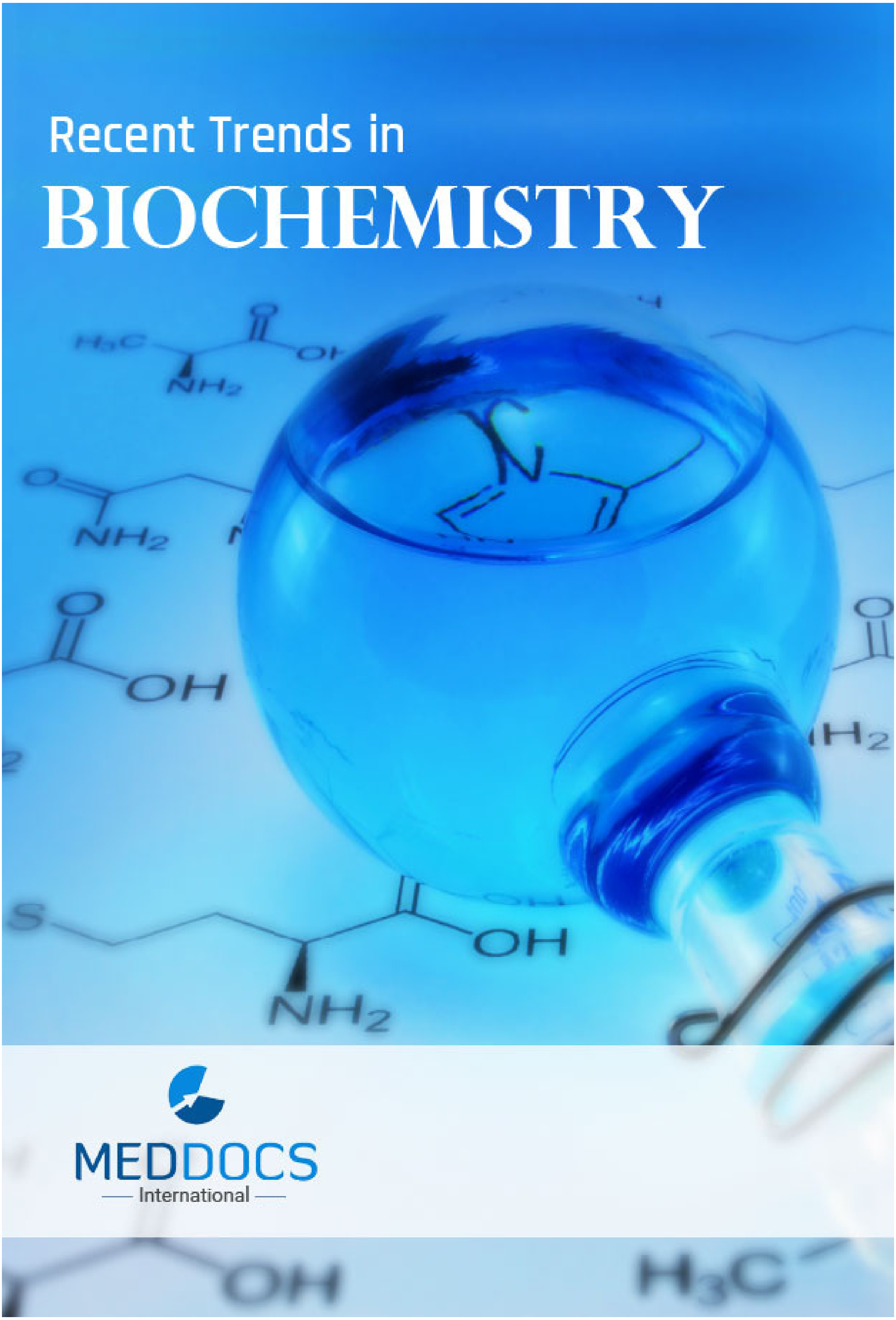
Recent Trends in BIOCHEMISTRY



MEDDOCS
International

Genome Databases, Types and Applications: An overview

Jeyachandran Sivakamavalli[1]*; Kiyun Park[1]; Ihn-Sil Kwak[1,2]

[1]Fisheries Science Institute, Chonnam National University, Yeosu 59626, South Korea

[2]Faculty of Marine Technology, Chonnam National University, Chonnam 550-749, Republic of Korea

Corresponding Author: Jeyachandran Sivakamavalli

Fisheries Science Institute, Chonnam National University, Yeosu 59626, South Korea

Email: dr.jsvalli@gmail.com

Published Online: Jul 07, 2020

eBook: Recent Trends in Biochemistry

Publisher: MedDocs Publishers LLC

Online edition: http://meddocsonline.org/

Keywords: Genome; NCBI; PDB; UniProt; Proteome; Sequencing

Abstract

Genomic and proteomic databases are very important useful platform to store, share and compare the data across research purposes, between individuals and other organisms. Development of molecular biological techniques and computational approaches such as genome sequencing, trancriptomics, proteomics and metabolic studies unravels the history of biomolecules, interactions inside the cell. This kind of enormous big data (experimental data) are difficult for analysis, hence to store this data genomic database, proteomic databases is much needed. Here, this chapter displays the numerous databases are existing especially for molecular biology, amongst genome and proteome databases such National Centre for Biological Information (NCBI), UniProtKB and Protein Data Bank (PDB) plays the vital role in research environment and medical purposes. Genomic and proteomic databases such as NCBI and Protein Data Bank (PDB) are very helpful to know research history about the genome of any organism, protein function, proteome nature etc. This existing databases will assist to understand the new data and also easy for the comparison and new novel data conclusions.

Introduction

Nowadays researchers explored the uniqueness of DNA sequencing to release the genetic code of numerous diverse organisms to reveal the function of the every organ inside the animal model. From the development of DNA researchers attempted to find the sequencing of complete DNA of many organisms, in some organisms and plants already the whole DNA sequencing genome was established such as human, mouse, rat, bacterial, and plant genomes [1-3]. Form this findings scientist conclude that the most of the biological functions are genetically conserved within and between species, this informs the by gaining the knowledge will helpful to understand the more information about human genome. Sequencing the genomes of diverse organisms brings the greater the intellectual yield DNA sequencing provides significant clue regarding the genes and proteins that are obligatory to generate and sustain related species [4,5].

Sequencing of the genome for all organisms is not possible because of its high cost and time consuming process, for instance obtaining a draft sequence of a mammalian genome costs as much as 100 million dollars. For commercial purposes many kind of animals genomics informations are explored rapidly domestic oriented animals such as pigs, sheep, chickens, cattle, horses, and companion animals such as dogs and cats. Sequencing of animals offers a great potential for move forward in human and animal health knowledge, improving animal production practices, and which brings the economic benefits. For instance, gene or genes that confers the disease resistance in plants [6] or animals reducing health improvement in animal and plants which outputs the animal production industries. Furthermore, some instances these animals have a sentimental value that distinguishes them from other organisms. This kind of benefits majorly occurs in agriculture and aquaculture industries and companion animal science, evolutionary biology, and human health with respect to the creation of models for genetic disorders, the National Academies have the plan to organize the public workshop towards the: (1) Assess these contributions; (2) Identify potential research directions for existing genomics programs; and (3) Highlight the opportunities of a coordinated, multi-species genomics effort for the science and policymaking communities. Their efforts culminated in a workshop sponsored by the U.S. Department of Agriculture, Department of Energy, National Science Foundation, and the National Institutes of Health. The workshop was convened on February 19, 2002. The goal of the workshop was to focus on domestic animal genomics and its integration with other genomics and functional genomics projects [7]. One can frame the issue in terms of access to data, "When it comes to data access," there are two ways to think about it. In order to empower all of the users that are interested in getting a hold of these data, are far better databases and tools to really exploit the information [8]. And I think this is an area that so far has been more of an afterthought with these projects than it should have been.

The result of some genomics researchers ends up having easier access to the data than others. "We are seeing a bit of a genomics-divide being created between those groups that are involved in generating the data and have been forced to build the tools in order to manipulate it, and the more typical user who doesn't necessarily have access to the same tools, (and) who expertise at his or her university. Several genome projects generally make no grant for taking care of the data generate once the project is finished [9]. For the most part, even for sequencing projects with bioinformatics support throughout the term of the project, that supports ends when the sequence is completed [10]. There's been no sketch put in position for how to preserve and update all of this information.

