Functional Annotation & Beyond

with Integrated Microbial Genomes (IMG)

Rekha Seshadri, Ph.D.
MGM-23 Workshop
Walnut Creek, CA (9/27/16)

Functional Annotation: Sequence to Biology

- "Feature prediction" - location of genes and other features - noncoding (e.g., rRNA, tRNA) and protein coding sequences (CDS), and more (e.g., regulatory sites, repeats, frameshifts)

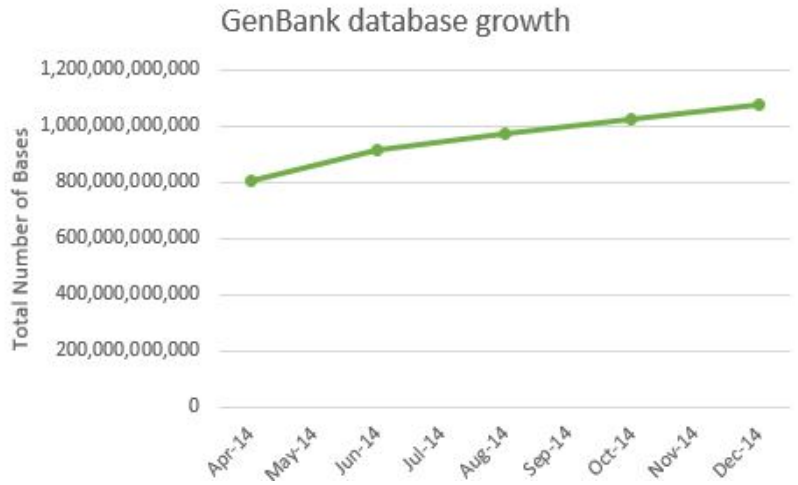
- Next, attempt to "name" and interpret function/role by comparing against functional "databases"

- Archives of accumulated biological data and knowledge
  - Genome, gene sequences, mutations
  - Gene Regulation, expression, splice variants
  - Protein sequence, post-translational modifications
  - Protein tertiary structure, localization, networks
  - Enzyme kinetics, metabolites, metabolic networks
  - E.g., nr, swissprot, pdb, Pfam
- List of databases:
  - http://www.oxfordjournals.org/our_journals/nar/database/c

Searching databases

- Similarity is the primary predictor of homology, which is the predictor of function *(sort of)*
- How to search for sequence similarity? *(Sensitivity versus speed)*

GenBank surpasses one trillion total bases of publicly available sequence data
Thursday, January 22, 2015

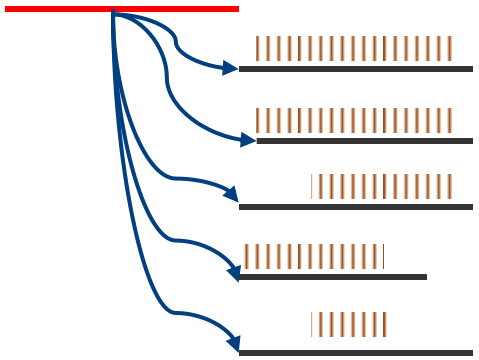


Search Method: Pairwise

- Find related or similar sequences by mapping letters of two sequences, with some spacers (indels),

```
76 GGMLKPIEGGTYEVNEAMVEDLKIGVQGPHASNLGGILSNEIAKEIGKRAFIVDPVVVDE 135
   ||||||||||||||||||||||||||  |||| ||||||||||||  ||||||||||||||
61 GGMLKPIEGGTYEVNEAMVEDLKIGFEGPHAXNLGGILSNEIAKKLGKRAFIVDPVVVDX 120
```

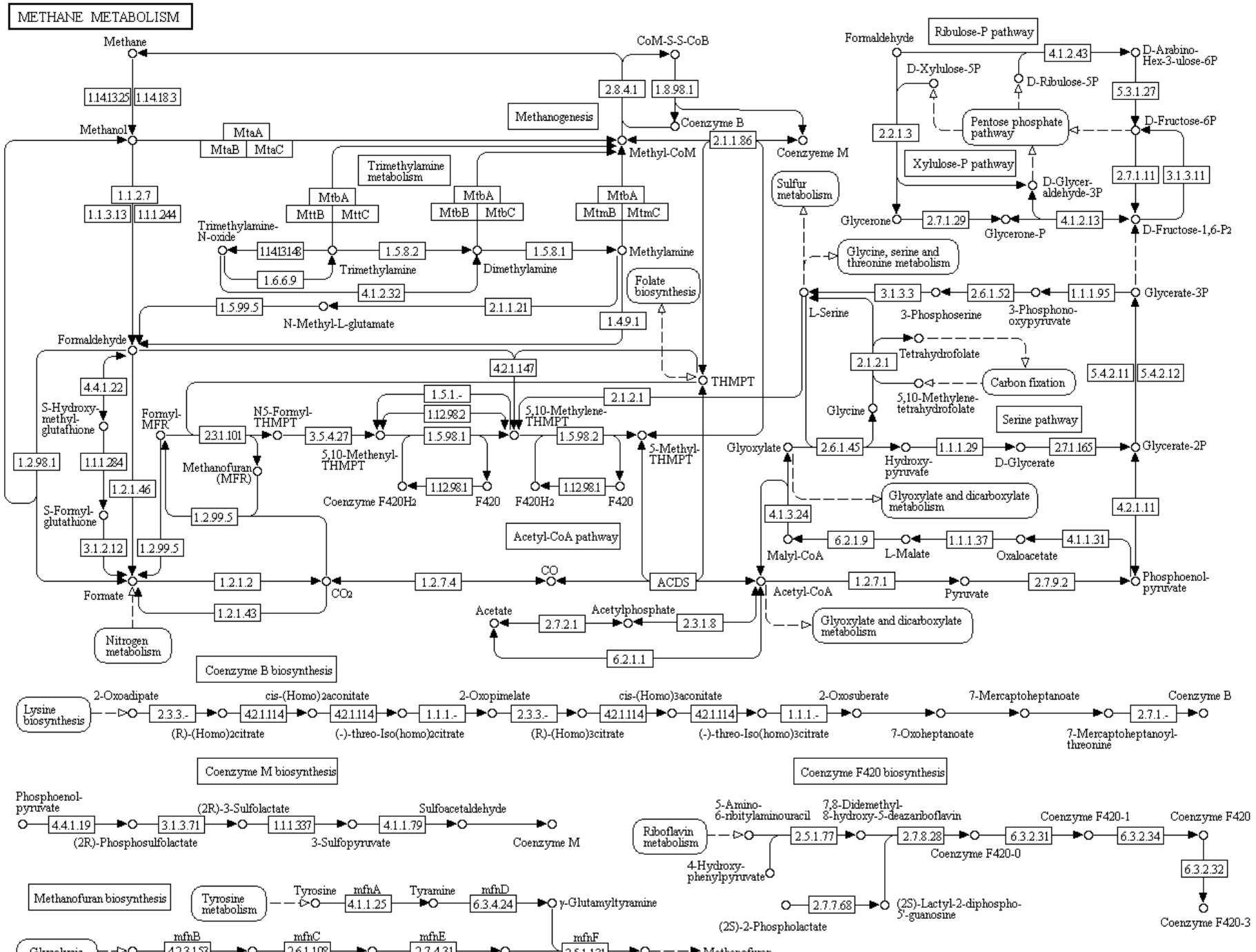
- Parse similarities, determine “best hit”



- Examples of pairwise search tools - Basic Local Alignment Search Tool or BLAST, LAST, etc.

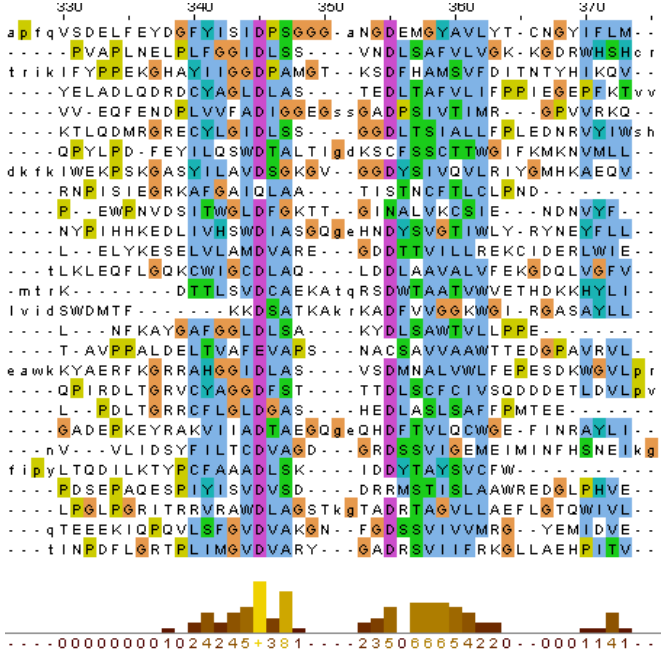
Sequence databases

- Non-redundant database (nr), Refseq nr (curated), Uniref
- Kyoto Encyclopedia of Genes and Genomes (KEGG) – integrates functional information into biological pathways
  - genomic, chemical (compound, reaction, enzyme), systems (pathway), health (disease, drug)
  - 10,371 KO terms
  - Pathway database provides maps, e.g.,


METHANE METABOLISM
Methane
CoM-S-S-CoB
1.14.13.25
1.14.18.3
2.8.4.1
1.8.98.1
Methanogenesis
Coenzyme B
Methanol
MtaA
MtaB
MtaC
Methyl-CoM
2.1.1.86
Coenzyme M
Trimethylamine metabolism
1.1.2.7
1.1.3.13
1.1.1.244
MtbA
MttB
MttC
MtbA
MtbB
MtbC
MtbA
MtmB
MtmC
Trimethylamine-N-oxide
1.14.13.148
1.5.8.2
1.5.8.1
Methylamine
Trimethylamine
Dimethylamine
1.6.6.9
4.1.2.32
1.5.99.5
2.1.1.21
N-Methyl-L-glutamate
1.4.9.1
Folate biosynthesis
Formaldehyde
4.2.1.147
THMPT
4.4.1.22
S-Hydroxy-methyl-glutathione
Formyl-MFR
23.1.101
N5-Formyl-THMPT
3.5.4.27
5,10-Methenyl-THMPT
1.5.1.-
1.12.98.2
1.5.98.1
2.1.2.1
5,10-Methylene-THMPT
1.5.98.2
5-Methyl-THMPT
1.2.98.1
1.1.1.284
1.2.1.46
Methanofuran (MFR)
1.2.99.5
Coenzyme F420H2
1.12.98.1
F420
F420H2
1.12.98.1
F420
S-Formyl-glutathione
3.1.2.12
1.2.99.5
Acetyl-CoA pathway
Formate
1.2.1.2
CO2
1.2.7.4
CO
ACDS
1.2.1.43
Nitrogen metabolism
Acetate
2.7.2.1
Acetylphosphate
2.3.1.8
6.2.1.1
Acetyl-CoA
Sulfur metabolism
Formaldehyde
Ribulose-P pathway
4.1.2.43
D-Arabino-Hex-3-ulose-6P
D-Xylulose-5P
D-Ribulose-5P
5.3.1.27
2.2.1.3
Pentose phosphate pathway
D-Fructose-6P
Xylulose-P pathway
2.7.1.11
3.1.3.11
D-Glycer-aldehyde-3P
Glycerone
2.7.1.29
Glycerone-P
4.1.2.13
D-Fructose-1,6-P2
Glycine, serine and threonine metabolism
3.1.3.3
2.6.1.52
1.1.1.95
Glycerate-3P
L-Serine
3-Phosphoserine
3-Phosphono-oxypyruvate
Tetrahydrofolate
2.1.2.1
Carbon fixation
5.4.2.11
5.4.2.12
Glycine
5,10-Methylene-tetrahydrofolate
Serine pathway
Glyoxylate
2.6.1.45
1.1.1.29
2.7.1.165
Glycerate-2P
Hydroxy-pyruvate
D-Glycerate
Glyoxylate and dicarboxylate metabolism
4.2.1.11
4.1.3.24
6.2.1.9
1.1.1.37
4.1.1.31
Malyl-CoA
L-Malate
Oxaloacetate
1.2.7.1
2.7.9.2
Pyruvate
Phosphoenol-pyruvate
Glyoxylate and dicarboxylate metabolism
Coenzyme B biosynthesis
Lysine biosynthesis
2-Oxoadipate
2.3.3.-
(R)-(Homo)2citrate
4.2.1.114
cis-(Homo)2aconitate
4.2.1.114
(-)-threo-Iso(homo)2citrate
1.1.1.-
2-Oxopimelate
2.3.3.-
(R)-(Homo)3citrate
4.2.1.114
cis-(Homo)3aconitate
4.2.1.114
(-)-threo-Iso(homo)3citrate
1.1.1.-
2-Oxosuberate
7-Oxoheptanoate
7-Mercaptoheptanoate
2.7.1.-
7-Mercaptoheptanoyl-threonine
Coenzyme B
Coenzyme M biosynthesis
Phosphoenol-pyruvate
4.4.1.19
(2R)-Phosphosulfolactate
(2R)-3-Sulfolactate
3.1.3.71
3-Sulfopyruvate
Sulfoacetaldehyde
1.1.1.337
4.1.1.79
Coenzyme M
Coenzyme F420 biosynthesis
Riboflavin metabolism
5-Amino-6-ribitylaminouracil
2.5.1.77
7,8-Didemethyl-8-hydroxy-5-deazariboflavin
2.7.8.28
Coenzyme F420-0
6.3.2.31
Coenzyme F420-1
6.3.2.34
Coenzyme F420
6.3.2.32
Coenzyme F420-3
4-Hydroxy-phenylpyruvate
2.7.7.68
(2S)-Lactyl-2-diphospho-5'-guanosine
(2S)-2-Phospholactate
Methanofuran biosynthesis
Tyrosine metabolism
Tyrosine
mfnA
4.1.1.25
Tyramine
mfnD
6.3.4.24
γ-Glutamyltyramine
mfnB
mfnC
mfnE
mfnF

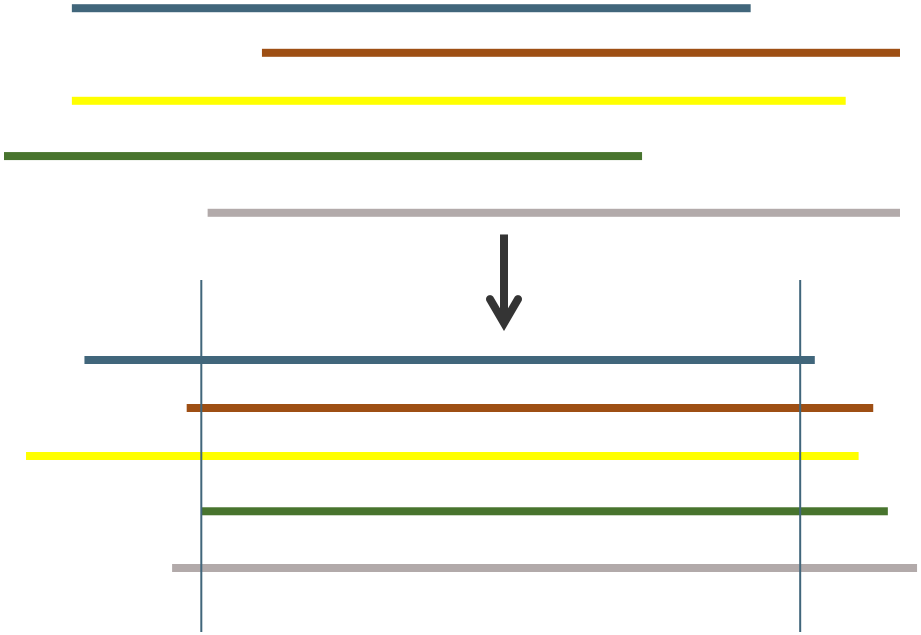
Search Method : Profiles/Models

- MSA of known sequences – detect “regions of similarity” – build consenus
- Use structural and mechanistic information (catalytic sites)
- Generate profiles using Hidden Markov Models (HMMs) or Position-Specific Scoring Models (PSSMs)
- More sensitive than pairwise – detect distant relationships
- Example of profile/HMM search tools: RPS-BLAST, Hmmer



Search Method: Profile based

Collect “seed” proteins



Generate & Trim Alignment

Region of good alignment and closest similarity

Generate Profile with HMM or PSSM

Compute statistical probabilities for amino acid patterns in the seed

Search New Model against all proteins

*Choose “noise” and **“trusted”** cutoff scores based on “known” versus “unknown” protein scores*

Profile/HMM databases

- COGs - protein clusters from at least 3 complete prokaryotic genomes - 4631 COGs from 711 A & B genomes
- KOGs - eukaryotic clusters – 4249 KOGs from 59,838 proteins from 7 euk genomes
- Tigrfam – manually curated collection of 4,488 relatively full-length multiple sequence alignments for annotation → Genome properties
- Pfam – large (16,230), widely used curated collection of protein families and domains

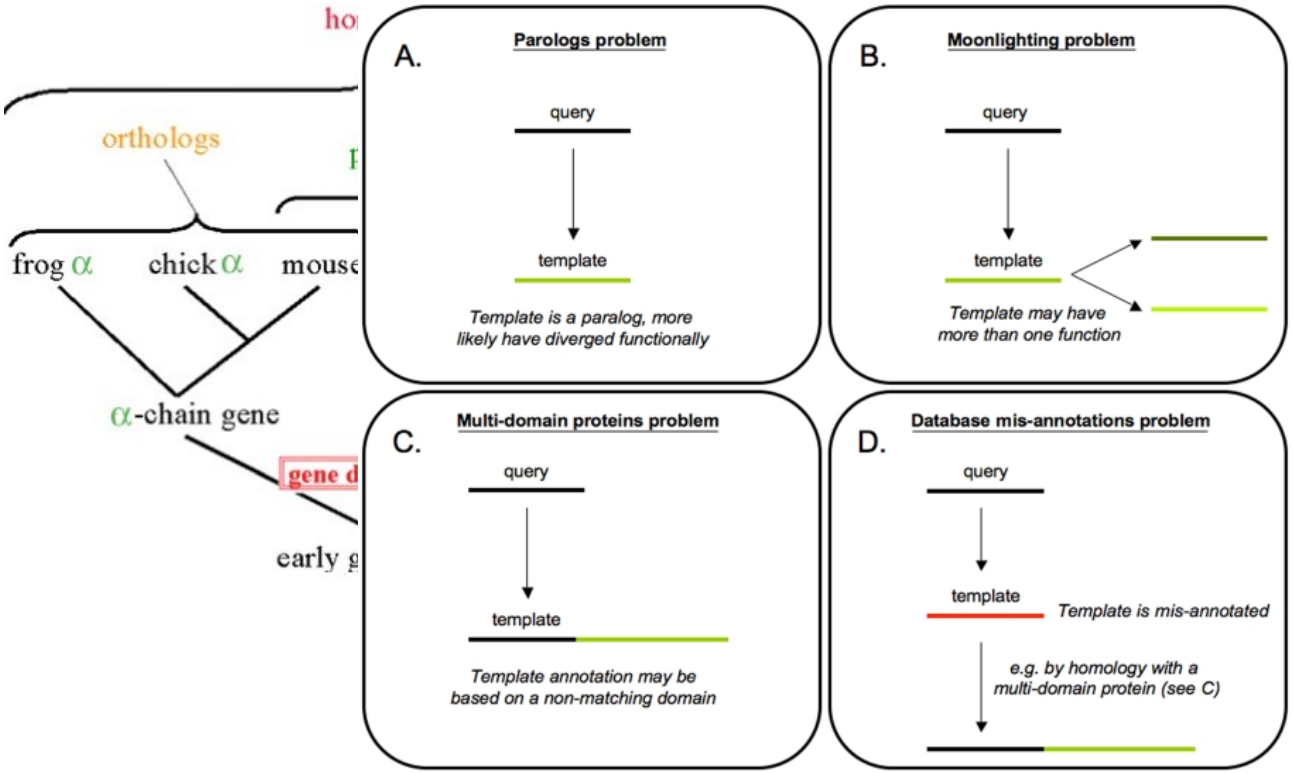
"Local" versus "Global" HMMs, e.g., Tigrfam "definitions"

- Equivalog – Full-length, all members share the same function
- Superfamily – full-length similarity, same domain architecture, but not same function
- Domain – a shorter region of homology (10-100 residuues) Seq similarity with or without associated function (e.g., ATP-binding site)
- (Pfam has different "definitions" or classifications)

- Limited organization into functional hierarchies or classification systems

Homology ≠ similarity of function.

- No scoring scheme provides “biological truth” – any pair of sequences can be aligned – finding meaning is up to you!
- Other heuristics – context-based, best reciprocal hit



Punta & Ofran. PLOS Comp Biol. 2008

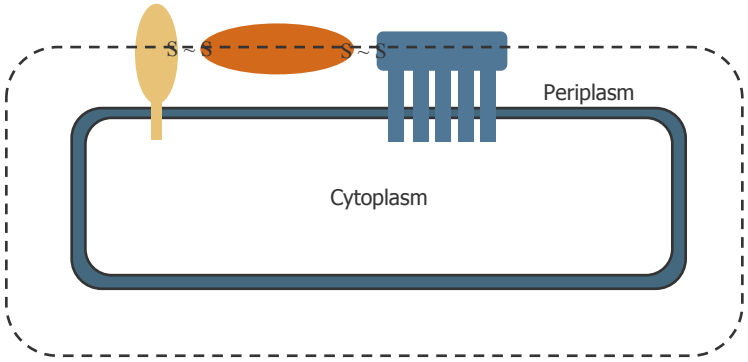
Errors in databases

There are many errors in the primary sequence databases:

- In the sequences themselves:
  - sequencing errors.
  - cloning vectors sequences.
- Gene Calling errors
- In the annotations:
  - “genome rot” - inaccuracies, omissions, mistakes.
  - inconsistencies between fields.

When there is “no similarity”?

- Gene Context
- Sub-cellular localization
- Topological features
- Prediction of binding residues



img.jgi.doe.gov

JGI JOINT GENOME INSTITUTE

JGI IMG/ER INTEGRATED MICROBIAL GENOMES / EXPERT REVIEW

Quick Genome Search: Go

Hi Rekha Seshadri | Logout
(JGI SSO) 21632

My Analysis Carts: 0 Genomes | 0 Scaffolds | 0 Functions | 0 G

Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | OMICS | Workspace | My IMG | Da

IMG/ER Content

	Datasets JGI	All
Bacteria	6341	43428
Archaea	373	1175
Eukarya	31	257
Plasmids	1	1220
Viruses		5185
Genome Fragments		1196
Metagenome & Metatranscriptome	4886	10333
Total Datasets		62794
My Private Datasets		12294

Last Datasets Added On:

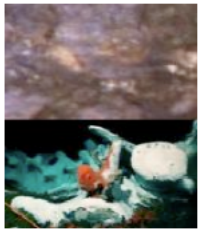
Genome	2016-05-18
Metagenome	2016-06-07

Project Map
Metagenome Projects Map
System Requirements

Hands on training available at the

Microbial Genomics & Metagenomics Workshop

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available metagenome datasets.



IMG/ER provides users with tools (IMG/ER UI Map) for analyzing their private (password protected access) genome datasets (http://nar.oxfordjournals.org/content/42/D1/D560) and/or metagenome datasets (http://nar.oxfordjournals.org/content/42/D1/D568) in the context of all public (free access) genome and metagenome datasets in IMG.

IMG/ER Statistics | **Data Submission Site**

Data management system for comparative analysis of -omic data

Metagenome and Metatranscriptome dataset distribution:

	Engineered		Environmental		Host-associated	
Sequenced at:	JGI	All	JGI	All	JGI	All
Metagenome	506	911	3087	4716	456	3140
Metatranscriptome	103	153	633	960	100	294

IMG contains 249 public studies, 5606 public metagenome datasets (5106 unique samples) distributed as follows:
(Public Metagenome count / Public Metatranscriptome count)

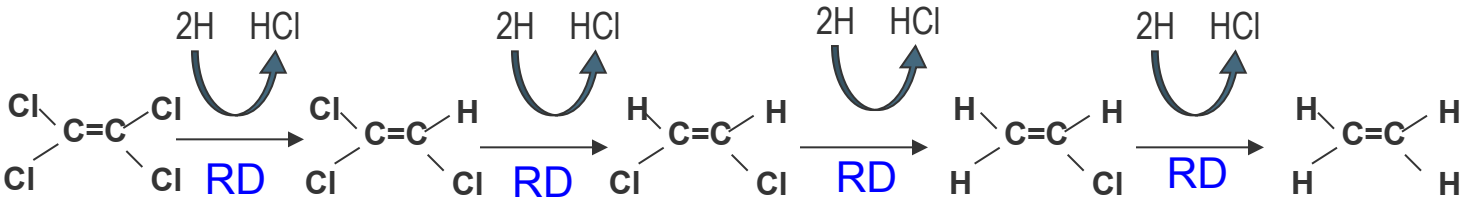
Engineered	474 / 117	Environmental	2892 / 623	Host-associated	1397 / 102
Bioreactor	15 / 1	Air	31 / 0	Algae	52 / 0
Bioremediation	45 / 0	Aquatic	1832 / 383	Animal	1 / 0
Biotransformation	21 / 8	Terrestrial	1029 / 240	Annelida	80 / 0
Food production	2 / 1			Arthropoda	79 / 2
Lab enrichment	97 / 6			Birds	10 / 0
Lab synthesis	3 / 0			Cnidaria	3 / 24
Modeled	1 / 0			Human	875 / 0
Solid waste	29 / 0			Mammals	32 / 3

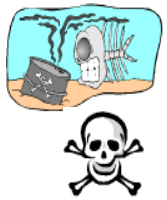
IMG – Integrated Microbial Genomes

https://img.jgi.doe.gov/mer/

CASE STUDY: Organohalide Respirers

- Major groundwater contaminant – PCE, TCE (industrial degreaser)
- Electron acceptor for **anaerobic dehalorespiration** – incomplete however
- **Serial dechlorination by** reductive dehalogenases (RD) **with cobalamin cofactor**



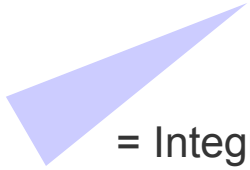


Known human carcinogens

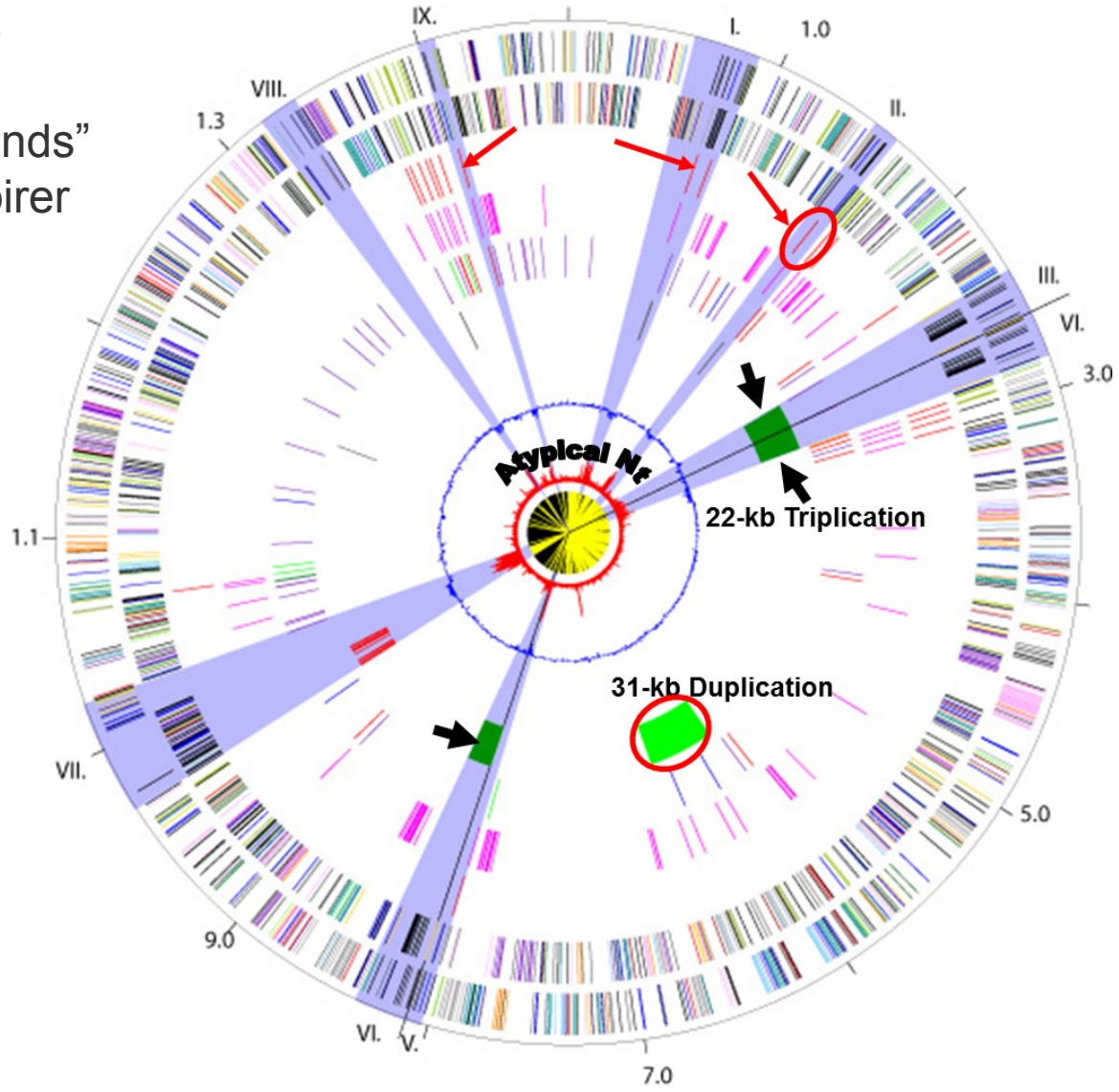
- *Dehalococcoides mccartyi* strain 195 from Ithaca Sewage Treatment Plant - **COMPLETE dechlorination** *(Maymo-Gatell, Science 1997: 276(5318):1568-1571)*

Genome Size: 1.46 Mbp

- Streamlined genome
- 13.6% integrated “islands”
- Dedicated dehalorespirer



= Integrated element

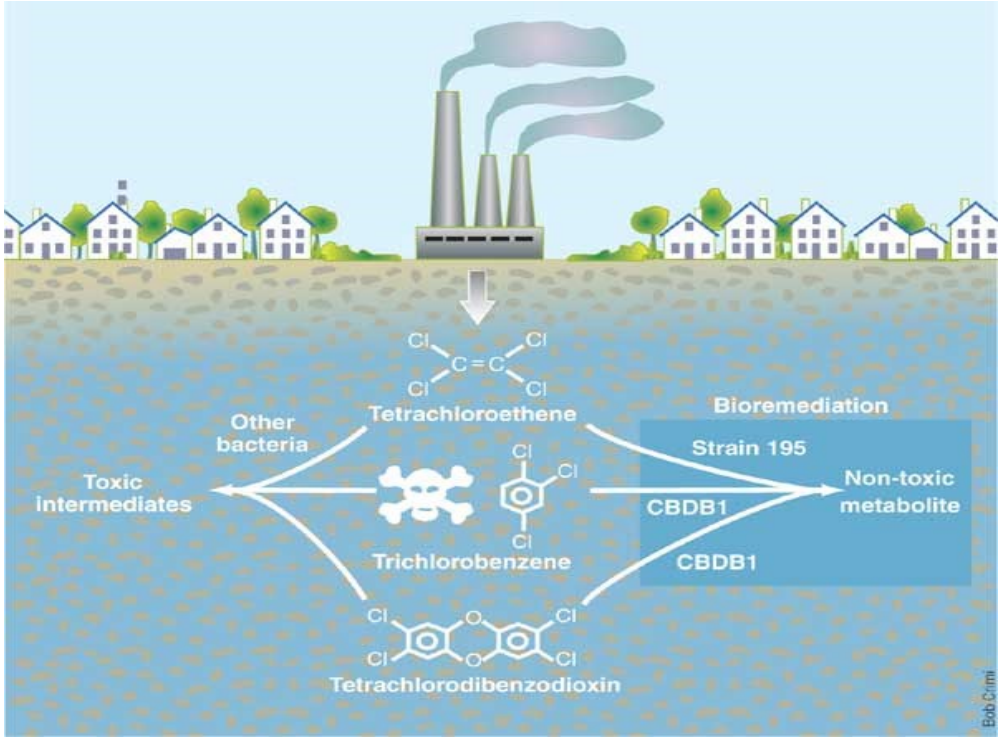


Seshadri et. al., Science 2005 : 307:105–108

Objective: Compare 195 vs CBDB1

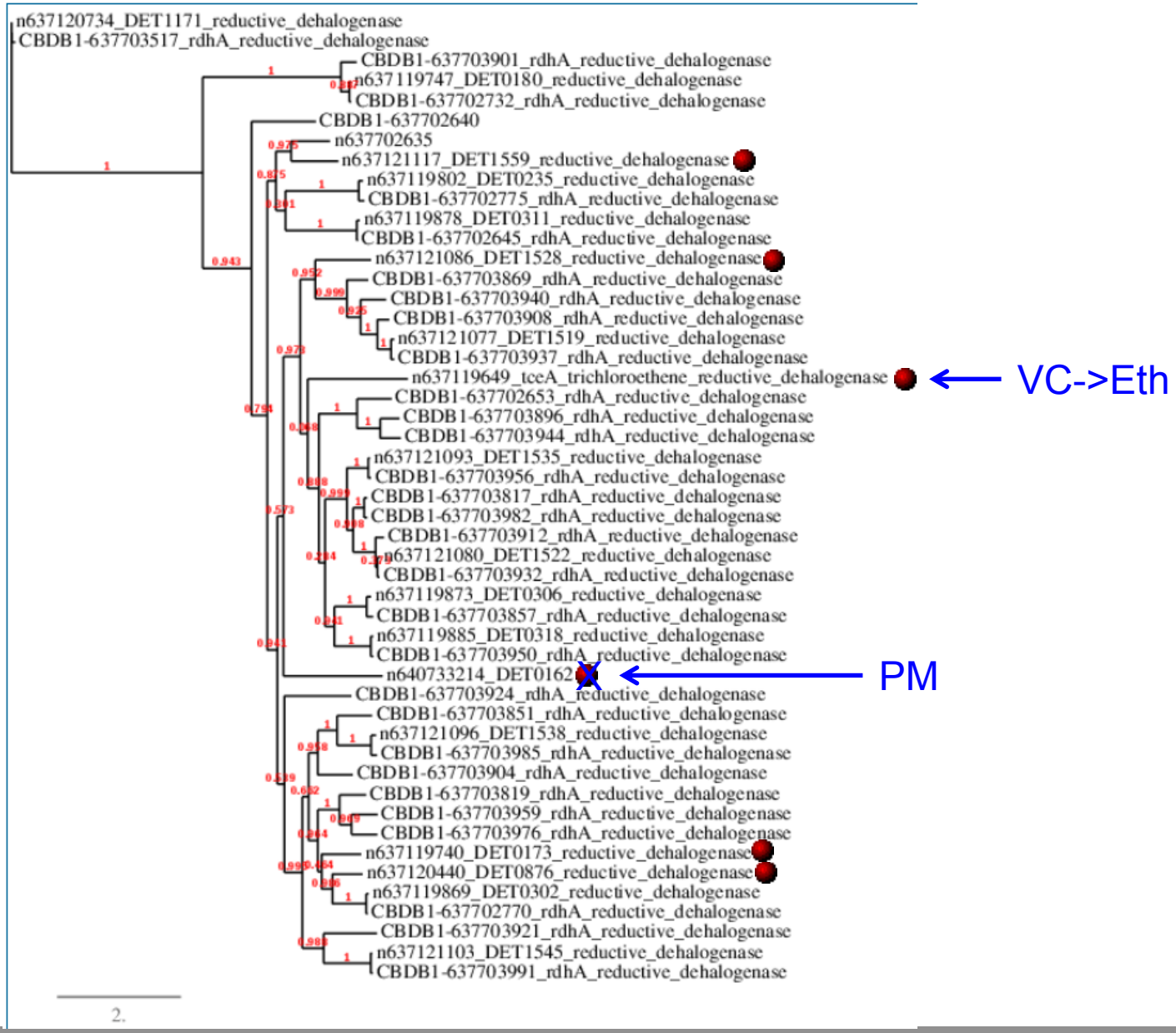
Strain 195	Strain CBDB1
1.467 Mbp (1590 CDS)	1.395 Mbp (1458 CDS)
Chloroethylenes, etc	Chlorobenzenes, etc
PCE->TCE->DCE**->VC->Eth**	PCE->TCE->DCE
17(+2) RDs	32 RDs

Nature Biotech. 23, 1269-73 (2005) Kube et. al.



- Is there any synteny?
- What proportion of genes are shared?
- How many unique?
- Can we find the RDs responsible for the terminal step(s)?
- For step-by-step solutions with screenshots, see: Seshadri_MGM-23_gene_discovery_demo.pdf

Tree of RDs from 195 and CBDB1



Objective: Compare str 195 vs CBDB1

Strain 195	Strain CBDB1
1.467 Mbp (1590 CDS)	1.395 Mbp (1458 CDS)
Chloroethylenes, etc	Chlorobenzenes, etc
PCE->TCE->DCE**->VC->Eth**	PCE->TCE->DCE
17(+2) RDs	32 RDs

Step-by-step solutions: Seshadri_MGM-23_gene_discovery_demo.pdf

- **Is there any synteny?** *Compare Genomes > Synteny Viewers > Dot Plot*
  - Yes, extensive, some rearrangement involving RDs, several gaps
- **What proportion of genes are conserved?** *Compare Genomes > Genome Genes Best Hmlgs*
  - ~80%
- **How many unique to 195?** *Find Genes > Phylogenetic Profilers > Single Genes*
  - ~20% including RDs and Nitrogen Fixation
- **Can we find the RDs responsible for the terminal step(s)?** *Compare Genomes > Abundance Profile Tools > Overview (All Functions) > Add RD genes to gene cart > Align Sequences & View Phylogram*
  - Yes, TceA, for example