments in the case of scores equal to or superior to the alignment that is being investigated (expectation value), and the normalized score (bit-score), are indispensible for the interpretation of the results. The smaller the value of "E", the smaller the chance of such a comparison being found merely by chance. Therefore, one can infer greater homology between the sequence being investigated and the data base [87].

2) PFAM

Proteins generally are composed of one or more functional domains. Different combinations of domains result in the large variety of proteins found in nature. Identification of the domains that are found in proteins can, therefore, provide insight about protein function [88]. In sequences with an identity of less than 70%, without end to end similarity, the protein domains are searched through the Pfam database (http://pfam.sanger.ac.uk) [89]. In Pfam, the sequences that are in full alignment are identified through a search for a hidden profile using a hidden Markov model algorithm generated using the software HMMER, based on the UniProt database (http://www.uniprot.org/).

6.3. Challenges Ahead

Although a number of current efforts have been devoted to data integration, none have yet become preeminent. As NGS data are grow at an exponential rate, the need for data integration is continually demanding (**Figure 2**).

Low-cost and high-throughput NGS technologies can generate huge amounts of data in a relatively short period. To keep pace with sequencing technologies, genome sequencing projects have transitioned from classical model organisms (e.g., fly, mouse, yeast), to other organisms (e.g., dog, panda) and even to sequencing individuals within populations. Examples of this are the 1000 Genomes Project, a collection of the genomes of 1000 humans (http://www.1000genomes.org), and the Genome 10 K Project, a genomic zoo of genome sequences of 10,000 vertebrate species (http://www.genome10k.org). In addition, we are in the era when personal genome sequencing will cost a few hundred dollars is approaching and will accumulate unparalleled amounts of large-scale data (**Figure 2**). It is necessary to establish an efficient way to data exchange among these distributed and heterogeneous data sources. The growing volume of biological data also requires "computer-readable" approaches for data integration. Data sources should not only provide data for human reading via web interfaces; they should also provide data for computer interoperability.

6.3.1. Standards for Biological Data

There are a wide variety of biological data types such as sequence, gene expression, protein-protein interaction, and pathway data [89]. Data sources store different data types in different formats like flat files, FASTA sequence files, structure files, and XML files that are often incompatible [90]. Complications in data exchange and integration arising from format heterogeneity can be resolved by using standards for data formats. BioPAX [91] has been developed to deliver a compatible standard, facilitating integration, exchange, visualization and analysis of biological pathway data. Standard data formats, in general, facilitate data analysis and visualization as well as downstream software development.

Equally important, data integration requires standardizing nomenclature and ontologies for biological data [92]. For example, if two data sources need to exchange gene annotations, they must share a standard regarding gene names. Otherwise, any ambiguity or inconsistency in nomenclature would burden the integration process.

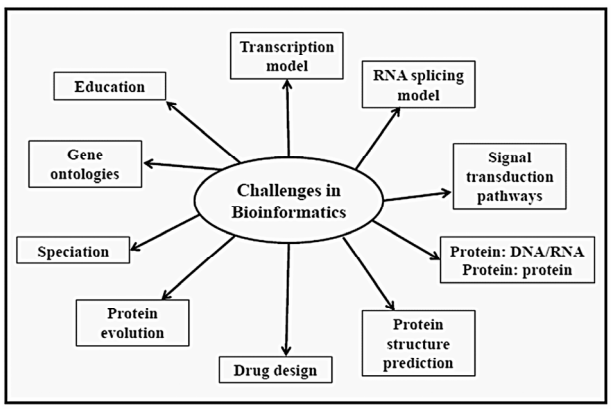

Figure 2. Applications and challenges for bioinformatics.

The following are efforts for standardizing nomenclature and ontologies for biological data: BioPortal (http://bioportal.bioontology.org) for integrating and sharing biomedical ontologies in National Center for Biomedical Ontology, Gene Ontology (GO) (http://www.geneontology.org/) for standardizing the representation of gene and gene product attributes, HGNC (http://www.genenames.org/guidelines.html) for standardizing human gene symbols and names, and Open Biomedical Ontologies (OBO) (http://www.obofoundry.org/) for creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. A wiki-based system might be promising way to collaboratively and efficiently develop standards with all communities' efforts.

6.3.2. Web Services (WS)-Based Pipelines

The goal of data integration is to combine information from different resources in an automated fashion without human intervention, accommodating for increasing accumulation of biological data [93]. Towards this goal, data to be integrated should be re-defined in a broader manner, which include not merely sequences and other raw data, but also methods, tools, algorithms, analyzed results, discovered knowledge [94] and even connections among people [48]. A pipeline with a combination of multiple WS can achieve data integration. Any user may easily create WS-based pipelines (adding value), publish them online, and subscribe to pipelines created by other users. Consequently, pipelines may be widely shared, reused and even integrated into other pipelines. As a result, communications and collaborations among users in Scientific Social Community can be greatly increased, making knowledge discovery through collective intelligence possible.

7. CONCLUSIONS

Deep-sequencing data analysis is a growing field. The overflow of available bioinformatics tools for each of the optional analysis steps represents a challenge for the researcher aiming to evaluate and interpret deep sequencing data. The field is rapidly evolving both in sequencing platform technology and in computational tools. The development of high throughput technologies has not only increased the amount of data, but also the types of data available, opening new prospects for investigations. Due to automatic approaches for data analysis, disciplines such as bioinformatics and computational biology are able to combine the expertise of biologists and computer scientists in a synergism of human knowledge and efficiency.

Bioinformatics has given us the first "complete" catalogues of genomes and proteomes of organisms across the entire Tree of Life; it has furnished the requisite software to help analyze biological data on an unprecedented scale; it has hence yielded the possibilities to understand more about evolutionary processes, and ultimately, a great deal more about health, disease and disease processes. A detailed in this report, the evolution and broader impact of bioinformatics is evidenced by the fact that bioinformatics has enabled systems level approach to analyze complex biological networks in a wide range of biological systems, bringing life science data to local communities and making available computing software tools for modeling and data analysis, while providing on-line training of bioinformatics databases and software to users.

REFERENCES

[1] Swindells, M., Rae, M., Pearce, M., Moodie, S., Miller, R. and Leach, P. (2002) Application of high throughput computing in bioinformatics. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, **360**, 1179-1189. doi:10.1098/rsta.2002.0987

[2] Kann, M.G. (2010) Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Brief Bioinformatics*, **11**, 96-110. doi:10.1093/bib/bbp048

[3] Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., *et al*. (2002) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 200866-2000872. doi:10.1038/nature07485

[4] Isakov, O., Modai, S. and Shomron, N. (2011) Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics*, **27**, 2027-2030. doi:10.1093/bioinformatics/btr349

[5] Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133-141. doi:10.1016/j.tig.2007.12.007

[6] Koboldt, D.C., Ding, L., Mardis, E.R. and Wilson, R.K. (2010) Challenges of sequencing human genomes. *Brief Bioinformatics*, **11**, 484-498. doi:10.1093/bib/bbq016

[7] Clarke, S.C. (2005) Pyrosequencing: Nucleotide sequencing technology with bacterial genotyping applications. *Expert Review of Molecular Diagnostics*, **5**, 947-953. doi:10.1586/14737159.5.6.947

[8] Claesson, M.J., O'Sullivan, O., Wang, Q., Nikkilä, J., Marchesi, J.R., Smidt, H., de Vos, W.M., Ross, R.P., and O'Toole, P.W. (2009) Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One*, **20**, e6669. doi:10.1371/journal.pone.0006669

[9] Hamady, M., Lozupone, C. and Knight, R. (2010) Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *International Society for Microbial Ecology Journal*, **4**, 17-27. doi:10.1038/ismej.2009.97

[10] Margulies, M., Egholm, M., Altman, W.E., Attiya, S.,

Bader J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y-J. and Chen, Z. (2005a) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380. doi:10.1038/nature03959

[11] Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59. doi:10.1038/nature07517

[12] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303. doi:10.1038/nature07517

[13] McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., *et al*. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**, 1527-1541. doi:10.1101/gr.091868.109

[14] Eid, J., Fehr, A., Gray J., Luong, K., Lyle, J., *et al*. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138. doi:10.1126/science.1162986

[15] Chan, E.Y. (2009) Next-generation sequencing methods: Impact of sequencing accuracy on SNP discovery. *Methods in Molecular Biology*, **578**, 95-111. doi:10.1007/978-1-60327-411-1_5

[16] Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., *et al*. (2010) Multi-platform next generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biology*, **8**, e1000475. doi:10.1371/journal.pbio.1000475

[17] Nothnagel, M., Herrmann, A., Wolf, A., Schreiber, S., Platzer, M., Siebert, R., Krawczak, M. and Hampe, J. (2011) Technology-specific error signatures in the 1000 Genomes Project data. *Human Genome*, **130**, 505-516. doi:10.1007/s00439-011-0971-3

[18] Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175-185. doi:10.1101/gr.8.3.175

[19] Castellana, S., Romani, M., Valente, E.M. and Mazza, T.A. (2012) Solid quality-control analysis of AB SOLiD short-read sequencing data. *Brief Bioinformatics*, **13**, 1-12. doi:10.1093/bib/bbs048

[20] Parkinson, N.J., Maslau, S., Ferneyhough, B., Zhang, G., Gregory, L., Buck, D., Ragoussis, J., Ponting, C.P. and Fischer, M.D. (2012) Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Research*, **22**, 125-133. doi:10.1101/gr.124016.111

[21] Allen, J.E., Pertea, M. and Salzberg, S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Research*, **14**, 142-148. doi:10.1101/gr.1562804

[22] Sleator, R.D. (2010) An overview of the current status of eukaryote gene prediction strategies. *Gene*, **461**, 1-4. doi:10.1016/j.gene.2010.04.008

[23] Tompa, M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *International Conference on Intelligent Systems for Molecular Biology*, **1999**, 262-271.

[24] Tompa, M., Li, N., Bailey, T.L., Church G.M., Moor B.D., *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**, 137-144. doi:10.1038/nbt1053

[25] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410. doi:10.1016/S0022-2836(05)80360-2

[26] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., *et al*. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*, **10**, 421. doi:10.1186/1471-2105-10-421

[27] Flicek, P. and Birney, E. (2009) Sense from sequence reads: Methods for alignment and assembly. *Nature Methods*, **6**, S6-S12. doi:10.1038/nmeth.1376

[28] Lassmann, T., Hayashizaki, Y. and Daub C.O. (2011) SAMStat: Monitoring biases in next generation sequencing data. *Bioinformatics*, **27**, 130-131. doi:10.1093/bioinformatics/btq614

[29] Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S. and Robinson, P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722-729. doi:10.1093/bioinformatics/btq027

[30] Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 754-1760. doi:10.1093/bioinformatics/btp324

[31] Paşaniuc, B., Zaitlen, N. and Halperin, E. (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *Journal of Computational Biology*, **18**, 459-468. doi:10.1089/cmb.2010.0259

[32] Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073. doi:10.1038/nature09534

[33] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402. doi:10.1093/nar/25.17.3389

[34] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S. and Sigrist, C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Research*, **36**, D245-D249. doi:10.1093/nar/gkm977

[35] Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Research*, **38**, D211-222. doi:10.1093/nar/gkp985

[36] Pirovano, W. and Heringa, J. (2010) Protein secondary structure prediction. *Methods in Molecular Biology*, **609**, 327-348. doi:10.1007/978-1-60327-241-4_19