While accumulating the data the data storage and data transferability is the main issue, because sometimes the data collection and management are different sectors, while partition the data tax on genome projects that goes to fund a bioinformatics trust managed by an inter-agency group responsible for maintaining these databases [11].

Several contributors pointed out that in order to exploit the value of the information generated by domestic animal genome projects, researchers and information technology specialists will have to pay more attention to data handling. In particular, programs need to be designed not only to maintain the data and make it accessible to any researcher who needs it but also to make sure the information can be integrated with new data and new understandings. The Institute for Genomic Research (TIGR), made a similar point and National Center for Bioinformatics (NCBI) is doing a heroic job [12,13]. Both are doing an amazing job managing sequence data and publication data. That's a specific data type, and they have a fighting chance of scaling up for just the raw sequence information. But there's another data type that a lot of us are familiar with, which is annotation. Annotation is used to identify the functional genes and functional genome assignments which are recognized and structured in a database [14,15].

Structuring genome databases

Structuring of genome database is very important because when we structuring the data become much easier, and there is condensed reinvention of the wheel. Once the data format are assigned and structure it's easy to apply to new organisms." in addition, these data-specific centers able to expand easily adequate to accommodate ever-growing amounts of data. Suppose that some individual research center had developed a good way to represent expression information for the particular organism studied at that center. "Hopefully those are generalizing their services enough so they can apply them to another organism [16]. Then if those instantiate what the standard operational procedures are, they develop a relatively good training program, and they have a robust representation system going on in the database.

A member of the audience disagreed with this suggestion. however, it made more sense to keep smaller, individualized databases and develop standards so that the various databases could exchange information and work with each other almost as if who had a single database [17,18]. Try to create a level of information that can be exchanged among databases. In part, this goes along the lines of the discussions about whether the sequence in a center only or distribute the work in order to create local communities of scientists and train graduate students. This is particularly true in bioinformatics. If somebody have only centers for collecting information, who develop no local skills and no local students to use that information".

In plant biotechnology peoples generally have the more interest in plants secondary metabolism especially in legume plants the interest on studying secondary metabolism, symbiosis, and nitrogen fixation. According to the importance of the plant or cereal the data collection and database management might be differ, sometimes the cereals doesn't have the legumes features , that case both has to managed individually. Those are all functions that fit within community exploration of data and creation of data models and data mining mechanisms appropriate to those [19]. The concurrent development of molecular cloning techniques, DNA sequencing methods, rapid sequence comparison algorithms, and computer workstations has revolutionized the role of biological sequence comparison in molecular biology [20]. Today, the most powerful method for inferring the biological function of a gene is by sequence similarity searching on protein and DNA sequence databases [21,22]. Sequence alignment methodology used to compare two (pairwise alignment) or more sequences by searching for a series of individual sequences in the NCBI or PDB [23]. The most common comparative method in sequence alignment, which provides an explicit mapping between the residues of two or more sequences. In this activity, the similarities and differences at the level of individual bases or amino acids are analyzed, with the aim of inferring structural, functional and evolutionary relationships among the sequences under study (**Figure 1).** The schematic diagram explains about the database construction for proteins active site prediction.

