

Sequence Submission / Data Submission to GenBank or EMBL

Sequence data may be submitted to GenBank or EMBL using one of the methods described in the GenBank release notes. The relevant parts of the GenBank release notes are quoted below.

"To assist researchers in entering their own sequence data, GenBank provides a WWW submission tool called BankIt, as well as a stand-alone software package called Sequin. BankIt and Sequin are both easy-to-use programs that enable authors to enter a sequence, annotate it, and submit it to GenBank. Through the international collaboration of DNA sequence databases, GenBank submissions are forwarded daily for inclusion in the EMBL and DDBJ databases."

Sequin

"Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides you through the process of entering your sequence and providing biological and bibliographic annotation. Intended as an alternative to the older Authorin program, Sequin is designed to simplify the sequence submission process, and to provide increased data handling capabilities to accommodate very long sequences, complex annotations, and robust error checking. E-mail the completed submission file to: gb-sub@ncbi.nlm.nih.gov

Sequin is currently provided as a beta-test version, and runs on Macintosh, PC/Windows, UNIX and VMS computers. It is available by anonymous ftp from [ncbi.nlm.nih.gov](ftp://ncbi.nlm.nih.gov), login as anonymous and use your e-mail address as the password. It is located in the sequin directory."

BankIt

"BankIt provides a simple forms approach for submitting your sequence and descriptive information to GenBank. Your submission will be submitted directly to GenBank via the World Wide Web, and immediately forwarded for inclusion in the EMBL and DDBJ databases. BankIt may be used with Netscape clients for Unix, Macs, and PCs, the Mosaic client for Unix, and the MacWeb client for Macs. You can access BankIt from GenBank's home page: <http://www.ncbi.nlm.nih.gov/>"

Webin

www.rbehera.in

It is the European Bioinformatics Institute submitting program which guides users via a sequence checklist and their forms to allow the interactive as well as descriptive submission information. All the information required to create a databases access could be amassed during this process, i.e.:

1. Submitter data
2. Launch date information
3. Sequence statistics, description, and source information
4. Reference quotation information

This program is used to enter the data as in single as well as multiple entries.

Data Retrieval Systems

**Text-based Database
Searching**

www.rbehera.in

The amount of biologically relevant data accessible via the WWW is increasing at a very rapid rate.

It is important for Scientists to have **easy and efficient** ways of wading through the data and finding what is important for their research.

Knowing how to access and search for information in the database is essential.

Depending on the type of data at hand, there are **two** basic ways of searching:

- using descriptive words to search **text databases**
- using a nucleotide or protein sequence to search **sequence databases**

Text-based Database Searching

There are three important data retrieval systems of particular relevance to molecular biologists:

- ◆ **Entrez** (at NCBI)
- ◆ **Sequence Retrieval System, SRS** (at EBI)
- ◆ **DBGET/LinkDB** (at Japan)

The advantage of these retrieval systems is that they not only **return matches** to a query, but also **provide handy pointers** to additional important information in related databases

Text-based Database Searching

The three systems **differ** in the databases they **search** and the **links** they provide to other information.

In using any of these systems, queries can be as simple as entering the **accession number** of a newly published sequence or as complex as searching **multiple database fields for specific terms**

Text-based Database Searching

Basic Search Concepts

- **Boolean Search** – An advanced query search using two or more terms, using Boolean operator AND, OR, NOT, default – AND
- **Broadening the Search** – If the results of a search produce no useful entries, change or remove terms
- **Narrowing the Search** – If the results of a search produce too many entries, change or add terms
- **Proximity Searching** – To search with multiword terms or phrases, place quotes around the terms
- **Wild Card** – The character * prepended or appended to a search term make a search less specific., e.g., to look for all authors with last name Zav, search using Zav*

Entrez

Entrez - is a molecular biology database and retrieval system developed by the National Center for Biotechnology Information (NCBI)

It is an entry point for exploring distinct but integrated databases.

www.rbehera.in

Reference:

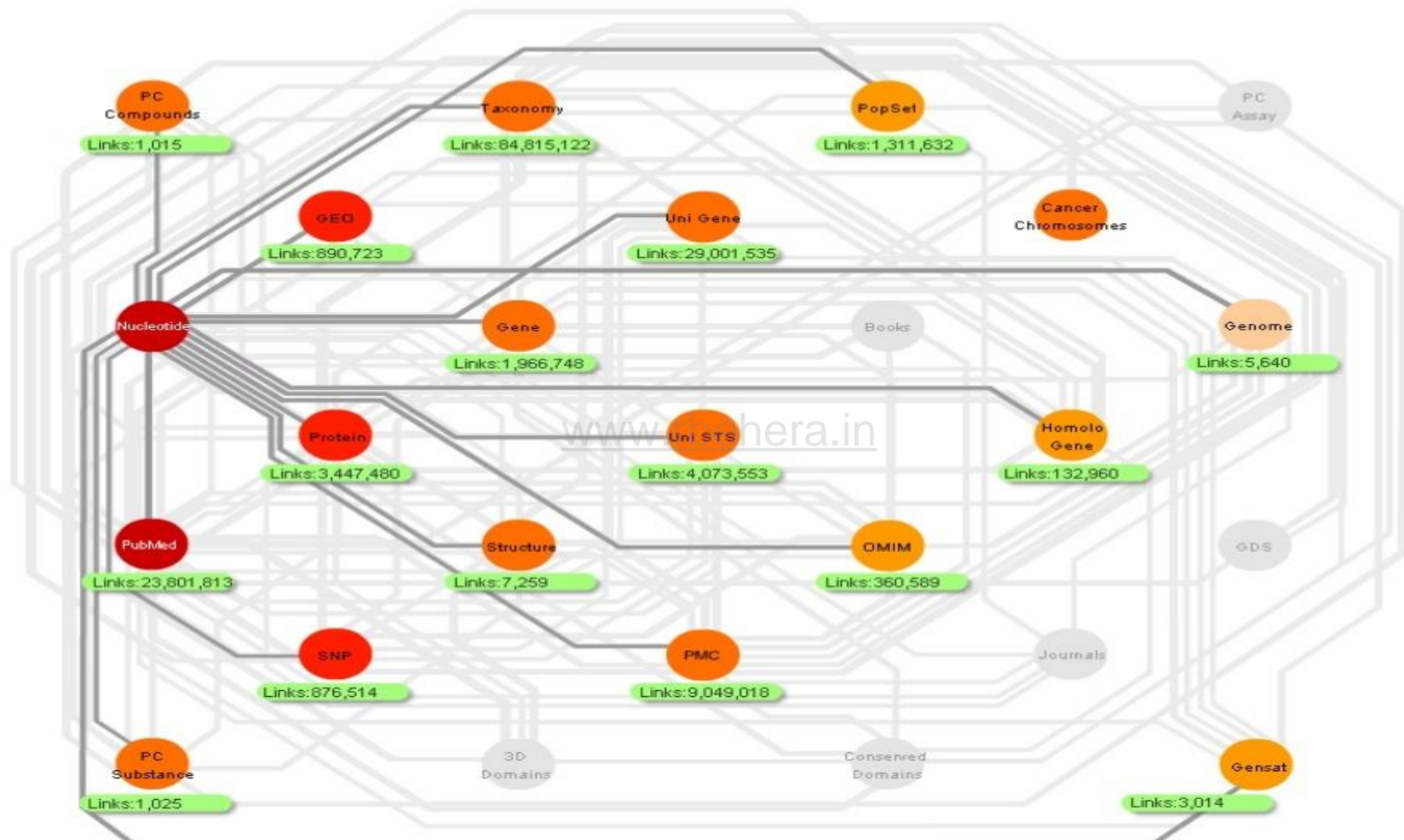
Entrez: Molecular Biology Database and Retrieval System, Schuler GD, Epstein JA, Ohkawa H, Kans JA, *Methods Enzymol.* 266, 141-62, 1996.

(<http://www.ncbi.nlm.nih.gov/Entrez/>)

Entrez

The Entrez system provides access to:

- **Nucleotide sequence databases** – GenBank/DDBJ/EBI
- **Protein sequence databases** - Swiss-Prot, PIR, PRF, PDB, and translated protein sequences from DNA sequence databases
- **Genome and chromosome mapping data**
- **Molecular Modeling 3-D structures Database**
- **Literature database**, PubMed - provides excellent and easy access to MEDLINE and pre-MEDLINE articles.
- **Taxonomy database** - allows retrieval of DNA and protein sequences for any taxonomic group.
- **Specialized Databases** – OMIM, dbSNP, UniSTS, etc.





Search across databases

GO

Clear

Help

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts	Books: online books
PubMed Central: free, full text journal articles	OMIM: online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites	OMIA: online Mendelian Inheritance in Animals

Nucleotide: Core subset of nucleotide sequence records	dbGaP: genotype and phenotype
EST: Expressed Sequence Tag records	UniGene: gene-oriented clusters of transcript sequences
GSS: Genome Survey Sequence records	CDD: conserved protein domain database
Protein: sequence database	3D Domains: domains from Entrez Structure
Genome: whole genome sequences	UniSTS: markers and mapping data
Structure: three-dimensional macromolecular structures	PopSet: population study data sets
Taxonomy: organisms in GenBank	GEO Profiles: expression and molecular abundance profiles
SNP: single nucleotide polymorphism	GEO DataSets: experimental sets of GEO data
Gene: gene-centered information	Cancer Chromosomes: cytogenetic databases
HomoloGene: eukaryotic homology groups	PubChem BioAssay: bioactivity screens of chemical substances
GENSAT: gene expression atlas of mouse central nervous system	PubChem Compound: unique small molecule chemical structures
Probe: sequence-specific reagents	PubChem Substance: deposited chemical substance records
Genome Project: genome project information	Protein Clusters: a collection of related protein sequences

Journals: detailed information about the journals indexed in PubMed and other Entrez databases

NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections

MeSH: detailed information about NLM's controlled vocabulary

<http://www.ncbi.nlm.nih.gov/Entrez/>

Search Nucleotide for Go Clear

Limits

Preview/Index

History

Clipboard

Details

The Entrez Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of April 2006, there are over 130 billion bases in GenBank and RefSeq alone.

Human Genome

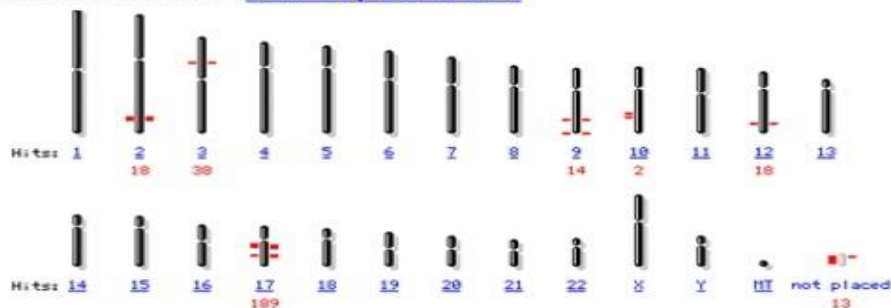
Explore [human genome resources](#) or browse the human genome sequence using the [Map Viewer](#).

Building the human genome


The Human Genome Reference DNA Sequence was completed in April 2003. The current version is listed as a build number on the [Genome View](#) page and includes an accompanying set of [statistics](#) and [release notes](#).

Homo sapiens (human) genome view BLAST search the human genome

Build 36.2 statistics [Switch to previous build](#)



The chromosomal locations of several genes believed to be associated with the human BRCA1 gene implicated in breast cancer, highlighted

 About Entrez 

 Entrez Nucleotide
 Help | FAQ

Entrez Tools

 Check sequence
 revision history

LinkOut

My NCBI (Cubby)

 Related resources
 BLAST

 Reference sequence
 project

Search for Genes

Submit to GenBank

 Search for full length
 cDNAs

[PubMed](#)
[Entrez](#)
[BLAST](#)
[OMIM](#)
[Books](#)
[TaxBrowser](#)
[Structure](#)

Search for

NCBI

[Site Map](#)

Guide to NCBI
resources

Entrez Help

Help documentation for
the Entrez system

Entrez Tutorial**Entrez Global
Query**

Search a subset of
Entrez databases

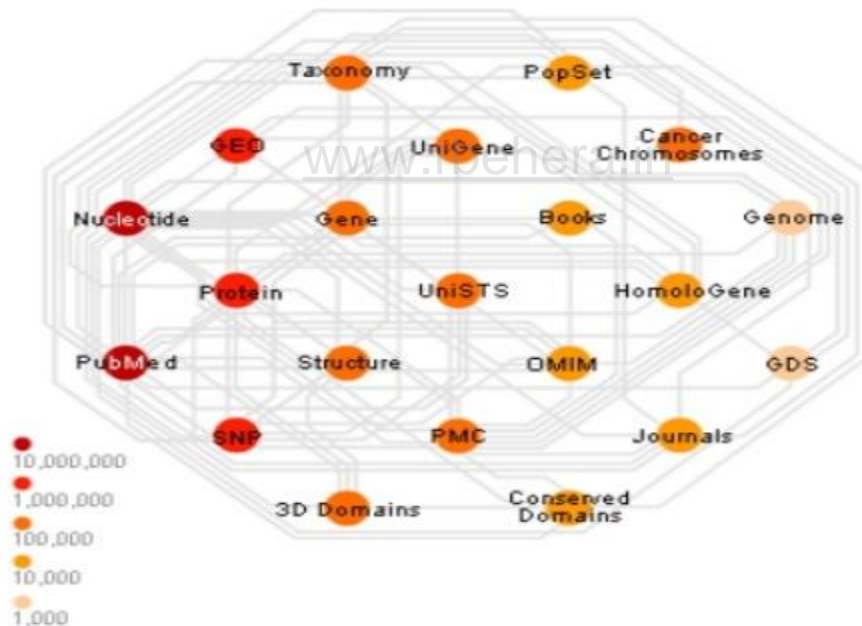
Entrez Tools

Links to advanced
Entrez tools such as
Batch Entrez and
E-Utilities

NCBI Handbook

In-depth guide to NCBI
resources

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. Click on the graphic below for a more detailed view of Entrez integration.



PubMed

Entrez

BLAST

OMIM

Books

TaxBrowser

Structure

Search Entrez



for

Go

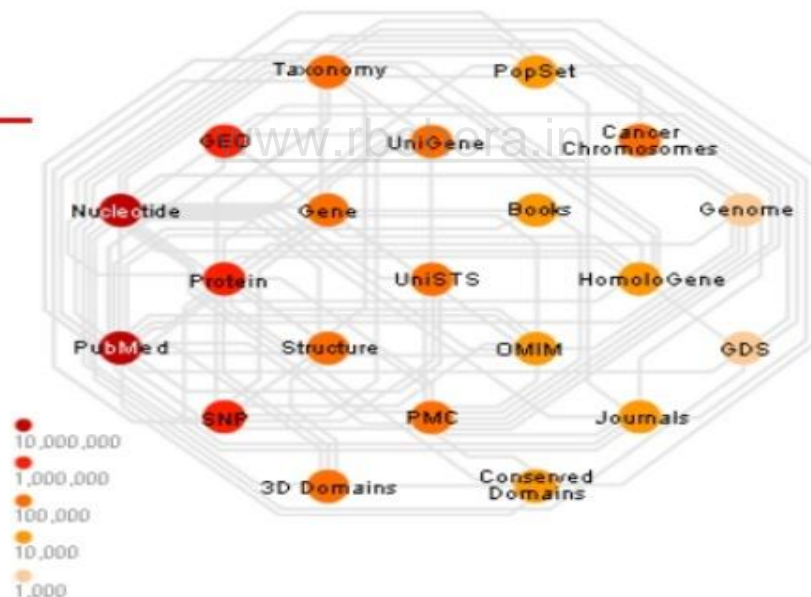
NCBI

Site Map

Guide to NCBI
resources**Entrez Help**Help documentation for
the Entrez system**Entrez Tutorial****Entrez Global
Query**Search a subset of
Entrez databases**Entrez Tools**Links to advanced
Entrez tools such as
Batch Entrez and
E-Utilities**NCBI Handbook**In-depth guide to NCBI
resources

LinkOut

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. Click on the graphic below for a more detailed view of Entrez integration.



Entrez Tutorial

Entrez: Making use of its power

This tutorial is designed to show users how to make use of the full power of the Entrez data retrieval system. Using a human gene as an example, it demonstrates the variety of information that can be gathered for a single gene. The number of records retrieved will change over time as the databases grow. However, the same techniques shown in the tutorial can be used for any topic of interest.

Learn how to:

- identify a representative, well-annotated mRNA sequence record from the millions of sequences in the Entrez Nucleotide data domain
- retrieve associated literature and protein records
- identify conserved domains within the protein
- identify known mutations within the gene or protein
- find a resolved three-dimensional structure for the protein, or, in its absence, identify structures with homologous sequence
- view the genomic context of the gene and download the sequence region

Go to the pdf version of [Entrez: Making use of its power](#).

Geer, R.C. and Sayers, E.W. Entrez: Making use of its power. *Briefings in Bioinformatics*. 2003 June;4(2):1779-184.

PubMed

Entrez

BLAST

OMIM

Books

TaxBrowser

Structure

Search

Entrez



for

Go

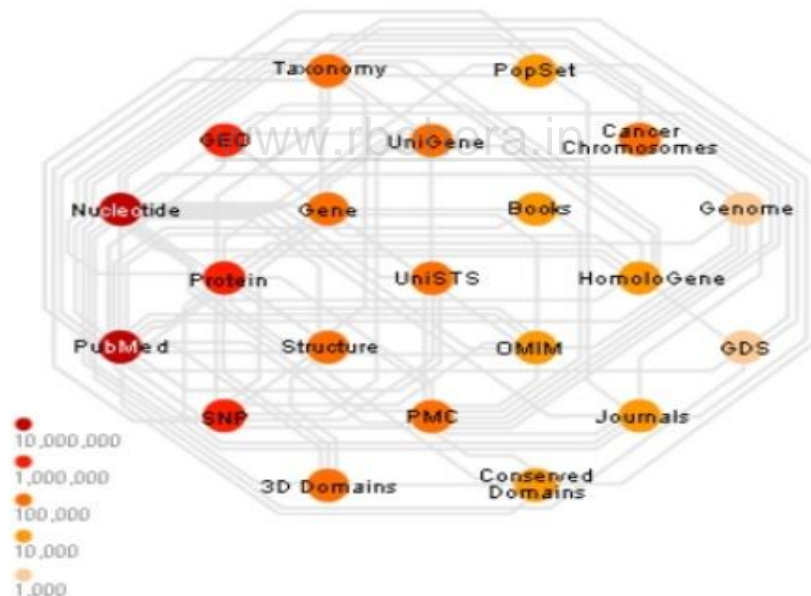
NCBI

Site Map

Guide to NCBI
resources**Entrez Help**Help documentation for
the Entrez system**Entrez Tutorial****Entrez Global
Query**Search a subset of
Entrez databases**Entrez Tools**Links to advanced
Entrez tools such as
Batch Entrez and
E-Utilities**NCBI Handbook**In-depth guide to NCBI
resources

LinkOut

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. Click on the graphic below for a more detailed view of Entrez integration.



[Entrez](#)[PubMed](#)[Nucleotide](#)[Protein](#)[Genome](#)[Structure](#)[OMIM](#)[PMC](#)[Journals](#)[Books](#)

Advanced Entrez Tools

Web Tools:

[Batch Entrez](#) - Upload a file of GI or accession numbers to retrieve sequences.

PubMed [Batch Citation Matcher](#) - Send citation information to Entrez and retrieve PubMed IDs for linking, citation display, or other applications.

[Advanced Entrez Searching](#) - Advanced searching techniques for Web Entrez.

[My NCBI](#) - includes automatic e-mailing of search updates and filters for search results.

Programming Tools:

[E-Utilities](#) - Run Entrez queries and download data from your own scripts over the Web.

[Linking to Entrez](#) - Link to specific Entrez pages from your own web pages or applications.

[Entrez Client/Server](#) - C language library for embedding Entrez calls into your programs.

Entrez

The most valuable feature of Entrez is

- its exploitation of the concept of ‘neighbouring’
- which allows related articles in different databases to be linked to each other, whether or not they are cross-referenced directly.

www.rbehera.in

Neighbours and links are listed in the order of **similarity** to the query.

The **similarity** is based on pre-computed analyses of sequences, structures and the literature.

