Bioinformatics in a Nutshell

Zarrin Minuchehr[a]*, Samira Kheitan[a]

[a] Systems Biotechnology Group, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

Abstract. We are now confronting an overwhelmingly increase in biological data in public archives. With the complete genome sequence of human and other species available, there is now a constant need for bioinformatics databases and tools to analyze these data. Bioinformatics has recently focused on the manipulation of information and is playing a significant role in biology and human health and disease leading to a personalized medicine approach. Wet lab biologists use bioinformatics tools in a daily manner, sometimes without even noticing its existence and importance. It is unlikely for a research to be designed without prior information obtained using these powerful and time saving tools. This data will undoubtedly increase our knowledge in complex biological systems. Here we review different aspects in this interdisciplinary science by presenting a historical overview and introducing major databases and tools along with a glimpse in newly emerging fields including next-generation sequencing (NGS) technologies and systems biology.

Keywords: Bioinformatics; Databases; NGS; Systems biology.

1. Introduction

Human genome project was performed in the early 1980's [1]. Its aim was to create the human genetic map in order to cure unknown genetic and complex diseases, the result of this attempt led to an overwhelming data and information, emerging the computational biology science and bioinformatics. In this review we give an overview on the historical background of this exciting field of science. We also present a collection of major biological databases and tools, along with an introduction to NGS technologies and systems biology.

1.1. What really is bioinformatics?

Bioinformatics is said to be a shotgun marriage between biology, information science, mathematics and statistics (Fig. 1) [2].

There are several definitions for bioinformatics such as application of computer technology to the management of biological information, or the science of developing and utilizing computational tools to enhance biological research in many ways. Designing different biological databases and search engines, along with obtaining knowledge on the structure and properties of biological macromolecules (proteins, nucleic acids and complex molecules), all in one provide an explanation for bioinformatics. Another challenge for bioinformatics is to explore the data and to uncover biologically relevant interactions and pathways. Such knowledge is of crucial importance in many different areas, ranging from computer science and mathematics to medical, pharmaceutical sciences and plant technology. As far as we are concerned the best definition for bioinformatics is a union of all of these areas of scientific enterprise that evolves both practical and conceptual tools for the generation, dissemination, representation, analysis and understanding biological problems [3]. In this review we will give an introduction to bioinformatics centers, databases, sequence analysis, structure and function tools and techniques and as a whole the application of computer science in different biological aspects.

1.2. Historical overview

We now introduce some events which are of great value in bioinformatics. Margaret Dayhoff, a professor in Georgetown university, was the first scientist to write an atlas of protein sequence in 1960, this atlas was eventually maintained as an electronic database known as PIR (Protein Identification Resource) [4]. One of the most important events in bioinformatics, soon thereafter a program was developed for global sequence alignment by Needleman and Wunsch [5]. They compared two sequences starting at one end and

* Corresponding author.
E-mail address: minuchhr@nigeb.ac.ir

moving towards the other, this method was called dynamic programming. Finding local alignments between sequences was developed in 1981 by Smith and Waterman, they recognized that the most biologically significant regions in DNA and protein sequences were local parts, these parts align well and the remaining regions of less-related sequences were less significant [6]. The method was called local alignment algorithm for sequence alignment. The next important event in bioinformatics evolution was finding the concept of sequence motifs designing [7]. Soon afterwards the first DNA database was designed with the name of GenBank which was first founded in 1982. Many types of databases are introduced each year in the first issue of Nucleic Acids Research [8]. The challenge for different sequence database searching techniques was then started and the fast sequence similarity searching was developed [9, 10]. Three years later the National Center for Biotechnology Information (NCBI) at the National Library of Medicine of NIH (http://www.ncbi.nlm.nih.gov/) and the EMBnet network for database distribution were created (http://www.embnet.org/). Blast or the Basic Local Alignment Search Tool was then developed for finding local similarities between sequences [11]. In 1995 the Haemophilus influenzea genome was completely sequenced, Mycoplasma genitalium genome was sequenced in the same year [12, 13]. These helped to provide a model of the minimum number of genes needed for independent existence. The genome for E. coli and Caenorhabitis elegans, were published in 1997 and 1998 respectively. The first human chromosome was sequenced completely in 1999 and the genome for Pseudomonas aeruginosa was published in 2000. The Human Genome Project was completed in 2003 which was a start point for challenges in creating different databases and improving searching techniques.