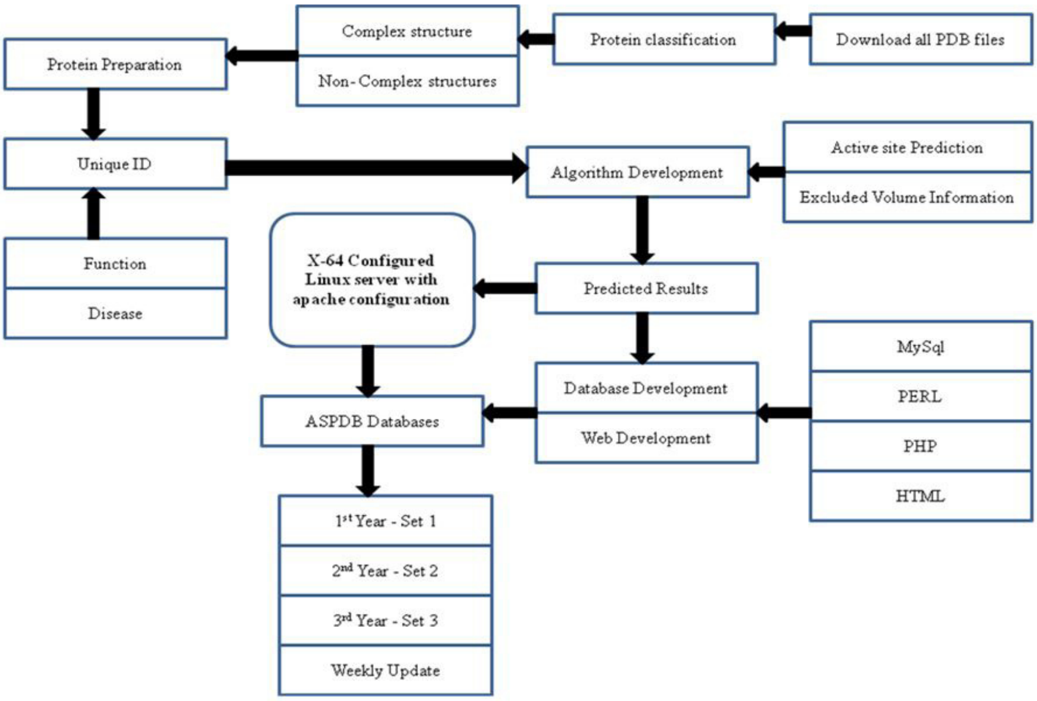


Figure 1: The schematic diagram explains about the database construction for proteins active site prediction.

An alignment between two sequences is simply a pair wise match between the characters of each sequence. Sequence similarity alignment of nucleotide or amino acid sequences provides the evolutionary connection between two or more homo logs. Homology refers to a conclusion drawn from these data that two genes share a common evolutionary history. Although it is presumed that homologues sequences have diverged from a common ancestral sequence through iterative molecular changes. The changes that occur during divergence from the common ancestor can be categorized as substitutions, insertions and deletions. Regions where the residues of one sequence correspond to nothing in the other would be interpreted as either an insertion into one sequence or a deletion from the other. These gaps are usually represented in the alignment as consecutive dashes aligned with letters.

The UniProt Consortium encompass the European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB), Protein Information Resource (PIR). EBI located at the Welcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy (Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the National Biomedical Research Foundation (NBRF) at the Georgetown University Medical Center in Washington, DC, USA, is heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure. In 2002, EBI, SIB, and PIR joined forces as the UniProt Consortium.

The roots of UniProt databases

Each consortium affiliate is a great deal with protein database maintenance and annotation. Until lately, EBI and SIB jointly fashioned Swiss-Prot and TrEMBL, while PIR shaped the Protein Sequence Database (PIR-PSD). These databases coexisted with conflicting protein sequence coverage and annotation priorities. Swiss-Prot is documented as the gold standard of protein annotation, with extensive cross-references, literature citations, and computational analyses provided by expert curators [24,25]. Recognizing that sequence data were being generated at a pace exceeding Swiss-Prot's capability to keep up, TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was fashioned to afford computerized interpretation for those proteins not in Swiss-Prot. In the meantime, PIR maintain the PIR-PSD and connected databases, includes iProClass, a database of protein sequences and curated families. The consortium members-all devoted to the similar objective of provided that expansive and meaningful protein annotation, and all with solid foundations stemming from decades of activity-decided to pool their overlapping (and, importantly, their complementary) resources, efforts, and expertise. The UniProt databases build upon these solid foundations.

Organization of UniProt databases

UniProt provides four core database:

- The UniProt Knowledgebase (UniProtKB) is a key database for protein sequences with accurate, consistent, rich sequence and functional annotation.

Similarly, UniProt Reference Clusters (UniRef) databases provide non-redundant reference data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions.

The UniProt Metagenomics and Environmental Sequences database (UniMES) repository particularly developed for metagenomic and environmental sequence data [26].

The UniProt Archive (UniParc) provides a stable, comprehensive sequence collection without redundant sequences by storing the complete body of publicly available protein sequence data [14].

