

Databases in Bioinformatics and Systems Biology

Carsten O. Daub
Omics Science Center
RIKEN, Japan
May 2008

Overview

- Introduction
- Nucleotide sequences
- Protein sequences
- Protein families and interactions
- Non coding RNA
- TFBS, splicing
- Genome browsers

Introduction

- Bioinformatics and Systems Biology
- Internet resources develop
 - Evolution of databases
 - Constant change
- Databases are more: Web resources
- Web resources as “superstructures” of databases
- What are the standard databases?

Nucleotide Sequences – DNA and RNA

- International Nucleotide Sequence Database Collaboration
- Genbank
 - National Institute of Health, US
 - <http://www.ncbi.nlm.nih.gov/Genbank/>
- EMBL Nucleotide Sequence Database (EMBL-Bank)
 - Several institutes in Europe, e.g. Heidelberg, Hinxton
 - <http://www.ebi.ac.uk/embl/>
- DDBJ (DNA Databank of Japan)
 - National Institute of Genetics, Japan
 - <http://www.ddbj.nig.ac.jp/>

Nucleotide Sequences – DNA and RNA

- Genbank, EMBL, DDBJ
- Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis

What goes into these Databases?

- DNA and RNA sequence
 - Submitted by scientists directly
- Annotation to sequences
 - Details in tomorrow's lecture *Genome Assembly and Annotation*
 - What is “Annotation”?
- There will be more comments about these resources later on in the lecture!

Protein Sequences

- UniProt
 - <http://www.uniprot.org>
- Protein Informartion Resource - International Protein Sequence Database (PIR-PSD)
 - <http://pir.georgetown.edu/>

Protein Sequences

- UniProt is the standard protein sequence repository
 - New URL: <http://beta.uniprot.org/>
- Derived from
 - SwissProt
 - Manually annotated and reviewed
 - TrEMBL
 - Automatically annotated and NOT reviewed
 - Translations from EMBL nucleotide sequences

Protein Structure – 3D

- Protein Data Bank (PDB)
 - <http://www.wwpdb.org>
- SCOP
 - <http://scop.mrc-lmb.cam.ac.uk/scop/>

Protein Families

- What do you need to characterize protein families?

Protein Families

- Pfam
 - <http://pfam.sanger.ac.uk/>
 - Hidden Markov Models for protein sequence multiple alignments
 - Pfam A: manually curated models
 - Pfam B: automatically generated models

Protein Families

- Prosite
- <http://www.expasy.ch/prosite/>
- Started with regular expression for families
- Later extended to profiles

Protein Families

- ProDom
 - <http://prodom.prabi.fr/prodom.html>
 - a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases

InterPro

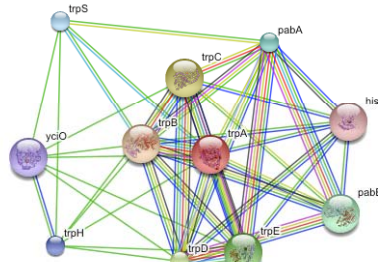
- <http://www.ebi.ac.uk/interpro/>
- EBI's approach to **integrate** many protein databases

Member database information

Signature Database	Version	Signatures*	Integrated Signatures
PANTHER	6.1	30127	2070
Pfam	21.0	8956	8956
PIRSF	2.70	2742	2691
PRINTS	38.0	1900	1898
ProDom	2005.1	3538	1056
PROSITE patterns	20.0	1331	1330
PROSITE profiles	20.0	675	654
SMART	5.1	724	721
TIGRFAMs	7.0	3423	3393
GENE3D	3.0.0	2147	1012
SUPERFAMILY	1.69	1538	1072

Protein Interaction

- String – EMBL
- Systems Biology style
- <http://string.embl.de/>



Your Input:

- trpA Tryptophan synthase alpha chain (EC 4.2.1.20) (268 aa) (*Escherichia coli K12*)

Predicted Functional Partners:

Gene	Description	Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Database	Textmining	Homology	Score
trpB	Tryptophan synthase beta chain (EC 4.2.1.20) (397 aa)	●	●	●	●	●	●	●	0.999
trpC	Tryptophan biosynthesis protein trpCF [Includes- Indole-3-glycerol phosphate sy	●	●	●	●	●	●	●	0.999
trpD	Anthraniolate synthase component II (EC 4.1.3.27) [Includes- Glutamine amidotr	●	●	●	●	●	●	●	0.999
trpE	Anthraniolate synthase component I (EC 4.1.3.27) (Anthraniolate synthase compo	●	●	●	●	●	●	●	0.999
pabB	Para-aminobenzoate synthase component 1 (EC 6.3.5.8) (Para- aminobenzoate	●	●	●	●	●	●	●	0.914
pabA	Para-aminobenzoate synthase glutamine amidotransferase component II (EC 6.	●	●	●	●	●	●	●	0.903
trpS	Tryptophanyl-tRNA synthetase (EC 6.1.1.2) (Tryptophan--tRNA ligase) (TrpRS) (●	●	●	●	●	●	●	0.864
trpH	Protein trpH {UniProtKB/Swiss-Prot-P77766} (293 aa)	●	●	●	●	●	●	●	0.845
yciO	Protein yciO {UniProtKB/Swiss-Prot-P0A9F4} (206 aa)	●	●	●	●	●	●	●	0.841
hisC	Histidinol-phosphate aminotransferase (EC 2.6.1.9) (Imidazole acetyl- phosphat	●	●	●	●	●	●	●	0.830

Views:



Non Coding RNA

- Why is non coding RNA important?
- What would you want to have in databases?

Non Coding RNA

- Rfam
 - <http://www.sanger.ac.uk/Software/Rfam/>
- RNADB
 - <http://research.imb.uq.edu.au/rnadb/>
- NONCODE
 - <http://www.noncode.org/>

Non Coding RNA – specific DBs

- miRNA DBs
- PicTar
 - <http://pictar.bio.nyu.edu/>
- miRBase
 - <http://microrna.sanger.ac.uk/>
- microRNA.org
 - <http://www.microrna.org/microrna/>

Gene Expression

- Gene Expression Omnibus (GEO) at NCBI
 - <http://www.ncbi.nlm.nih.gov/geo/>
- Tissue specific expression of genes
- Download expression datasets

Transcription Factor Binding Site

- FANTOM3 database
 - By RIKEN
 - Based on Cap Analysis of Gene Expression (CAGE)
 - <http://fantom.gsc.riken.jp/>
- DBTSS
 - DB for transcriptional starting sites
 - Based on cDNA
 - <http://dbtss.hgc.jp/>

Splicing

- Alternative splicing database project
 - <http://www.ebi.ac.uk/asd/>
- Alternative transcript diversity database
 - <http://www.ebi.ac.uk/astd>

Genome browsers

- Visualize
- UCSC browser
 - <http://genome.ucsc.edu/>
- ENSEMBL
 - <http://www.ensembl.org>
 - EMBL, EBI, Sanger joint project
- More in the Genome Browser lecture

Multipurpose Portals

<http://www.ncbi.nlm.nih.gov/sites/gquery>

The screenshot displays the NCBI Entrez search engine interface. At the top, the NCBI logo and the Entrez logo are visible, along with the text "Entrez, The Life Sciences Search Engine". Below this, there are navigation tabs for "PubMed", "All Databases", "Human Genome", "GenBank", "Map Viewer", and "BLAST". The search bar contains the query "piwi" and a "GO" button. Below the search bar, a message states: "Result counts displayed in gray indicate one or more terms not found".

The search results are organized into a grid of database-specific results. Each result includes a count, a database icon, the database name, a brief description, and a "GO" button. The results are as follows:

Count	Database	Description
180	PubMed	PubMed: biomedical literature citations and abstracts
303	PubMed Central	PubMed Central: free, full text journal articles
none	Site Search	Site Search: NCBI web and FTP sites
3	Books	Books: online books
8	OMIM	OMIM: online Mendelian Inheritance in Man
none	OMIA	OMIA: online Mendelian Inheritance in Animals
172653	CoreNucleotide	Core subset of nucleotide sequence records
269	EST	EST: Expressed Sequence Tag records
none	GSS	GSS: Genome Survey Sequence records
1392	Protein	Protein: sequence database
22	Genome	Genome: whole genome sequences
5	Structure	Structure: three-dimensional macromolecular structures
none	Taxonomy	Taxonomy: organisms in GenBank
234	SNP	SNP: single nucleotide polymorphism
426	Gene	Gene: gene-centered information
18	HomoloGene	HomoloGene: eukaryotic homology groups
none	GENSAT	GENSAT: gene expression atlas of mouse central
none	dbGaP	dbGaP: genotype and phenotype
76	UniGene	UniGene: gene-oriented clusters of transcript sequences
13	CDD	CDD: conserved protein domain database
32	3D Domains	3D Domains: domains from Entrez Structure
10	UniSTS	UniSTS: markers and mapping data
3	PopSet	PopSet: population study data sets
872	GEO Profiles	GEO Profiles: expression and molecular abundance profiles
11	GEO DataSets	GEO DataSets: experimental sets of GEO data
none	Cancer Chromosomes	Cancer Chromosomes: cytogenetic databases
none	PubChem BioAssay	PubChem BioAssay: bioactivity screens of chemical substances
none	PubChem Compound	PubChem Compound: unique small molecule chemical

<http://www.ebi.ac.uk/>

The screenshot displays the EMBL-EBI website interface. At the top, the EMBL-EBI logo is followed by the 'EB-eye Search' section, which includes a dropdown menu for 'All Databases', a search input field with the placeholder 'Enter Text Here', a 'Go' button, a 'Reset' button with a help icon, and a 'Give us feedback' button. Below the search bar is a horizontal navigation menu with links for 'Databases', 'Tools', 'EBI Groups', 'Training', 'Industry', 'About Us', and 'Help'. On the right side of this menu, there are links for 'Site Index' and social media icons for RSS and Facebook. The main content area features a section titled 'Data Resources & Tools' with a list of sub-links:

- EMBL-BANK
- UniProt
- ArrayExpress
- Ensembl
- InterPro
- PDB-EBI
- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules
- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes
- Literature
- Taxonomy
- Ontologies
- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads