Gulshan Wadhwa · P. Shanmughavel
Atul Kumar Singh · Jayesh R. Bellare
Editors

Current trends in Bioinformatics: An Insight

Springer

Current trends in Bioinformatics: An Insight

Gulshan Wadhwa • P. Shanmughavel
Atul Kumar Singh • Jayesh R. Bellare
Editors

Current trends in Bioinformatics: An Insight

Springer

Editors
Gulshan Wadhwa
Department of Biotechnology
Apex Bioinformatics Centre
Ministry of Science & Technology
New Delhi, India

Atul Kumar Singh
Central Research Facility
Indian Institute of Technology Delhi
New Delhi, India

Centre for Research in Nanotechnology and Sciences
Indian Institute of Technology Bombay
Mumbai, India

P. Shanmughavel
Department of Bioinformatics
Bharathiar University
Coimbatore, Tamil Nadu, India

Jayesh R. Bellare
Department of Chemical Engineering
Indian Institute of Technology Bombay
Mumbai, India

ISBN 978-981-10-7481-3 ISBN 978-981-10-7483-7 (eBook)
https://doi.org/10.1007/978-981-10-7483-7

Library of Congress Control Number: 2018943304

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

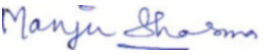
Foreword

It gives us immense pleasure to present this edited book to the biotechnological research communities. Bioinformatics and computational biology – the science of using biological data to develop algorithms and relations among various biological systems – are the cutting edge areas for research. Computational sciences have their roots in the development of increasingly powerful computers over the last few decades. Rather rapidly, the instrumentation and the newly developed methodology with the underlying algorithms became widely appreciated and used as novel research strategies serving in many different fields of academic investigation, particularly in natural, engineering, social sciences, and humanities. Computational sciences have been recognized for their invaluable contributions to data collection, storage, handling, and analysis, thus leading to efficient strategies of modeling, prediction, and design of molecular structures and of their functional properties that are often of immediate relevance for the medical sciences. Computational comparisons of DNA sequences from different organisms provide invaluable insights into past evolutionary developments, and this has become a powerful new tool in the systematics of living organisms.

The growth in high-throughput full genomic sequencing, structural genomics, proteomics, epigenetics, etc., would be rather limited without bioinformatics. In order to give concise information on basic concept and advances in bioinformatics, authors have thought of bringing out an edited volume "Current Trends in Bioinformatics" for the benefit of the students and researchers working in the field of life science, medicine, and pharmaceutical science. It also focuses on reviews on advances in computational molecular/structural biology, encompassing areas such as computing in biomedicine and genomics, computational proteomics and systems biology, and metabolic pathway engineering. Developments in these fields have direct impact on key issues related to health care, medicine, genetic disorders, development of agricultural products, renewable energy, environmental protection, etc. The book has 18 chapters, divided into two sections.

The overview of important aspects of bioinformatics would further contribute to strengthen international contacts and serve as a testament to such a fruitful development for the basic as well as applied sciences. The Department of Biotechnology considering the great significance of this field established a countrywide network of bioinformatics centers in academic institutions. These have paid rich dividends.

We hope that scientific community especially students, in particular, would enjoy reading, learn and make best use of this book.



(Dr. Manju Sharma)

Former Secretary, Department of Biotechnology Manju Sharma
New Delhi, India

Principal Advisor to the Deptartment of Science and Technology
Gandhinagar, Gujarat, India

Distinguished Women Scientist Chair, NASI
Allahabad, India

Preface

Bioinformatics has become a frontline applied science and is of vital importance to study new biology, which is widely recognized as the new scientific endeavor of the twenty-first century. The growth in full genomic sequencing, structural genomics, proteomics, and microarray will be very slow without application of bioinformatics. In fact the very high importance of bioinformatics comes from its usefulness in these areas to solve complex biological problems. So up-to-date information in the field of bioinformatics is the most needed one. The proposed book *Current Trends in Bioinformatics* fulfills these requirements.

Current Trends in Bioinformatics aims to publish all the latest and outstanding developments in bioinformatics. The book contains a series of timely, in-depth reviews, drug clinical trial studies, and biodiversity informatics and thematic issues written by leaders in the field, covering a wide range of the integration of biology with computer and information science.

It also focuses on reviews on advances in computational molecular/structural biology, encompassing areas such as computing in biomedicine and genomics, computational proteomics and systems biology, and metabolic pathway engineering. Developments in these fields have direct implications on key issues related to health care, medicine, genetic disorders, development of agricultural products, renewable energy, and environmental protection.

This book is an ideal foundation for teaching at the undergraduate and graduate levels. It is also highly suited for self-instruction by research investigators interested in applying bioinformatics methods of analysis and information technologists associated with academic and industrial laboratories.

It is supposed that the nonspecialists would be the principal readers of the book. So, before embarking on the bioinformatics, some fundamental aspects of molecular evolution, taxa-related studies, some core concepts of genomics and some of the important genomic techniques were discussed in this book, to make the readers conceptualize the bioinformatics analysis.

The author would also like to thank colleagues for their encouragement, enthusiasm, and support for the success of this project.

Last but not the least, the author is grateful to the Staff of Springer for making this project a reality, helping to bring it to successful completion, and always being available whenever help and advice were needed.

New Delhi, India — Gulshan Wadhwa
Mumbai, India — Jayesh R. Bellare
Coimbatore, India — P. Shanmughavel
New Delhi, India — Atul Kumar Singh
Mumbai, India

Contents

Contributors

C. Awasthi Department of Biotechnology, Gobind Ballabh Pant Engineering College, Pauri Garhwal, Uttarakhand, India

Jayesh R. Bellare Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai, India

Vaibhav D. Bhatt Department of Pharmaceutical Sciences, Saurashtra University, Rajkot, Gujarat, India

Chandan Chakraborty School of Medical Science & Technology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

Dev Kumar Das School of Medical Science & Technology, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

Abhijit Datta Department of Botany, Jhargram Raj College, Medinipur, West Bengal, India

Rajat K. De Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

Rohit Farmer School of Biosciences, University of Birmingham, Birmingham, UK

Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

J. Fermin Angelo Selvin Department of Physics, Nadar Mahajana Sangam S. Vellaichamy Nadar College, Madurai, Tamil Nadu, India

Sayak Ganguli Theoretical and Computational Biology Division, Amplicon Institute of Interdisciplinary Science and Technology, Palta, West Bengal, India

Budhayash Gautam Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Kavita Goswami Plant RNAi Biology Group, International Center for Genetic Engineering and Biotechnology, New Delhi, India

T. Hema Thanka Christlet Department of Physics, Dr. Ambedkar Government Arts College, Chennai, Tamil Nadu, India

Chakresh Kumar Jain Department of Biotechnology, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

Yogesh K. Jaiswal School of Studies in Biochemistry, Jiwaji University, Gwalior, India

Qazi Mohammad Sajid Jamal Department of Health Information Management, College of Applied Medical Sciences, East Qassim University, Al Qassim-Buraydah, Kingdom of Saudi Arabia

A. Jasmine Department of Physics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

Chaitanya G. Joshi Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand, Gujarat, India

Sukhdeep Kaur Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Kavindra Kumar Kesari Department of Applied Physics, and Department of Bioproduct & Biosystem, Aalto University, Espoo, Finland

K. Konda Reddy Department of Pharmacy, National University of Singapore, Singapore, Singapore

Ajit Kumar Centre for Bioinformatics, Maharshi Dayanand University, Rohtak, India

Sunil Kumar Bioinformatics Centre, Institute of Life Sciences, Bhubaneswar, Odisha, India

ICAR-NBAIM, Mau, Uttar Pradesh, India

Sharad Singh Lodhi School of Studies in Biochemistry, Jiwaji University, Gwalior, India

Ashok K. Maiti Medipath Clinic (P) Ltd, West Medinipur, West Bengal, India

Jayakanthan Mannu Department of Plant Molecular Biology and Bioinformatics, Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

Soma S. Marla Indian Council of Agricultural Research, National Bureau of Plant Genetic Resources, New Delhi, India

Premendu P. Mathur Department of Biochemistry and Molecular Biology, School of Life Sciences, Pondicherry University, Pondicherry, India

Ranjeet Maurya Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Neelofar Mirza Indian Council of Agricultural Research, National Bureau of Plant Genetic Resources, New Delhi, India

Neeti Sanan Mishra Plant RNAi Biology Group, International Center for Genetic Engineering and Biotechnology, New Delhi, India

Pallavi Mishra Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Rashmi Mukherjee RNLKWC, Vidyasagar University, Midnapur, West Bengal, India

K. D. Nadella Directorate of Knowledge Management Units (DKMU), ICAR, New Delhi, India

Genetics division, ICAR, IARI, New Delhi, India

H. P. Pandey Department of Biochemistry, Nepalgunj Medical College, Chisapani Campus, Kathmandu University, Nepalgunj, Nepal

Monika Patel Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand, Gujarat, India

Anjali Rao Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

P. Shanmughavel Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu, India

Mohd. Haris Siddiqui Department of Bioengineering, Faculty of Engineering, Integral University, Lucknow, Uttar Pradesh, India

Atul Kumar Singh Central Research Facility, Indian Institute of Technology Delhi, New Delhi, India

Centre for Research in Nanotechnology and Sciences, Indian Institute of Technology Bombay, Mumbai, India

Satendra Singh Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Manish Sinha Laureate Institute of Pharmacy, Kangra, Himachal Pradesh, India

Gitanjali Tandon Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Christopher M. Thomas School of Biosciences, University of Birmingham, Birmingham, UK

Namrata Tomar Department of BioMedical Engineering, Medical College of Wisconsin, Milwaukee, WI, USA

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

Mobashar Hussain Urf Turabe Fazil Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore

K. Veluraja Department of Physics, School of Advanced Sciences, VIT University, Vellore, Tamil Nadu, India

Jyoti Verma Department of Biotechnology, Gobind Ballabh Pant Engineering College, Pauri Garhwal, Uttarakhand, India

Gulshan Wadhwa Department of Biotechnology, Apex Bioinformatics Centre, Ministry of Science & Technology, New Delhi, India

Peter J. Winn School of Biosciences, University of Birmingham, Birmingham, UK

Arvind Kumar Yadav Department of Computational Biology and Bioinformatics, Jacob Institute of Biotechnology and Bioengineering, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

About the Editors

Gulshan Wadhwa is currently the Joint Director in the Department of Biotechnology, Ministry of Science & Technology, Government of India. He has developed the Bioinformatics program in India (BTISnet) and established teaching and training programs and a super computing facility in Bioinformatics in India. He has undergone trainings from WIPO, Geneva, and NUS, Singapore.

Jayesh R. Bellare is currently Institute Chair Professor and Professor with Chemical Engineering Department & Center for Research in NanoTechnology and Science, Indian Institute of Technology-Bombay, Mumbai, India. He was a Post Doctoral Fellow, from M.I.T., Cambridge, USA, and has over 33 years of experience.

P. Shanmughavel is currently working as Associate Professor at Bharathiar University, Coimbatore, India. He has published 7 books in bioinformatics (two books published in Germany), 6 book chapters, 5 conference proceedings, and 31 research papers in reputed national and international journals in Bioinformatics. He has established the Centre of Bioinformatics-"DBT-BIF" in 2006.

Atul Kumar Singh is currently working as Senior Scientist at Indian Institute of Technology-Delhi. He has completed his Doctorate from Indian Institute of Technology-Bombay, Mumbai, in Nanotechnology. He has published 11 research papers in reputed international journals and has 6 patents.

Part I

Overview

An Insight of Biological Databases Used in Bioinformatics

1

Vaibhav D. Bhatt, Monika Patel, and Chaitanya G. Joshi

Abstract

Collections of life sciences information from scientific investigations, high-throughput experiment technology, available literature, and computational analysis are called biological databases. It contains information from research areas comprising genomics, microarray gene expression, proteomics, phylogenetics, metabolomics, gene function, structure, localization and similarities of biological sequences. In a nutshell, databases are libraries for storage and representation of biological data obtained from the scientific community which converts data into knowledge. Utmost biological databases are available from websites that categorize data which operators can browse through the data online. Due to the vast amount of data generated by high-throughput DNA sequencers in the investigation of genome, transcriptome, and exome sequences of various organisms in current times, the biological data has stored with an exponential rate. The availability of enormous amount of biological data (sequences as well as structural) has generated a need for managing, storing, and retrieving this huge data. This chapter reviews current knowledge of the different types of databases available with examples of their file formats.

Keywords

Biological sequences · High-throughput DNA sequencers · Transcriptome and exome sequences

V. D. Bhatt (✉)
Department of Pharmaceutical Sciences, Saurashtra University, Rajkot, Gujarat, India

M. Patel · C. G. Joshi
Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand, Gujarat, India

G. Wadhwa et al. (eds.), *Current trends in Bioinformatics: An Insight*,
https://doi.org/10.1007/978-981-10-7483-7_1

1.1 Introduction

Databases are the convenient system to properly store, search, and recover several types of data. A database helps to easily handle and share large amount of data and supports large-scale analysis by easy access and data update (Liu and Özsu 2009).

Due to the vast amount of data generated in experiments of genome, transcriptome, and exome sequences of various organisms in current times, the biological data has stored with an exponential rate. The availability of enormous amount of biological data (sequences as well as structural data) has generated a need for managing, storing, and retrieving this huge data.

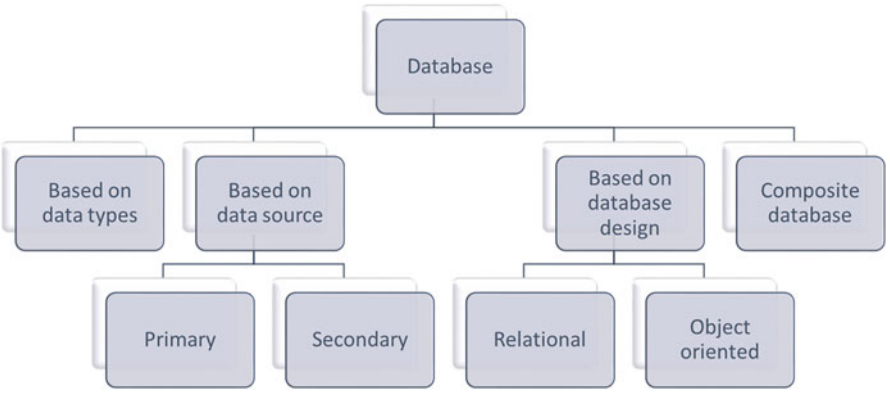
Therefore the biological databases have come into existence as invaluable sources for the biological community. In a nutshell, databases are libraries for storage and representation of biological data obtained from the scientific community which converts data into knowledge.

1.2 History

A book published in 1965, *Atlas of Protein Sequences and Structures*, *was the first biological database* by Margaret Dayhoff and colleagues, and further they have published other editions of the book in the 1970s; however the first edition was limited to 65 sequences only (Dayhoff and Foundation 1973, 1976; Foundation 1972).

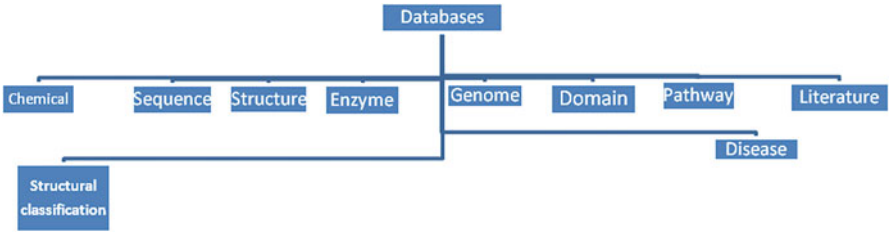
With the discovery of the integrated circuit, the powerful and reliable third generation computers are became the choice of storage of biological databases for scientists. An English scientist Tim Berners-Lee in 1989 invented the "World Wide Web" (WWW) which is the primary tool people use to interact on the Internet and is the way to access all biological databases. Production of high throughput sequencing machines leads production of data rich science, needs an interdisciplinary arena to develop software tools which is used to understand biological data. The field of science with the involvement of computer, statistics and engineering to study biological data is called Bioinformatics.

1.3 Classification of Biological Databases



1.3.1 Databases Based on Data Types

This database was divided into several databases; some of the databases were discussed below in detail.



1.3.1.1 Sequence Databases

Sequence databases contain both nucleic acid and protein sequences. First we will discuss about nucleotide sequence repositories.

(I) Nucleic Acid Sequence Database

There are three main nucleotide sequence repositories:

(A) GenBank
(B) European Molecular Biology Laboratory (EMBL)
(C) DNA Data Bank of Japan (DDBJ)

Raw nucleic acid sequences are stored in these databases and make available through Internet sources. Initially, these databases worked independently, but later the *International Nucleotide Sequence Database Collaboration* (INSDC, http://insdc.org) was developed to maintain collaboration between DDBJ, GenBank, and EMBL (Fig. 1.1). These databases started exchanging their data through constant communication between the team at each collaborating organizations in order to access the sequences present in all three different formats.

(A) *GenBank*

GenBank is a collection of raw and annotated nucleotide as well as protein information. GenBank is maintained and accessed through the National Center for Biotechnology Information (NCBI). Every 2 months a new release is made. It is maintained by NCBI as part of the INSDC (Benton 1990). There are approximately 137384889783 bases, from 149819246 sequence records in the GenBank release 188.0 on February 15, 2012. Type "insulin" in the search tab on the GenBank home page to view list of sequences of insulin gene, partial or complete from different organisms (Fig. 1.2).

Example of GenBank Format

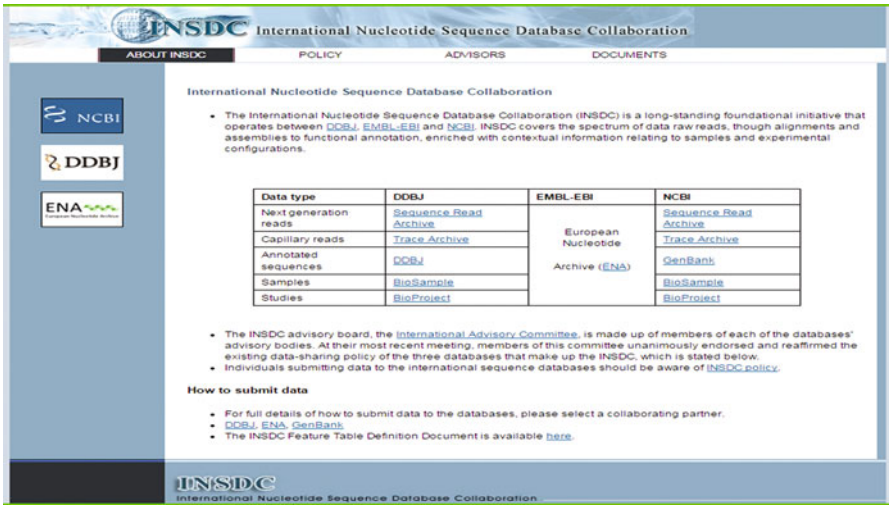


Fig. 1.1 The home page of International Nucleotide Sequence Database Collaboration (INSDC) (http://insdc.org)

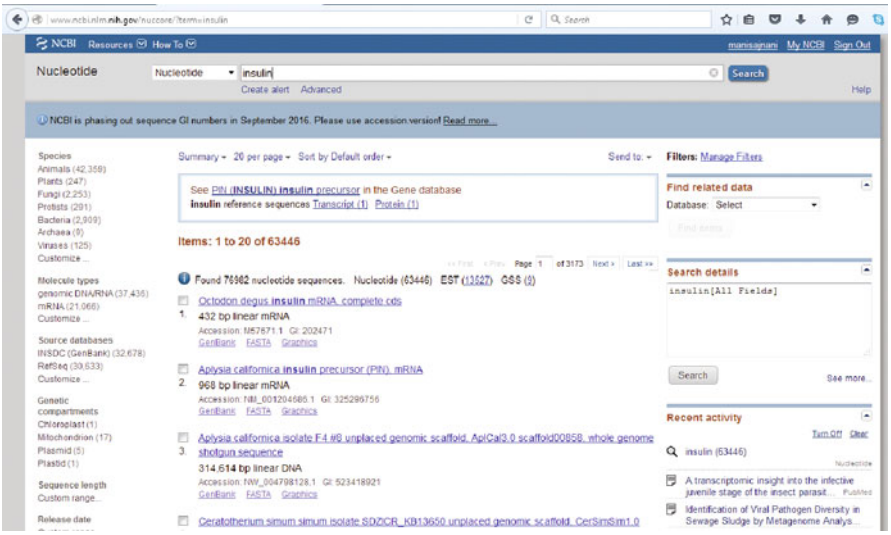


Fig. 1.2 Using GenBank to query insulin sequences (http://www.ncbi.nlm.nih.gov/nuccore/?term=insulin)