RefSeq

The Reference Sequence (**RefSeq**) databases an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein translations. This database is associated with the NCBI and GenBank, give biological molecule nature from viruses to bacteria to eukaryotes [27,28]. *RefSeq* aim to give divide and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. *RefSeq* is inadequate to major organisms for which sufficient data is available [29].

GeneRIF

GeneRIFs provide a functional annotation of genes n the Entrez Gene database For example, GeneRIFs confers the role of a gene in a disease, structure of a gene and also gene function. GeneRIFs are always associated with specific entries in the Entrez Gene database. Each GeneRIF has a pointer to the PubMed ID (a type of document identifier) of a scientific publication that provides evidence for the statement made by the GeneRIF. GeneRIFs are frequently extracted directly from the document PubMed ID.

1. A published paper relating that function, executed through PubMed ID of a citation in PubMed;

2. A valid e-mail address (confidential).

Ensembl

Ensembl is a joint systematic scientific project between the European Bioinformatics Institute and the Welcome Trust Sanger Institute, which was launched in 1999 after completion of Human Genome Project [30]. Researchers could easily able to access the centralized resources of genetics, molecular biology, biochemistry, metabolic pathways and the whole genome function and structure of all species including vertebrates [31] NCBI and invertebrates are revealed through this kind of curated databases [32,33]. Retrieval of genomic information from Ensembl is very easy, accurate and convenient to update with time periods. Various databases are available to access the gnomic information, from this information we can able to annotate the gene, location, inter linkages and its relationships with other genes, human genome consists of 3 billion base pairs, which code for approximately 20,000-25,000 genes [34]. Such a kind of predicted and annotated data are very helpful to find the experimental evidences, publications references and paves the way to find the novel new drugs against the contagious diseases. However this is a slow, scrupulous task, so Ensembl used to do the complex pattern-matching of protein to DNA through supercomputers. Sequence data is fed into a software "pipeline" (written in Perl) which creates a set of predicted gene locations and saves them in a MySQL database for subsequent analysis and display. An important aspect of the Ensembl freely accessible to the world research community, available to download, and remote access. In addition, the Ensembl website provides computer-generated visual displays of much of the data [35].

Databases

Entrez searches the following databases:

PubMed: Biomedical literature citations and abstracts, including Medline - articles from (mainly medical) journals, often including abstracts. Links to PubMed Central and other full-text resources are provided to articles from the 1990s.

- PubMed Central: Free, full text journal articles
- Site Search: NCBI web and FTP web sites
- Books: Online books
- OMIM: Online Mendelian Inheritance in Man
- OMIA: Online Mendelian Inheritance in Animals
- Nucleotide: Sequence database (GenBank-Pennisi, 1599)
- Protein: Sequence database
- Genome: Whole genome sequences and Mapping [36]
- Structure: Three-dimensional macromolecular structures
- Taxonomy: Organisms in GenBank Taxonomy
- SNP: Single Nucleotide Polymorphism
- Gene: Gene-centered information
- HomoloGene: Eukaryotic homology groups
- PubChem Compound: Unique small molecule chemical structures
- PubChem Substance: Deposited chemical substance records
- Genome Project: Genome project information
- UniGene: Gene-oriented clusters of transcript sequences
- CDD: Conserved protein Domain Database
- 3D Domains: Domains from Entrez Structure
- UniSTS: Markers and mapping data
- PopSet: Population study data sets (epidemiology)
- GEO Profiles: Expression and molecular abundance profiles [37]
- GEO DataSets: Experimental sets of GEO data [38]
- Cancer Chromosomes: Cytogenetic databases
- PubChem BioAssay: Bioactivity screens of chemical substances
- GENSAT: Gene expression atlas of mouse central nervous system
- Probe: Sequence-specific reagents
- NLM Catalog: NLM bibliographic data for over 1.2 million journals, books, audiovisuals, computer software, electronic resources, and other materials resident in LocatorPlus (updated every weekday).

References

1. Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, et al. Gramene: A growing plant comparative genomics resource. Nucleic Acids Research. 2007; 36: D947-953.

2. Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ, et al. The Rat Genome Database, update 2007-easing the path from disease to data and back again. Nucleic acids research. 2007; 35: D658-662.

3. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, et al. The Consensus Coding Sequence (CCDS) project: Identifying a com-

mon protein-coding gene set for the human and mouse genomes. Genome research. 2009; 19: 1316-1323.

4. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, Journal of molecular biology. 1977; 94: 441-448.

5. Watson JD, Crick FH. THE CLASSIC: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid, Clinical Orthopaedics and Related Research®. 2007; 462: 3-5.

6. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, et al. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic acids research. 2007; 36: D1009-1014.

7. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science. 2008; 322: 44-48.

8. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. 2003; 422: 835-847.

9. Siva N. 1000 Genomes Project, Nature Biotechnology. 2008; 26: 256.

10. Waldrop M. Wikiomics, Nature. 2008; 455: 22.

11. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, et al. The Integrated Microbial Genomes (IMG) system. Nucleic acids research. 2006; 34: D344-348.

12. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. Database resources of the national center for biotechnology information, Nucleic acids research. 2007; 36: D13-21.

13. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. The complete genome of an individual by massively parallel DNA sequencing, Nature. 2008; 452: 872-876.

14. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. The UCSC Table Browser data retrieval tool. Nucleic acids research. 2004; 32: D493-496.

15. Brinley E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE Pilot Project (14 June 2007), Nature. 2007; 447: 799-816.

16. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA. Mouse Genome Database Group. The Mouse Genome Database (MGD): Mouse biology and model systems. Nucleic acids research. 2008; 36: D724-728.

17. Rogers A, Antoshechkin I, Bieri T, Blasiar D, Bastiani C, et al. Worm Base 2007, Nucleic acids research. 2007; 36: D612-617.

18. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, et al. The Zebra fish Information Network: The zebra fish model organism database, Nucleic acids research. 2006; 34: D581-585.

19. Fernández-Suárez XM, Birney E. Advanced genomic data mining, PLoS computational biology. 2008; 4.

20. Maxam AM, Gilbert W. A new method for sequencing DNA, Proceedings of the National Academy of Sciences. 1977; 74: 560-564.

21. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. The Pfam protein families database, Nucleic acids research. 2007; 26: D281-288.

22. Cooper E, Patterson I. The legacy of GenBank: The DNA sequence database that set a precedent. 1663: The Los Alamos Science and Technology Magazine.2008.

23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein data bank, Nucleic Acids Research. 2000; 28: 235-242.

24. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, et al. The future of biocuration. Nature. 2008; 455: 47-50.

25. International HapMap Consortium. The international HapMap project, Nature. 2003; 426: 789.

26. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. IMG/M: A data management and analysis system for metagenomes. Nucleic acids research. 2007; 36: D534-538.

27. Bilofsky HS, Christian B. The GenBank® genetic sequence data bank. Nucleic acids research. 1988; 16: 1861-1863.

28. Salzberg SZ. Genome re-annotation: A wiki solution?, Genome biology. 2007; 8: 102.

29. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research. 2005; 33: D501-514.

30. Flicek P, Aken BL, Beal K, Ballester B, Cáccamo M, et al. Ensembl 2008, Nucleic acids research. 2007; 36: D707-714.

31. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, et al. The Vertebrate Genome Annotation (Vega) database, Nucleic acids research. 2007; 36: D753-760.

32. Galperin MY. The molecular biology database collection: 2008 update, Nucleic acids research. 2008; 36: D2-4.

33. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology, Nature genetics. 2008; 40: 124-125.

34. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, et al. Gene Ontology annotations at SGD: New data sources and annotation methods. Nucleic acids research. 2007; 36: D577-581.

35. Wilson RJ, Goodman JL, Strelets VB. FlyBase Consortium. FlyBase: Integration and improvements to query tools, Nucleic acids research. 2008; 36: D588-593.

36. Couzin J. Whole-genome data not anonymous, challenging assumptions. 2008: 1278.

37. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. ArrayExpress-a public database of microarray experiments and gene expression profiles. Nucleic acids research. 2007; 35: D747-750.

38. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. NCBI GEO: Mining tens of millions of expression profiles-database and tools update. Nucleic acids research. 2007; 35: D760-765.