Entrez

One particularly **useful feature** in Entrez is –

The ability to retrieve large sets of data based on some criterion and to download them to a local computer – **Batch Entrez**

www.rbehera.in

- allowing these sequences to be worked on using analytical tools available on local computer.

Entrez Features

- **Entrez Global Query** - Search a subset of Entrez databases
- **Batch Entrez** - Upload a file of GI or accession numbers to retrieve sequences
- **Making Links Entrez** - Linking to PubMed and GenBank
- **E-Utilities** - Entrez programming utilities
- **LinkOut** - External links to related resources
- **Cubby** - provides with a Stored Search feature to store and update searches, allows to customize your LinkOut display

SRS

The Sequence Retrieval System (SRS) – a network browser for databases in molecular biology.

It is a powerful sequence information indexing, search and retrieval system (<http://srs.ebi.ac.uk/>)

Reference:

1. T. Etzold, A. Ulyanov, and P. Argos, SRS: Information Retrieval System for Molecular Biology Data Banks. *Methods in Enzymology*, 266, 114, 1996.
2. T. Etzold and P. Argos, SRS - An Indexing and Retrieval Tool for Flat File Data Libraries. *Comp. Appl. Biosciences (CABIOS)*, 9, 49-57, 1993.

Index

- # databank
- # faq
- # guides
 - Linking to SRS
 - SRS Command-line Usage
 - SRS Query Language Quick Guide
 - SRS URL API
- # tutorial
- # view
- # SRS@EBI

Trace: > [Linking to SRS](#)

Linking to SRS

This document describes how to add hyperlinks from your own pages to the EBI SRS server. We strive to maintain links to older versions of SRS continue to do so in the future. However, SRS is evolving all the time and users need to be aware of changes to the SRS syntax. This page describes changes.

Supported links and Services at EBI

First we would like to make you aware of the following services which should be used to obtain entries by id or accession number from datab

- UniProt - [uniprot.org](http://www.uniprot.org) service, see [documentation](#) for details. E.g. http://www.uniprot.org/entry/fos_human
- EBI Fetch Tools:
 - [dbfetch](#) - E.g. <http://www.ebi.ac.uk/cgi-bin/dbfetch?db=EMBL&id=hscfos>
 - [emblfetch](#) - get entries from EMBL
 - [medlinefetch](#) - get entries from MEDLINE
 - [pdbfetch](#) - get entries from PDB
 - [swissfetch](#) - get entries from UniProt
- [SOAP based Web Services](#) - <http://www.ebi.ac.uk/Tools/webservices/>

Program and Script Usage

Important: resources at EBI are not unlimited and many users are using the services concurrently. Therefore, we ask that you set a pause on scripts or programs. Occasionally we may have to black-list a user, host or IP address in order to ensure fair access to the service for all.

We also impose limits on memory and CPU usage. Currently `wget-z` jobs are limited to a maximum of 2GB of memory and 20 minutes CPU requests. So remember to check that you have the expected number of results.

Reset

Quick Search

Search Options

1. Select the **databanks** you want to search

2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

Standard Query Form

Extended Query Form

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:

Browse Entries

Tips

- ▶ bookmark this [link](#) to return to your project
- ▶ *Linking to SRS?*
- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

BookMarkLets

About BookmarkLets

Available Databanks

 Expand all
 Collapse all
Show databanks tooltips:
 Literature, Bibliography and Reference Databases

 all
 TAXONOMY
 GENETICCODE
 OMIM
 MEDLINE

 Patent Abstracts
 Karyn's Genomes

Literature, Bibliography and Reference Databases - subsections
 all
 MEDLINE (Updates)
 MEDLINE (Main Release 2007)
 MED2PUB

 Gene Dictionaries and Ontologies

 Nucleotide sequence databases

 all
 EMBL
 Patent DNA
 IMGT/LIGM-DB
 IMGT/HLA

 IPD-KIR
 EMBL (Contig)
 EMBL (Contigs expanded)
 EMBL (Annotated C

 EMBL (Coding Sequences)
 Genome Reviews
 RefSeq Genome
 LiveLists

 EMBL ID/Accession Mapping
 EMBL MGA

Nucleotide sequence databases - subsections
 all
 EMBL (Updates)
 EMBL (Release)
 EMBL (Whole Genome Shotgun Shotgun)

 EMBL (Whole Genome Shotgun release)
 EMBL (Whole Genome Shotgun updates)
 EMBL (Contig release)

 EMBL (Contig updates)
 EMBL (Contigs expanded release)
 EMBL (Contigs expanded updates)

 EMBL (Annotated Cons release)
 EMBL (Annotated Cons updates)
 RefSeq Genome (Release

 RefSeq Genome (Updates)
 EMBL (Whole Genome Shotgun Masters)

 Nucleotide related databases

 UniProt Univers

 all
 UniProtKB
 UniProtKB/Swiss-Prot
 UniProtKB/TrEMBL
 UniRef100
 UniRef90

 UniRef50

 Other protein s

UniProtKB: The Universal Protein Resource Knowledgebase- produced by SIB,PIR and EBI.

To obtain comprehensive information on

SRS

SRS is a homogeneous interface to over **80** biological databases developed at the European Bioinformatics Institute (EBI) at Hinxton, UK.

The types of databases included are sequence and sequence related, metabolic pathways, transcription factors, application results (e.g., BLAST), protein 3D-structure, genome, mapping, mutations, and locus-specific mutations.

One can access and query their contents and navigate among them.

SRS

The Web page listing all the databases contains a link to a description page about the database and includes the date of last update.

One can select one or more databases to search before entering the query.

Over **30 versions of SRS** are currently running on the WWW. Each includes a different subset of databases and associated analytical tools.

SRS

SRS Features:

- SRS databases are **well indexed**, thus reducing the search time for the large number of potential databases
- SRS allows any **flat file database** to be indexed to any other. The advantage being the derived indices may be rapidly searched, allowing users to retrieve, link and access entries from all the interconnected resources
- The system has the particular strength that it can be readily **customized** to use any defined set of databanks.

SRS

Simple SRS queries


- By accession number
- Query on accession number: J00231
- By a simple author **or** organism name
- Query on author **and** organism: Ausubel
AND Rhizobium
- Boolean relations between keywords: and, or,
but not

SRS

contd....

- **Searching by dates:** 01-Jan-2019: 31-Dec-2019
- **Searching by size:** 400:600
- **Using hypertext links in an entry:** Medline, Swiss-Prot, and PDB entries can be linked from within the EMBL database
- **Display of molecules via Rasmol plug-in**

Tools in SRS at EBI

EMBL-EBI  All Databases 

[Databases](#) [Tools](#) [Groups](#) [Training](#) [Industry](#) [About Us](#) [Help](#)

[Quick Search](#) [Library Page](#) [Query Form](#) [Tools](#) [Results](#) [Projects](#) [Views](#)

Quick Launch

Launch analysis tool:

Packages Information

[BLAST](#)[OTHER](#)[FASTA](#)[CLUSTAL](#)[HMMER](#)[EMBOSS](#)

Search Descriptions

Available Analysis Tools - listed by type

- Alignment Tools**
- Display Tools**
- Edit Tools**
- Information Tools**
- Nucleic Tools**
- Protein Tools**
- Phylogeny Tools**
- Similarity Search Tools**

DBGET

DBGET/LinkDB - is an integrated bioinformatics database retrieval system at GenomeNet, developed by the Institute for Chemical Research, Kyoto University, and the Human Genome Center of the University of Tokyo.

References:

www.rbehera.in

1. M. Kanehisa, Linking databases and organisms: GenomeNet resources in Japan. Trends Biochem Sci. 22, 442-444 (1997).
2. Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., & Kanehisa, M.; DBGET/LinkDB: an integrated database retrieval system. Pacific Symp. Biocomputing 1998, 683-694 (1997).



Search for

GenomeNet

[About GenomeNet](#)
[Announcements](#)
[Release notes](#)
[Acknowledgments](#)

KEGG

[Overview](#)
[DB release info](#)

DBGET

[Overview](#)
[DB release info](#)
[DB growth curve](#)

Community DBs

Bioinformatics tools

[Other tools](#)

Feedback

GenomeNet Database Resources

KEGG: Kyoto Encyclopedia of Genes and Genomes

[KEGG - Top page](#)
[KEGG2 - Table of contents](#)
[KEGG Atlas - Global maps interface](#)
[KEGG PATHWAY - Systems information: pathways](#)
[KEGG BRITE - Systems information: ontologies](#)
[KEGG ORTHOLOGY \(KO\) - Ortholog information](#)
[KEGG GENES - Genomic information](#)
[KEGG LIGAND - Chemical information](#)
[KEGG Organisms - Organism-specific entry points](#)
[KEGG DISEASE - Disease information resource](#)
[KEGG DRUG - Drug information resource](#)
[KEGG GLYCAN - Glycan information resource](#)
[KEGG PLANT - Plant information resource](#)

DBGET: Integrated Database Retrieval System

[DBGET search](#)
[LinkDB search](#)

Community Databases

[CYORF - Cyanobacteria annotation database](#)
[BSORF - Bacillus subtilis genome database](#)
[EXPRESSION - Gene expression profile database](#)

KEGG Update Notes

[KEGG Organisms](#)
[KEGG Organisms in the Taxonomy](#)

[KEGG Atlas Metabolism map](#)
[KEGG Atlas Cancer map](#)

Features of DBGET search

1. "All databases" search option
2. Combination of brite search and bfind search
3. Direct links to original sites
4. Direct program links by Java Web Start
5. Over 500 databases in LinkDB

GenomeNet Bioinformatics Tools

Sequence Analysis

[BLAST / FASTA - Sequence similarity search](#)
[MOTIF - Sequence motif search](#)
[CLUSTALW / MAFFT / PRRN - Multiple alignment](#)

Genome Analysis

[KAAS - KEGG automatic annotation server](#)
[EGassembler - EST consensus contigs](#)
[GENIES - Gene network prediction](#)
[GECS - Gene expression to chemical structure](#)

Chemical Analysis

[SIMCOMP - Chemical structure search](#)
[KCaM - Glycan structure search](#)
[e-zyme - Reaction prediction](#)

<http://www.genome.ad.jp/>

Search for

Go

Clear

GenomeNet

- About GenomeNet
- Announcements
- Release notes
- Acknowledgments

KEGG

- Overview
- DB release info

DBGET

- Overview
- DB release info
- DB growth curve

Community DBs**Bioinformatics tools**

- Other tools

Feedback

DBGET: Integrated Database Retrieval System

DBGET is an integrated database retrieval system for handling the web of molecular biology databases, which is used as a backbone system in GenomeNet and KEGG including the search box shown above. This web is a huge graph consisting of individual database entries as nodes and cross-reference links among databases as edges. Each database entry is identified by the combination of a database name and an entry name (or accession number), which can usually be converted to an URL, and the name space containing all cross-reference links forms the LinkDB database. DBGET/LinkDB is currently under a new development phase for integration of both Genomenet databases and outside databases.