2. Bioinformatics databases

2.1. Nucleotide databases

In 1970's scientists started to develop methods for nucleic acid sequencing with two sequencing methods; chain termination and chemical degradation [14, 15]. We already know that humans carry ~3,000 megabases (i.e., 3 × 10^9 bases or base pairs) of DNA in each of their cells. The Human Genome Project (HGP) started in 1990 with the collaboration of the US Department of Energy and National Institutes of Health as a 15-year program to determine the sequence of the complete nucleotide content of the human genome. Three major DNA databases were designed in different parts of the world, all three major DNA databases were interconnected and exchange data in a daily manner. These three major DNA databases are:

- Genbank (USA): This database can be accessed from the National Center for Biotechnology Information (NCBI) [16].
- EMBL (Europe): The European site for storage of DNA sequences is at the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) at Hinxton, United Kingdom [17].
- DDBJ (Japan): The DNA Data Bank of Japan (DDBJ) is also one of the major locations for storing DNA data in the world [18].

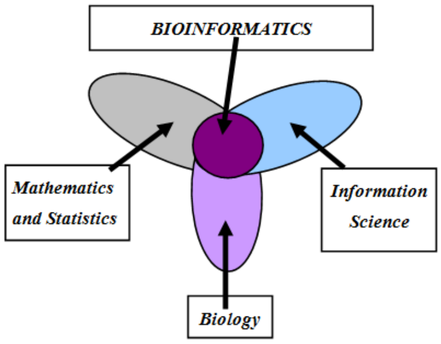

Fig. 1. Bioinformatics as a multidisciplinary science.

2.1. Protein databases

In a historical point of view the first complete sequence to be found was insulin [19] and the first complete enzyme sequence to be found was ribonuclease [20]. By 1965 only 20 proteins were sequenced, but today we have more 500000 protein sequences available in the Swiss-Prot protein sequence database [21]. Sequence databases, such as Swiss-Prot + TrEMBL [22] and PIR-PSD [23], coexisted as protein databases with differing sequence coverage and annotation priorities. In 2002, the Swiss-Prot + TrEMBL groups at the SIB (Swiss Institute of Bioinformatics) and EBI (European Bioinformatics Institute) and the PIR (Protein Information Resource) group at Georgetown University Medical Center and National Biomedical Research Foundation joined forces as the UniProt consortium [24].

The UniProt consortium maintains three database layers:

The UniProt Archive (UniParc) which is said to provide a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data [25]. Although most protein sequence data are derived from the translation of DDBJ/EMBL/GenBank sequences, primary protein sequence data are also submitted directly to UniProt or appear in patent applications or in entries from the Protein Data Bank (PDB) [26]. The UniParc is designed to capture all available protein sequence data—not just from the forementioned databases, but also from sources such as Ensembl [27], RefSeq [28], FlyBase [29] and WormBase [30]. This combination of sources is said to make UniParc the most comprehensive publicly accessible, non-redundant protein sequence database available.

The UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation. The UniProt Knowledgebase merges Swiss-Prot, TrEMBL and PIR-PSD to provide a central database of protein sequences with annotations and functional information.

The UniProt Reference (UniRef) databases provide

Table 1. A list of major protein databases.

Database	Description	URL
BLOCKS	Multiple alignments of conserved regions of protein families.	http://blocks.fhcrc.org/
CATH	Protein domain structures.	http://www.biochem.ucl.ac.uk/bsm/cath
ENZYME	Enzyme nomenclature database.	http://www.expasy.ch/enzyme/
FSSP	FSSP stands for "Fold classification based on Structure-Structure alignment of Proteins".	http://ekhidna.biocenter.helsinki.fi/dali/start
PDB	Structure data determined by X-ray crystallography and NMR.	http://www.rcsb.org
PDBsum	Summary of key information on structures in PDB.	http://www.biochem.ucl.ac.uk/bsm/pdbsum
Pfam	Multiple sequence alignments and hidden Markov models of common protein domains.	http://www.sanger.ac.uk/Pfam
PIR-PSD	Comprehensive, annotated, non-redundant protein sequence databases.	http://pir.georgetown.edu/
PRINTS	A compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
InterPro	Functional analysis of proteins by classifying them into families and predicting domains and important sites	http://www.ebi.ac.uk/interpro/
iPROCLASS	The iProClass database provides value-added information reports for UniProtKB and unique UniParc proteins, with links to over 90 biological databases.	http://pir.georgetown.edu/pirwww/dbinfo/iproclass .shtml
Prosite	Biologically-significant protein patterns and profiles.	http://www.expasy.ch/prosite
SCOP	Familial and structural protein relationships.	http://scop.mrc-lmb.cam.ac.uk/scop

non-redundant data collections based on the UniProt Knowledgebase and UniParc in order to obtain complete coverage of sequence space.

Automatic procedures have been developed to create three UniRef databases, such as UniRef100, UniRef90 and UniRef50, from the UniProt Knowledgebase and UniParc as representative protein sequence databases, with high information content. The database provides complete coverage of sequence space while hiding redundant sequences from view. The non-redundancy facilitates sequence merging in the UniProt Knowledgebase (based on UniRef100) and allows faster sequence similarity searches (by using UniRef90 and UniRef50). There are many other protein databases available worldwide, for a comprehensive overview see Minuchehr and Goliaei [31], Table 1 lists some major protein databases and their URL.

3. Protein sequence analysis

Much more protein sequences are determined compared to protein structures in a daily manner (UniProtKB/Swiss-Prot Release of 2016_07 included 551705 entries), 3-D structure of proteins (PDB Holdings List: 2016_07 included 121654 macromolecule entries) is significantly much more difficult to determine. We should value the amino acid sequence of proteins in determining the overall fold of proteins as mentioned earlier but the relationship between sequence and structure is only partially yet understood [32]. Protein propensities remain excellent descriptors of amino acid tendencies to belong to different secondary structures, alpha helices, beta strands, loops and turns. Examining the frequency of occurrence of different amino acids in protein secondary structures may give us an insight into the prediction of the three dimensional fold of the proteins or even de-novo design of a desired fold. Studies have been conducted on preferences of amino acids in different secondary structures residues. There has been seen that alanine, glutamate and leucine tend to be present in alpha helices whereas valine and isoleucine tend to be present in strands, valine and isoleucine tend to destabilize alpha helices due to the steric clashes in the branching at beta carbon atom, but they are at the same time abundant in beta strands. Since studying the propensity of amino acids in different secondary structures is an important task to perform, as the protein data bank tends to grow, scientists have performed more specific studies on the propensity of amino acids in different positions in the secondary structures such as alpha helices [33-35] and less extensively on beta strands [36, 37]. Loops are also of functional importance in biology and may have key roles in recognition (antibody hyper variable loops); ligand binding (e.g. Triosphosphate isomerase [38]) or forming enzyme active sites (e.g. Serine protease, [39]). It has been known that loops are the most difficult structures of globular proteins to be modeled [40] The most successful modeling occurs when loops from homologous structures are available [41]. Besides, we can define loops as segments that do not correspond to alpha-helical or beta-strand secondary structures. There have been many attempts to classify loops, particularly Strand-Loop-Strand classes [42-44]. different studies have thus far tried to predict these particular structures [45-47] and many scientists have calculated the amino acid propensities in different secondary structures [35, 48-50], but there is very little work done on loop regions [51].

4. Bioinformatics tools

A wide range of tools are needed to deal with large amount of data being generated due to the Human Genome Project and related biological projects worldwide. A number of tools have been developed to extract knowledge from the rich and complex data available. Here we describe some of the commonly used bioinformatics methods and tools and include examples of their applications.

4.1. Alignment similarity search tools

4.1.1. FASTA

FASTA (pronounced FAST-Aye) stands for FAST-All, is a program for rapid alignment of pairs of protein and DNA sequences. This program achieves a high level of sensitivity for similarity searching at high speed. The high speed of this program is achieved by using the observed pattern of word hits called k-tuples which the trade-off between speed and sensitivity is controlled by this parameter [9, 52]. Increasing the k-tuple decreases the number of background hits. FASTA is useful for database searches of different types.

4.1.2. Blast

BLAST stands for Basic Local Alignment Search Tool. This tool finds the regions of sequence similarity and compares a new sequence to sequences which are already deposited to sequence databases. The blast algorithm was developed in a way to be faster than FASTA. Access to this system is possible through https://blast.ncbi.nlm.nih.gov/Blast.cgi [11].

4.1.3. Clustal omega

Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) is the latest version of Clustal family that uses seeded guide trees and HMM profile-profile techniques. This program can generate alignments between hundreds of thousands of sequences in a superior quality and increased scalability over previous versions. Clustal Omega also has powerful features such as adding sequences to and exploiting information in existing alignments and applying precomputed HMMs from public databases like Pfam [53].

4.2. Molecule structure visualization tools

Molecular structure viewers usually show molecules (PDB files) in different shapes such as wireframe of carbon alpha backbone; space-filling and secondary structure ribbon. There are many viewers developed for this purpose, we here mention some most popular ones and the URL to access them via internet.

4.2.1. CN3D

This program provides the three dimensional structure from Entrez. Cn3D is a helper application for the web browser that allows viewing three dimensional structures from NCBI's Entrez retrieval service [54] Cn3D simultaneously displays structure, sequence, and alignment, it also allows the user to set display styles for features of interest. The program URL is: http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml.

4.2.2. Protein explorer

This program is an easy way to visualize macromolecular structures. Protein Explorer (PE) is built upon Chime a molecular graphics browser which is free [55]. PE is free and can operate on Windows, Macintosh computers and Linux or SGI/Irix platforms. In order to download and use Protein Explorer you can go to http://www.umass.edu/microbio/chime/pe/protexpl/frntdoor.htm.

4.2.3. Rasmol

This viewer is the most commonly used viewer worldwide. In order to access the program you should use http://www.umass.edu/microbio/rasmol/. Rasmol is a molecular graphics program for visualizing proteins, nucleic acids and small molecules [56]. Rasmol reads a number of file formats. Currently supported file formats are Brookhaven Protein Databank (PDB), Molecular Design Limited (MDL), Mol file format and many other formats. The molecules can be seen as wireframe, cylinder, spacefilling (CPK), ribbons and dot surface. This program can run on a wide range of systems including SGI, Sun, DEC, IBM RS/6000, Microsoft Windows, Apple Macintosh and Linux.

4.2.4. Swiss-PdbViewer

This program can be accessed viahttp://spdbv.vital-it.ch/. In addition to showing the molecule the program calculates angels and distances [57]. This application can allow analyzing many proteins at the same time and can also read electron density maps. The latest stable version of the program is 4.1 and is available for Mac, PC, SGI and Linux.

5. Evolution of bioinformatics software's and resources

In order to understand the cell machinery, identifying the proteins and their function would not be enough. Interactions between proteins or protein-protein interactions are crucial in understanding the role of the cell in its different biological functions [22]. Therefore many bioinformatics software's and resources have been merged to fulfill such needs. There are many bioinformatics tools developed by bioinformaticians to use the rapidly growing pool of molecular biology data.

For this purpose there are systems designed to integrate biological tools and data, experts divide these systems into two groups [58].

1 Systems based on a centralization or data warehousing strategy

2 Systems based on a federated or distributed strategy

Centralization has many benefits; the most important one is speed. Centralized strategy is the only realistic option; this is because tools can fetch data much faster from a local hard drive than from a source on the internet. There are also other benefits that locally installed tools have more control over updating processes.

Using the federated strategy maintenance, money is saved and the data and tools are accessed remotely. HTML web interfaces were widely used for bioinformatics purposes but are not suitable for programmatic access. Therefore, XML (eXtensible Mark-up Language) was designed to overcome the limitations of HTML for bioinformatics. XML is used for data description, service description and service discovery. The standards are, Simple Object Access Protocol (SOAP) which uses XML to create messaging framework that can exchange data over underlying protocols (http://www.w3.org/TR/soap/). Web Services Description Language (WSDL) an XML format for describing network services (http://www.w3.org/TR/wsdl), Universal Description Discovery and Integration protocol (UDDI) a standard to create service directories (http://www.uddi.org) that enable applications to dynamically find the use of web services. SOAP has already gained a dominant position in the bioinformatics community and services are now available such as Distributed Annotation System (DAS), (http://biodas.org) [59]. This software provides access to complete genome annotations using a SOAP web interface. Another service available is Pathway Database System and KEGG API which provide access to pathways using SOAP web interface [60]. PDBML [61] is an XML-based service for PDB data, xPSSSS is a tool which provides a SOAP based service to retrieve PDBML. XEMBL is another service in EBI [62]. There are several other services such as MAGE-ML [63], AGML [64], Jemboss [65] which all use SOAP as their web interface. There are also many bioinformatics projects based on these services such as BIOMOBY [66] myGrid [67], Discovery Net68 and caCore [69].

Although many bioinformatics projects and tools are available, it is not always easy to find the relevant ones. Searching the web using general search engines such as Google, Yahoo, AltaVista, etc. cannot always help you in finding the appropriate tool for your purpose. There are many bioinformatics resources which you can use in finding the best available research tool [70].

