

# Information Sources for Genomics

**Konstantinos Mavrommatis**  
**Genome Biology Program**



## *Databases*

- Databases used for the analysis of biological molecules.
- Databases contain information organized in a way that allows users/researchers to retrieve and exploit it.
- Goals:
  - ☒ Store information.
  - ☒ Organize data in a way that is easier to be described and studied.
  - ☒ Predict the functional role of an element.
  - ☒ Understand the way things work.

- Sequence databases

- ▣ Primary (contain "raw" data)

- Nucleotide
    - Protein

- ▣ Secondary (processed information)

- Genes
    - Proteins

- Classification databases

- ▣ Sequence classification

- ▣ Function classification

- ▣ Other methods

- Other specialized databases

## EMBL/GenBank/DBJ

(<http://www.ncbi.nlm.nih.gov>, <http://www.ebi.ac.uk/emb>)

- Archive containing all sequences from:

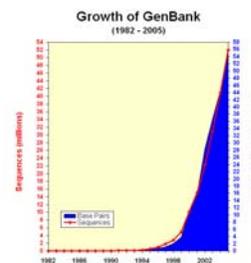
- ▣ genome projects
  - ▣ sequencing centers
  - ▣ individual scientists
  - ▣ patent offices

- The sequences are exchanged between the three centers on a daily basis.

- Database is doubling every 10 months.

- Sequences from >140,000 different species.

- 1400 new species added every month.



- Contain coding sequences derived from the translation of nucleotide sequences

- GenBank

- Valid translations (CDS) from nt GenBank entries.



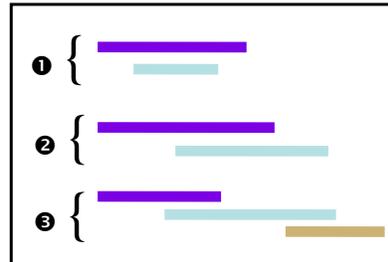
- UniProtKB/TrEMBL (1996)

- Automatic CDS translations from EMBL.
    - TrEMBL Release 37.6 (04-Dec-2007) contains 5072048 entries.



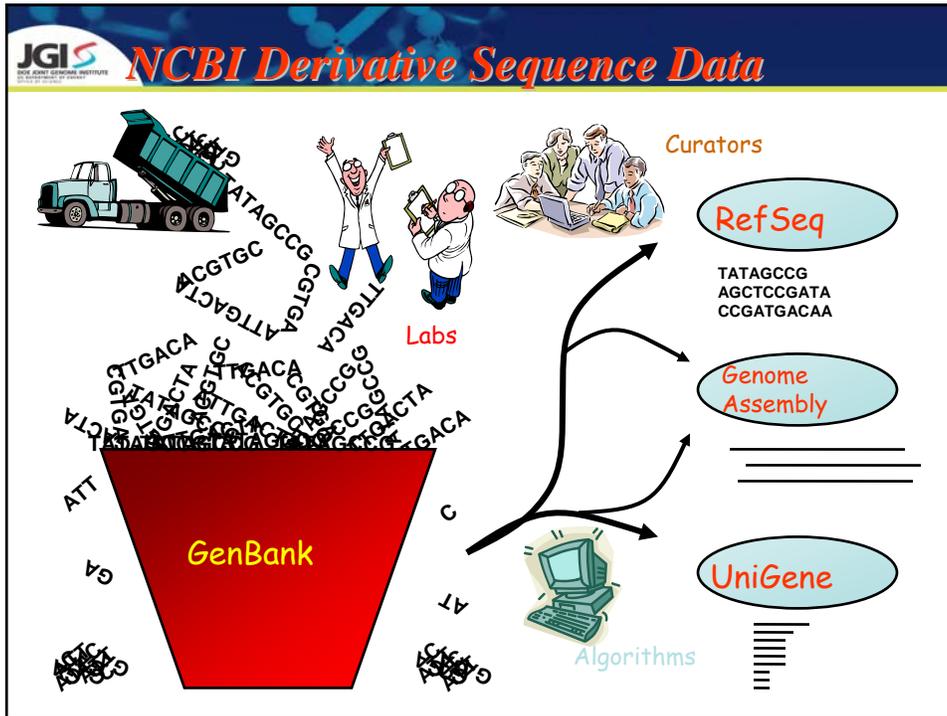
- There are a lot of errors in the primary sequence databases:
  - In the sequences themselves:
    - Sequencing errors.
    - Cloning vectors inserted.
  - For the annotations, the free submission of entries involves:
    - Inaccuracies, omissions, and even mistakes.
    - Inconsistencies between some fields.

- Another major problem is redundancy.
- A lot of entries are partially or entirely duplicated:
  - 20% of vertebrate sequences in GenBank.



Partial and complete sequence duplications

- Sequence databases
  - ▣ Primary (contain "raw" data)
    - Nucleotide
    - Protein
  - ▣ Secondary (processed information)
    - Genes
    - Proteins
- Classification databases
  - ▣ Sequence classification
  - ▣ Function classification
  - ▣ Other methods
- Other specialized databases



**JGIS** *RefSeq*

- Curated transcripts and proteins.
  - reviewed by NCBI staff.
- Model transcripts and proteins.
  - generated by computer algorithms.
- Assembled Genomic Regions (contigs).
- Chromosome records.

**Key Characteristics of GenBank versus RefSeq**

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

- **SWISS-PROT (1986)** (<http://ca.expasy.org/spro>)

- ☒ Best annotated, least redundant
- ☒ Contains entries from >10,000 species
- ☒ Cross-references with >60 other databanks

- **PIR (Protein Information Resource)** (<http://pir.georgetown.edu>)

- ☒ More automated annotation
- ☒ Collaborations with MIPS and JIPID

- **Uniprot (2003)**

- ☒ UniProt (Universal Protein Resource) is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

### Key Characteristics of UniProt versus GenBank and RefSeq

UniProt	GenBank and RefSeq
Produced by SIB, EBI & Georgetown U.	Produced by INSDC and NCBI
Protein data only	Protein and nucleotide data
Curated in Swiss-Prot, not in TrEMBL	Curated in RefSeq, not in GenBank

- **Sequence databases**

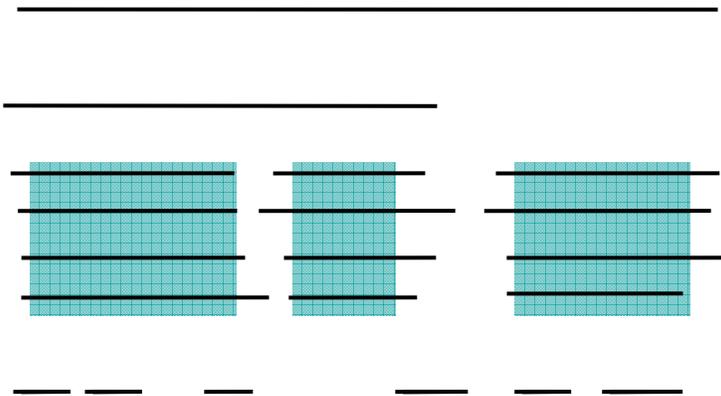
- ☒ Primary (contain "raw" data)
  - Nucleotide
  - Protein
- ☒ Secondary (processed information)
  - Genes
  - Proteins

- **Classification databases**

- ☒ Sequence classification
- ☒ Function classification
- ☒ Other methods

- **Other specialized databases**

- Groups (families/clusters) of proteins based on...
  - ▣ Overall sequence similarity.
  - ▣ Local sequence similarity.
  - ▣ Presence / absence of specific features.
  - ▣ Structural similarity.
  - ▣ ...
- These groups contain proteins with similar properties.
  - ▣ Specific function, enzymatic activity.
  - ▣ Broad function.
  - ▣ Evolutionary relationship.
  - ▣ ...



**JGIS** *Clusters of orthologous groups (COGs)*

- COGs were delineated by comparing protein sequences encoded in 43 complete genomes representing 30 major phylogenetic lineages.



```

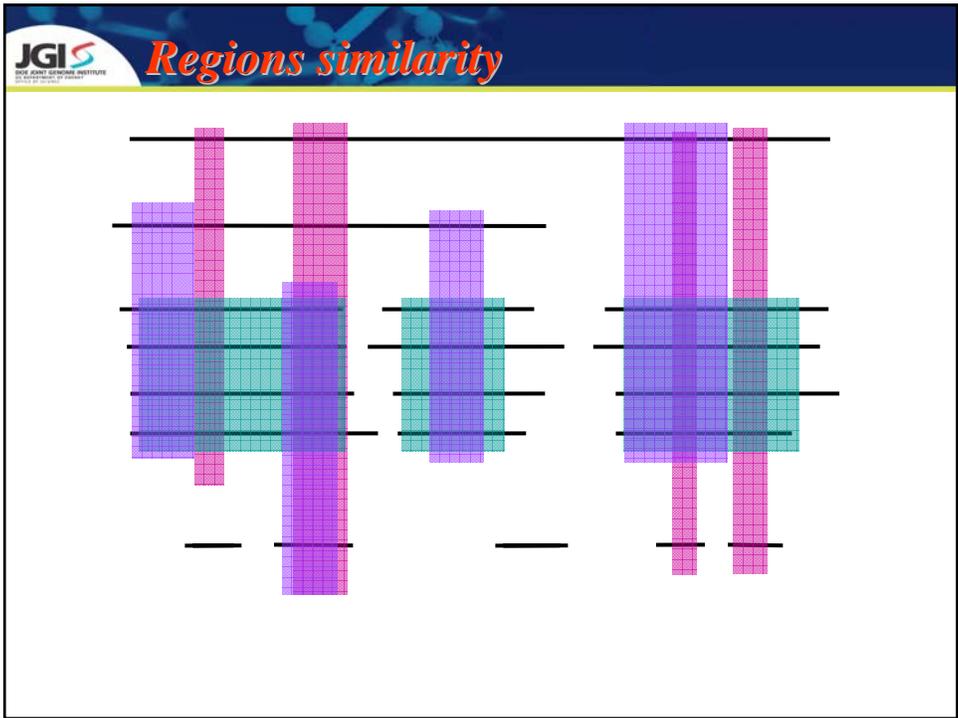
- 22 --m-k--vd-lb-efgh--j---- BacC [J] COG3813 16S RNA G1207 methylase RsmC
- 64 -----qvdrlbcefgshanujit- BacA [J] COG1187 16S rRNA uridine-516 pseudouridylylate synthase and related pseudouridylylate synthases
- 21 a-mpkz-qvd-b-ef-----j----- Lig7 [J] COG1514 2'-5' RNA ligase
3 59 a-mpkz-qvd-bcefgshanujit- H14d [J] COG0621 2-methylthioadenine synthetase
- 99 -c-pk-y-vdr-lb-efgh-nuix--y BimA [J] COG1670 Acetyltransferases, including N-acetylases of ribosomal proteins
4 43 a-mpkzqvdrlbcefgshanujitv Al4d [J] COG0013 Alanine-tRNA synthetase
- 5 -----dlr-----j----- Aph [J] COG3231 Aminoglycoside phosphotransferase
- 9 a-mpkz----- BceT [J] COG2511 Archaeal Glu-tRNAArgin amidotransferase subunit E (contains GAD domain)
5 45 a-mpkzqvdrlbcefgshanujitv ArgP [J] COG0018 Arginyl-tRNA synthetase
66 a-m--sqvdr-lbc-f--nujkitv ArgA [J] COG0154 Asp-tRNAAsn/Glu-tRNAArgin amidotransferase A subunit and related amidases
32 a-m--sqvdr-lbc-f--nujkitv ArgB [J] COG0054 Asp-tRNAAsn/Glu-tRNAArgin amidotransferase B subunit (PET112 homolog)
30 a-m--z-qvdr-lbc-f--nujkitv ArgC [J] COG0721 Asp-tRNAAsn/Glu-tRNAArgin amidotransferase C subunit
2 34 -----qvdrlbcefgshanujitv ArgJ [J] COG0173 Aspartyl-tRNA synthetase
5 40 a-mpkzy--d-lbce-ghs-----tv ArgS [J] COG0017 Aspartyl/asparaginyl-tRNA synthetases
4 43 a-mpkzqvdrlbcefgshanujitv Cys3 [J] COG0115 Cysteinyl-tRNA synthetase
2 17 -----qvdr-lb-efgh-n----- DeA [J] COG1490 D-Tyr-tRNATyr deacylase
2 12 a-mpkzy----- DYS1 [J] COG1899 Deoxyhypusine synthase
4 49 a-mpkzqvdrlbcefgshanujitv RspA [J] COG0030 Dimethyladenosine transferase (rRNA methylation)
2 10 a-mpkzy----- DPH5 [J] COG1798 Diphthamide biosynthesis methyltransferase DPH5
4 11 a-mpkzy----- DPH2 [J] COG1746 Diphthamide synthase subunit DPH2
2 10 a-mpkzy----- EIF6 [J] COG1976 Eukaryotic translation initiation factor 6 (EIF6)
3 10 a-mpkzy----- MOP1 [J] COG1889 Fibrillar-like rRNA methylase
2 78 a-mpkzqvdrlbcefgshanujitv TufB [JE] COG0050 GTPases - translation elongation factors
6 71 a-mpkzqvdrlbcefgshanujitv Glu3 [J] COG0008 Glutamyl- and glutamyl-tRNA synthetases
- 25 -----qv- lbcefgshanujit- G19 [J] COG0752 Glycyl-tRNA synthetase, alpha subunit
2 19 a-mpkzy--dr-----tv G19 [J] COG0751 Glycyl-tRNA synthetase, beta subunit
2 49 a-mpkzqvdrlbcefgshanujitv His3 [J] COG0423 Glycyl-tRNA synthetase, class II
- 3 a-m----- His5 [J] COG0124 Histidyl-tRNA synthetase
5 44 a-mpkzqvdrlbcefgshanujitv His6 [J] COG0491 Homolog of the eukaryotic argonaute protein, possible role in translation
4 45 a-mpkzqvdrlbcefgshanujitv Leu3 [J] COG0060 Isoleucyl-tRNA synthetase
- - - - - Leu5 [J] COG0495 Leucyl-tRNA synthetase

```

Class of Each COG is shown in the left margin. The color key indicates the phylogenetic lineage: Bacteria (B), Eukarya (E), Archaea (A), and other groups (C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF, BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV, BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG, CH, CI, CJ, CK, CL, CM, CN, CO, CP, CQ, CR, CS, CT, CU, CV, CW, CX, CY, CZ, DA, DB, DC, DD, DE, DF, DG, DH, DI, DJ, DK, DL, DM, DN, DO, DP, DQ, DR, DS, DT, DU, DV, DW, DX, DY, DZ, EA, EB, EC, ED, EE, EF, EG, EH, EI, EJ, EK, EL, EM, EN, EO, EP, EQ, ER, ES, ET, EU, EV, EW, EX, EY, EZ, FA, FB, FC, FD, FE, FF, FG, FH, FI, FJ, FK, FL, FM, FN, FO, FP, FQ, FR, FS, FT, FU, FV, FW, FX, FY, FZ, GA, GB, GC, GD, GE, GF, GG, GH, GI, GJ, GK, GL, GM, GN, GO, GP, GQ, GR, GS, GT, GU, GV, GW, GX, GY, GZ, HA, HB, HC, HD, HE, HF, HG, HH, HI, HJ, HK, HL, HM, HN, HO, HP, HQ, HR, HS, HT, HU, HV, HW, HX, HY, HZ, IA, IB, IC, ID, IE, IF, IG, IH, II, IJ, IK, IL, IM, IN, IO, IP, IQ, IR, IS, IT, IU, IV, IW, IX, IY, IZ, JA, JB, JC, JD, JE, JF, JG, JH, JI, JJ, JK, JL, JM, JN, JO, JP, JQ, JR, JS, JT, JU, JV, JW, JX, JY, JZ, KA, KB, KC, KD, KE, KF, KG, KH, KI, KJ, KK, KL, KM, KN, KO, KP, KQ, KR, KS, KT, KU, KV, KW, KX, KY, KZ, LA, LB, LC, LD, LE, LF, LG, LH, LI, LJ, LK, LL, LM, LN, LO, LP, LQ, LR, LS, LT, LU, LV, LW, LX, LY, LZ, MA, MB, MC, MD, ME, MF, MG, MH, MI, MJ, MK, ML, MM, MN, MO, MP, MQ, MR, MS, MT, MU, MV, MW, MX, MY, MZ, NA, NB, NC, ND, NE, NF, NG, NH, NI, NJ, NK, NL, NM, NN, NO, NP, NQ, NR, NS, NT, NU, NV, NW, NX, NY, NZ, OA, OB, OC, OD, OE, OF, OG, OH, OI, OJ, OK, OL, OM, ON, OO, OP, OQ, OR, OS, OT, OU, OV, OW, OX, OY, OZ, PA, PB, PC, PD, PE, PF, PG, PH, PI, PJ, PK, PL, PM, PN, PO, PP, PQ, PR, PS, PT, PU, PV, PW, PX, PY, PZ, QA, QB, QC, QD, QE, QF, QG, QH, QI, QJ, QK, QL, QM, QN, QO, QP, QQ, QR, QS, QT, QU, QV, QW, QX, QY, QZ, RA, RB, RC, RD, RE, RF, RG, RH, RI, RJ, RK, RL, RM, RN, RO, RP, RQ, RR, RS, RT, RU, RV, RW, RX, RY, RZ, SA, SB, SC, SD, SE, SF, SG, SH, SI, SJ, SK, SL, SM, SN, SO, SP, SQ, SR, SS, ST, SU, SV, SW, SX, SY, SZ, TA, TB, TC, TD, TE, TF, TG, TH, TI, TJ, TK, TL, TM, TN, TO, TP, TQ, TR, TS, TT, TU, TV, TW, TX, TY, TZ, UA, UB, UC, UD, UE, UF, UG, UH, UI, UJ, UK, UL, UM, UN, UO, UP, UQ, UR, US, UT, UU, UV, UW, UX, UY, UZ, VA, VB, VC, VD, VE, VF, VG, VH, VI, VJ, VK, VL, VM, VN, VO, VP, VQ, VR, VS, VT, VU, VV, VW, VX, VY, VZ, WA, WB, WC, WD, WE, WF, WG, WH, WI, WJ, WK, WL, WM, WN, WO, WP, WQ, WR, WS, WT, WU, WV, WW, WX, WY, WZ, XA, XB, XC, XD, XE, XF, XG, XH, XI, XJ, XK, XL, XM, XN, XO, XP, XQ, XR, XS, XT, XU, XV, XW, XX, XY, XZ, YA, YB, YC, YD, YE, YF, YG, YH, YI, YJ, YK, YL, YM, YN, YO, YP, YQ, YR, YS, YT, YU, YV, YW, YX, YY, YZ, ZA, ZB, ZC, ZD, ZE, ZF, ZG, ZH, ZI, ZJ, ZK, ZL, ZM, ZN, ZO, ZP, ZQ, ZR, ZS, ZT, ZU, ZV, ZW, ZX, ZY, ZZ.

**JGIS** *Profiles & Pfam*

- A method for classifying proteins into groups exploits **region similarities**, which contain valuable information (domains/profiles).
- These domains/profiles can be used to detect distant relationships, where only few residues are conserved.



**JGI S**  
JOINT GENOME INSTITUTE  
AN INTERNATIONAL COLLABORATION

# Pfam

HOME | SEARCH | BROWSE | FTP | HELP

Pfam 22.0 (July 2007, 9318 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). [More...](#)

**USING PFAM** YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

- SEQUENCE SEARCH** Analyze your protein sequence for Pfam matches
- VIEW A PFAM FAMILY** View Pfam family annotation and alignments
- VIEW A CLAN** See groups of related families
- VIEW A SEQUENCE** Look at the domain organisation of a UniProt sequence
- VIEW A STRUCTURE** Find the domains on a PDB structure
- KEYWORD SEARCH** Query Pfam by keywords

Or view the [help](#) pages for more information

**A**

**B**

[Back to Pfam](#)

**Citing Pfam**

If you find Pfam useful, please consider [citing](#) the reference that describes this work:

[Pfam: clans, web tools and services](#) © P.D. Finn, J. Mistry, B. Schuster-Bödlér, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Phanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman

**Nucleic Acids Research** (2006) Database Issue  
34:D247-D251

**Mirrors**

The following are official Pfam [mirror](#) sites:

- [WTSI, UK](#)
- [SBC, Sweden](#)
- [JRC, USA](#)
- [INBA, France](#)
- × [CCRI, South Korea](#)

<http://pfam.sanger.ac.uk>

- Full length alignments.
- Domain alignments.
- Equivalogs: families of proteins with specific function.
- Superfamilies: families of homologous genes.

### TIGRFAMs Ordered by Role Category ③

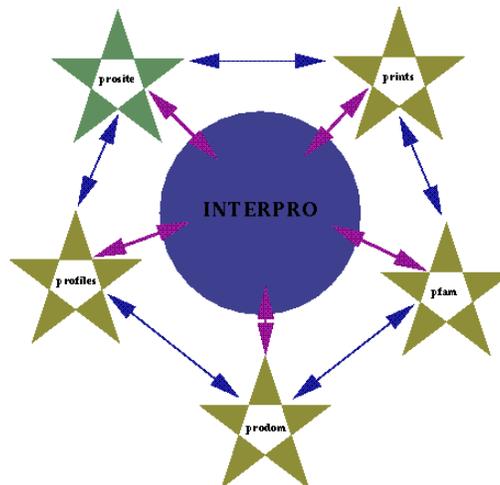
- ☑ Amino acid biosynthesis
- ☑ Biosynthesis of cofactors, prosthetic groups, and carriers
- ☑ Cell envelope
- ☑ Cellular processes
- ☑ Central intermediary metabolism
- ☑ DNA metabolism
  - ☑ DNA replication, recombination, and repair
  - ☑ Restriction/modification
  - ☑ Degradation of DNA
  - ☑ Chromosome-associated proteins
  - ☑ Other
- ☑ Energy metabolism
- ☑ Fatty acid and phospholipid metabolism
- ☑ Hypothetical proteins
- ☑ Mobile and extrachromosomal element functions
- ☑ Protein fate
- ☑ Protein synthesis
- ☑ Purines, pyrimidines, nucleosides, and nucleotides
- ☑ Regulatory functions
- ☑ Signal transduction
- ☑ Transcription
- ☑ Transport and binding proteins
- ☑ Unclassified
- ☑ Unknown function

<http://www.tigr.org/TIGRFAMs/>

- To simplify sequence analysis, the family databases are being integrated to create a unified annotation resource – **InterPro**
  - Release 16.1 (October 07) contains **14768** entries
  - Central annotation resource, with pointers to its satellite dbs

#### Member database information

Signature Database	Version	All Signatures	Integrated Signatures
PANTHER	6.1	26313	2072
Pfam	21.0	8957	8957
PIRSF	2.70	2877	1034
PRINTS	38.0	1900	1898
ProDom	2005.1	3538	1054
PROSITE patterns	20.0	1331	1331
PROSITE profiles	20.0	675	654
SMART	5.1	724	721
TIGRFAMs	6.0	2949	2933
GENE3D	3.0.0	2147	837
SUPERFAMILY	1.69	1538	692