Octodon degus insulin mRNA, complete cds

GenBank: M57671.1

FASTA Graphics

Go to:

```
LOCUS       OCOINS                  432 bp    mRNA    linear   ROD 27-APR-1993
DEFINITION  Octodon degus insulin mRNA, complete cds.
ACCESSION   M57671
VERSION     M57671.1
KEYWORDS    insulin; insulin alpha-chain; insulin beta-chain; insulin
            connecting peptide.
SOURCE      Octodon degus (degu)
  ORGANISM  Octodon degus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Hystricognathi; Octodontidae; Octodon.
REFERENCE   1  (bases 1 to 432)
  AUTHORS   Nishi,M. and Steiner,D.F.
  TITLE     Cloning of complementary DNAs encoding islet amyloid polypeptide,
            insulin, and glucagon precursors from a New World rodent, the degu,
            Octodon degus
  JOURNAL   Mol. Endocrinol. 4 (8), 1192-1198 (1990)
   PUBMED   2293024
COMMENT     Original source text: Octodon degus pancreas, cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..432
                     /organism="Octodon degus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10160"
                     /tissue_type="pancreas"
     gene            1..432
                     /gene="insulin"
     CDS             42..371
                     /gene="insulin"
                     /codon_start=1
                     /product="insulin"
                     /protein_id="AAA40590.1"
                     /translation="MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCG
                     RSGFYRPHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFNQL
                     QNYCNVP"
     sig_peptide     42..113
                     /gene="insulin"
     mat_peptide     114..200
                     /gene="insulin"
                     /product="insulin B-chain"
     mat_peptide     207..293
                     /gene="insulin"
                     /product="insulin C-peptide"
     mat_peptide     300..368
                     /gene="insulin"
                     /product="insulin A-chain"
     regulatory      414..419
                     /regulatory_class="polyA_signal_sequence"
                     /gene="insulin"
     polyA_site      432
                     /gene="insulin"
ORIGIN
        1 gcattctgag gcattctcta acaggttctc gaccctccgc catggccccg tggatgcatc
       61 tcctcaccgt gctggccctg ctggccctct ggggacccaa ctctgttcag gcctattcca
      121 gccagcacct gtgcggctcc aacctagtgg aggcactgta catgacatgt ggacggagtg
      181 gcttctatag accccacgac cgccgagagc tggaggacct ccaggtggag caggcagaac
      241 tgggtctgga ggcaggcggc ctgcagcctt cggccctgga gatgattctg cagaagcgcg
      301 gcattgtgga tcagtgctgt aataacattt gcacatttaa ccagctgcag aactactgca
      361 atgtccctta gacacctgcc ttgggcctgg cctgctgctc tgccctggca accaataaac
      421 cccttgaatg ag
//
```

Format Explanation

GenBank format includes *locus name* which is similar to the accession number and unique to the entry, and it is followed by sequence length. In our example sequence length is 587 bp. Definition includes description of source organism, gene/protein name, and other details about sequence.

- *Accession number* is the unique identifier of the sequence (NM_013564).
- *Version* is similar to accession number, but whenever a change occurs in sequence data, the version increases by 1. In our example, version is NM_013564.7; this indicates that sequence has been changed seven times.
- *GI (GenInfo Identifier)* number also runs parallel to the accession number and version system. A new GI is allotted, if the sequence has been changed and the version has increased by unity. In our example, GI is 365192585.
- *Keywords* are words or expressions about sequence. The keyword field contains a dot if nothing is provided.
- *Source* contains name of the organism from which the sequence has been derived.
- *Organism* is a related sub-keyword of source and contains the scientific name of the organism along with the lineage as described in NCBI taxonomy database.
- *Reference* contains the publication by the authors of the sequence.
- *Authors* contain list of authors in the same order as appears in publication.
- *Title* shows the title of published/unpublished work.
- *Journal* contains MEDLINE abbreviations of the journal name where the work is published.
- *PubMed* field provides the PubMed identifier (PMID) of that article.
- *Comment* points out the change occurred in the submitted sequence.
- *Features* provide information about genes and their products, segment of biological significance in the submitted sequence, as well as other characteristics.
- *Gene* provides gene length and gene name and its function and synonyms. CDS represents coding sequence which codes for protein sequence.
- *Origin* contains the sequence data. Finally, GenBank record ends with // sign.

Sequence Submission to GenBank

Sequence submission is done by using different tools available at NCBI. Few of them are:

BankIt: direct submissions are made to GenBank using it (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank).
Sequin: it is a stand-alone submission platform (www.ncbi.nlm.nih.gov/Sequin/).
tbl2asn: it is a command-line program, used for submission of large batches of sequences and complete genomes (www.ncbi.nlm.nih.gov/genbank/tbl2asn2).

Table 1.1 Various databases and software tools of NCBI for sequence analysis

NCBI			
Tools			Databases
Sequence Submission	Sequence	Data mining	Literature
	Analysis		Nucleotide
Sequin	BLAST	Entrez	Protein
BankIt	Blink	My NCBI	Structure
tbl2asn	Stand-alone BLAST	LinkOut	Genome
			OMIM
			SNP
Barcode Submission Tool	e-PCR	Citation	Books
		Matcher	Domain
	ORF Finder		Chemical
			Expression
	Map viewer		Other databases
	Tax plot		
	Trace archive		

Barcode Submission Tool: it is a WWW-based tool for the submission of sequences and trace read data (http://www.ncbi.nlm.nih.gov/WebSub/?tool=barcode).

National Center for Biotechnology Information (NCBI)

NCBI was started in 1988, as a part of the US National Library of Medicine (NLM) located at Bethesda, Maryland. It is a division of the National Institutes of Health and is directed by David Lipman. The responsibility of NCBI is to make available the GenBank nucleotide sequence database since 1992. NCBI is playing a very remarkable role for biological scientists by making available various public databases and software tools for sequence analysis (Table 1.1). GenBank manages with individual laboratories and other sequence databases like those of the EMBL and the DDBJ. Meanwhile in 1992, NCBI has developed to run other databases in addition to GenBank ((US) 2013). The home page of NCBI is shown in Fig. 1.3.

Databases and Tools of NCBI

Database Retrieval Tool

Entrez (www.ncbi.nlm.nih.gov/Entrez/) in Fig. 1.4 is a primary text search engine which comprises of 40 molecular and literature databases. It extracts huge information from the PubMed database, such as DNA and protein sequences and structure, gene, genome, genetic variation, and gene expression.

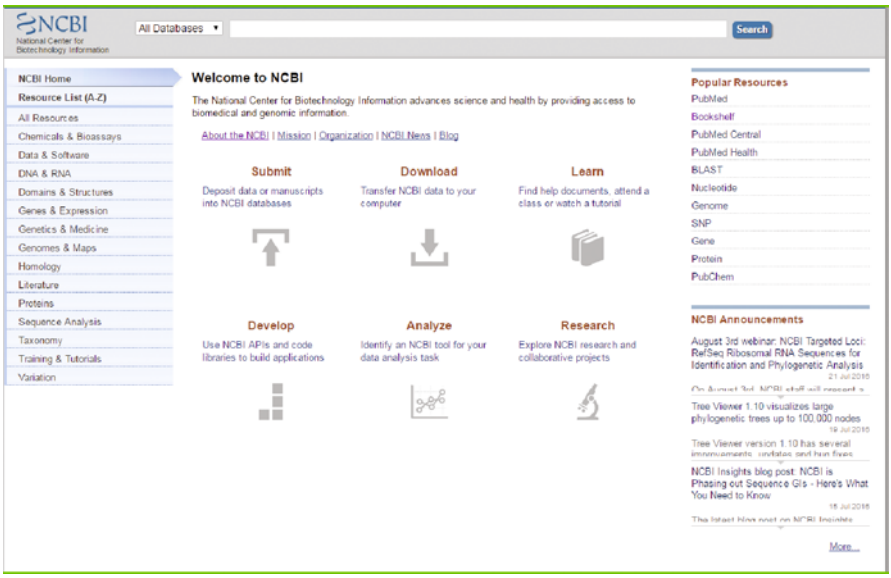


Fig. 1.3 The home page of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)

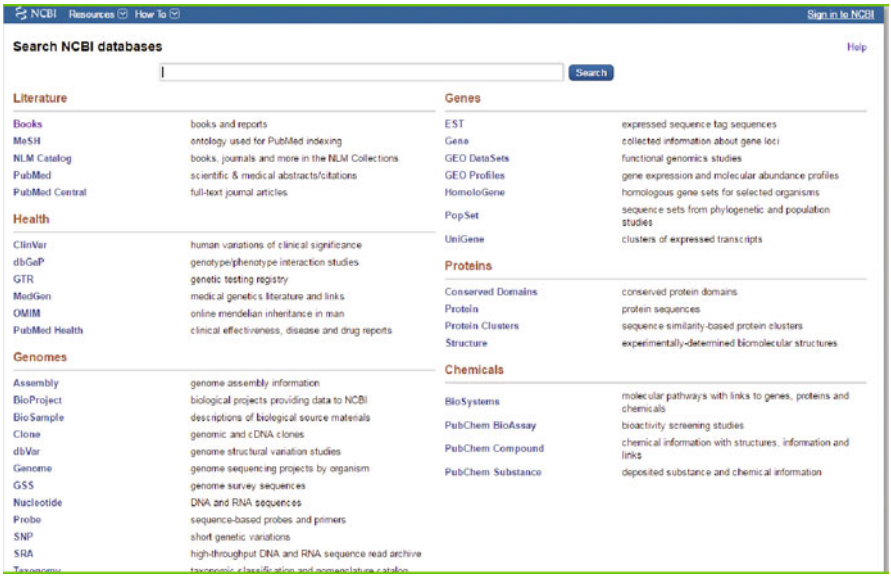


Fig. 1.4 The home page of Entrez (www.ncbi.nlm.nih.gov/Entrez/)

Fig. 1.5 The home page of European molecular biology laboratory (http://www.embl.org/)

(B) *European Molecular Biology Laboratory (EMBL)*

The *European Molecular Biology Laboratory* (*EMBL*) (http://www.embl.org/) in Fig. 1.5 is a molecular biology organization which is maintained by 20 European countries, with Australia as associate member state. It is an intergovernmental organization created in 1974. It develops and maintains a large number of databases, and scientists can access the data free of cost. This research laboratory functions from five different locations, the main laboratory, the European Bioinformatics Institute (EBI), Heidelberg, Germany, is a hub for bioinformatics research and services, directed by Dr. Rolf Apweiler and Dr. Ewan Birney. It is a part of INSDC, which includes DDBJ and GenBank. Typing insulin gene at EMBL search engine produced a result in Fig. 1.6.

EMBL File Format

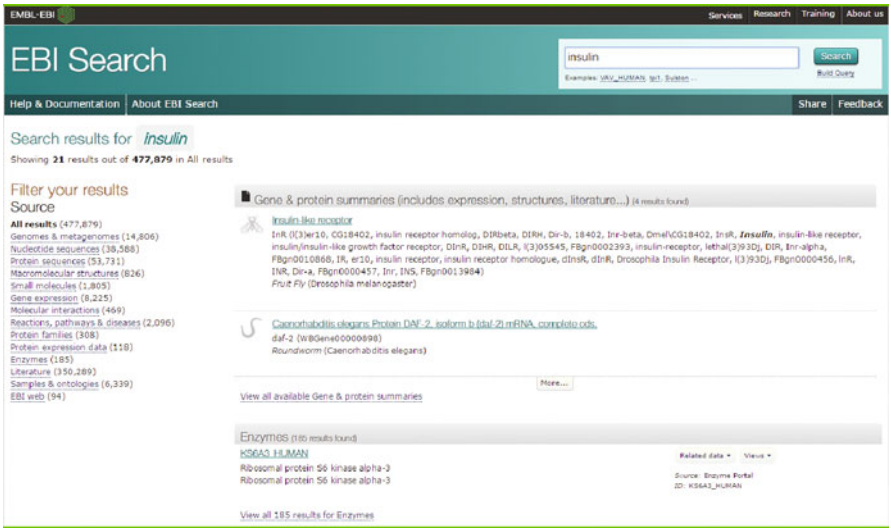


Fig. 1.6 Insulin gene search at European molecular biology laboratory website (https://www.ebi.ac.uk/ebisearch/search.ebi?query=insulin&db=allebi&requestFrom=searchBox)

```
ID   AH002190; SV 2; linear; genomic DNA; STD; ROD; 782 BP.
XX
AC   AH002190; M25583; M25585;
XX
DT   13-JUN-2016 (Rel. 129, Created)
DT   13-JUN-2016 (Rel. 129, Last updated, Version 1)
XX
DE   Rattus norvegicus insulin 2 (INS2) gene, complete cds.
XX
KW   insulin.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Rattus.
XX
RN   [1]
RP   1-782
RX   DOI; 10.1111/j.1749-6632.1980.tb47271.x.
RX   PUBMED; 6249167.
RA   Lomedico P.T., Rosenthal N., Kolodner R., Efstratiadis A., Gilbert W.;
RT   "The structure of rat preproinsulin genes";
RL   Ann. N. Y. Acad. Sci. 343:425-432(1980).
XX
DR   MD5; 2b03b65970e00e00d50a5054fad8125c.
XX
CC   On or before Jun 10, 2016 this sequence version replaced gi:204949,
CC   gi:204950, gi:204951.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..782
FT                   /organism="Rattus norvegicus"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:10116"
FT   gene            1..739
FT                   /gene="INS2"
FT   exon            <1..46
FT                   /gene="INS2"
FT                   /number=1
FT   intron          47..165
FT                   /gene="INS2"
FT                   /number=1
FT   CDS             join(180..366,541..686)
FT                   /codon_start=1
FT                   /gene="INS2"
FT                   /product="insulin 2"
FT                   /note="precursor"
FT                   /protein_id="AAA41440.1"
FT                   /translation="MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGE
FT                   RGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQ
FT                   LENYCN"
FT   sig_peptide     180..251
FT                   /gene="INS2"
FT   exon            180..366
FT                   /gene="INS2"
FT                   /number=2
FT                   /note="first expressed exon"
FT   mat_peptide     252..341
FT                   /gene="INS2"
FT                   /product="beta chain"
FT   mat_peptide     join(348..366,541..614)
FT                   /gene="INS2"
FT                   /product="insulin 2 connecting peptide"
FT   intron          367..>410
FT                   /gene="INS2"
FT                   /number=2
FT   gap             411..510
FT                   /estimated_length=unknown
```

```
FT   intron          <511..540
FT                   /gene="INS2"
FT                   /number=2
FT   exon            541..739
FT                   /gene="INS2"
FT                   /number=3
FT   exon            541..>686
FT                   /gene="INS2"
FT                   /number=3
FT                   /note="preproinsulin 2"
FT   mat_peptide     621..683
FT                   /gene="INS2"
FT                   /product="insulin 2"
FT                   /note="alpha chain"
XX
SQ   Sequence 782 BP; 136 A; 212 C; 173 G; 161 T; 100 other;
     cccagcccta agtgaccagc tacagtcgga aaccatcagc aagcaggtat gtactctcca        60
     aggtgggcct agcttcccca gtcaagactc caaggatttg agggacgctg tgggctcttc       120
     tcttacatgt acctttttgct agcctcaacc ctgactatct tccaggtcat tgttccaaca       180
     tggccctgtg gatccgcttc ctgcccctgc tggccctgct catcctctgg gagccccgcc       240
     ctgcccaggc ttttgtcaaa cagcaccttt gtggttctca cttggtggaa gctctctacc       300
     tggtgtgtgg ggagcgtgga ttcttctaca cacccatgtc ccgccgcgaa gtggaggacc       360
     cacaaggtaa gctctgctcc tgaattctat cccaagtgct aactaccctg nnnnnnnnnn       420
     nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn       480
     nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn tgccctgtgc tgacatgacc tccctggcag       540
     tggcacaact ggagctgggt ggaggcccgg gggccggtga ccttcagacc ttggcactgg       600
     aggtggcccg gcagaagcgc ggcatcgtgg atcagtgctg caccagcatc tgctctctct       660
     accaactgga gaactactgc aactaggccc accactaccc tgtccacccc tctgcaatga       720
     ataaaacctt tgaaagagca ctacaagttg tgtgtacatg cgtgcatgtg catatgtggt       780
     gc                                                                     782
//
```

Sequence Retrieval System (SRS)

SRS (http://srs.ebi.ac.uk/) (Fig. 1.7) is a powerful searching tool to retrieve sequences (and other types of data) and also to perform various operations on retrieved information for EMBL. It is similar to Entrez of NCBI, a search engine for extracting all sort of information available at EMBL.

Sequence Submission at EMBL

There are mainly three tools available for submitting data at EMBL.

1. Webin: for nucleotide sequence submission
2. Sequin: a stand-alone tool for submitting nucleotide sequences to GenBank, EMBL, and DDBJ developed by NCBI
3. Webin-Align: a tool for sequence alignment submission

(C) *DNA Data Bank of Japan (DDBJ)*

DDBJ, (http://ddbj.sakura.ne.jp/) (Fig. 1.8) part of *INSDC*, was established at the National Institute of Genetics (NIG), Japan, in 1986 with the support of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

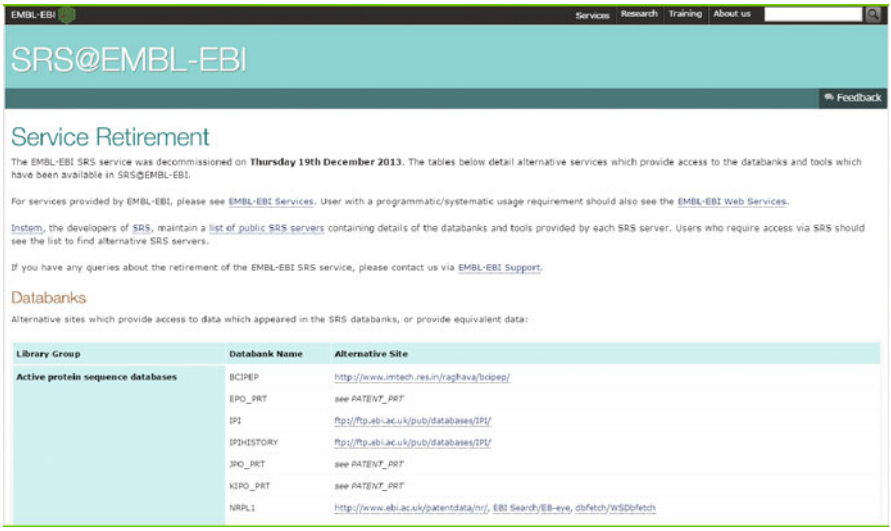


Fig. 1.7 The home page of Sequence Retrieval System (http://srs.ebi.ac.uk/)

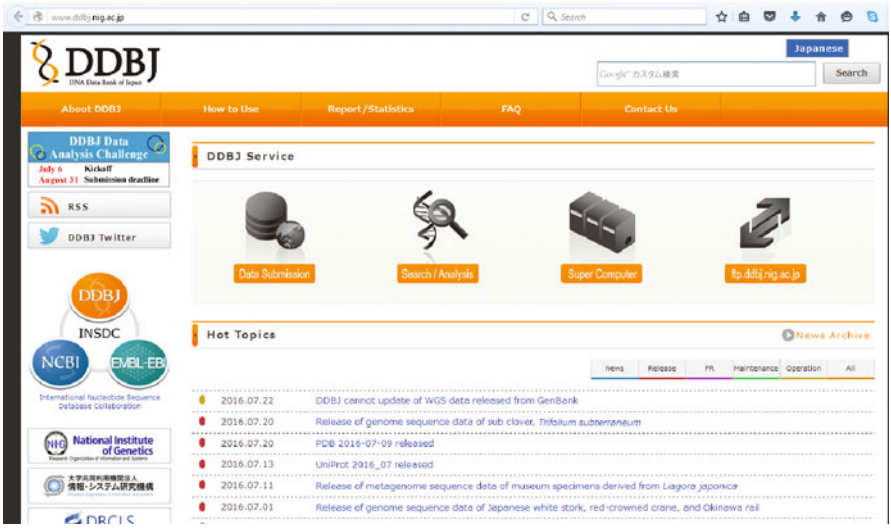


Fig. 1.8 The home page of DNA Data Bank of Japan (http://ddbj.sakura.ne.jp/)

SAKURA

SAKURA (http://sakura.ddbj.nig.ac.jp/top-e.html) is a source for data (nucleotide sequence) submission system through the WWW-based server where one can enter and submit nucleotide sequences and translated amino acid sequences. Since 1995 it is open to the public and scientists community.