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

# Current trends in Bioinformatics: An Insight



Current trends in Bioinformatics: An Insight

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

# Current trends in Bioinformatics: An Insight



*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

#### © Springer Nature Singapore Pte Ltd. 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.

Manju Sharom

(Dr. Manju Sharma)

Former Secretary, Department of Biotechnology New Delhi, India

Manju Sharma

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

# Contents

#### Part I Overview

1	An Insight of Biological Databases Used in Bioinformatics Vaibhav D. Bhatt, Monika Patel, and Chaitanya G. Joshi	3
2	<b>Bioinformatics in Next-Generation Genome Sequencing</b> Satendra Singh, Anjali Rao, Pallavi Mishra, Arvind Kumar Yadav, Ranjeet Maurya, Sukhdeep Kaur, and Gitanjali Tandon	27
3	The Role of Bioinformatics in Epigenetics Budhayash Gautam, Kavita Goswami, Neeti Sanan Mishra, Gulshan Wadhwa, and Satendra Singh	39
4	Three Dimensional Structures of Carbohydrates andGlycoinformatics: An OverviewK. Veluraja, J. Fermin Angelo Selvin, A. Jasmine,and T. Hema Thanka Christlet	55
5	<b>Epigenome: The Guide to Genomic Expression</b>	89
Par	t II Bioinformatics Approaches	
6	Molecular Modeling and Drug Design: A Contemporary Analysis in <i>Vibrio cholerae</i>	107
7	Modelling Polyketide Synthases and Similar Macromolecular Complexes	121
8	In Silico Studies on Colon Cancer	145

9	Tools, Databases, and Applications of Immunoinformatics Namrata Tomar and Rajat K. De	159
10	Metabolic Pathway Analysis Employing Bioinformatic Software Soma S. Marla, Neelofar Mirza, and K. D. Nadella	175
11	<b>The Interactomics of the RNA-Induced Silencing Complex</b> Abhijit Datta and Sayak Ganguli	193
12	Computational Tools: RNA Interference in Fungal Therapeutics Chakresh Kumar Jain and Gulshan Wadhwa	207
13	<b>Genome-Wide Essential Gene Identification in Pathogens</b> Budhayash Gautam, Kavita Goswami, Satendra Singh, and Gulshan Wadhwa	227
14	Disease Informatics	245
15	Development in Malaria and Anemia Screening: Medical Imaging Informatics Approach	263
16	Role of Bioinformatics in Drug Resistance Predictionfor HIV/AIDSJayakanthan Mannu and Premendu P. Mathur	277
17	<b>Bioinformatics Approaches for Animal Breeding and Genetics</b> Satendra Singh, Budhayash Gautam, Anjali Rao, Gitanjali Tandon, and Sukhdeep Kaur	287
18	<b>α-Amylase Inhibitor's Performance in the Control of </b> <i>Diabetes</i> <i>Mellitus</i> : An Application of Computational Biology Jyoti Verma, C. Awasthi, Qazi Mohammad Sajid Jamal, Mohd. Haris Siddiqui, Gulshan Wadhwa, and Kavindra Kumar Kesari	307

# 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

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

#### Abstract

Collections of life sciences information from scientific investigations, highthroughput 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  $\cdot$  High-throughput DNA sequencers  $\cdot$  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

<sup>©</sup> Springer Nature Singapore Pte Ltd. 2018

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).

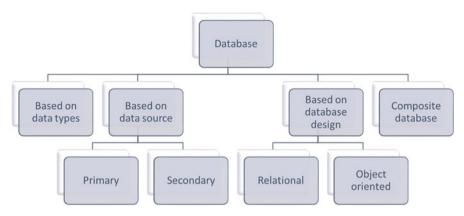
Due to the vast amount of data generated in experiments of genome, transcriptome, and exome sequences of various organisms in current times, the biological data has stored with an exponential rate. The availability of enormous amount of biological data (sequences as well as structural data) has generated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

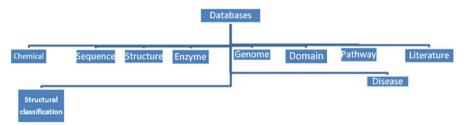
With the discovery of the integrated circuit, the powerful and reliable third generation computers are became the choice of storage of biological databases for scientists. An English scientist Tim Berners-Lee in 1989 invented the "World Wide Web" (WWW) which is the primary tool people use to interact 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

NSDC	POLICY	ADVISORS	DOCUME	NTS
• Ti 0	ne International Nucleotic perates between DDBJ, E	sence Database Collab de Sequence Database C <u>IMBL-EBI</u> and <u>NCBI</u> INSD nnotation, enriched with c	ollaboration (INSDC) is a C covers the spectrum of	data raw reads, tho
	Data type	DDBJ	EMBL-EBI	NCBI
	Next generation reads	Sequence Read Archive	-	Sequence Read
	Capillary reads	Trace Archive	European	Trace Archive
	Annotated sequences	1900	Archive (ENA)	GenBank
	Samples	BioSample		BioSample
	Studies	BioProject	_	BioProject
a e • In	he INSDC advisory board dvisory bodies. At their m kisting data-sharing polic	BioProtect the International Advisor iost recent meeting, memt y of the three databases t a to the international sequ	hat make up the INSDC, v	of members of eac animously endorse which is stated belo
• D	DBJ. ENA. GenBank	ibmit data to the database Definition Document is av		prating partner.

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

www.ncbi.nlm.nih.gov/nu	ccore/iterm=insulin	C Q Search		☆ 自 ♡	+ + 9
🗟 NCBI 🛛 Resources 🖾	How To 🗵			manisajnani M	y NCBI Sign
Nucleotide	Nucleotide	• insulin		O Search	
		Create alert Advanced			1
NCBI is phasing out seq	uence GI numbers	in September 2016. Please use accession version! Read more			
Species Animals (42,359)	Summary -	20 per page + Sort by Default order +	Send to: +	Filters: Manage Filters	
Plants (247) Fundi (2.253)	See PIN	(INSULIN) insulin precursor in the Gene database		Find related data	
Protists (291)		eference sequences Transcript (1) Protein (1)		Database: Select	
Bacteria (2,909) Archaea (0)					
fruses (125)	Items: 1 t	to 20 of 63446			
Dustomize		Contract CPress Page 1 of 3173 1	iext x Last xx	Search details	
tiolecule types		76982 nucleotide sequences. Nucleotide (63446) EST (13527) GSS (9)		insulin[All Fields]	
penomic DNA/RNA (37,436) mRNA (21,066)	Cctode	on degus insulin mRNA, complete ods		insulin[All Fields]	
Customize		p linear mRNA			
Source databases		ion:N57671.1 GL202471 nk FASTA Graphica			
NSDC (GenBank) (32,678)					
RefSeq (30,633) Dustomize		a californica insulin precursor (PIN). mRNA		Search	See mo
Senetic	000 00	Dinear mRNA ion: NM_001204686.1 GE 325286756			
compartments		ns FASTA Grachica		Recent activity	
Chloroplast(1) Mitchondrion (17)	E Anhois	a californica isolate F4 #8 unplaced genomic scalfold, AplCal3.0 scalfold00858, w			Tam Off O
Plasmid (5)		in sequence	nove Benance	Q insulin (63446)	
Plastd (1)	314,61	14 bp linear DNA		- <u>S</u>	N/de
Sequence length		ion: NWV_004798128.1 Git 523418921		A transcriptomic insight in javenile stage of the insect	
Dustom range	GenEar	nk EASTA Graphics		Identification of Viral Patho	
Release date	Cerato	otherium simum simum isolate SDZICR_KB13650 unplaced genomic scatfold, Cert	SimSim1.0	Sewage Sludge by Metage	

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.
FFATURES
                     Location/Qualifiers
                     1..432
     source
                     /organism="Octodon degus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10160"
                     /tissue_type="pancreas"
                     1..432
     gene
                     /gene="insulin"
     CDS
                     42..371
                     /gene="insulin"
                     /codon_start=1
                     /product="insulin"
                     /protein_id="AAA40590.1"
                     /translation="MAPWMHLLTVLALLALWGPNSVOAYSSOHLCGSNLVEALYMTCG
                     RSGFYRPHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFNQL
                     QNYCNVP"
                     42..113
     sig_peptide
                     /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
11
```

#### 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).

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		

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

#### 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.

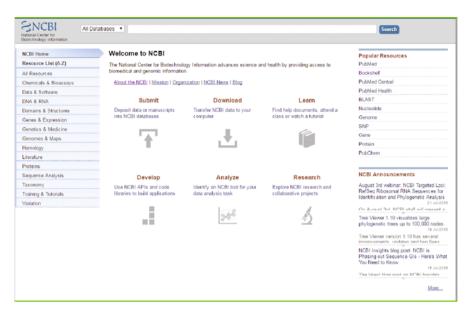


Fig. 1.3 The home page of National Center for Biotechnology Information (http://www.ncbi. nlm.nih.gov/)

Search NCBI data			Sign in to NG
	I		Search
Literature		Genes	
Books	books and reports	EST	expressed sequence tag sequences
MeSH	ontology used for PubMed indexing	Gene	collected information about gene loci
NLM Catalog	books, journals and more in the NLM Collections	GEO DataSets	functional genomics studies
PubMed	scientific & medical abstracts/citations	GEO Profiles	gene expression and molecular abundance profiles
ubMed Central	full-text journal articles	HomoloGene	homologous gene sets for selected organisms
lealth		PopSet	sequence sets from phylogenetic and population studies
linVer	human variations of clinical significance	UniGene	clusters of expressed transcripts
bGeP	genotype/phenotype interaction studies	Proteins	
TR IodGen	genetic testing registry	Conserved Domains	conserved protein domains
	medical genetics literature and links	Protein	protein sequences
MIM	online mendelian inheritance in man	Protein Clusters	sequence similarity-based protein clusters
ubMed Health	clinical effectiveness, disease and drug reports	Structure	experimentally-determined biomolecular structures
Senomes			adventurely and the second second
		Chemicals	
Assembly BioProject	genome assembly information biological projects providing data to NCBI	BioSystems	molecular pathways with links to genes, proteins and chemicals
Bio Sample	descriptions of biological source materials	PubChem BioAssav	bioactivity screening studies
lone IbVar	genomic and cDNA clones genome structural variation studies	PubChem Compound	chemical information with structures, information and links
ienome	genome sequencing projects by organism	PubChem Substance	deposited substance and chemical information
SS	genome survey sequences		
lucleotide	DNA and RNA sequences		
robe	sequence-based probes and primers		
NP	short genetic variations		
RA	high-throughput DNA and RNA sequence read archive		
Тахолоти	taxonomic classification and extended taxonomic classification		

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

EMBL-EBI 🧶		Services Research Training About us
EBI Search		insuln Scarch Exercises 300_452405 3ct. Builto
Help & Documentation About EBI Search		Share Feedback
Search results for <i>insulin</i> Showing 21 results out of 477,879 in All results		
Filter your results Source	Gene & protein summaries (includes expression, structure	os, literature) (4 results found)
All results (477,879) Genomes & metagenomes (14,806) hudestide secuences (13,568) Protein sequences (53,731) Macromolecular intructures (826) Small melocular (1456) Gene expression (8,225) Molecular interactions (469)	insulin/insulin-like growth factor receptor, DInR, DIHR, DILR, I(3)055	n-b. 18402, Inn-beta, DmahQ018403, Insk, <b>Enselfe</b> , Inskin-Neereseptor, 45, FBp00002933, Inskin-reseptor, Jethol(2)3303, DID, Enrapha, , dinsk, dink, Drosophila Insvin Receptor, (3)9303, FBpn0000456, Ink,
Reactions, pathways & diseases (2.096) Protein families (308) Protein expression data (118) Enzymes (185)	Ganorhabdilis elegans: Protein DAF-2, isoform b (dal-2) mRNA, com daf-2 (W8Genet0000898) Roundworm (Caenorhabditis elegans)	plete ods.
LEetature (350,289) Samples & ontologies (6,339) EBI web (94)	View all available Gene & protein summaries	E.c.
	Enzymes (no results lound)	
	KSGA3 HJMAN Ribosomal protein 56 kinase alpha-3 Ribosomal protein 56 kinase alpha-3	Related data • Viros • Source: Brayme Fortal 20: KS643_HUMAN
	View all 185 results for Enzymes	

**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
     Rattus norvegicus (Norway rat)
05
OC
     Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
oc
     Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
     Muridae; Murinae; Rattus.
OC
XX
RN
     [1]
RP
     1-782
RX
     DOI; 10.1111/j.1749-6632.1980.tb47271.x.
RX
     PUBMED: 6249167.
     Lomedico P.T., Rosenthal N., Kolodner R., Efstratiadis A., Gilbert W.;
"The structure of rat preproinsulin genes";
RΔ
RT
RL
     Ann. N. Y. Acad. Sci. 343:425-432(1980).
XX
     MD5; 2b03b65970e00e00d50a5054fad8125c.
DR
XX
CC
     On or before Jun 10, 2016 this sequence version replaced gi:204949,
cc
     gi:204950, gi:204951.
ΧХ
EH
                      Location/Qualifiers
     Key
FH
FT
                      1..782
     source
FT
                      /organism="Rattus norvegicus"
FT
                      /mol_type="genomic DNA"
/db xref="taxon:10116"
FT
FT
     gene
                      1..739
FT
                      /gene="INS2"
FT
     exon
                       (1. 46
                      /gene="INS2"
FT
FT
                       /number=1
FT
     intron
                      47..165
FT
                      /gene="INS2"
FT
                       /number=1
FΤ
     CDS
                       join(180..366,541..686)
FT
                       /codon_start=1
FT
                      /gene="INS2"
FT
                       /product="insulin 2"
FT
                       /note="precursor"
                      /protein id="AAA41440.1"
FT
FT
                       /translation="MALWIRFLPLLALLILWEPRPAQAFVKQHLCGSHLVEALYLVCGE
FT
                      RGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCTSICSLYQ
FT
                      LENYCN"
FΤ
     sig_peptide
                      180..251
FT
                      /gene="INS2"
FT
                       180..366
     exon
FT
                       /gene="INS2"
FT
                       /number=2
FT
                       /note="first expressed exon"
FT
     mat_peptide
                      252..341
                       /gene="INS2"
FT
FT
                       /product="beta chain"
FΤ
     mat_peptide
                      join(348..366,541..614)
                       /gene="INS2"
FT
FT
                       /product="insulin 2 connecting peptide"
FT
                      367..>410
     intron
                       /gene="INS2"
FT
FT
                       /number=2
FT
                      411..510
     gap
FT
                      /estimated length=unknown
```

```
FT
    intron
                    <511..540
FT
                   /gene="INS2"
FT
                    /number=2
FT
    exon
                    541..739
                    /gene="INS2"
FT
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"
                    /product="insulin 2"
FT
FT
                    /note="alpha chain"
XX
SO
   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 accttttgct agcctcaacc ctgactatct tccaggtcat tgttccaaca
                                                                          180
    tggccctgtg gatccgcttc ctgcccctgc tggccctgct catcctctgg gagccccgcc
                                                                          240
                                                                          300
     ctgcccaggc ttttgtcaaa cagcaccttt gtggttctca cttggtggaa gctctctacc
    tggtgtgtgg ggagcgtgga ttcttctaca cacccatgtc ccgccgcgaa gtggaggacc
                                                                          360
                                                                          420
    cacaaggtaa gctctgctcc tgaattctat cccaagtgct aactaccctg nnnnnnnn
    480
    nnnnnnnn nnnnnnnn nnnnnnnn tgccctgtgc tgacatgacc tccctggcag
                                                                          540
    tggcacaact ggagctgggt ggaggcccgg gggccggtga ccttcagacc ttggcactgg
                                                                          600
    aggtggcccg gcagaagcgc ggcatcgtgg atcagtgctg caccagcatc tgctctctt
                                                                          660
    accaactgga gaactactgc aactaggccc accactaccc tgtccacccc tctgcaatga
                                                                          720
    ataaaaacctt tgaaagagca ctacaagttg tgtgtacatg cgtgcatgtg catatgtggt
                                                                          780
                                                                          782
    gc
```

#### 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.

		Services Research Training About us
SRS@EMBL-I	EDI	
		🗣 Feedba
Service Retiremen	+	
	-	
he EMBL-EBI SRS service was decommiss ave been available in SRS@EMBL-EBI.	sioned on Thursday 19th	h December 2013. The tables below detail alternative services which provide access to the databanks and tools which
or services provided by EMBL-EBL please	e see EMBL-EBI Services.	User with a programmatic/systematic usage requirement should also see the EMBL-EBI Web Services.
nstem, the developers of SRS, maintain a see the list to find alternative SRS servers		rs containing details of the databanks and tools provided by each SRS server. Users who require access via SRS should
f you have any queries about the retirem	ent of the EMBL-EBI SRS	service, please contact us via EMBL-EBI Support.
Databanks		
Alternative sites which provide access to o	data which appeared in th	he SRS databanks, or provide equivalent data:
	data which appeared in ti Databank Name	he SRS databanks, or provide equivalent data: Alternative Site
Library Group		
Library Group	Databank Name	Alternative Site
Library Group	Databank Name	Alternative Site http://eww.intech.resi.ny/ragilaxu/bdpeg/
Library Group	Databank Name BCIPEP EPO_PRT	ARemative Site http://www.intech.res.in/raghava/bdpeg/ see PATEUT_PRT
Library Group	Databank Name BCIPEP EPO_PRT EP1	ARemative Site http://www.intech.res.in/raghava/bdipep/ see PATEINT_PRT ftp://ftp.4bi.ac.uk/pub/dstabases/TPL/
Uternative sites which provide access to o Library Group Active protein sequence databases	Databank Nome BCIPEP EPO_PRT IPI IPIHISTORY	Alternative Site           http://mww.imtech.ws.in/roghava/holpep/           see Pattive_Per           fg://ftp.vbiac.us/pub/databases/PEr/           fg://ftp.vbiac.us/pub/databases/PEr/

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

🛞   www.ddbj <b>nig.ac.jp</b>								-	_	_
				Google".	スタム検索			Ja	panes	e Search
				Cor						
DDBJ Data Analysis Challenge July 6 Kickaff August 31 Submission deadline	DDBJ Service									
A 855		60					-	1		
		-	2							
DDBJ Twitter	Data Submissi	on Search / Ar	nalysis	Super Computer			Rp.ddbj.r	ig ac ja		
DDBJ INSDC	Data Submissi	on Search / A	ulysis	Super Computer				ON		rchive
DDBJ	Hot Topics			News	Release	R	Haintenar			rchive At
DDBJ INSDC	Hot Topics	DDBJ cannot update of WGS		News			Haintenar	ON		rchive Al
INSDC NCRIMEL-EB	Hot Topics	DDBJ cernot update of WGS Release of genome sequence	deta released from GenBe data of sub clover, Trifale	nens ink im subterraneum			Haintenar	ON		rchive At
INSDC NCRIMEL-EB	Hot Topics	DDBJ cernot update of WGS Release of genome sequence PDB 2016-07-09 released	data released from GenBa data of sub clover, Trifalu	nens ink im subterraneum			Haintenar	ON		rchive Al
INSDC INSDC NCBI Presidential Rudenton Research Sectore Contemportune Contemportune Sectore Contemportune Contemportune Sectore Contemportune Sectore Cont	Hot Topics	DDBJ cernot update of WGS Release of genome sequence PDB 2016-07-09 released	data released from GenBa	nens Ink Im subberraneum			Haintonar	ON:	stion	Al
INSOC NCB Demandar lucicides Factures	Hot Topics  2016.07.22 2016.07.20 2016.07.20 2016.07.13 2016.07.13	DDBJ cernot update of WoS Release of genome sequence POB 2016-07-09 released Uniting 2016_07 released	data released from GenBa data of sub clover, Trifaiu usence data of museum spe	news Ink In subterraneum	i from Liagon	a japon	Maintenar	ON4	stion	Al

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.