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

**JGIS**  
 JOY KUMAR GOSWAMI INSTITUTE  
 UNIVERSITY OF CALICUT

\* It is up to the user to decide if the annotation is correct \*

**JGIS**  
 JOY KUMAR GOSWAMI INSTITUTE  
 UNIVERSITY OF CALICUT

## ENZYME

<a href="#">1. - . - .</a>	Oxidoreductases.		
<a href="#">1. 1. - .</a>	Acting on the CH-OH group of donors.		
<a href="#">1. 1. 1. -</a>	With NAD(+) or NADP(+) as acceptor.	<a href="#">4. - . - .</a>	Lyases.
<a href="#">1. 1. 1. 1. -</a>	With a cytochrome as acceptor.	<a href="#">4. 1. - .</a>	Carbon-carbon lyases.
<a href="#">1. 1. 1. 2. -</a>	With oxygen as acceptor.	<a href="#">4. 1. 1. -</a>	Carboxy-lyases.
<a href="#">1. 1. 1. 3. -</a>	With a disulfide as acceptor.	<a href="#">4. 1. 2. -</a>	Aldehyde-lyases.
<a href="#">1. 1. 1. 4. -</a>	With a quinone or similar compound as acceptor	<a href="#">4. 1. 3. -</a>	Oxo-acid-lyases.
<a href="#">1. 1. 1. 5. -</a>	With other acceptors.	<a href="#">4. 1. 99. -</a>	Other carbon-carbon lyases.
<a href="#">1. 1. 99. -</a>	With other acceptors.		
<a href="#">1. 2. - . -</a>	Acting on the aldehyde or oxo group of donors.	<a href="#">5. - . - .</a>	Isomerases.
		<a href="#">5. 1. - .</a>	Racemases and epimerases.
<a href="#">2. - . - .</a>	Transferases.	<a href="#">5. 1. 1. -</a>	Acting on amino acids and derivatives.
<a href="#">2. 1. - . -</a>	Transferring one-carbon groups.	<a href="#">5. 1. 2. -</a>	Acting on hydroxy acids and derivatives.
<a href="#">2. 1. 1. -</a>	Methyltransferases.	<a href="#">5. 1. 3. -</a>	Acting on carbohydrates and derivatives.
<a href="#">2. 1. 1. 1. -</a>	Hydroxymethyl-, formyl- and related	<a href="#">5. 1. 99. -</a>	Acting on other compounds.
<a href="#">2. 1. 1. 2. -</a>	Carboxyl- and carbamoyltransferases.	<a href="#">5. 2. - . -</a>	Cis-trans-isomerases.
<a href="#">2. 1. 1. 3. -</a>	Carboxyl- and carbamoyltransferases.		
<a href="#">2. 1. 1. 4. -</a>	Aminotransferases.		
<a href="#">3. - . - .</a>	Hydrolases.	<a href="#">6. - . - .</a>	Ligases.
<a href="#">3. 1. - . -</a>	Acting on ester bonds.	<a href="#">6. 1. - .</a>	Forming carbon-oxygen bonds.
<a href="#">3. 1. 1. -</a>	Carboxylic ester hydrolases.	<a href="#">6. 1. 1. -</a>	Ligases forming aminoacyl-tRNA
<a href="#">3. 1. 1. 1. -</a>	Thiolester hydrolases.	<a href="#">6. 2. - . -</a>	Forming carbon-sulfur bonds.
<a href="#">3. 1. 1. 2. -</a>	Phosphoric monoester hydrolases.	<a href="#">6. 2. 1. -</a>	Acid--thiol ligases.
<a href="#">3. 1. 1. 3. -</a>	Phosphoric monoester hydrolases.	<a href="#">6. 3. - . -</a>	Forming carbon-nitrogen bonds.
<a href="#">3. 1. 1. 4. -</a>	Phosphoric diester hydrolases.		



**JGIS** *Overview*

- Sequence databases
  - ☒ Primary (contain "raw" data)
    - Nucleotide
    - Protein
  - ☒ Secondary (processed information)
    - Genes
    - Proteins
- Classification databases
  - ☒ Sequence classification
  - ☒ Function classification
  - ☒ Other methods
- Other specialized databases

**JGIS** *Sequencing projects*

- GOLD
  - ☒ Information for ongoing and finished genomic projects.
  - ☒ Information about the phylogeny and phenotype of genomes.

Contact: Genomesonline	Last Update: January 03, 2008	Location www.genomesonline.org
Archaeal Tree	<b>GOLD Tables</b>	Bacterial Tree
Right-click to save all data: DOWNLOAD		

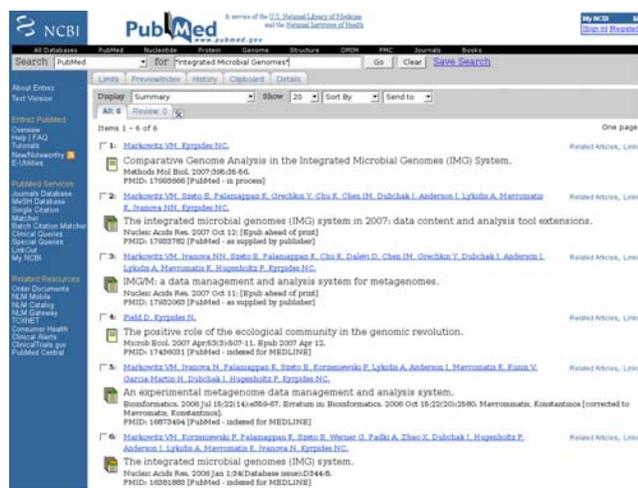
Comparative Analysis of 3639 Genomes at the: <b>HGS system</b>	Comparative Analysis of 24 Metagenomes at the: <b>HGS system</b>
---	---

<http://www.genomesonline.org>

Bacterial Ongoing Genome Projects: 950												
GOLDSTAMP	Organism	ID	Information	Size	GC content	DATA-search	Institution	Funding	Genome Database	Status	Contact	Availability
G0072R	Acaryochloris marina NBRC11047	☒	CYANOBACTERIA Taxonomy Biology Entry	8000 Kb			Arizona State Univ IGem	NSF	IGem	Incomplete	Blankenship RE	Public
G01399	Acaryochloris sp. CCMEE 5410	☒	CYANOBACTERIA Taxonomy Information	4200 Kb			Univ of Oregon J. Craig Venter Institute	Moses Foundation	ICVI	Incomplete	Wood M	Public
G00091B	Acidiphilium cryotum JF-5	☒	PROTEOBACTERIA-ALPHA Taxonomy Entry	2460 Kb			Joint Genome Institute	DOE		Incomplete	Richardson P	Public
G0129B	Acidiphilium macDermotii ATCC 392	☒	CHLOROFLEXI Taxonomy				NITE		NITE	Incomplete		Public

- There is a large number of databases devoted to specific organisms.
- For some model organisms there are often concurrent systems.
- These databases are associated to sequencing or mapping projects.

### PubMed



The screenshot shows the PubMed search results page for the query "Integrated Microbial Genomes". The search results are displayed in a list format, showing the title, authors, journal, and PMID for each entry. The first result is "Comparative Genome Analysis in the Integrated Microbial Genomes (IMG) System" by Markowitz V.M., Sato R., Palanisappan P., Orzechka Y. Chu X., Chen H., Dubchak I., Anderson I., Lykidis A., Marmorato F., Ivanova N.I., Kyrpides N.C., published in Methods Mol Biol. 2007;98:69-84. PMID: 17605066 [PubMed - in process].

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

**JGIS** *Other specialized databases*

- Signal
  - TRAP
  - BRIS
  - Exp
  - DIP
  - BIN
  - BioC
- Biochem
  - KLO
  - BRE
  - LIG
- Gene o
  - STR
- Gene expression
  - GXD (Mouse Gene Expression Database)
  - The Stanford Microarray Database
- Mapping
  - GDB (Genome Data Base)
  - EMG (Encyclopedia of Mouse Genome)
  - MGD (Mouse Genome Database)
  - INE (Integrated Rice Genome Explorer)
- Protein quantification
  - SWISS-2DPAGE
  - PDD (Protein Disease Database)
  - Sub2D (B. subtilis 2D Protein Index)

**JGIS** *Databanks interconnection*

Not all databases are updated regularly.  
Changes of annotation in one database are not reflected in others.

## Concluding remarks

- We have main archives (Genbank), and curated databases (Refseq, SwissProt), and protein classification database (COG, Pfam).
- This is the tip of the iceberg.
- They help *predict* the function, or the network of functions.
- Systems that integrate the information from several databases, visualize and allow handling of data in an intuitive way *are required*

Thank you for your attention

- Profiles are scoring tables derived from domain alignments
  - they define which residues are allowed at given positions
  - which positions are conserved & which degenerate
  - which positions can tolerate insertions

