Functional annotation

Mavrommatis Konstantinos PhD
Genome Biology Program
DOE-JGI

Gene annotation

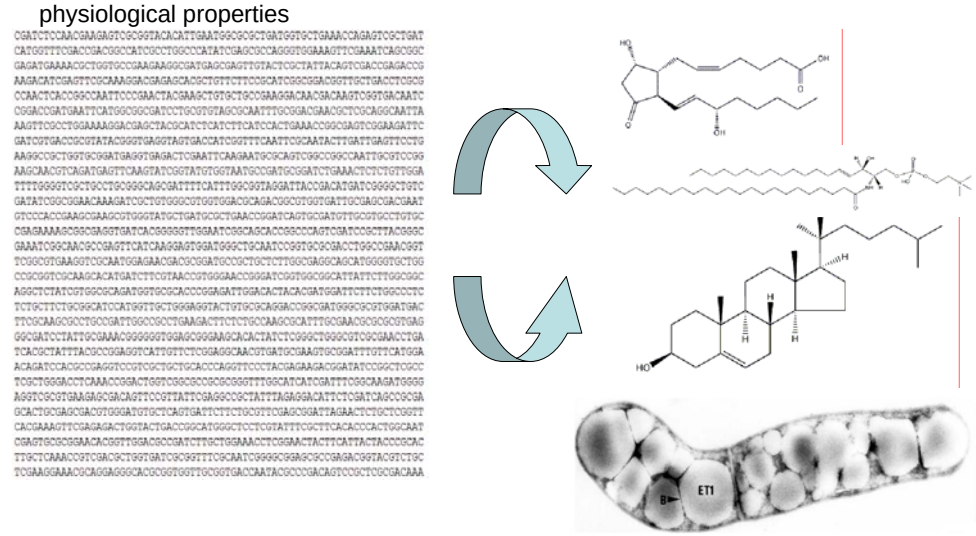
- Find the genes.
  - *ab initio* gene prediction (GeneMark, Glimmer etc).
  - evidence based gene prediction (BLAST based).
- Understand (predict) what these genes are doing (functional annotation).
  - Computationally
    - sequence comparison
      - BLAST, protein family (Pfam, COG, InterPro)
    - chromosomal context
    - fusion
    - other source of information
  - Experimentally
    - enzyme assays
    - mutants

Functional annotation

- **Why do we need functional annotation?**
  - **Metabolic reconstruction**. What can an organism do and how?
  - **Evolutionaly analysis**. How does an organism compare to the rest?

Metabolic reconstruction

- Predict the biochemistry and physiology of an organism based on its genome sequence
- Explain known biochemical and physiological properties



Metabolic reconstruction

- Knowing the function does not mean that we can predict what an organism looks like and what it can do. A butterfly and a caterpillar share the same genome, but they are not alike.
- Knowledge of other parameters, such as positional and temporal regulation of genes, is needed.

Before we start …

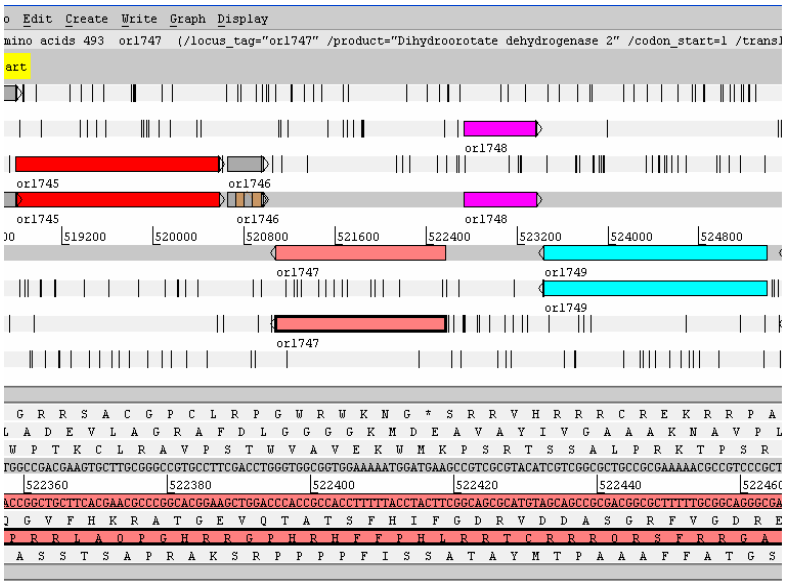
- Correct gene prediction
  - find missing genes
  - delete wrongly predicted genes
  - improve gene coordinates

Find missing genes

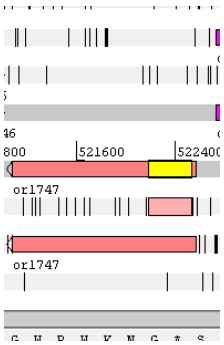
… and wrong ORFs

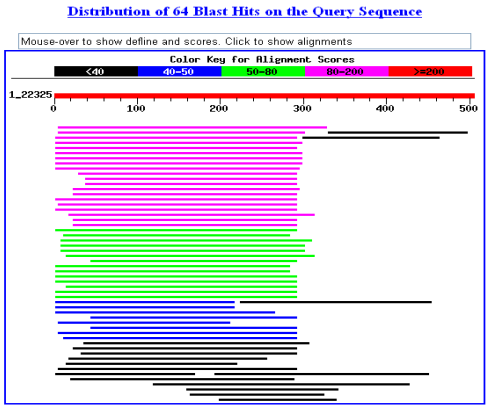
or2654 is unique and hides a real CDS which is acyl carrier protein

Everything looks perfect in this area of *Nitrobacter winogradskyi*, but …



… hides a real ORF





Function prediction

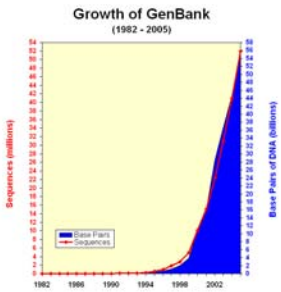
- Similarity searches (BLAST).
- Domain identification(Pfam).
- Small sequence identification(PROSITE).
- Localization (membrane, intracellular, extracellular).
- Other sequence features.
- Similarity to known structures.

Primary nucleotide/protein databases

EMBL/GenBank/DDBJ

(http://www.ncbi.nlm.nih.gov/,http://www.ebi.ac.uk/embl)

- 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.
- Database name nt / nr



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

Secondary 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.

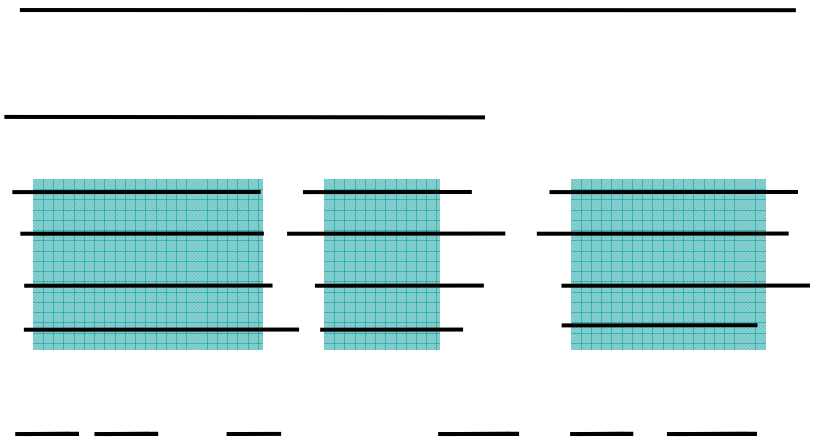
Key Characteristics of UniProt *versus* GenBank and RefSeq

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

Classification 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.
  - ...

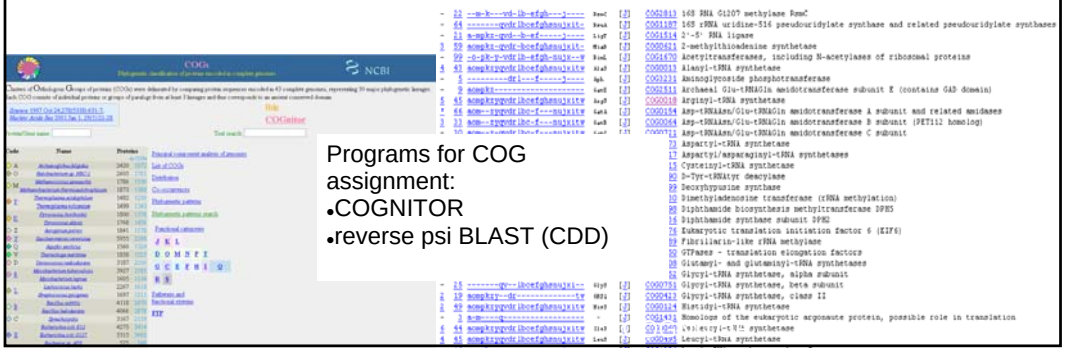
Overall sequence similarity



Similarity spans the full length of the protein

Clusters of orthologous groups (COGs)

- COGs were delineated by comparing protein sequences encoded in 43 complete genomes representing 30 major phylogenetic lineages.
  - Each Cluster has representatives of at least 3 lineages
- A function (specific or broad) has been assigned to each COG. Functions grouped to functional categories.
- Lineage information of organisms harboring genes assigned to COGs allow to identify non homologous replacement events.

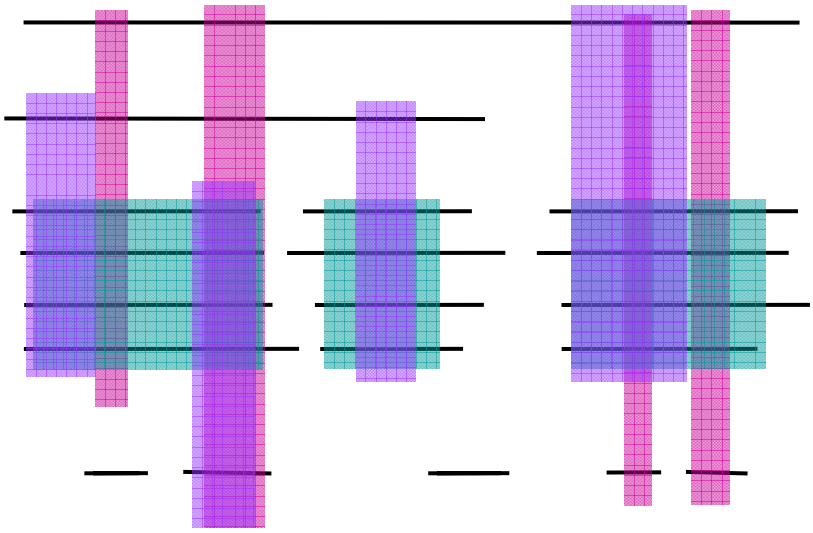


Programs for COG assignment:
- COGNITOR
- reverse psi BLAST (CDD)

Profiles & Pfam

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

Regions similarity



Small regions are common between several genes

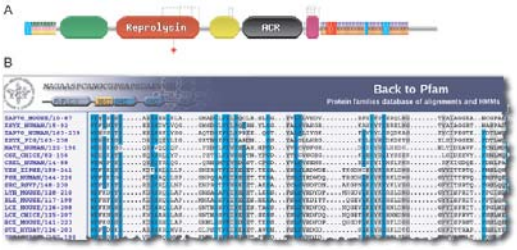
Pfam

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



Citing Pfam

If you find Pfam useful, please consider citing the reference that describes this work:

Pfam: clans, web tools and services: R.D. Finn, J. Mistry, B. Schuster-Bückler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, 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
- JFRC, USA
- INRA, France
- CCBB, South Korea

Methods to assign genes to Pfams
- HMMs (hmmer).
- reverse psi blast (CDD)

http://pfam.sanger.ac.uk

TIGRfam

- Full length alignments.
- Domain alignments.
- Equivalogs: families of proteins with specific function.
- Superfamilies: families of homologous genes.
- Grouping of TIGRfams to functional categories (roles).

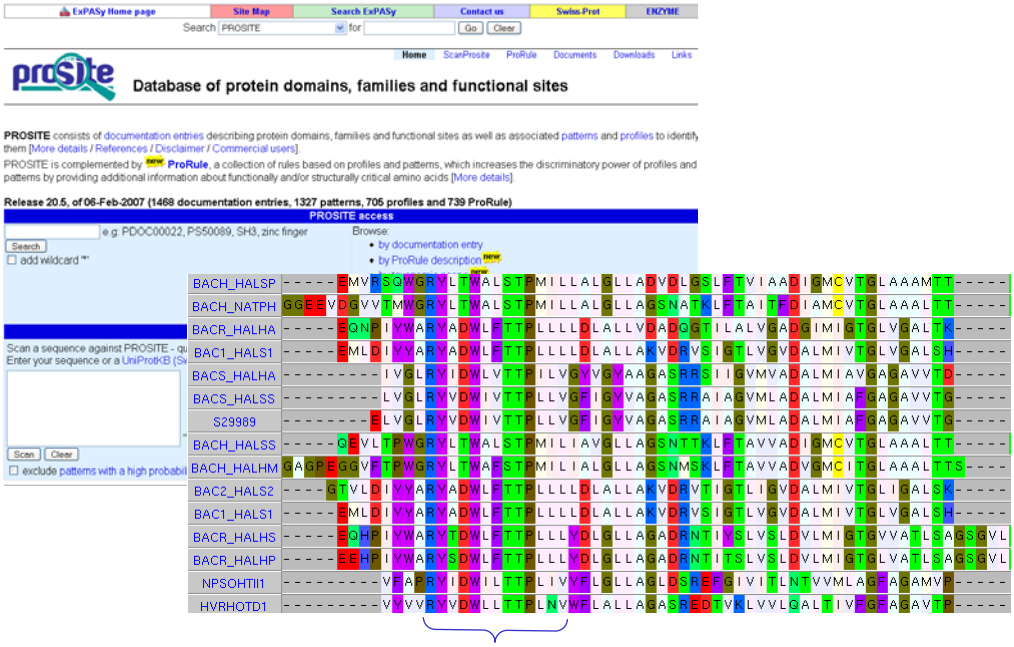
Methods to assign genes to TIGRfams
•HMMs (hmmer).

TIGRFAMs Ordered by Role Category

- ⊞ 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/

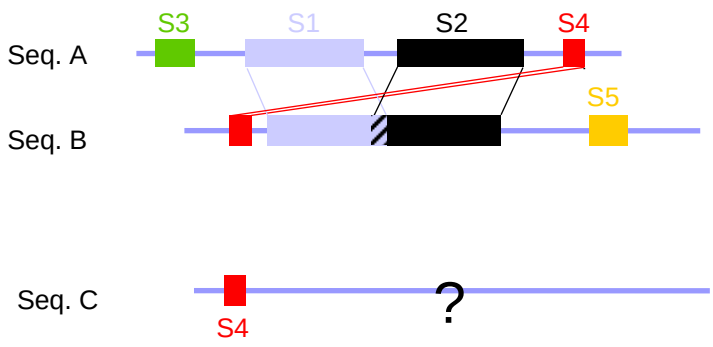
PROSITE



R-Y-x-[DT]-W-x-[LIVM]-[ST]-T-P-[LIVM](3)

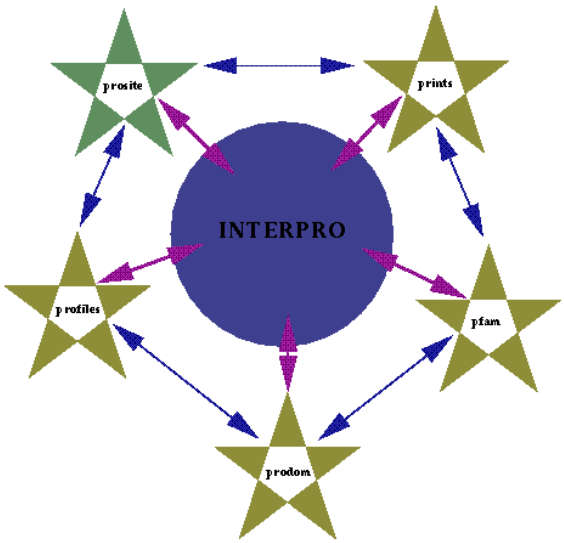
Finding a domain is not enough…

Identification of a domain in a protein is not enough to predict the function. Only some general predictions can be made.
Can we predict the function of C if we know the functions of A and B?



Composite pattern databases

- 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/

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

ENZYME

1. -. -.- Oxidoreductases.
1. 1. -.- Acting on the CH-OH group of donors.
1. 1. 1.- With NAD(+) or NADP(+) as acceptor.
1. 1. 2.- With a cytochrome as acceptor.
1. 1. 3.- With oxygen as acceptor.
1. 1. 4.- With a disulfide as acceptor.
1. 1. 5.- With a quinone or similar compound as acceptor
1. 1.99.- With other acceptors.
1. 2. -.- Acting on the aldehyde or oxo group of donors.

2. -. -.- Transferases.
2. 1. -.- Transferring one-carbon groups.
2. 1. 1.- Methyltransferases.
2. 1. 2.- Hydroxymethyl-, formyl- and related
2. 1. 3.- Carboxyl- and carbamoyltransferases.
2. 1. 4.- Amidinotransferases.

3. -. -.- Hydrolases.
3. 1. -.- Acting on ester bonds.
3. 1. 1.- Carboxylic ester hydrolases.
3. 1. 2.- Thiolester hydrolases.
3. 1. 3.- Phosphoric monoester hydrolases.
3. 1. 4.- Phosphoric diester hydrolases.

4. -. -.- Lyases.
4. 1. -.- Carbon-carbon lyases.
4. 1. 1.- Carboxy-lyases.
4. 1. 2.- Aldehyde-lyases.
4. 1. 3.- Oxo-acid-lyases.
4. 1.99.- Other carbon-carbon lyases.

5. -. -.- Isomerases.
5. 1. -.- Racemases and epimerases.
5. 1. 1.- Acting on amino acids and derivatives.
5. 1. 2.- Acting on hydroxy acids and derivatives.
5. 1. 3.- Acting on carbohydrates and derivatives.
5. 1.99.- Acting on other compounds.
5. 2. -.- Cis-trans-isomerases.

6. -. -.- Ligases.
6. 1. -.- Forming carbon-oxygen bonds.
6. 1. 1.- Ligases forming aminoacyl-tRNA
6. 2. -.- Forming carbon-sulfur bonds.
6. 2. 1.- Acid--thiol ligases.
6. 3. -.- Forming carbon-nitrogen bonds.

Grouping of proteins based on their function.

ENZYME

ENZYME
Enzyme nomenclature database

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [More details / References / Disclaimer].

Release 33, October 2003, and updates up to 29-Nov-2003 (4208 entries)

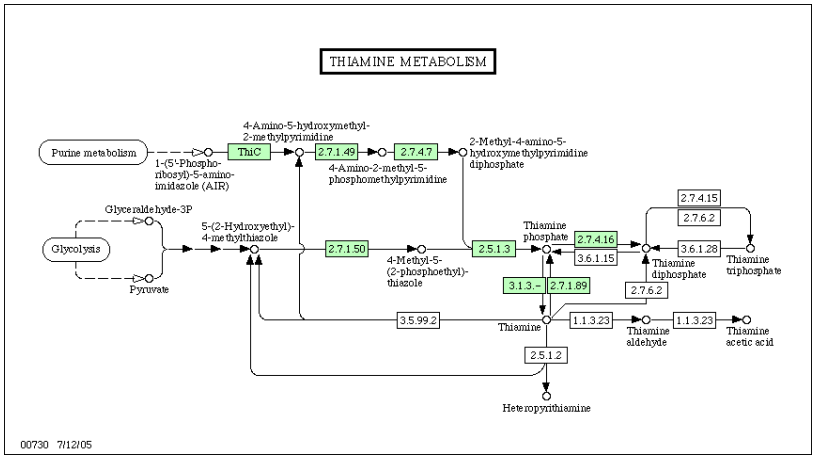
Access to ENZYME

- by EC number: . . . . Search
- by enzyme class
- by description (official name) or alternative name(s): Search
- by chemical compound
- by cofactor
- by search in comments lines
- SRS - Sequence Retrieval System

http://ca.expasy.org/enzyme/

KEGG

- Contains information about biochemical pathways, and protein interactions
  - Information is organized in maps that simplify the flow of metabolites in the cell



http://www.kegg.com

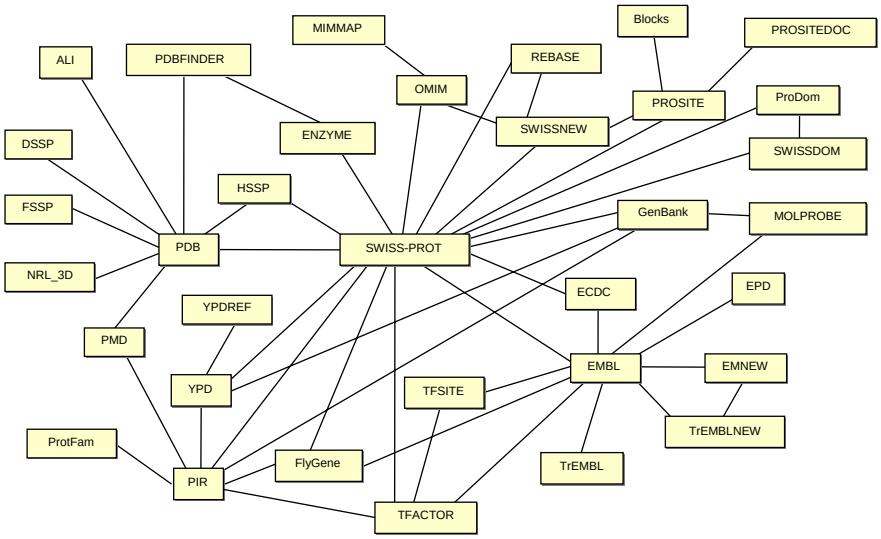
Specialized databanks

- There is a large number of databases devoted to specific organisms and specific protein families/ functions.
- For some model organisms there are often concurrent systems. These databases are associated to sequencing or mapping projects.
- Functional databanks contain extensive information about a group of functions. Good source of classification and current knowledge of this group of proteins.

Other specialized databases

- Signal transduction, regulation, protein-protein interactions
  - TRANSFAC (Transcription Factor database)
  - BRITE (Biomolecular Relations in Information Transmission and Expression database)
  - DIP (I
  - BIND
  - BioCa
- Biochem
  - KLOT
  - BREN
  - LIGAI
- Gene ord
  - STRII
- 3D structures
  - PDB (Protein Data Bank)
  - MM
  - NR
    Str
  - SC
- Polymo
  - ALI
- Molecu
  - DIP
  - BIN
    Da
- 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)

Databanks interconnection



Databases use common identifiers (e.g. Genbank acc.number, swissprot id)
Not all databases are updated regularly.
Changes of annotation in one database are not reflected in others.

Summary

- We have main archives (Genbank), and currated databases (Refseq, SwissProt), and protein classification database (COG, Pfam etc).
- There are many more databases.
- They help *predict* the function, or the network of functions, but *not prove*.
- Systems that integrate the information from several databases, visualize and allow handling of data in an intuitive way *are required.*

Function prediction

- Similarity searches (BLAST).
- Domain identification(Pfam).
- Small sequence identification(PROSITE).
- Localization (membrane, intracellular, extracellular).
- Other sequence features.
- Similarity to known structures.

Annotation based on best hits

Select	Ortholog	Product Name	Percent Identity	Alignment On Query Gene	Alignment On Subject Gene	Length	E-value	Bit Score	Cons. Region Score[1]	D	C	Genome Name
☐	640229721	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	74.21			122aa	1.0e-62	239	4	B	D	Escherichia coli B
☐	638664970	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	72.56			122aa	1.0e-62	239	4	B	D	Escherichia coli E22
☐	637523567	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS10394
☐	637991098	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS10270
☐	637090213	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MI GAS
☐	638453633	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23			117aa	9.0e-06	50	0	B	D	Bacillus cereus G9241
☐	640207913	diacylglycerol kinase	27.45			118aa	1.0e-05	50	0	B	D	Thermoanaerobacter carboxydovorans Nor1
☐	639732576	diacylglycerol kinase	26.17			236aa	1.0e-05	50	0	B	F	Clostridium novyi NT
☐	641042897	hypothetical protein	24.59									
☐	638991165	probable diacylglycerol kinase	28.57									
☐	638926147	probable diacylglycerol kinase	35.42			108aa	1.0e-05	50	0	B	D	Cellulophaga sp. MED134
☐	638838924	putative diacylglycerol kinase	31.31			184aa	1.0e-05	50	0	B	D	Roseovarius nubinhibens ISM
☐	640079875	Prokaryotic diacylglycerol kinase	29.17			136aa	1.0e-05	50	0	B	F	Prochlorococcus marinus MIT 9515
☐	640632535	probable diacylglycerol kinase	27.59			124aa	1.0e-05	50	0	B	D	Flavobacteria bacterium BAL38
☐	639005848	diacylglycerol kinase	23.21			126aa	1.0e-05	50	0	B	D	Polaribacter irgensii 23-P
☐	638345788	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23			117aa	1.0e-05	50	0	B	D	Bacillus anthracis A2012
☐	637449177	Prokaryotic diacylglycerol kinase	28.12			138aa	1.0e-05	50	0	B	F	Prochlorococcus marinus pastoris CCMP1986

We usually predict a gene to have the same function as its best hit, especially if the e-value is very low.

Don't forget

A functional prediction has to make sense in the context of metabolism

Gene Information	
Gene Object ID	637017366
Gene Symbol	dgkA
Locus Tag	b4042
Product Name	diacylglycerol kinase
Genome	Escherichia coli K12
DNA Coordinates	4254660..4255028 (+)(369bp)
Scaffold Source	Escherichia coli K12: NC_000913 (4639675bp)
IMG ORF Type	
GC Content	0.52
Accession	NP_418466
External Links	GI:16131868; PUMA/GI:16131868; VIMSS/GI:16131868; GeneID:948543
Protein Information	
Amino Acid Sequence Length	122aa
COG	COG0818 - Diacylglycerol kinase
IMG Term	diacylglycerol kinase (EC 2.7.1.107) (THANOS 24-APR-06)
Transmembrane Helices	No
Signal Peptide	No
Statistics	peptide
Pathway Information	
Enzymes	EC:2.7.1.107 - Diacylglycerol kinase
KO Term	K00901 - diacylglycerol kinase
KEGG Pathway	Glycerolipid metabolism Glycerophospholipid metabolism Phosphatidylinositol signaling system
IMG Pathways	Phosphatidic acid biosynthesis via kinase (asserted)

Select	Ortholog	Product Name	Percent Identity	Alignment On Query Gene	Alignment On Subject Gene	Length	E-value	Bit Score	Cons. Region Score[1]	D	C	Genome Name
☐	640229721	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	D	Escherichia coli B
☐	638664970	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	D	Escherichia coli E22
☐	637336319	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Shigella flexneri 2a 301
☐	637425330	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Shigella flexneri 2a 2457T
☐	637809694	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Shigella boydii Sb227
☐	637003978	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Escherichia coli W3110
☐	637085686	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Escherichia coli O157:H7 Sakai
☐	637066833	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Escherichia coli O157:H7 EDL933
☐	637358875	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Escherichia coli CFT073
☐	638077808	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	F	Shigella flexneri 5 8401
☐	638669437	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	100			122aa	1.0e-62	239	4	B	D	Escherichia coli E110019

☐	638475713	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	D	Streptococcus pyogenes M49 591
☐	638423316	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	31.43			131aa	9.0e-08	57	0	B	D	Listeria monocytogenes 1/2a F6854
☐	637674930	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS6180
☐	637676923	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS5005
☐	637306861	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS315
☐	637993185	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS2096
☐	637995143	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS10750
☐	637523567	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS10394
☐	637991098	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes MGAS10270
☐	637090213	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	34.19			135aa	9.0e-08	57	0	B	F	Streptococcus pyogenes M1 GAS
☐	640977819	diacylglycerol kinase (EC 2.7.1.107) (IMGterm)	31.43			131aa	9.0e-08	57	0	B	D	Listeria monocytogenes 10403S