DBGET search**LinkDB search**

Other entry points

- Links Diagram (to be discontinued)
- Advanced search (to be discontinued)

Document

- How to use DBGET ←
- URLs for making DBGET queries

Release information

- Database Release Information (Daily updated)
- Growth of Major Databases (Since 1982)

DBGET

DBGET - is used to search and extract entries from a wide range of molecular biology databases

LinkDB - is used to compute links between entries in different databases.

It is designed to be a **network distributed database system** with an open architecture, which is suitable for incorporating local databases or establishing a server environment.

<http://www.genome.ad.jp/dbget/>



DBGET Search

DBGET

LinkDB

KEGG2

 Search for

Go

Clear

DBGET is an integrated database retrieval system for major biological databases, which are classified into five categories:

Category	Main commands			Remark
	bget	bfind	blink	
1. KEGG databases in DBGET	yes	yes	yes	Mirrored at GenomeNet
2. Other DBGET databases	yes	yes	yes	
3. Searchable databases on the Web	no	yes	yes	Used as Web resources
4. Link-only databases on the Web	no	no	yes	
5. PubMed database	yes	no	yes	

Databases in the third category are integrated for keyword search, but the actual data are to be obtained from the original sites. Databases in the fourth category are available only in the LinkDB system. PubMed is a link-only database, but the bget page is generated using the NCBI service in order to better integrate with KEGG and other DBGET databases.

1. KEGG Databases in DBGET

Database name		Abbreviation	Content	Remark	
kegg	brite	br	KEGG functional hierarchies	See KEGG BRITE	
	pathway	path	KEGG pathways	See KEGG PATHWAY	
	module	md	KEGG modules		
	disease	ds	KEGG diseases	See KEGG DISEASE	
	orthology	ko	KEGG orthology	See KEGG ORTHOLOGY	
	genes	Individual organisms	org code	Gene catalogs in high-quality genomes	See KEGG GENES
	dgenes			Gene catalogs in draft genomes	
	egenes			Gene catalogs generated as EST contigs	
	genome	gn	KEGG organisms		
	ligand	compound	cpd	Chemical compounds	See KEGG LIGAND
		drug	dr	Drugs	
		glycan	gl	Glycans	
		reaction	rn	Chemical Reactions	
		rpair	rp	Reactant pairs	
	enzyme	ec	Enzyme nomenclature		
	expression	ex	Gene expression profiles	Submitted by authors	
	vgenome	vgnm	Viral genomes	Computationally generated from RefSeq	
	vgenes	vg	Viral gene catalogs		
	ogenes	og	Organelle gene catalogs		

2. Other DBGET Databases

Database name		Abbreviation		Content	Original site
refseq	refnuc	rs	rsnt	NCBI Reference Sequence Database	NCBI
	refpep		rsaa		
uniprot	swissprot	up	sp	UniProt (Universal Protein Resource) protein sequence database	ExPASy / EBI
	trembl		tr		
pir		pir	PIR (Protein Information Resource) protein sequence database (final release of Dec 04)	NBRF	
prf		prf	PRF peptide/protein sequence database	PRF	
pdb		pdb	PDB (Protein Data Bank) 3D structure database	RCSB	
epd		epd	Eukaryotic promoters	ISREC	
transfac		tf	Transcription factors (warning: release of Dec 99)	BIOBASE	
motifdic	prosite	ps	Protein domains and families	ExPASy	
	blocks	bl		FHCRC	
	prodom	pd		PRABI	
	prints	pr		UMBER	
	pfam	pf		Sanger	
pmd		pmd	Protein mutants	DDBJ	
aaindex	aaindex1	aax1	Amino acid indices	Kyoto	
	aaindex2	aax2			
	aaindex3	aax3			
pdbstr		str	Protein sequences generated from PDB	Kyoto	
carbbank		ccsd	Carbohydrate structures	Teikyo U / U Georgia	
litdb		lit	PRF peptide/protein literature	PRF	
prosdoc		pdoc	Prosite literature	ExPASy	

3. Searchable Databases on the Web

Database name	Abbreviation	Content	Original site
insdc	genbank	Nonredundant database of International Nucleotide Sequence Database Collaboration	NCBI
	embl		EBI
	ddbj		DDBJ
ncbi-gene		NCBI Entrez Gene database	NCBI
unigene		NCBI UniGene (EST clusters) database	NCBI
ensembl		Eukaryotic genome annotation database	Ensembl
hgnc		Human gene nomenclature	HGNC
go		Gene Ontology	GO
ipi		EBI IPI (reference sequence) database	EBI
interpro		Protein domains and families	EBI
omim		Genetic diseases	NCBI
pubchem		NCBI PubChem (small molecules) database	NCBI
chebi		EBI ChEBI (small molecules) database	EBI
pdb-ccd		PDB Chemical Component Dictionary	PDB
lipidmaps		LIPID Metabolites And Pathways Strategy	LIPIDMAPS
knapsack		Secondary metabolite database	KNAPsACK
3dmet		3D structures of natural metabolites	3DMET
drugbank		Drug and target information resource	DrugBank

4. Link-only Databases on the Web

The databases in this category can be found in LinkDB

5. PubMed Database

Database name	Abbreviation	Content	Original site
pubmed		Biomedical literature	NCBI

DBGET

DBGET/LinkDB is integrated with other search tools, such as BLAST, FASTA and MOTIF to conduct further retrievals instantly.

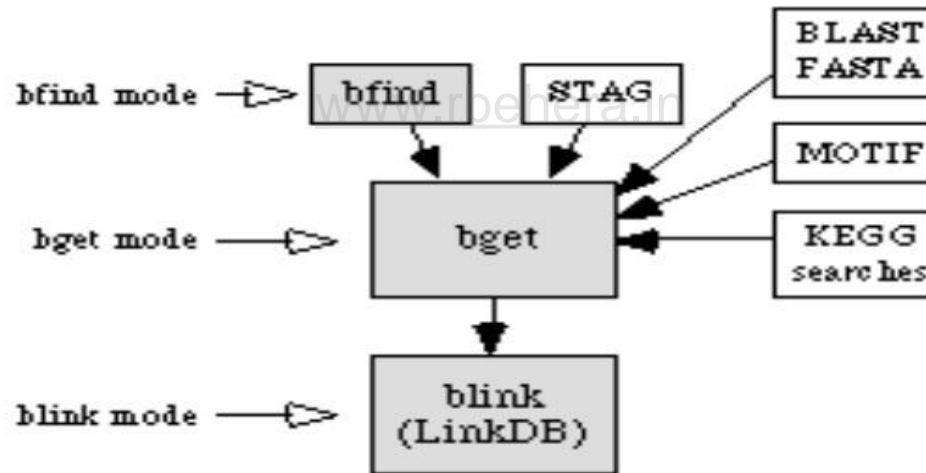
DBGET provides access to about 20 databases, which are queried one at a time. After querying one of these databases, DBGET presents links to associated information in addition to the list of results

A unique feature of DBGET is its connection with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database - a database of metabolic and regulatory pathways

DBGET

DBGET has simpler, but more limited search methods than either SRS or Entrez.

The architecture of the DBGET system:



DBGET

DBGET has three basic commands (or three basic modes in the Web version), **bfind**, **bget**, and **blink**, to search and extract database entries.

bget – performs the retrieval of database entries specified by the combination of **dbname:identifier**

www.rbehera.in

bfind – is used for searching entries by keywords

Notable feature of DBGET, different from other text search systems, is that **no keyword indexing is performed** when a database is installed or updated.

DBGET

Selected fields are extracted and stored in separate files for bfind searches.

- an advantage for rapid database updates, but sometimes a disadvantage for elaborate searching

To supplement bfind, the full text search **STAG** is provided

blink - the LinkDB search. Once entries of interest are found, it can be used to retrieve related entries in a given database or all databases in GenomeNet

Example

Let's consider an example to show how each system can be used to retrieve the SwissProt entry P04391, an ornithine carbamoyltransferase protein in *Escherichia coli*

In **Entrez**, enter the name P04391 in the protein database query form and view the entry and associated links and neighbours

Example - Entrez



Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy

Search Protein for P04391 Go Clear

Limits Preview/Index History Clipboard Details

www.rbehera.in

Display Summary Show: 20 Send to Text

1: [P04391](#)

[BLink](#), [Domain](#)

Ornithine carbamoyltransferase chain I (OTCase-1)

gi|129264|sp|P04391|OTC1_ECOLI[129264]

About Entrez

Entrez Protein

Help | FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve

sequences

Example - Entrez

Entrez	PubMed	Nucleotide	Protein	Genome	Structure	PMC	Taxonomy	Bo
Search	Protein	for		Go	Clear			
Display	default	Show:	20	Send to	File	Get Subsequence	Features	
Limits		Preview/Index		History		Clipboard		Details
<input type="checkbox"/> 1: P04391 . Ornithine carbamo...[gi:129264] BLink, Domains,								

LOCUS P04391 334 aa linear BCT 15-MAR-2004

DEFINITION Ornithine carbamoyltransferase chain I (OTCase-1).

ACCESSION P04391

VERSION P04391 GI:129264

DBSOURCE swissprot: locus OTC1_ECOLI, accession P04391;

class: standard.

created: Mar 20, 1987.

sequence updated: Oct 1, 1989.

annotation updated: Mar 15, 2004.

xrefs: gi: [145343](#), gi: [145344](#), gi: [40961](#), gi: [40962](#), gi: [1263172](#),

gi: [537095](#), gi: [2367366](#), gi: [1790703](#), gi: [66492](#), pdb accession

1AKM, pdb accession 2OTC, pdb accession 1DUV

xrefs (non-sequence databases): SWISS-2DPAGEP04391,

ECO2DBASEF039.0, EcoGeneEG10069, HAMAPMF_01109, InterProIPRO06130,

InterProIPRO02292, InterProIPRO06131, InterProIPRO06132,

PfamPFO0185, PfamPFO2729, PRINTSPRO0100, TIGRFAMsTIGR00658,

PROSITEPS00097

KEYWORDS Arginine biosynthesis; Transferase; 3D-structure; Complete proteome.

SOURCE Escherichia coli

ORGANISM [Escherichia coli](#)

Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;

Enterobacteriaceae; Escherichia.

REFERENCE 1 (residues 1 to 334)

AUTHORS Kuo,L.C., Miller,A.W., Lee,S. and Kozuma,C.

TITLE Site-directed mutagenesis of Escherichia coli ornithine transcarbamoylase: role of arginine-57 in substrate binding and

Cross-
references

Example - Entrez



My NCBI

[Sign In] [R]

All Databases

PubMed

Nucleotide

Protein

Genome

Structure

PMC

Taxonomy

Bo

Search Protein for P04391 [Save Search](#)

Display Related Sequences Show 20 Sort by Relevance Send to

All: 1

 1:

- Summary
- ASN.1
- FASTA
- XML
- GenPept
- GI List
- Graphics
- TinySeq XML
- INSDSeq XML
- LinkOut
- Related Sequences**
- Identical Proteins
- Protein (UniProtKB)
- Protein (RefSeq)
- Conserved Domain Links
- Concise Conserved Domain Links
- 3D Domain Links
- Gene Links
- Genome Links
- Genome Project Links

Structures: 1

ase chain I, AltName: Full=OTCase-1
[264]

[BLink](#), [Conserved Domains](#), [Links](#)

**Pre-computed
BLAST results**

[Write to the Help Desk](#)

[NCBI](#) | [NLM](#) | [NIH](#)

[Department of Health & Human Services](#)

[Privacy Statement](#) | [Freedom of Information Act](#) | [Disclaimer](#)

Recent Activity

[Turn Off](#)

[P04391](#) (0)

Conserved D

[P04391](#) (1)

Example - SRS

In **SRS**, first select the SwissProt database, then enter P04391 in the query form and, once the entry is displayed, search for links to other related databases.

Example - SRS

1. Select the **databanks** you want to search

2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

[Standard Query Form](#)

[Extended Query Form](#)

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:

[Browse Entries](#)

Tips

- ▶ bookmark this [link](#) to return to your project
- ▶ [Linking to SRS?](#)
- Please read our [Linking to SRS](#) guide for important information regarding linking to our SRS server.

BookMarkLets

About BookmarkLets

- [Protein Seq](#)
- [DNA/RNA Seq](#)

Literature, Bibliography and Reference Databases

- [MEDLINE](#) [Taxonomy](#) [OMIM](#) [OMIM Morbid Map](#)
 [Patent Abstracts](#) [Karyn's Genomes](#)

Literature, Bibliography and Reference Databases - subsections

- [MEDLINE \(Updates\)](#) [MEDLINE \(Main Release 2009\)](#) [MEDLINE \(Main Release 2008\)](#) [MED2PUB](#)

Gene Dictionaries and Ontologies

Nucleotide sequence databases

- [EMBL](#) [Patent DNA](#) [EMBL \(Contig\)](#)
 [EMBL \(Contigs expanded\)](#) [EMBL \(Annotated Cons\)](#) [EMBL \(Coding Sequences\)](#)
 [EMBL ID/Accession Mapping](#) [EMBL MGA](#) [IMG/HLA](#)
 [IMG/HLA](#) [IPD-KIR](#) [Genome Reviews](#)
 [GR Genes](#) [GR Transcripts](#) [RefSeq Genome](#)
 [LiveLists](#)

Nucleotide sequence databases - subsections

- [EMBL \(Updates\)](#) [EMBL \(Release\)](#) [EMBL \(Whole Genome Shotgun\)](#)
 [EMBL \(Whole Genome Shotgun release\)](#) [EMBL \(Whole Genome Shotgun updates\)](#) [EMBL \(Contig release\)](#)
 [EMBL \(Contig updates\)](#) [EMBL \(Contigs expanded release\)](#) [EMBL \(Contigs expanded updates\)](#)
 [EMBL \(Annotated Cons release\)](#) [EMBL \(Annotated Cons updates\)](#) [EMBL \(Release, Deleted\)](#)
 [EMBL \(Whole Genome Shotgun Masters\)](#) [ENA Project](#) [RefSeq Genome \(Release\)](#)
 [RefSeq Genome \(Updates\)](#)

Nucleotide related databases

UniProt Universal Protein Resource

- [UniProtKB](#) [UniProtKB](#) [UniRef90](#)
 [UniRef50](#) [UniParc](#)

Other protein sequence data

Protein function, structure and

Enzymes, reactions and met

Database of protein sequences produced collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

Example - SRS


[Quick Search](#)
[Library Page](#)
[Query Form](#)
[Tools](#)
[Results](#)
[Projects](#)
[Views](#)
[Reset](#)

Query "[swissprot-ALLTEXT:P04391*]" found 24 entries

Apply Options to:

- selected results only
 unselected results only

Result Options

Launch analysis tool:

BlastP [Launch](#)

Show tools relevant to these results: [Tools](#)

Link to related information: [Link](#)

Save results: [Save](#)

Display Options

View results using:

UniProt/Swiss-Prot	Accession	Description	G
<input type="checkbox"/> UniProt/Swiss-Prot:OTC1_ECOLI	P04391 /	Ornithine carbamoyltransferase chain I (EC 2.1.3.3) (OTCase-1).	AP B-
<input type="checkbox"/> UniProt/Swiss-Prot:OTC1_LACLA	Q9CHD1	Ornithine carbamoyltransferase 1 (EC 2.1.3.3) (OTCase 1).	AP LL
<input type="checkbox"/> UniProt/Swiss-Prot:OTC2_ECOLI	P06960	Ornithine carbamoyltransferase chain F (EC 2.1.3.3) (OTCase-2).	AP BC
<input type="checkbox"/> UniProt/Swiss-Prot:OTC2_LACLA	Q9CEY4	Ornithine carbamoyltransferase 2 (EC 2.1.3.3) (OTCase 2).	AP LL
<input type="checkbox"/> UniProt/Swiss-Prot:OTCC_BACII		Ornithine	AP

Example - SRS


[Quick Search](#)
[Library Page](#)
[Query Form](#)
[Tools](#)
[Results](#)
[Projects](#)
[View](#)
[Text Entry](#)
[SwissEntry](#)
[NiceProt](#)
[iProClass](#)
[Reset](#)

 Entry **1 of 24** from [Query 1](#)
[Next Entry](#)

Entry Information

Entry from:



Entry Options

Launch analysis tool:

[Launch](#)

Link to related information:

[Link](#)
[General](#)
[Description](#)
[References](#)
[Comments](#)
[Links](#)

General information

Entry name	OTC1_ECOLI
Accession number	<u>P04391</u>
Created	Rel. 04, 20-MAR-1987
Sequence update	Rel. 12, 1-OCT-1989
Annotation update	Rel. 43, 15-MAR-2004

Description and origin of the Protein

Description	Ornithine carbamoyltransferase chain I (EC 2.1.3.3) (OTC)
Gene name(s)	ARGI OR B4254.
Organism source	Escherichia coli.
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia
NCBI TaxID	562

References

- [1] Kuo,L.C., Miller,A.W., Lee,S., Kozuma,C.,
Site-directed mutagenesis of Escherichia coli ornithine 1 binding and catalysis.
 (1988) *Biochemistry* **27**:8823-8832

Example - SRS

Database cross-references

EMBL	J02842 ; AAA23483.1 ; -. X00210 ; CAA25037.1 ; -. U14003 ; AAA97150.1 ; -. AE000496 ; AAC77211.1 ; -.
PIR	A31314 ; OWECI.
PDB	1AKM ; 27-MAY-98. 2OTC ; 17-JUN-98. 1DUV ; 04-JUL-00.
SWISS-2DPAGE	P04391 ; COLI.
ECO2DBASE	F039.0 ; 6TH EDITION.
EcoGene	EG10069 ; argI.
HAMAP	MF_01109; -; 1.
InterPro	IPR006130 ; Asp/Orn_COtranf. IPR002292 ; Orn_carbtransf. IPR006131 ; OTCace_O. IPR006132 ; OTCace_P.
Pfam	PF00185 ; OTCace ; 1. PF02729 ; OTCace_N ; 1.
PRINTS	PR00100 ; AOTCASE.
TIGRFAMs	TIGR00658 ; orni_carb_tr; 1.
PROSITE	PS00097 ; CARBAMOYLTRANSFERASE;

Example - LinkDB

However, the fastest way of gathering the related information for this entry is to search LinkDB.

By simply entering swissprot:P04391, a list of all links to all the related databases is displayed

Example - LinkDB

Single Entry to Database

example) hsa:126

www.rbehera.in

From : (db:entry)

To : ▼

Example - LinkDB

LinkDB Search Result

Database: LinkDB

Database of Link Information
Release 04-01-29, Jan 04
Institute for Chemical Research, Kyoto University
112,039,792 entries

From: SWISS-PROT:P04391

To : All

218 hits from 143+ databases

1. [GenBank \(4\)](#)
2. [EMBL \(4\)](#)
3. [PIR \(1\)](#)
4. [PRF \(1\)](#)
5. [PDB \(3\)](#)
6. [PDBSTR \(15\)](#)
7. [PROSITE \(1\)](#)
8. [Blocks \(4\)](#)
9. [PRINTS \(3\)](#)
10. [Pfam \(2\)](#)
11. [PubMed \(14\)](#)
12. [PMD \(5\)](#)
13. [ENZYME \(1\)](#)
14. [PATHWAY \(2\)](#)
15. [HSA \(1\)](#)
16. [MMU \(1\)](#)
17. [RNO \(1\)](#)

www.rbehera.in

Example - LinkDB

```
113. BBU (1)
114. LIL (1)
115. BTH (1)
116. SYN (1)
117. SYW (1)
118. TEL (1)
119. GVI (1)
120. ANA (1)
121. PMA (1)
122. PMM (1)
123. PMT (1)
124. CTE (1)
125. DRA (1)
126. AAE (1)
127. TMA (1)
128. MJA (1)
129. MAC (1)
130. MMA (1)
131. MTH (1)
132. MKA (1)
133. AFU (1)
134. HAL (1)
135. TAC (1)
136. TVO (1)
137. PHO (1)
138. PAB (1)
139. PFU (1)
140. APE (1)
141. SSO (1)
142. STO (1)
143. PAI (1)
144. OTHERS (4)
145. All databases
```

BLAST (Basic Local Alignment Search Tool)

BLAST: A simplification of Smith-Waterman

The BLAST algorithm uses a word based heuristic similar to that of FASTA to approximate a simplification of the Smith-Waterman algorithm known as the maximal segment pairs algorithm. Maximal segment pairs alignments do not allow gaps and are specified by an equation that includes only the first and fourth terms of the Smith-Waterman equation presented above.