[37] Raghava, G.P., Searle, S.M., Audley, P.C., Barber, J.D. and Barton, G.J. (2003) OXBench: A benchmark for

evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47. doi:10.1186/1471-2105-4-47

[38] Stebbings, L.A. and Mizuguchi, K. (2004) HOMSTRAD: Recent developments of the homologous protein structure alignment database. *Nucleic Acids Research*, **32**, D203-D207. doi:10.1093/nar/gkh027

[39] Edgar, R.C. (2004b) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792-1797. doi:10.1093/nar/gkh340

[40] Thompson, J.D., Koehl, P., Ripp, R. and Poch, O. (2005) BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127-136. doi:10.1002/prot.20527

[41] Van Walle, I., Lasters, I. and Wyns, L. (2005) SABmark-a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267-1268. doi:10.1093/bioinformatics/bth493

[42] Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M. and Morgenstern, B. (2005) DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics, **6**, 66. doi:10.1186/1471-2105-6-66

[43] Stinchcombe, J.R. and Hoekstra, H.E. (2008) Combining population genomics and quantitative genetics: Finding the genes underlying ecologically important traits. *Heredity*, **100**, 158-170. doi:10.1038/sj.hdy.6800937

[44] Fridman, E. and Pichersky, E. (2005) Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Current Opinion in Plant Biology*, **8**, 242-248. doi:10.1016/j.pbi.2005.03.004

[45] Middleton, F.A., Rosenow, C., Vailaya, A., Kuchinsky, A., Pato, M.T. and Pato, C.N. (2007) Integrating genetic, functional genomic, and bioinformatics data in a systems biology approach to complex diseases: Application to schizophrenia. *Methods in Molecular Biology*, **401**, 337-364. doi:10.1007/978-1-59745-520-6_18

[46] Lahdesmakia, H., Hautaniemia, S., Shmulevichc, I. and Yli-Harja, O. (2006) Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. *Signal Processing*, **86**, 814-834. doi:10.1016/j.sigpro.2005.06.008

[47] Goble, C. and Stevens, R. (2008) State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, **41**, 687-693. doi:10.1016/j.jbi.2008.01.008

[48] Zhang, Z., Cheung, K.H. and Townsend, J.P. (2009) Bringing Web 2.0 to bioinformatics. *Brief Bioinformatics*, **10**, 1-10. doi:10.1093/bib/bbn041

[49] Shah, S.P., Huang, Y., Xu, T., Yuen, M.M.S., Ling, J. and Ouellette B.F.F. (2005) Atlas—A data warehouse for integrative bioinformatics. *BMC Bioinformatics*, **6**, 34. doi:10.1186/1471-2105-6-34

[50] Lee T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D.W.J., Tenenbaum, J.D. and Karp, P.D. (2006) Biowarehouse: A bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**, 170. doi:10.1186/1471-2105-7-170

[51] Birkland, A. and Yona, G. (2006) BIOZON: A hub of heterogeneous biological data. *Nucleic Acids Research*, **34**, D235-D242. doi:10.1093/nar/gkj153

[52] Trissl, S., Rother, K., Müller, H., Steinke, T., Koch, I., Preissner, R., Frömmel, C. and Leser, U. (2005) Columba: An integrated database of proteins, structures, and annotations. *BMC Bioinformatics*, **6**, 81. doi:10.1186/1471-2105-6-81

[53] Hariharaputran, S., Töpel, T., Brockschmidt, B. and Hofestädt, R. (2007) VINEdb: A data warehouse for integration and interactive exploration of life science data. *Journal of Integrative Bioinformatics*, **4**, 63.

[54] Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P. and Kasprzyk, A. (2009) BioMart central portal-unified access to biological data. *Nucleic Acids Research*, **37**, W23-W27. doi:10.1093/nar/gkp265

[55] Haas, L.M., Schwarz, P.M., Kodali, P., Kotlar, E., Rice, J.E. and Swope, W.C. (2001) DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, **40**, 489-511. doi:10.1147/sj.402.0489

[56] Chung, S.Y., Wong, L. (1999) Kleisli: A new tool for data integration in biology. *Trends in Biotechnology*, **17**, 351-355. doi:10.1016/S0167-7799(99)01342-6

[57] Hekkelman, M.L. and Vriend, G. (2005) MRS: A fast and compact retrieval system for biological data. *Nucleic Acids Research*, **33**, W766-W769. doi:10.1093/nar/gki422

[58] Crasto, C.J. and Shepherd, G.M. (2007) Managing knowledge in neuroscience. *Methods in Molecular Biology*, **401**, 3-21. doi:10.1007/978-1-59745-520-6_1

[59] Bota, M. and Swanson, L.W. (2010) Collating and curating neuroanatomical nomenclatures: Principles and use of the brain architecture knowledge management system (BAMS). *Frontier in Neuroinformatics*, **4**, 3. doi:10.3389/fninf.2010.00003

[60] Cheung, K.H., White, K., Hager, J., Gerstein, M., Reinke, V., Nelson, K., *et al*. (2002) YMD: A microarray database for large-scale gene expression analysis. *AMIA Annual Symposium Proceedings*, **2002**, 140-144.

[61] Zdobnov, E.M., Lopez, R., Apweiler, R. and Etzold T. (2002) The EBI SRS server-recent developments. *Bioinformatics*, **18**, 368-373. doi:10.1093/bioinformatics/18.2.368

[62] Sigrist, C.J.A., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, **38**, D161-D166. doi:10.1093/nar/gkp885

[63] BioMoby Consortium, Wilkinson, M.D., Senger, M., Kawas, E., Bruskiewich, R., *et al*. (2008) Interoperability with Moby 1.0—It's better than sharing your toothbrush. *Briefings in Bioinformatics*, **9**, 220-231. doi:10.1093/bib/bbn003

[64] Jenkinson, A.M., Albrecht, M., Birney, E., Blankenburg H., Down, T., *et al*. (2008) Integrating biological data—The Distributed Annotation System. *BMC Bioinformatics*, **9**, S3. doi:10.1186/1471-2105-9-S8-S3

[65] Messina, D.N. and Sonnhammer, E.L. (2009) DASher: A stand-alone protein sequence client for DAS, the Distributed Annotation System. *Bioinformatics*, **25**, 1333-1334. doi:10.1093/bioinformatics/btp153

[66] Olason, P.I. (2005) Integrating protein annotation re-

sources through the Distributed Annotation System. *Nucleic Acids Research*, **33**, W468-W470. doi:10.1093/nar/gki463

[67] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., *et al*. (2004) Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045-3054. doi:10.1093/bioinformatics/bth361

[68] Hendler, J. (2003) Science and the semantic web. *Science*, **299**, 520-521. doi:10.1126/science.1078874

[69] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P. and Morissette, J. (2008) Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, **41**, 706-716. doi:10.1016/j.jbi.2008.03.004

[70] Cheung, K.H., Yip, K.Y., Smith, A., Deknikker, R., Masiar, A., Gerstein, M. (2008) YeastHub: A semantic web use case for integrating data in the life sciences domain. *Bioinformatics*, **21**, 85-96. doi:10.1093/bioinformatics/bti1026

[71] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., *et al*. (2007) Advancing translational research with the semantic web. *BMC Bioinformatics*, **8**, S2. doi:10.1186/1471-2105-8-S3-S2

[72] Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. (2010) Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, **11**, 647-657. doi:10.1038/nrg2857

[73] Wilkinson, M.D., McCarthy, L., Vandervalk, B., Withers, D., Kawas, E. and Samadian, S. (2010) SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics*, **11**, S7. doi:10.1186/1471-2105-11-S12-S7

[74] Lee, T.L. (2008) Big data: Open-source format needed to aid wiki collaboration. *Nature*, **455**, 461. doi:10.1038/455461c

[75] Potthast, M., Stein, B. and Gerling, R. (2008) Automatic vandalism detection in Wikipedia. *Advances in Information Retrieval*, **4956**, 663-668. doi:10.1007/978-3-540-78646-7_75

[76] Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., Conley, A.B., *et al*. (2010) A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics*, **26**, 1819-1826. doi:10.1093/bioinformatics/btq284

[77] Li, L., Shiga, M., Ching, W.K. and Mamitsuka, H. (2010) Annotating gene functions with integrative spectral clustering on microarray expressions and sequences. *Genome Information*, **22**, 95-120. doi:10.1142/9781848165786_0009

[78] Lorenzi, H.A., Puiu, D., Miller, J.R., Brinkac, L.M., Amedeo, P., Hall, N. and Caler, E.V. (2010) New assembly, reannotation and analysis of the entamoeba histolytica genome reveal new genomic features and protein content information. *PLoS Neglected Tropical Diseases*, **4**, e716. doi:10.1371/journal.pntd.0000716

[79] Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., *et al*. (2003) Gendb—An open source genome annotation system for prokaryote genomes. *Nucleic Acids Research*, **31**, 2187-2195. doi:10.1093/nar/gkg312

[80] Stothard, P. and Wishart, D.S. (2006) Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*, **9**, 505-510. doi:10.1016/j.mib.2006.08.002

[81] Stein, L. (2001) Genome annotation: From sequence to biology. *Nature Review in Genetics*, **2**, 493-503. doi:10.1038/35080529

[82] Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., *et al*. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, **33**, 5691-5702. doi:10.1093/nar/gki866

[83] Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S. and Ouzounis, C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641-1649. doi:10.1093/bioinformatics/18.12.1641

[84] Prosdocimi, F. (2003) Bioinformática: Manual do usuario. *Biotecnologia Ciência & Desenvolvimento*, **2**, 2.

[85] Pareja, E., Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Bonal, J. and Tobes, R. (2006) Extratrain: A database of extragenic regions and transcriptional information in prokaryotic organisms. *BMC Microbiology*, **6**, 29. doi:10.1186/1471-2180-6-29

[86] Lerat, E. and Ochman, H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research*, **33**, 3125-3132. doi:10.1093/nar/gki631

[87] Baxevanis, A.D. and Ouellette, F.F. (2001) A practical guide to the analysis of genes and proteins. *Wiley*: *Bioinformatics*, **2**, 260-262.

[88] Mazumder, R. and Vasudevan, S. (2008) Structure-guided comparative analysis of proteins: Principles, tools, and applications for predicting function. *PLoS Computational Biology*, **4**, e1000151. doi:10.1371/journal.pcbi.1000151

[89] Karasavvas, K.A., Baldock, R. and Burger, A. (2004) Bioinformatics integration and agent technology. *Journal of Biomedical Informatics*, **37**, 205-219. doi:10.1016/j.jbi.2004.04.003

[90] Li, A. (2006) Facing the challenges of data integration in biosciences. *Engineering Letter*, **13**, 3.

[91] Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer C., *et al*. (2010) The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, **28**, 935-942. doi:10.1038/nbt.1666

[92] Rubin, D.L., Shah, N.H. and Noy, N.F. (2008) Biomedical ontologies: A functional perspective. *Brief Bioinformatics*, **9**, 75-90. doi:10.1093/bib/bbm059

[93] Sarkar, I.N., Egan, M.G., Coruzzi, G., Lee, E.K. and DeSalle, R. (2008) Automated simultaneous analysis phylogenetics (ASAP): An enabling tool for phlyogenomics. *BMC bioinformatics*, **9**, 103. doi:10.1186/1471-2105-9-103

[94] Clark, T. (2007) Knowledge integration in biomedicine: Technology and community. *Brief Bioinformatics*, **8**, E1-E3. doi:10.1093/bib/bbm019