6. Next generation sequencing (NGS)

Next generation sequencing or NGS is a newly developed technique for sequencing. It is also called high throughput sequencing as well, there are different kinds of NGS depending on the company in which it is developed such as Illumina sequencing, Roche 454 sequencing, Ion DNA nanoball sequencing, Ion torrent sequencing, SOLiD sequencing and Heliscope single molecule sequencing. NGS allows RNA and DNA to be sequenced more rapidly and much cheaper and has revolutionized the molecular biology studies worldwide. NGS technology is playing a crucial role in today's approach in personalized medicine despite its young age. The mechanisms of complex diseases such as heart diseases, diabetes, neurological disorders and cancer can be identified using NGS technologies, these mechanisms include; genetics variations, post translational modification variations, epigenetic variation, single nucleotide variations (SNVs) and indel mutations. NGS methods have made genomics one of the first scientific areas entering Big Data era.

7. Systems biology

Systems biology is an interdisciplinary field of study closely related to bioinformatics sometimes said to be a branch of it which models biological systems using mathematical and computational approaches and is a systemic thinking and solving biological problems. Systems biology and bioinformatics are the key elements of modern biology and medical sciences. Systems biology as a whole combines the wet lab concept with computational modeling of the biological processes in our body, studying complex diseases and their mechanisms which require an integrated and multidisciplinary approach.

Since human is a very complex system, progress towards understanding diseases is a very difficult task to perform [71] and therefore systems biology is a concept to recognize the emergent properties of life.

The platforms of the systems level analysis of biology are generally referred to as OMICS such as; Genomics [72], Proteomics [73], Transcriptomics [74], Metabolomics [75], Lipidomics [76] etc. approaches, in which they all have applications in the study of life complex systems, drug discovery and generation of a deeper insight into the biological causes of diseases. Network biology has generated much more knowledge than gene expression individual studies, yielding us to personalized medicine. Systems biology approaches in drug discovery both increases the drug efficiency and decreases its side effects [77], therefore systems biology now a days has been a center of attention in bioinformatics.

8. Bioinformatics journals

If you are interested in reading new articles in bioinformatics or even interested in writing papers in the field of computational biology and bioinformatics you can find some useful journals and their related links in Table 2.

At the end we should mention that the application of bioinformatics as a science has been reviewed in different aspects such as functional genomics [78], nanobiosciences [79], proteomics [80], cellular signaling [81], regulatory elements identification [82], alternative splicing [83], and even industry [84] which could be of special interest for scientists in different fields.

Table 2. Highly impacted journals in bioinfomatics.

Full Journal Title	Impact Factor	Link
Molecular Systems Biology	10.872	http://msb.embopress.org/
Briefings in Bioinformatics	9.617	http://bib.oxfordjournals.org/
Bioinformatics	4.981	http://bioinformatics.oxfordjournals.org/
Epigenomics	4.649	http://www.futuremedicine.com/loi/epi
Pharmacogenomics Journal	4.229	http://www.nature.com/tpj/index.html
BMC Genomics	3.986	https://bmcgenomics.biomedcentral.com/
Journal of Proteomics	3.888	http://www.journals.elsevier.com/journal-of-proteomics/
Metabolomics	3.855	http://link.springer.com/journal/11306
Briefings in Functional Genomics	3.67	http://bfg.oxfordjournals.org/
Journal of Genetics and Genomics	3.585	http://www.journals.elsevier.com/journal-of-genetics-and-genomics
Pharmacogenetics and Genomics	3.481	http://journals.lww.com/jpharmacogenetics/pages/default.aspx
Protein & Cell	3.247	http://link.springer.com/journal/13238
Molecular BioSystems	3.21	http://www.rsc.org/journals-books-databases/about-journals/molecular-biosystems/
Wiley Interdisciplinary Reviews-Systems Biology and Medicine	3.205	http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1939-005X
Current Protein & Peptide Science	3.154	http://benthamscience.com/journals/current-protein-and-peptide-science/
Advances in Protein Chemistry and Structural Biology	3.036	http://www.sciencedirect.com/science/bookseries/18761623
Protein Science	2.854	http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1469-896X
Biochimica et Biophysica Acta-Proteins And Proteomics	2.747	http://www.sciencedirect.com/science/journal/15709639
Proteins-Structure Function and Bioinformatics	2.627	http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0134
BMC Bioinformatics	2.576	https://bmcbioinformatics.biomedcentral.com/
Protein Engineering Design & Selection	2.537	http://peds.oxfordjournals.org/
BMC Systems Biology	2.435	https://bmcsystbiol.biomedcentral.com/

Acknowledgments

We would like to thank the founders of Bioinformatics Society in Iran for their valuable contribution to the field.

References

1. Wright, J. T., and T. C. Hart. “The genome projects: implications for dental practice and education.” Journal of Dental Education 66.5 (2002): 659-671.
2. Spengler, Sylvia J. “Bioinformatics in the information age.” Science 287.5456 (2000): 1221-1223.
3. Attwood, T. K., and C. J. Miller. “Progress in bioinformatics and the importance of being earnest.” Biotechnology annual review 8 (2002): 1-54.
4. Dayhoff, Margaret O. “Computer aids to protein sequence determination.” Journal of theoretical biology 8.1 (1965): 97-112.
5. Needleman, Saul B., and Christian D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” Journal of molecular biology 48.3 (1970): 443-453.
6. Smith, Temple F., and Michael S. Waterman. “Identification of common molecular subsequences.” Journal of molecular biology 147.1 (1981): 195-197.
7. Doolittle, Russell F. “Similar amino acid sequences: chance or common ancestry.” Science 214.4517 (1981): 149-159.
8. Rigden, Daniel J., Xosé M. Fernández-Suárez, and Michael Y. Galperin. “The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection.” Nucleic acids research 44.D1 (2015): D1-D6.
9. Wilbur, W. John, and David J. Lipman. “Rapid similarity searches of nucleic acid and protein data banks.” Proceedings of the National Academy of Sciences 80.3 (1983): 726-730.
10. Lipman, David J., and William R. Pearson. “Rapid and sensitive protein similarity searches.” Science 227.4693 (1985): 1435-1441.
11. Altschul, Stephen F., et al. “Basic local alignment search tool.” Journal of molecular biology 215.3 (1990): 403-410.
12. Fleischmann, Robert D., et al. “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.” Science 269.5223 (1995): 496.
13. Fraser, Claire M., et al. “The minimal gene complement of Mycoplasma genitalium.” Science 270.5235 (1995): 197.
14. Maxam, Allan M., and Walter Gilbert. “A new method for sequencing DNA.” Proceedings of the National Academy of Sciences 74.2 (1977): 560-564.
15. Sanger, Frederick, Steven Nicklen, and Alan R. Coulson. “DNA sequencing with chain-terminating inhibitors.” Proceedings of the national academy of sciences 74.12 (1977): 5463-5467.
16. Benson, Dennis A., et al. “GenBank.” Nucleic acids research 41.D1 (2012): D36-D42.
17. Kanz, Carola. “The EMBL nucleotide sequence database: Nucleic Acids Res.” January 1 (2005): 33.
18. Mashima, Jun, et al. “DNA data bank of Japan (DDBJ) progress report.” Nucleic acids research 44.D1 (2015): D51-D57.
19. Ryle, A. P., et al. “The disulphide bonds of insulin.” Biochemical Journal 60.4 (1955): 541.
20. Hirs, C. H. W. “The structure of ribonuclease.” Annals of the New York Academy of Sciences 88.3 (1960): 611-641.
21. Bairoch, Amos, et al. “Swiss-Prot: juggling between evolution and stability.” Briefings in bioinformatics 5.1 (2004): 39-55.
22. Boeckmann, Brigitte, et al. “The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.” Nucleic acids research

31.1 (2003): 365-370.

23. Wu, Cathy H., et al. “The protein information resource.” Nucleic acids research 31.1 (2003): 345-347.
24. UniProt Consortium. “UniProt: a hub for protein information.” Nucleic acids research (2014): gku989.
25. Leinonen, Rasko, et al. “UniProt archive.” Bioinformatics 20.17 (2004): 3236-3237.
26. Berman, Helen M., et al. “The Protein Data Bank, 1999–.” International Tables for Crystallography Volume F: Crystallography of biological macromolecules. Springer Netherlands, 2006. 675-684.
27. Yates, Andrew, et al. “Ensembl 2016.” Nucleic acids research (2015): gkv1157.
28. Tatusova, Tatiana, et al. “RefSeq microbial genomes database: new representation and annotation strategy.” Nucleic acids research (2013): gkt1274.
29 Attrill, Helen, et al. “FlyBase: establishing a Gene Group resource for Drosophila melanogaster.” Nucleic acids research 44.D1 (2015): D786-D792.
30. Howe, Kevin L., et al. “WormBase 2016: expanding to enable helminth genomic research.” Nucleic acids research (2015): gkv1217.
31. Minuchehr, Zarrin, and Bahram Goliaei. “Protein Databases.” Iranian Journal of Pharmacology & Therapeutics 3.1 (2004): 1-11.
32. Baker, David, and Andrej Sali. “Protein structure prediction and structural genomics.” Science 294.5540 (2001): 93-96.
33. Kumar, Sandeep, and Manju Bansal. “Dissecting a-helices: position-specific analysis of a-helices in globular proteins.” Proteins-Structure Function and Genetics 31.4 (1998): 460-476.
34. Aurora, R. and G.D. Rose, Helix capping. Protein Sci. 7(1): 21-38, 1998.
35. Goliaei, Bahram, and Zarrin Minuchehr. “Exceptional pairs of amino acid neighbors in α-helices.” FEBS letters 537.1-3 (2003): 121-127.
36. Pal, Debnath, and Pinak Chakrabarti. “β-Sheet propensity and its correlation withparameters based on conformation.” Acta Crystallographica Section D: Biological Crystallography 56.5 (2000): 589-594.
37. Ghamkhar, M., Z. Minuchehr, and B. Goliaei, Propensity calculation for amino acids in different positions in beta strands. in8th Iranian Congress of Biochemistry, 2005.
38. Joseph, Diane, Gregory A. Petsko, and Martin Karplus. “Anatomy of a conformational change: hinged” lid” motion of the triosephosphate isomerase loop.” Science 249.4975 (1990): 1425-1429.
39. Wlodawer, Alexander, et al. “Crystal Structure of Synthetic HIV-Protease.” Science 245 (1989): 6I6.
40. Sali, Andrej. “Comparative protein modeling by satisfaction of spatial restraints.” Molecular medicine today 1.6 (1995): 270-277.
41. Martin, Andrew CR, Malcolm W. MacArthur, and Janet M. Thornton. “Assessment of comparative modeling in CASP2.” Proteins: Structure, Function, and Bioinformatics 29.S1 (1997): 14-28.
42. Burke, David F., Charlotte M. Deane, and Tom L. Blundell. “Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure.” Bioinformatics 16.6 (2000): 513-519.
43. Efimov, A. V. “Structure of α-α-hairpins with short connections.” Protein engineering 4.3 (1991): 245-250.
44. Wintjens, René T., Marianne J. Rooman, and Shoshana J. Wodak. “Automatic classification and analysis of αα-turn motifs in proteins.” Journal of molecular biology 255.1 (1996): 235-253.
45. Burke, David F., and Charlotte M. Deane. “Improved protein loop prediction from sequence alone.” Protein engineering 14.7 (2001): 473-478.
46. Kuhn, Michael, Jens Meiler, and David Baker. “Strand-loop-strand motifs: Prediction of hairpins and diverging turns in proteins.” PROTEINS: Structure, Function, and Bioinformatics 54.2 (2004): 282-288.
47. van Vlijmen, Herman WT, and Martin Karplus. “PDB-based protein loop prediction: parameters for selection and methods for optimization.” Journal of molecular biology 267.4 (1997): 975-1001.
48. Kumar, Sandeep, and Manju Bansal. “Structural and sequence characteristics of long alpha helices in globular proteins.” Biophysical journal 71.3 (1996): 1574-1586.
49. Penel, Simon, Eleri Hughes, and Andrew J. Doig. “Side-chain structures in the first turn of the α-helix.” Journal of molecular biology 287.1 (1999): 127-143.
50. Penel, Simon, et al. “Periodicity in α-helix lengths and C-capping preferences.” Journal of molecular biology 293.5 (1999): 1211-1219.
51. Minuchehr, Z., and B. Goliaei. “Propensity of amino acids in loop regions connecting beta-strands.” Protein and peptide letters 12.4 (2005): 379-382.
52. Pearson, William R. “[5] Rapid and sensitive sequence comparison with FASTP and FASTA.” Methods in enzymology 183 (1990): 63-98.
53. Sievers, Fabian, et al. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.” Molecular systems biology 7.1 (2011): 539.
54. Wang, Yanli, et al. “Cn3D: sequence and structure views for Entrez.” (2000): 300-302.
55. Martz, Eric. “Protein Explorer: easy yet powerful macromolecular visualization.” Trends in biochemical sciences 27.2 (2002): 107-109.
56. Sayle, Roger A., and E. James Milner-White. “RASMOL: biomolecular graphics for all.” Trends in biochemical sciences 20.9 (1995): 374-376.
57. Guex, Nicolas, and Manuel C. Peitsch. “SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling.” electrophoresis 18.15 (1997): 2714-2723.
58. Neerincx, Pieter BT, and Jack AM Leunissen. “Evolution of web services in bioinformatics.” Briefings in bioinformatics 6.2 (2005): 178-188.
59. Dowell, Robin D., et al. “The distributed annotation system.” BMC bioinformatics 2.1 (2001): 7.
60. Krishnamurthy, Larkshmi, et al. “Pathways Database System: An integrated set of tools for biological pathways.” Proceedings of the 2003 ACM symposium on Applied computing. ACM, 2003.
61. Westbrook, John, et al. “PDBML: the representation of archival macromolecular structure data in XML.” Bioinformatics 21.7 (2004): 988-992.
62. Wang, Lichun, Jean-Jack Riethoven, and Alan Robinson. “XEMBL: distributing EMBL data in XML format.” Bioinformatics 18.8 (2002): 1147-1148.
63. Tjandra, Donny, et al. “An XML message broker framework for exchange and integration of microarray data.” Bioinformatics 19.14 (2003): 1844-1845.
64. Stanislaus, Romesh, et al. “AGML Central: web based gel proteomic infrastructure.” Bioinformatics 21.9 (2005): 1754-1757.
65. Carver, Tim, and Alan Bleasby. “The design of Jemboss: a graphical user interface to EMBOSS.” Bioinformatics 19.14 (2003): 1837-1843.
66. Wilkinson, Mark, et al. “BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case.” Plant physiology 138.1 (2005): 5-17.
67. Stevens, Robert D., Alan J. Robinson, and Carole A. Goble. “myGrid: personalised bioinformatics on the information grid.” Bioinformatics 19.suppl 1 (2003): i302-i304.
68. Rowe, Anthony, et al. “The discovery net system for high throughput bioinformatics.” Bioinformatics 19.suppl 1 (2003): i225-i231.
69. Covitz, Peter A., et al. “caCORE: a common infrastructure for cancer informatics.” Bioinformatics 19.18 (2003): 2404-2412.
70. Gilbert, Don. “Bioinformatics software resources.” Briefings in bioinformatics 5.3 (2004): 300-304.
71. Gallagher, Iain J., et al. “Omics/systems biology and cancer cachexia.” Seminars in cell & developmental biology. Vol. 54. Academic Press, 2016.
72. Swaminathan, Rajeswari, et al. “A review on genomics APIs.” Computational and structural biotechnology journal 14 (2016): 8-15.
73. Eckhard, Ulrich, et al. “Positional proteomics in the era of the human proteome project on the doorstep of precision medicine.” Biochimie 122 (2016): 110-118.
74. Li, Shuzhao, Andrei Todor, and Ruiyan Luo. “Blood transcriptomics and metabolomics for personalized medicine.” Computational and structural biotechnology journal 14 (2016): 1-7.

75. Marcinkiewicz-Siemion, M., et al. "Metabolomics—A wide-open door to personalized treatment in chronic heart failure?." International Journal of Cardiology 219 (2016): 156-163.

76. Zhao, Ying-Yong, et al. "Lipidomics: Novel insight into the biochemical mechanism of lipid metabolism and dysregulation-associated disease." Chemico-Biological Interactions 240 (2015): 220-238.

77. Wierling, Christoph, et al. "Network and systems biology: essential steps in virtualising drug discovery and development." Drug Discovery Today: Technologies 15 (2015): 33-40.

78. Singh, Gautam B., and Harkirat Singh. "Databases, models, and algorithms for functional genomics." Molecular biotechnology 29.2 (2005): 165-183.

79. Oakley, Barbara A., and Darrin M. Hanna. "A review of nanobioscience and bioinformatics initiatives in North America." IEEE transactions on nanobioscience 3.1 (2004): 74-84.

90. Blueggel, Martin, Daniel Chamrad, and Helmut E. Meyer. "Bioinformatics in proteomics." Current pharmaceutical biotechnology 5.1 (2004): 79-88.

81. Papin, Jason, and Shankar Subramaniam. "Bioinformatics and cellular signaling." Current opinion in biotechnology 15.1 (2004): 78-81.

82. Wasserman, Wyeth W., and Albin Sandelin. "Applied bioinformatics for the identification of regulatory elements." Nature Reviews Genetics 5.4 (2004): 276-287.

83. Lee, Christopher, and Qi Wang. "Bioinformatics analysis of alternative splicing." Briefings in Bioinformatics 6.1 (2005): 23-33.

84. Zolg, J. Werner, and Hanno Langen. "How industry is approaching the search for new diagnostic markers and biomarkers." Molecular & Cellular Proteomics 3.4 (2004): 345-354.