BLAST, or Basic Local Alignment Search Tool , is an alignment tool that uses a measure of local similarity to score sequence alignments in such a way as to identify regions of good local alignment. The basic BLAST algorithm can be implemented in DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences.

There are 5 BLAST options on the NCBI web site

BLASTP compares any amino acid query sequence against a protein sequence database.

BLASTN compares a nucleotide sequence against a nucleotide sequence database.

BLASTX takes a nucleotide query sequence and translates it in all reading frames for comparison against a protein sequence database.

TBLASTN compares a protein sequence against a nucleotide sequence database, translating the nucleotide database sequences in all possible reading frames.

TBLASTX compares translations of a nucleotide query sequence against translations of a nucleotide sequence database.

Steps followed by BLAST :

BLAST searches the database in 2 phases:

1. It looks for short subsequences that are likely to have significant matches,
2. then it tries to extend these matched regions (subsequences) on both sides in order to obtain maximum sequence similarity.

The BLAST algorithm works in the following steps:

1. the sequence is optionally filtered to remove low complexity regions that are not useful for producing meaningful sequence alignment.
2. A list of words of length 3 in the query protein sequence is made starting with position 1, 2, 3; then 2, 3, 4; etc; until the last 3 available positions in the sequence are reached.
3. Using the BLOSUM62 (Block Substitution Matrix) substitution scores, the query sequence words of step 2 are evaluated for an exact match with a word in any database sequence. The words are also evaluated for matches with any other combination of 3 amino acids, the object being to find the scores for aligning the query word with any other 3-letter word found in a database sequence.

There are a total of $20 \times 20 \times 20 = 8000$ possible match scores for this one sequence position.

4. A cutoff score called neighborhood word score threshold (T) is selected to reduce the number of possible matches to a particular word to the most significant ones. The list of possible matching words is thereby shortened from 8000 of all possible to the highest scoring number of approximately 50.
5. The above procedure is repeated for each 3-letter word in the query sequence. (For a sequence of length 250 amino acids, the total no. of words to search for is approximately $50 \times 250 = 12500$)
6. The remaining high scoring words that comprise possible matches to each 3-letter position in the query sequence are organized into an efficient search tree for comparing them rapidly to the database sequences.
7. Each database sequence is scanned for an exact match to one of the 50 words corresponding to the first query sequence position, for the words of the second position, and soon. If a match is found, this match is used to seed a possible ungapped alignment between the query and the database sequences.
8. (a) In the original BLAST method an attempt was made to extend an alignment from the matching words in each direction along the sequences, continuing for as long as the score continued to increase.

The extension process in each direction was stopped when the accumulated score stopped increasing and had just begun to fall a small amount below the best score found for shorter extensions.

At this point, a larger stretch of sequence (Called the HSP high scoring segment pair), which has a larger score than the original word, may have been found.

(b) In the later version of BLAST (BLAST2 or Gapped BLAST) a different and much more time efficient method is used.

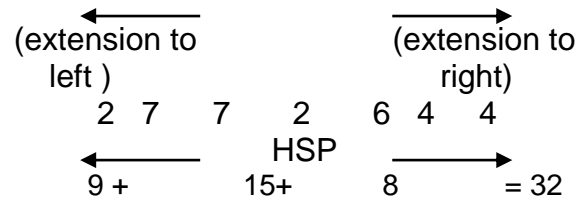
The method starts by making a list of high scoring matching words, as in Step 1-4 above, with the exception that a lower value of T, the word cutoff score, is also used.

This change results in a longer word list and matches to lower scoring words in the database sequences.

Example:

Query seq.	L	P	P	Q	G	L	L
Database seq.	M	P	P	E	G	L	L
3 letter word found initially			← word →				
BLOSUM62 scores			7	2	6		

Word scoer = 15



9. Determine whether each HSP score by one of the above methods is greater in value than a cutoff score S .

A suitable value for S is determined empirically by examining the range of scores found by comparing random sequences, and by choosing a value that is significantly greater.

The HSPs matched in the database are identified and listed.

10. BLAST next determines the statistical significance of each HSP score.
11. Sometimes, two or more HSP regions that can be made into a longer alignment will be found, thereby providing additional evidence that the query and database sequences are related. In such cases, a combined assessment of the significance will be made.
12. Smith Waterman local alignments are shown for the query seq. with each of the matched sequences in the database.
13. When the expected score for a given database sequence satisfies the user selectable threshold parameter E , the match is reported.

Statistical Estimators

E-value

The BLAST E-value is the number of expected hits of similar quality (score) that could be found just by chance.

E-value of 10 means that up to 10 hits can be expected to be found just by chance, given the same size of a random database.

E-value can be used as a first quality filter for the BLAST search result, to obtain only results equal to or better than the number given by the `-evalue` option. Blast results are sorted by E-value by default (best hit in first line).

```
blastn -query genes.ffn -subject genome.fna -evalue 1e-10
```

The smaller the E-value, the better the match.

```
-evalue 1e-50
```

small E-value: low number of hits, but of high quality

Blast hits with an E-value smaller than $1e^{-50}$ includes database matches of very high quality.

```
-evalue 0.01
```

Blast hits with E-value smaller than 0.01 can still be considered as good hit for homology matches.

```
-evalue 10 (default)
```

large E-value: many hits, partly of low quality

E-value smaller than 10 will include hits that cannot be considered as significant, but may give an idea of potential relations.

The E-value (expectation value) is a corrected bit-score adjusted to the sequence database size. The E-value therefore depends on the size of the used [sequence database](#). Since large databases increase the chance of false positive hits, the E-value corrects for the higher chance. It's a correction for multiple comparisons. This means that a sequence hit would get a better E-value when present in a smaller database.

$$E = m \times n / 2^{\text{bit-score}}$$

m - query sequence length

n - total database length (sum of all sequences)

Bit-score

The higher the bit-score, the better the sequence similarity

The bit-score is the requires size of a sequence database in which the current match could be found just by chance. The bit-score is a \log_2 scaled and normalized raw-score. Each increase by one doubles the required database size ($2^{\text{bit-score}}$).

Bit-score does not depend on database size. The bit-score gives the same value for hits in databases of different sizes and hence can be used for searching in an constantly increasing database.

Z-score

The Z-score is an old, yet commonly used statistical estimator for the validity of statistical results, including alignment scores. It is defined by the number of standard deviations that separate an observed score from the average random score. In other words, it is the difference between the observed score and the average random score, normalized by the standard deviation of the distribution. A higher Z-score means that the score can be trusted with a higher confidence level.

P-value

Once we have calculated the E-value, E , for a certain score, we can go one step further. The P-value is the probability of the observed score – the probability that a certain score occurred by chance. To find a formula for the P-value, let us define a random variable Y_E as the number of random records achieving an E-value of E or better. This random variable has a Poisson distribution with the parameter $\lambda=E$. The probability that no random events have a lower score than our score, i.e. that $Y_E = 0$, decreases exponentially with our score - s . Therefore, that probability that at least one random record achieved a better score than our E-value can be computed using the following simple formula: $P = 1 - e^{-E}$

Like the E-value, this value is dependent on the size of the database. A lower P-value means that the score has a higher confidence level. This estimator is not widely used for determining the validity of sequence alignment scores.

FASTA:

FASTA is a heuristic i.e an empirical method of computer programming like BALST. These methods are reliable in statistical sense, and usually provide a reliable alignment.

FASTA (and BLAST) use the word or k- tupe method. They align two sequences very quickly, by first searching for identical short stretches of sequences (called or word or k-tupes) and by then joining these words into an alignment by the dynamic programming method.

Rather than comparing individual residues in the 2 sequences, FASTA searches for matching sequence patterns or words, called k-tuples or k-tupes. These pattern comprises k consecutive matches in both sequences.

The program then attempts to build a local alignment based on these word matches.

For sequence fragments, FASTA is as good as Smith Waterman methods. For DNA searches, FASTA is theoretically better able than BLAST to find matches because a k-tupe smaller than the minimum obligatory one of 7 (default size 11) for BLAST may be used.

FASTA - algorithm details

In the initial stages of search for regions of similarity, FASTA uses an algorithmic method known as hashing.

Hashing: In this method, a lookup table showing the positions of each word of length k, or k- tupe is constructed for each sequence.

The relative positions of each word in the 2 sequences are then calculated by subtracting the position in the 1st sequence from that in the 2nd sequence.

Words that have he same off-set positions are in phase and reveal a region of alignment between the 2 sequences.

(Using hashing, the number of comparisons increases linearly in proportion to average sequence length.)

Example:

```

Position      1 2 3 4 5 6 7 8 9 10 11
Seq.1.       n c s p t a . . . . .
Position      1 2 3 4 5 6 7 8 9 10 11
Seq.2.       . . . . . a c s p r k
  
```

Amino Acid	Position in Seq1	Position in Seq2	Off set	
			Pos 1	Pos2
a	6	6	0	
c	2	7	5	
k	-	11	-	
n	1	-	-	
p	4	9	5	
r	-	10	-	
s	3	8	5	
t	5	-	-	

So, common offset for 3 amino acids c, s and p is 5.

A possible alignment is thus quickly found:

```

Protein 1      n c s p t a
                  | | |
Protein 2      a c s p r k
  
```


Common words or k-tuples, in 2 sequences are found by this method in a number of steps proportional to the sequence lengths).