638453633	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	9.0e-06	50	0	B	D	Bacillus cereus G9241
640207913	diacylglycerol kinase	27.45	118aa	1.0e-05	50	0	B	D	Thermosinus carboxydivorans Nor1
639732576	diacylglycerol kinase	26.17	236aa	1.0e-05	50	0	B	F	Clostridium novyi NT
641042897	hypothetical protein	24.59	123aa	1.0e-05	50	0	B	D	Parabacteroides merdae ATCC 43184
638991165	probable diacylglycerol kinase	28.57	121aa	1.0e-05	50	0	B	D	Flavobacteriales bacterium HTCC2170
638926147	probable diacylglycerol kinase	35.42	108aa	1.0e-05	50	0	B	D	Cellulophaga sp. MED134
638838924	putative diacylglycerol kinase	31.31	184aa	1.0e-05	50	0	B	D	Roseovarius nubinhibens ISM
640079875	Prokaryotic diacylglycerol kinase	29.17	136aa	1.0e-05	50	0	B	F	Prochlorococcus marinus MIT 9515
640632535	probable diacylglycerol kinase	27.59	124aa	1.0e-05	50	0	B	D	Flavobacteria bacterium BAL38

When the BLAST hits are not very good, we could misannotated a gene

638345788	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	1.0e-05	50	0	B	D	Bacillus anthracis A2012
637449177	Prokaryotic diacylglycerol kinase	28.12	138aa	1.0e-05	50	0	B	F	Prochlorococcus marinus pastoris CCMP1986
637515456	Undecaprenol kinase	25.23	117aa	1.0e-05	50	0	B	F	Bacillus thuringiensis

Although this gene in Bacillus, would make sense in the context of the metabolism to be annotated as diacylglycerol kinase…

637548263	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	1.0e-05	50	0	B	F	Bacillus cereus E33L
637507437	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	1.0e-05	50	0	B	F	Bacillus anthracis Sterne
637791221	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	1.0e-05	50	0	B	F	Bacillus anthracis Ames Ancestor
637304722	Undecaprenol kinase (EC 2.7.1.-) (IMGterm)	25.23	117aa	1.0e-05	50	0	B	F	Bacillus anthracis Ames

Gene Information	
Gene Object ID	637030809
Gene Symbol	yerQ
Locus Tag	BSU06720
Product Name	hypothetical protein
Genome	Bacillus subtilis 168
DNA Coordinates	735775..736686 (+)(912bp)
Scaffold Source	Bacillus subtilis subsp. subtilis str. 168: NC_000964 (4214630bp)
IMG ORF Type	
GC Content	0.51
Accession	NP_388554
External Links	GI:16077740; PUMA/GI:16077740; VIMSS/GI:16077740; GeneID:936061; UniProtKB/TrEMBL:O31502

there is another gene in Bacillus that has this function.

Protein Information	
Amino Acid Sequence Length	303aa
COG	COG1597 - Sphingosine kinase and enzymes related to eukaryotic diacylglycerol kinase
IMG Term	diacylglycerol kinase (EC 2.7.1.107) (THANOS 02-JUL-07)
Families	- IPR001206 Diacylglycerol kinase, catalytic region - IPR005218 Conserved hypothetical protein 147 - TIGR00147 conserved protein of unknown function cotranscribed with Bmr (bmrU)
Transmembrane Helices	No
Signal Peptide	No
Statistics	peptide
Pathway Information	
Enzymes	
KO Term	K07029 -
KEGG Pathway	
IMG Pathways	Phosphatidic acid biosynthesis via kinase

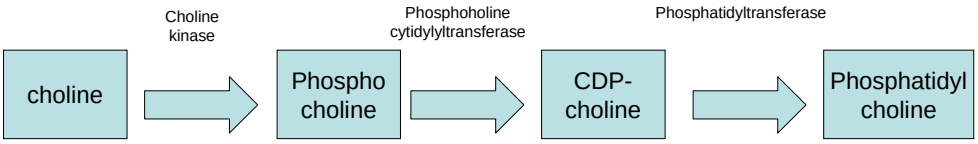
Based on the BLAST hits we get a hint that TP0671 in *T.pallidum* is a Phosphatidyl transferase (CEPT). This enzyme is frequently found in Eukarya. In order to accept this annotation in *T. pallidum* we have to verify that other related functions are present.

CDP-Cho	+	DAG	$\xrightarrow{\text{CPT}}$	PtdCho	Eukarya
CDP-Etn	+	DAG	$\xrightarrow{\text{EPT}}$	PtdEtn	Eukarya

Pathway

What is a pathway?

A sequence of reactions transforming one metabolite to another



Take known sequences of kinase and cytidylyltransferase and check whether they exist in the organism (*T. pallidum*)

All genomes selected

Quick Genome Search : GO

INTEGRATED MICROBIAL GENOMES

IMG Home | Find Genes | Find Genomes | Find Functions | Compare Genomes | MyIMG | Analysis Carts | About IMG | Using IMG | News

Search Terms and Pathways | COG Browser | KEGG Browser

Search Terms and Pathways

Find functions in selected genomes by keyword.

Keyword: choline kinase

Filters: Pfam

Select from genomes below or
assume current genome browser selections.
Domains(D): B = Bacteria, A = Archaea, E = Eukarya, V = Viruses.

Treponema denticola ATCC 35405 (B)
Treponema pallidum Nichols (B)
Trichodesmium erythraeum IMS101 (B)
Tropheryma whipplei TW08/27 (B)
Tropheryma whipplei Twist (B)
Ureaplasma parvum serovar 3 ATCC 700970 (B)
Vibrio cholerae O1 biovar eltor N16961 (B)
Vibrio fischeri ES114 (B)
Vibrio parahaemolyticus RIMD 2210633 (B)
Vibrio vulnificus CMCP6 (B)

Go | Reset

Examples

- "kinase" as a Function Name gets "Shikimate kinase", etc.
- "EC:1.2.3.3" or "1.2.3.3" as an EC Number (exact) gets only genes associated with EC 1.2.3.3.
- "1.2.3" as an EC Number (inexact) returns genes associated with EC 1.2.3.3 and EC 1.1.2.3.
- "IPR0044662" (exact) or "kinase" or "Acetylglutamate" (inexact) gets "IPR004462 Acetylglutamate kinase".
- "lipid" return "Blood group glycolipid biosynthesis-lactoseries", "Lipid Metabolism" for KEGG Pathway search.

All genomes selected

Quick Genome Search : GO

INTEGRATED MICROBIAL GENOMES

IMG Home | Find Genes | Find Genomes | Find Functions | Compare Genomes | MyIMG | Analysis Carts | About IMG | Using IMG | News

Search Terms and Pathways | COG Browser | KEGG Browser

Function Search Results

1 functions retrieved.

The number of genes is shown in parentheses.

Add Selected to Pfam Cart | Select All | Clear All

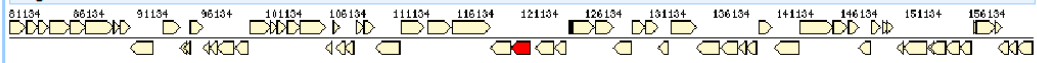
☐ pfam01633 - Choline_kinase - Choline/ethanolamine kinase. **Choline kinase** catalyses the committed step in the synthesis of phosphatidylcholine by the CDP-choline pathway (1)

hint: *The Pfam cart allows for phylogenetic profile comparisons.*

Version 1.4 March 1, 2006
Questions/Comments
©2006 The Regents of the University of California
Disclaimer

Evidence for Function Prediction

Neighborhood



red = Current Gene
green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene
Show ortholog neighborhood regions in user-selected genomes
Show in Chromosome Viewer

COG

COG ID	Consensus Sequence Length	Description	Percent Identity	Alignment on Conensus Sequence	E-value	Bit Score
COG4750	231	[M] Cell wall/membrane/envelope biogenesis CTP:phosphocholine cytidylyltransferase involved in choline phosphorylation for cell surface LPS epitopes	56.58		5.0e-99	355

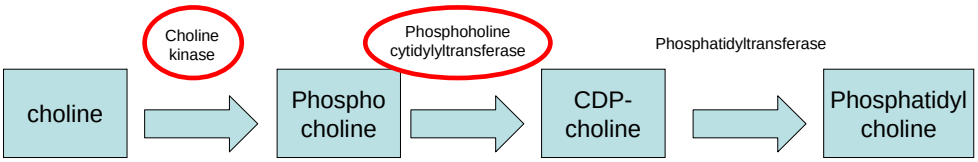
Pfam

Pfam Domain	CDD Pfam Hit	Description	Percent Identity	Alignment on Gene	E-value	Bit Score
APH	pfam01636	Phosphotransferase enzyme family	20		5.4e-07	49.38
Choline_kinase	pfam01633	Choline/ethanolamine kinase	17.67		2.1e-05	44.11
NTP_transferase	pfam00483	Nucleotidyl transferase	24.56		1.6e-04	41.34

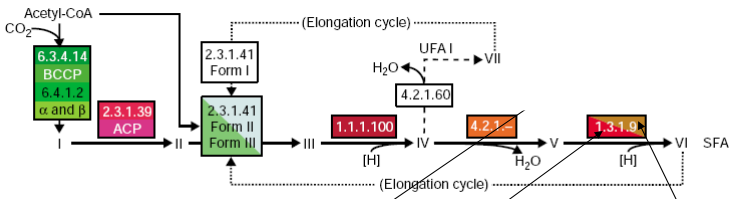
InterPro

IPR008266 = Tyrosine protein kinase, active site
IPR011009 = Protein kinase-like

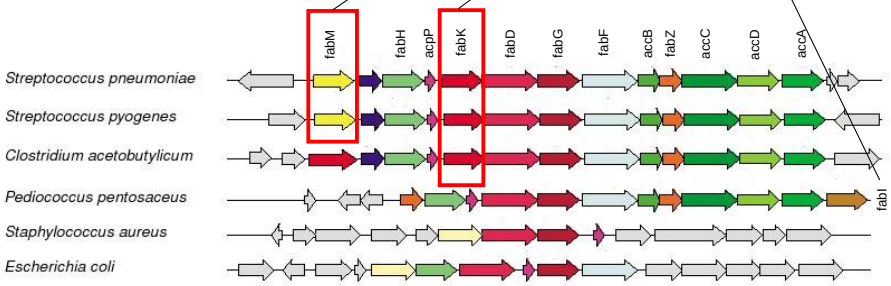
All genes of the pathway are present



Gene function prediction. Chromosomal neighborhood



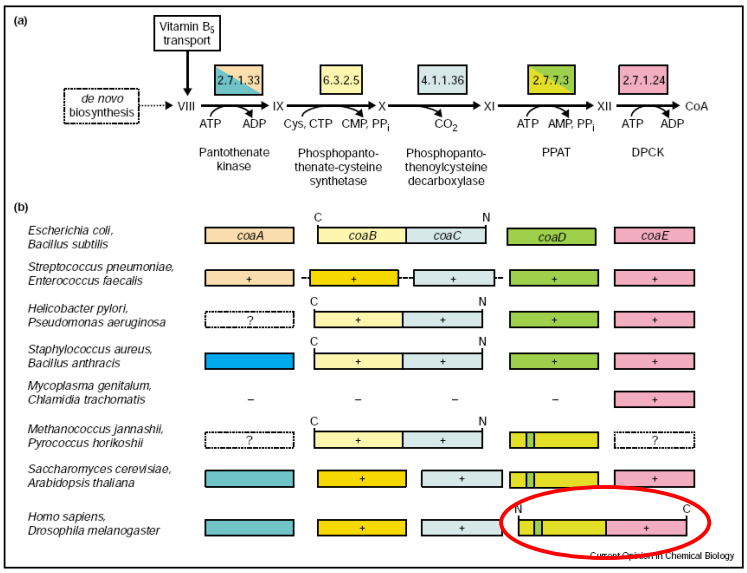
The presence of genes in the chromosomal neighborhood of genes of a pathway is an indication of their implication to the same pathway. They are candidates for missing functions from this pathway.



Osterman & Overbeek. Curr. Op. in Chemical Biology. 7(2003) 238-251

Function prediction. Fusions

Fused genes are functionally related. The fusion of a known gene(pink) with an unknown function(yellow) suggests a function for the latter. The function must complement the pathway that the pink gene participates.



Osterman & Overbeek. Curr. Op. in Chemical Biology. 7(2003) 238-251

Current Status of annotation

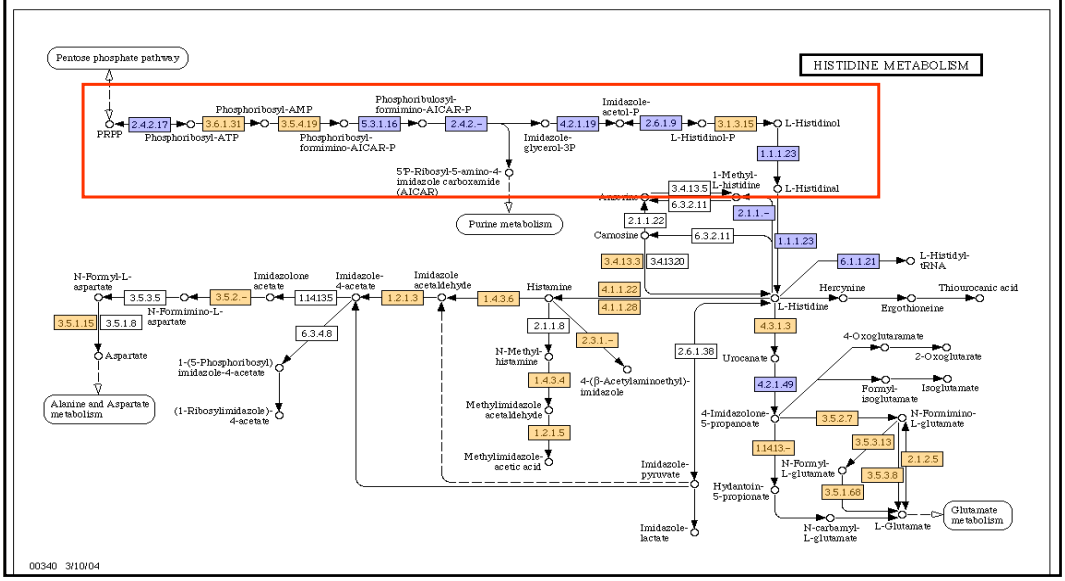
- ~ **50-80% precise, accurate prediction**
- ~ **10-30% "twilight zone" predictions**
- ~ **10-30% genome specific genes**

Poor quality data can lead to incorrect metabolic reconstruction

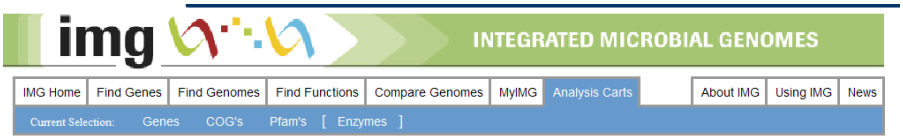
KEGG Map

Loaded.

Genes in Bacillus anthracis Ames.
EC equivalogs (genes found in other organisms that have the same EC number).



Is *Bacillus anthracis* Ames a histidine auxotroph?

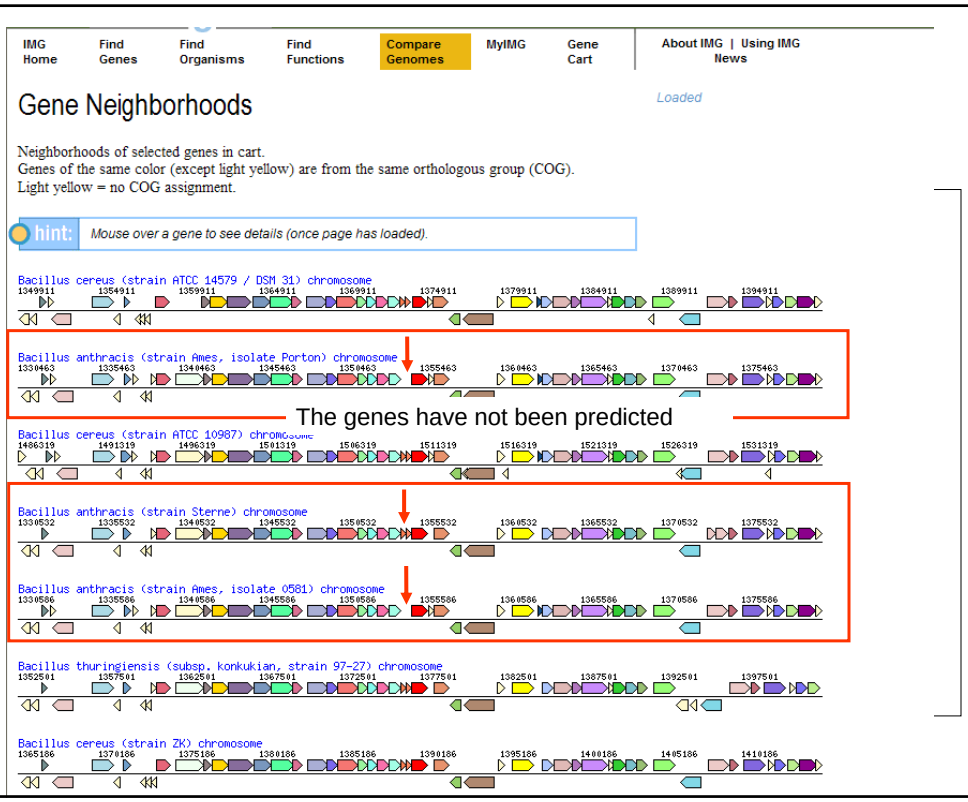


Enzyme Profile

Loaded.

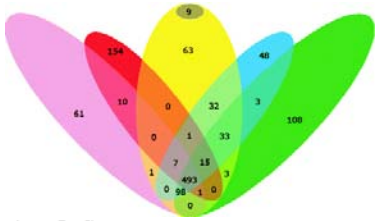
Mouse over function ID to see name.
(Cell coloring is highlighting of gene counts: white = 0, bisque = 1-4, yellow >= 5.)
Domains(D): B = Bacteria, A = Archaea, E = Eukarya, V = Viruses.

Genome/Bin Name	D	2.4.2.17	3.5.4.19	3.6.1.31	5.3.1.16
Bacillus anthracis Ames	B	1	0	0	1
Bacillus anthracis Ames 0581	B	1	0	0	1
Bacillus anthracis Sterne	B	1	1	1	1
Bacillus cereus ATCC 10987	B	1	1	1	1
Bacillus cereus ATCC 14579	B	1	1	1	1
Bacillus cereus E33L	B	1	1	1	1
Bacillus cereus NVH391-98	B	1	1	0	0
Bacillus subtilis 168	B	1	1	1	1



Genome comparison. Profiles

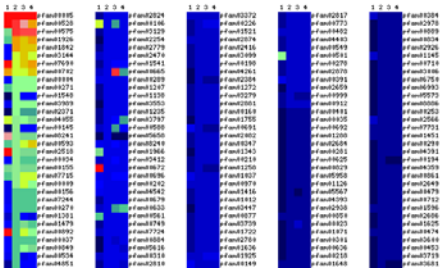
- Common/unique genes
  - Speficy a group of organisms.
  - Find genes that are shared between genomes.
  - Find genes that are unique in an organism.
- Profiles
  - Find abundant genes.



Abundance Profile

Mouse over labels to see additional information.
Clicking on the column number will sort rows for that column in descending gene count order.
Clicking on row cluster ID will add the cluster to the appropriate analysis cart (if cart is supported)
Mouse over heat map to see gene counts. Clicking on gene count will take you to the gene list.

1 - Chromobacterium violaceum ATCC 12472
2 - Neisseria gonorrhoeae FA 1090
3 - Neisseria meningitidis MC58
4 - Neisseria meningitidis Z2491



Finding unique genes

Obligate parasite of horses

☐	Buchnera aphidicola (subsp. Baizongia pistaciae)			
☐	Buchnera aphidicola (subsp. Schizaphis graminum)			
☐	Burholderia cepacia strain 383 ATCC 17660 (R-18194)	B	Draft	JGI
☐	Burkholderia cenocepacia HI2424	B	Draft	JGI
☑	Burkholderia mallei (strain ATCC 23344)	B	Finished	TIGR
☑	Burkholderia pseudomallei (strain K96243)	B	Finished	Sanger Institute; Porton Down
☐	Burkholderia vietnamiensis strain G4 (R1808)	B	Draft	JGI
☐	Burkholderia xenovorans LB400			
☐	Campylobacter jejuni (strain NCTC 11168)			
☐	Candida glabrata CBS138 (Torulopsis glabrata)			
☐	Candidatus Blochmannia floridanus			
☐	Caulobacter crescentus (strain CB15 / ATCC 19089)	B	Finished	TIGR

Causes human disease in tropical areas (melioidosis)

IMG Home | Gene Search | Organism Browser | **Phylogenetic Profiler** | Organism Stats | Gene Cart | Preferences | About IMG | Using IMG | News

Phylogenetic Profiler

Find genes in organism of interest qualified by similarity to sequences in other organisms (based on BLAST alignments). Only use

Find Genes In*	With Homologs In	Without Homologs In	Ignoring	Organism Name	Max. E-value	Min. % Identity
○	○	◉	○	Burkholderia mallei (strain ATCC 23344)	1e-5	30
◉	○	○	○	Burkholderia pseudomallei (strain K96243)	1e-5	30

Go | Reset

Find unique genes based on similarity

Phlyogenetic Profiler Results

1324 gene(s) retrieved

5855 genes found for organism of interest, Burkholderia pseudomallei (strain K96243)
1324 genes remaining after subtracting genes with homologs in Burkholderia mallei (strain ATCC 23344)

Add Selected to

Select	Gene Obje
☐	5359420
☐	5359430
☐	5359440
☐	5359460
☐	5359470
☐	5359600
☐	5359640
☐	5359710
☐	5359760
☐	5359770
☐	5359780
☐	5359790
☐	5359800
☐	5359810
☐	5359830
☐	5359840
☐	5359940
☐	5359950
☐	5359960

☐	5377420	BPSL1798	Putative exported protein
☐	5377540	BPSL1810	Putative membrane protein
☐	5377550	BPSL1811	Putative membrane protein
☐	5377560	BPSL1812	Putative membrane protein
☐	5377570	BPSL1813	Putative fimbriae-related membrane protein
☐	5377580	BPSL1814	Putative membrane protein
☐	5377590	BPSL1815	Putative fimbriae assembly-related protein
☐	5377600	BPSL1816	Putative fimbriae assembly-related protein
☐	5377610	BPSL1817	Putative lipoprotein
☐	5377630	BPSL1819	Putative fimbriae-assembly related protein
☐	5377640	BPSL1820	Putative fimbriae assembly-related protein
☐	5377650	BPSL1821	Putative fimbriae assembly-related protein
☐	5377680	BPSL1824	Putative ABC transport system, substrate-binding protein
☐	5377700	BPSL1826	Hypothetical protein
☐	5377710	BPSL1827	Hypothetical protein
☐	5377720	BPSL1828	Hypothetical protein
☐	5377980	BPSL1853	Hypothetical protein
☐	5377990	BPSL1854	Hypothetical protein
☐	5378040	BPSL1859	Putative membrane protein
☐	5378330	BPSL1885	Putative exported protein
☐	5378380	BPSL1889	Putative lipoprotein
☐	5378410	BPSL1892	Putative fimbriae-related outer membrane protein
☐	5378420	BPSL1893	Putative type II/IV secretion system ATP-binding protein
☐	5378460	BPSL1897	Putative membrane protein
☐	5378480	BPSL1899	Putative fimbriae assembly related protein

Genes involved in adherence and invasion

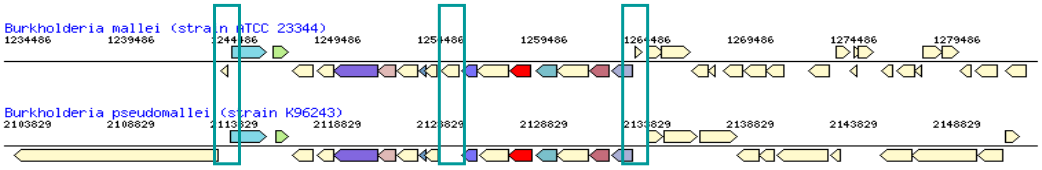
Don't trust the computer…

Gene Ortholog Neighborhoods

Loaded

Neighborhoods of orthologs in user-selected organisms.
Genes of the same color (except light yellow) are from the same orthologous group (COG).
Light yellow = no COG assignment.

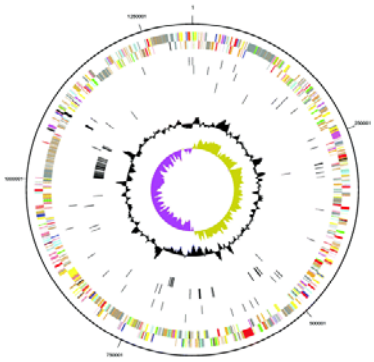
hint: *Mouse over a gene to see details (once page has loaded).*

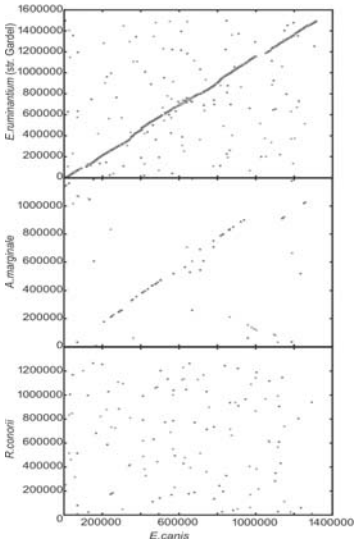


- **Phylogenetic profiler finds 548 unique genes in *B. mallei***
- **However, 497 of them in fact exist in B. pseudomallei, but they have not been called as real genes.**
- **The difference in gene models reveals** 89.2% error rate **in unique genes**

Genome comparison. Synteny

- Identify syntenous regions
  - Pathogenicity islands
  - Insertion of transposons, phages.
  - ...





Mavromatis etal. J.Bact. 188:11(2006). 4015-4023.

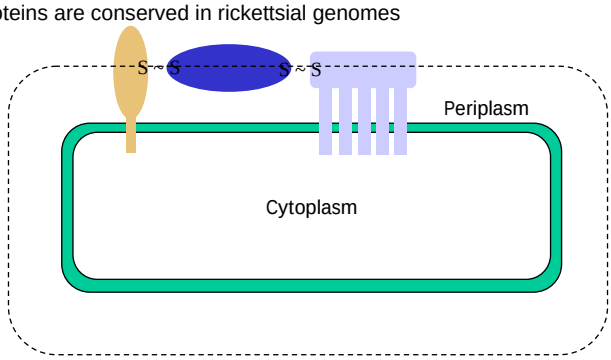
General predictions can be made even without similarity information

Ehrlichia canis lacks a peptidoglycan cell wall.

- Secreted/Membrane proteins create cell envelope
- Cell wall structure is enhanced by disulfide bonds

Candidates proteins for the formation of the cell wall:

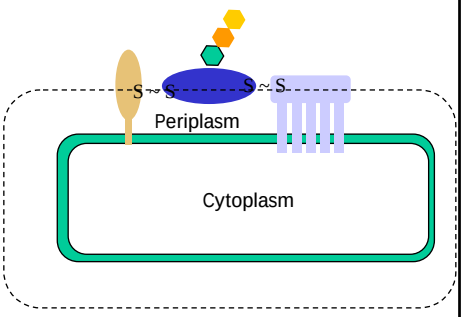
- 36 Cys rich proteins 27 of unknown function
- 16 membrane associated (transmembrane, or secreted)
- These proteins are conserved in rickettsial genomes



Mavromatis etal. J.Bact. 188:11(2006). 4015-4023.

Ser/Thr rich proteins

- Six proteins have been verified to be glycosylated
- O-linked glycans (attached to Ser/Thr residues)
- Function
  - Cell wall formation
  - Pathogen – host cell interaction
  - Antigens
- 180 proteins exhibit increased Ser/Thr content
- 67 are membrane associated
  - All identified glycoproteins
  - 9 Vir related proteins
  - Proteins of unknown function
  - 7 Membrane associated proteins unique to rickettsial/ehrlichial genomes exhibit increased Cys, Ser/Thr content



Mavromatis etal. J.Bact. 188:11(2006). 4015-4023.