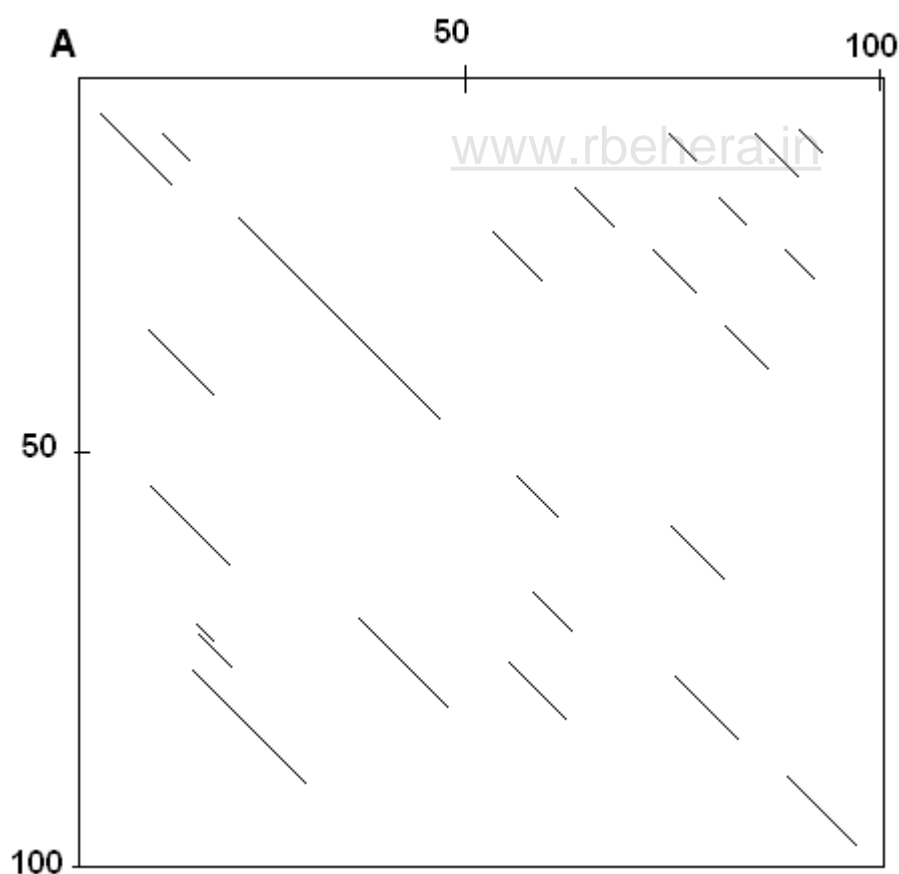
4 stages of FASTA algorithm:

- 1) find initial regions in search sequence
- 2) re-score to find top 10 initial regions (init1)
- 3) attempt to join initial regions together (initn)
- 4) optimize around initial region to find best fit (opt)

Step 1:

The 10 best matching regions in each sequence pair are located by a rapid screen. First, all sets of k consecutive matches are found by rapid method. (For DNA sequence k is usually 4-6 and for protein sequences 1-2). Second, those matches within a certain distance of each other (for proteins, 32 for k=1 and 16 for k=2) are joined along with the region between them into a longer matching region without gaps. The regions with the highest density of matches are identified. The calculation is very much like a dot matrix analysis, but is calculated in fewer steps.

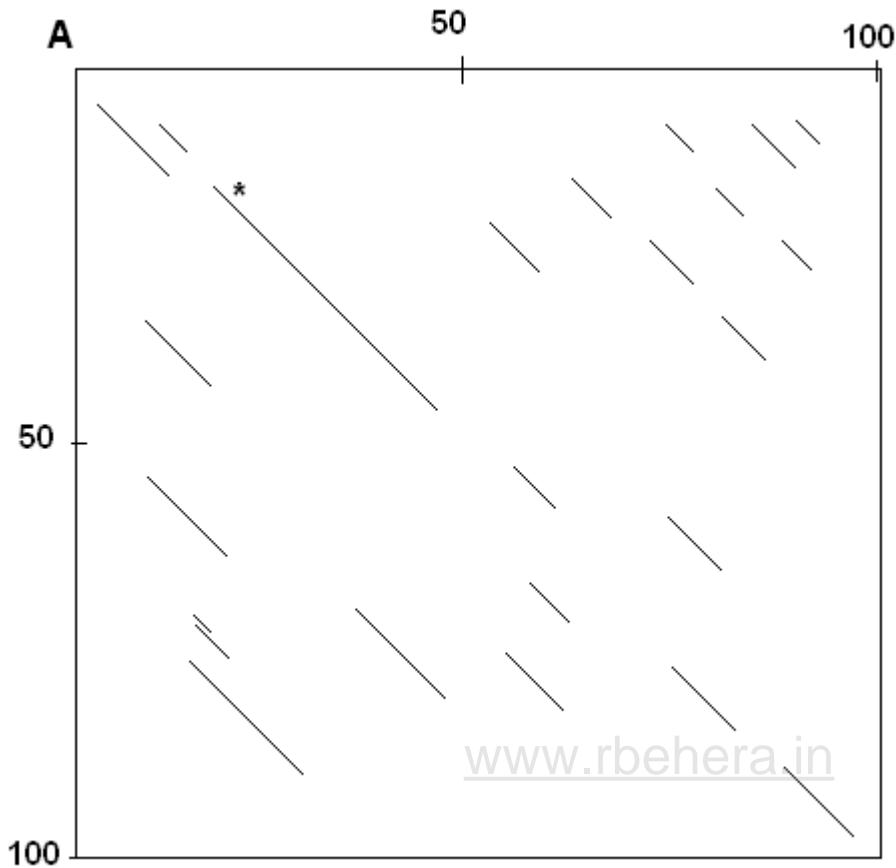
The diagonals shown in **A** represents the locations of these common patterns initially found in the 2 sequences.



Step 2:

The highest density regions of protein sequences identified in **A** are evaluated using an amino acid substitution matrix such as a PAM or BLOSUM scoring matrix. A corresponding matrix may also be used for DNA sequences.

The heights scoring regions, called the **best initial regions (INIT 1)** are identified and used to rank the matches for further analysis. The best scoring INIT 1 regions are shown marked by an asterisk in **B**.

**Step 3:**

Longer regions of identity of score (INIT N) are generated by joining initial regions with scores greater than a certain threshold. The INIT N score is the sum of the scores of the aligned individual regions less a constant gap penalty score for each gap introduced between the regions.

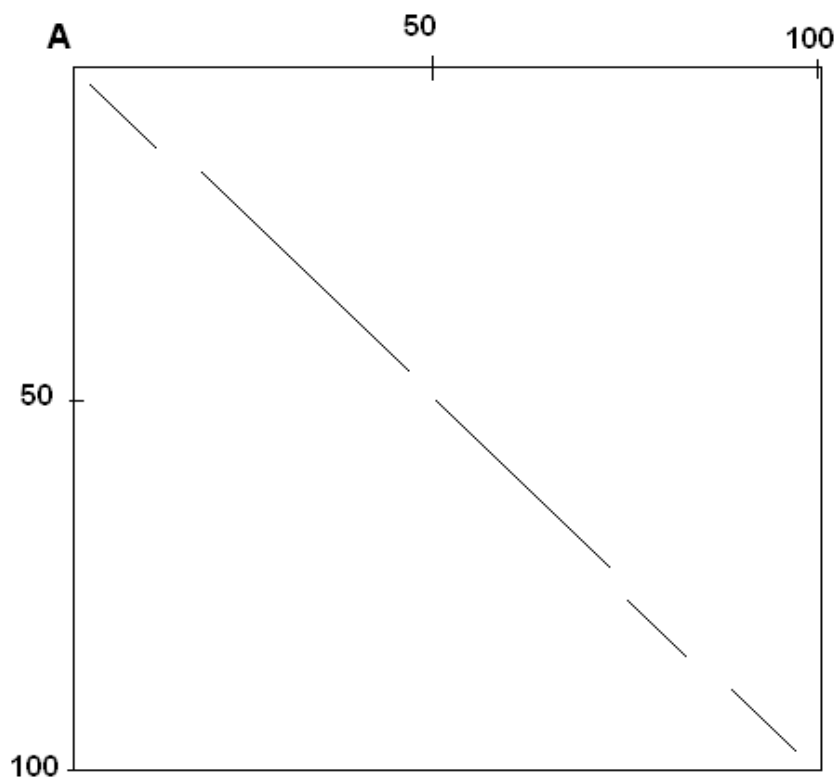
Step 4:

Later versions of FASTA included an optimization step.

When the INIT N score reaches a certain threshold value, the score of the region is recalculated to produce an OPT score by performing a full local alignment of the region using the Smith Waterman dynamic programming algorithm.

By improving the score, this step increases the sensitivity but decreases the selectivity of a search (Pearson 1990).

INIT N and OPT scores are used to rank database matches. Finally, an optimal local alignment between the input query sequence and the best scoring database sequences is performed based on the Smith Waterman dynamic programming algorithm.



Some tips for FASTA results:

1. **INIT 1 = INIT N = OPT => 100% homology over the matched stretch.**
2. **INIT N > INIT 1 => more than 1 matching regions was found in the database with poorly matching separating regions.**
3. **OPT > INIT N => the matching regions are greatly improved by adding gaps in 1 or both of the sequences.**

Z-Score: Evaluates the significance of the "OPT Score" by generating a score distribution from the alignment of many random pairs of sequences having the same length as the 2 compared sequences. From this distribution, the number of standard deviations from the mean for the alignment score of interest is calculated as the Z-score.

Better the match, higher the Z-Score.

GENOME ANNOTATION

The first step in genome annotation involves the integration of features revealed by the DNA and protein sequences into a systematic view of the organism's molecular machinery. Annotation can be divided into two processes; i) direct analysis of DNA sequences to locate coding regions and repeated elements, and ii) prediction of function and structure of the proteins encoded in the genome. In all organisms, coding regions are differentiated from neighbouring non-coding regions by specific features. Detecting these features is essential to transforming sequence data into a fully annotated genome. Once a gene is identified or predicted, the next step is to assign a putative function, identify possible homologs in other organism's gene, and to postulate its role in the biology of the organism. By comparing the genetic complement and genome organization of related organisms, novel insights may be realized regarding their evolutionary relationships.

Estimating complete annotation of a genome includes information regarding gene location and organization, transcripts and products of those genes, as well as regulation and control of expression, translation and degradation. This process included boundaries between coding and non-coding sequence, identification of DNA features associated with gene structures, and translation of protein coding genes into protein sequence. The following subsections describe two of the major challenges in genome annotation; repeat prediction and gene prediction.

Sequence Patterns: Pattern => conserved sequence motifs

Nucleotide and protein sequences contain **patterns or motifs** that have been preserved through evolution because they are important to the structure or function of the molecule. In proteins, these conserved sequences may be involved in the binding of the protein to its substrate or to another protein, may comprise the active site of an enzyme or may determine the three dimensional structure of the protein. Nucleotide sequences outside of coding regions in general tend to be less conserved among organisms, except where they are important for function, that is, where they are involved in the regulation of gene expression. Discovery of motifs in protein and nucleotide sequences can lead to determination of function and to elucidation of evolutionary relationships among sequences.

Pattern description notations:

Several notations for describing motifs are in use but most of them are variants of standard notations for regular expressions and use these conventions:

- there is an alphabet of single characters, each denoting a specific amino acid or a set of amino acids;
- a string of characters drawn from the alphabet denotes a sequence of the corresponding amino acids;
- any string of characters drawn from the alphabet enclosed in square brackets matches any one of the corresponding amino acids; e.g. [abc] matches any of the amino acids represented by a or b or c.

The fundamental idea behind all these notations is the matching principle, which assigns a meaning to a sequence of elements of the pattern notation:

a sequence of elements of the pattern notation matches a sequence of amino acids if and only if the latter sequence can be partitioned into subsequences in such a way that each pattern element matches the corresponding subsequence in turn.

Thus the pattern [AB] [CDE] F matches the six amino acid sequences corresponding to ACF, ADF, AEF, BCF, BDF, and BEF.

Different pattern description notations have other ways of forming pattern elements. One of these notations is the PROSITE notation, described in the following subsection.

PROSITE pattern notation

The PROSITE notation uses the IUPAC one-letter codes and conforms to the above description with the exception that a concatenation symbol, '-', is used between pattern elements, but it is often dropped between letters of the pattern alphabet.

PROSITE allows the following pattern elements in addition to those described previously:

- The lower case letter 'x' can be used as a pattern element to denote any amino acid.
- A string of characters drawn from the alphabet and enclosed in braces (curly brackets) denotes any amino acid

except for those in the string. For example, {ST} denotes any amino acid other than S or T.

- If a pattern is restricted to the N-terminal of a sequence, the pattern is prefixed with '<'.
- If a pattern is restricted to the C-terminal of a sequence, the pattern is suffixed with '>'.
- The character '>' can also occur inside a terminating square bracket pattern, so that S[T>] matches both "ST" and "S>".
- If e is a pattern element, and m and n are two decimal integers with $m \leq n$, then:
 - e(m) is equivalent to the repetition of e exactly m times;
 - e(m,n) is equivalent to the repetition of e exactly k times for any integer k satisfying: $m \leq k \leq n$.

Some examples:

- x(3) is equivalent to x-x-x.
- x(2,4) matches any sequence that matches x-x or x-x-x or x-x-x-x.

The signature of the C2H2-type *zinc finger* domain is:

- C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

Sequence Motifs

- Protein sequence motifs are **signatures of protein families** and can often be used as tools for the prediction of protein function. The generalization and modification of already known motifs are becoming major trends in the literature, even though new motifs are still being discovered at an approximately linear rate. The emphasis of motif analysis appears to be shifting from metabolic enzymes, in which motifs are associated with catalytic functions and thus often readily recognizable, to structural and regulatory proteins, which contain more divergent motifs. The consideration of structural
- Information increasingly contributes to the identification of motifs and their sensitivity. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a particular organism. A systematically derived motif database is therefore feasible, allowing the classification of the majority of the newly appearing protein sequences into known families.
- In genetics, a **sequence motif** is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance.
- An example is the N-glycosylation site motif:
- *Asn, followed by anything but Pro, followed by either Ser or Thr, followed by anything but Pro*
- Where the three-letter abbreviations are the conventional designations for amino acids.

Overview of motif

When a sequence motif appears in the exon of a gene, it may encode the "structural motif" of a protein; that is a stereotypical element of the overall structure of the protein. Nevertheless, motifs need not be associated with a distinctive secondary structure. "Noncoding" sequences are not translated into proteins, and nucleic acids with such motifs need not deviate from the typical shape (e.g. the "B-form" DNA double helix).

Outside of gene exons, there exist **regulatory sequence motifs** and motifs within the "junk," such as satellite DNA. Some of these are believed to affect the shape of nucleic acids (see for example RNA self-splicing), but this is only sometimes the case. For example, many DNA binding proteins that have affinities for specific motifs only bind DNA in its double-helical form. They are able to recognize motifs through contact with the double helix's major or minor groove.

Short coding motifs, which appear to lack secondary structure, include those that label proteins for delivery to particular parts of a cell, or mark them for phosphorylation.

Within a sequence or database of sequences, researchers search and find motifs using computer-based techniques of sequence analysis, such as BLAST. Such techniques belong to the discipline of bioinformatics.

Repeat Prediction

The genomes of all organisms, particularly eukaryotic organisms, contain repetitive elements of varying lengths that can occupy a significant fraction of the total DNA content. For example, the human genome consists of more than 50% repeated sequences of various types. Repeats play a vital role in a number of regulatory functions and are responsible for instability of genomes. Many tandem repeats like the tri-nucleotide motifs, (e.g. CCG; CAG; AAG; CTG; GCG etc.) are associated with diseases such as fragile X, myotonic dystrophy, Huntington's, ataxia and others. Thus, identification of repeat elements is an important task in annotating a genome. Genomic repeat elements can be divided in two categories; i) tandem repeats which are usually confined to specific chromosomal regions, and ii) interspersed repeats mainly represented by inactive (pseudo-genes) copies of historically or contemporarily active transposable elements. Tandem repeats are grouped into three major subclasses; satellites, mini-satellites and microsatellites (Strachan and Read 1999). Satellite repeats are composed of very long tandem arrays of short units usually present at centromeres. Mini-satellites consists of tandem repeats of short units with lengths of about 7 to 64 bp located near telomeres, while microsatellite repeats are highly repetitive sequences consisting of 1 to 6 bp segments that are repeated up to 5 times the unit length as tandem arrays dispersed throughout all the chromosomes. Similarly, interspersed repeats can also be sub grouped into 5 types: SINEs (Short Interspersed Nuclear Elements) of 80-300 bp long units, LINEs (Long Interspersed Nuclear Elements) that are 6000-8000 bp long, LTRs (Long Terminal Repeats) that are 300 – 1000 bp long, and DNA transposons of variable lengths with two short inverted repeats flanking the element. Several repeat-finding algorithms have been developed to detect repeats, and these programs can be divided into two groups based on the type of repetitive DNA they identify; i) Tandem repeat finders and ii) interspersed repeat finders.

Gene Prediction

Correct predictions of gene location and structure are major challenges in the post genomic era, particularly for eukaryotic genomes. In the last decade a large number of computer programs have been developed for scanning genomic sequences to locate DNA segments that encode proteins. Prokaryotic genes may be predicted with considerable accuracy if one knows the codon usage pattern of the organism in question. A simple, long ORF (open reading frame) in a prokaryotic DNA sequence can be predicted as protein coding. The problem with gene prediction in prokaryotes lies in identifying the promoter and regulatory region.

Unlike prokaryotic genes, the eukaryotic genes are neither continuous nor contiguous. They are separated by long stretches of intergenic DNA and their coding sequences are interrupted by non-coding introns. Coding sequences occupy just a small fraction of a typical higher eukaryotic genome. Additionally, some eukaryotic genes are alternatively spliced – i.e. they have more than one possible exon assembly. The arrangement of genes in genomes is also prone to exceptions. Some genes are nested (overlapping) within each other (Dunham et al. 1999). The presence of pseudogenes further complicates the identification of protein coding regions. Regulatory sequences usually located upstream of coding sequences can sometimes be found downstream and within the introns of genes. In prokaryotic systems, genes are simple in structure where introns do not split protein-coding regions and they are comparatively easy to identify. However, finding genes in eukaryotic genomic sequences is far from being a trivial problem. Unlike prokaryotic genomes, the coding regions in eukaryotes represent only a small proportion of the eukaryotic genome and are mostly found to lie in non-repetitive regions of the genome.

Algorithms and software tools for gene identification

Some of tools perform gene prediction ab initio, relying only on the statistical parameters in the DNA sequence for gene identification.

Homology-based methods rely primarily on identifying homologous sequences in other genomes and/or in public databases using BLAST or Smith-Waterman algorithms.

Many of the commonly used methods combine these two approaches.

Ab initio Gene Prediction Programs (Possibly with Homology Integration)

Program	Organism	Algorithm*	Website	Homology
GeneID	Vertebrates , plants	DP	http://www1.imim.es/geneid.html	
FGENESH	Human, mouse, Drosophila, rice	HMM	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	
GeneParser	Vertebrates	NN	http://beagle.colorado.edu/~eesnyder/GeneParser.html	EST
Genie	Drosophila, human, other	GHMM	http://www.fruitfly.org/seq_tools/genie.html	protein
GenLang	Vertebrates , Drosophila, dicots	Grammar rule	http://www.cbil.upenn.edu/genlang/genlang_home.html	
GENSCAN	Vertebrates , Arabidopsis , maize	GHMM	http://genes.mit.edu/GENSCAN.html	
GlimmerM	Small eukaryotes, Arabidopsis , rice	IMM	http://www.tigr.org/tdb/glimmer/glmr_form.html	
GRAIL	Human, mouse, Arabidopsis , Drosophila	NN, DP	http://compbio.ornl.gov/Grail-bin/EmptyGrailForm	EST, cDNA
HMMgene	Vertebrates , <i>C. elegans</i>	CHMM	http://www.cbs.dtu.dk/services/HMMgene/	
AUGUSTUS	Human, Arabidopsis	IMM, WWAM	http://augustus.gobics.de/	
MZEF	Human, mouse, Arabidopsis , Fission yeast	Quadratic discriminant analysis	http://rulai.cshl.org/tools/genefinder/	

*DP, dynamic programming; NN, neural network; MM, Markov model; HMM, Hidden Markov model; CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM.