Introduction of Metabolic Pathway Analysis

Usman Sumo Friend Tambunan

Arli Aditya Parikesit

Bioinformatics Group

Department of Mathematics

Faculty of Mathematics and Science

University of Indonesia

KEGG:
Kyoto Encyclopedia of Genes and Genomes

Susan Seo
Intro to Bioinformatics
Fall 2004

KEGG

Purpose & Overview

Features

Example

Purpose

- Developed at the Kanehisa Laboratory
- Integrates:
  - current knowledge of molecular interaction networks
  - information about genes and proteins
  - information about chemical compounds and reactions

Overview

PATHWAY database

GENES / SSDB / KO databases

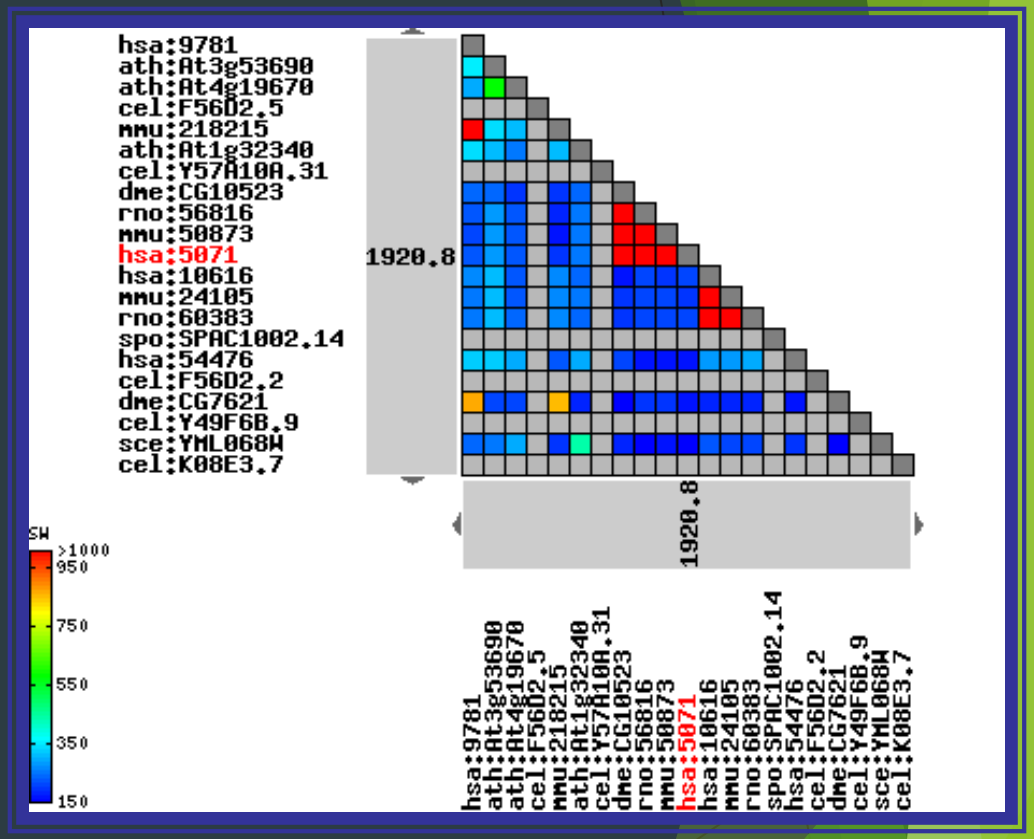
COMPOUND / GLYCAN / REACTION databases

Parkinson's Disease

- Search the Pathway database
- Explore a pathway linked with Parkinson's disease in humans
- Look more closely at the PARK2 gene

Features

- OC Viewer
- Gene catalog
- SSDB
- LinkDB
- Position
- Amino Acid sequence
- Nucleotide Sequence



EC system 0/2

- An Old, but still universally accepted system by biochemists
- EC system was developed long before protein sequence or structure information were available, so the system focuses on reaction, not sequence homology and structure
- Many biochemists and structural biologists try to harmonize newly available chemical, sequential, and structural data with traditional understanding of enzyme function.

Problems in EC system 1/2

- Inconsistency in the EC hierarchy
  - For each of the six top-level EC classes, subclasses and sub-subclasses may have different meanings.
  - e.g. EC1.* are divided by substrate type, but EC5.* by general isomerase type
- Problem with Multi-functional enzymes and multiple subunits involved in a function
  - EC presumes only a 1:1:1 relationship between gene, protein, and reaction.
- Different sequence/structure, but similar EC
  - Two enzymes with lower sequence identities sometimes belong to the same or very similar EC.
  - e.g. *o*-succinylbenzoate synthase across several bacteria have below the 40% sequence identity

Problems in EC system 2/2

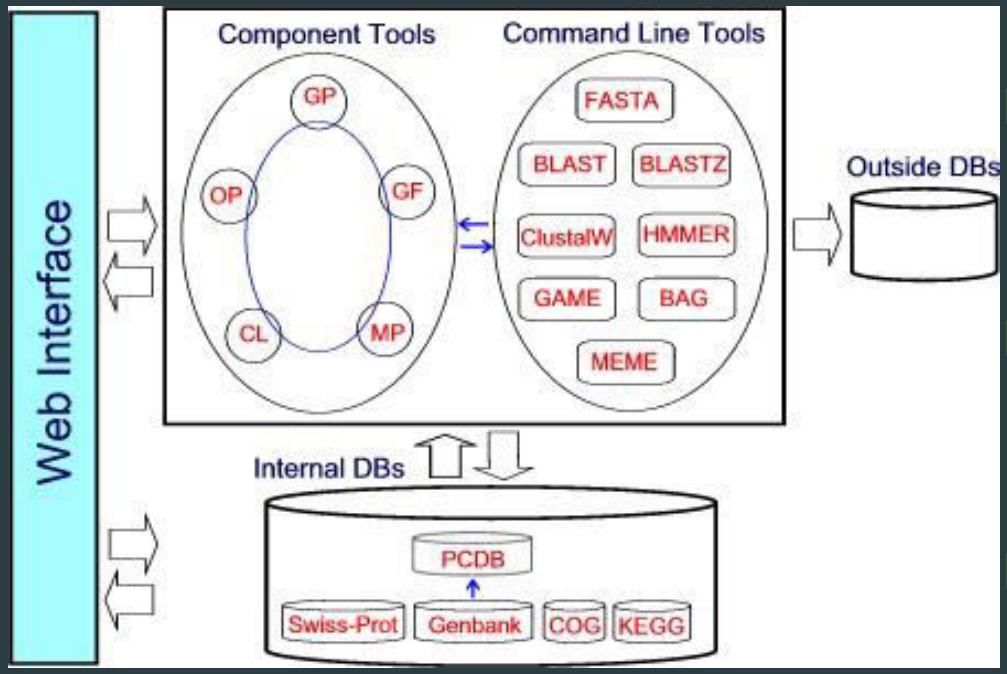
- Similar sequence/structure, but different EC
  - Even variation in the fourth digit of the EC number is rare above a sequence identity threshold of 40%.
  - However, exceptions to this rule are prevalent.
  - e.g. melamine deaminase and atrazine chlorohydrolase have 98% identical, but belong to different EC.
- No information on sequence/structure-mechanism relationship
  - EC system considers only overall transformation
  - Similarity among sequences is strongly correlated with similarities in the level of a common (structural domain-related) partial reaction, rather than overall transformation
  - How to combine enzyme structure data with partial reaction data?
- Research Goal
  - We provide a computational environment for enzyme analysis via genome comparison
  - And it will be built on PLATCOM system

Our Research Goal

- We provide a computational environment for enzyme analysis via multiple genome comparison
- And it will be built on PLATCOM system

PLATCOM

A Platform for Comparative Genomics



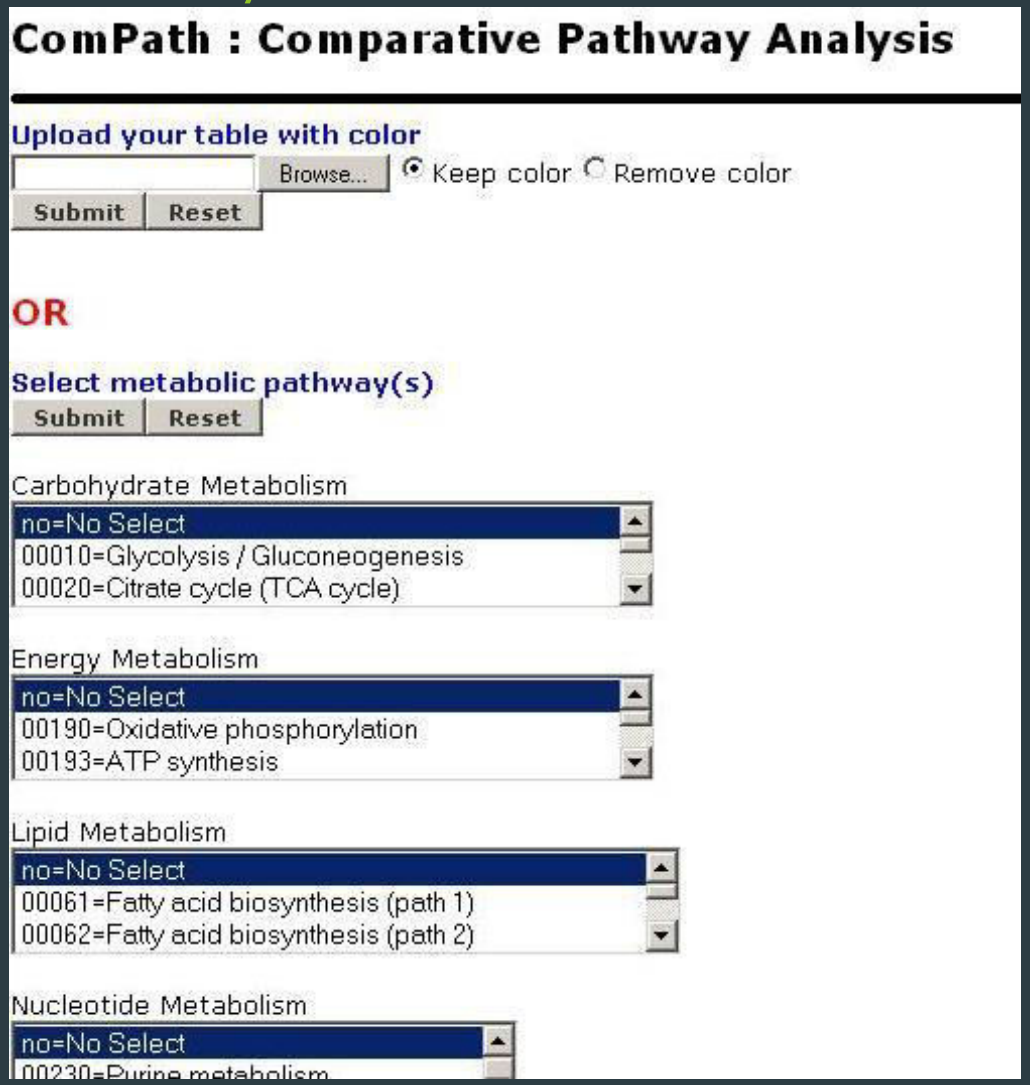
- Providing a platform for comparative genomics ON THE WEB
- Comparative analysis system for users to freely select any sets of genomes
- Scalable system interactively combining high-performance sequence analysis tools

CURRENT IMPLEMENTATION

ComPath

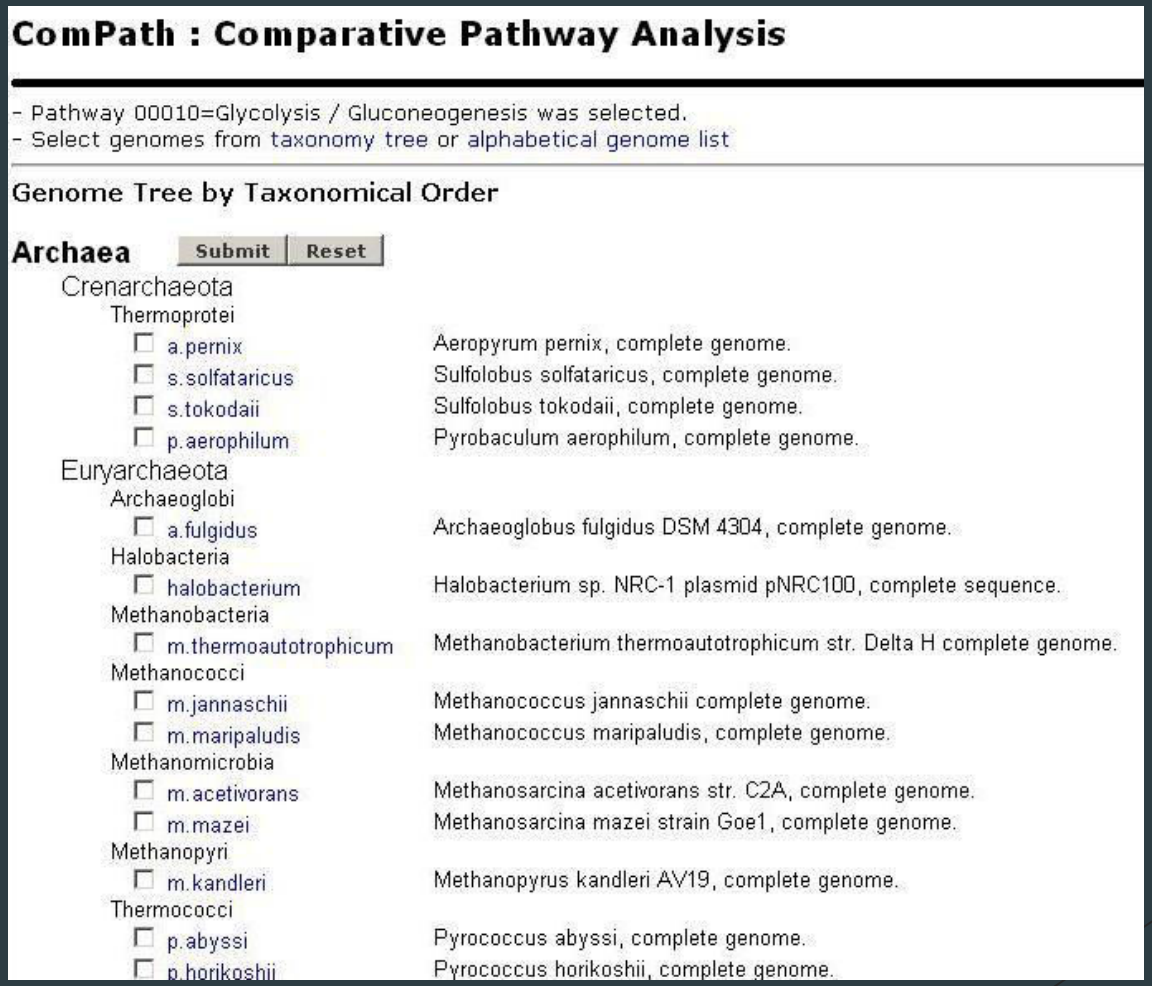
- ComPath = KEGG + PLATCOM
- Not just for retrieving information from Database,
- but focuses on analyzing enzymes *using the enzyme-genome table*
- Easy to use
  - **{Optional}** Upload a user sequence and/or a saved enzyme-genome table data
  - Select a metabolic pathway
  - Select any combination of genomes in KEGG
  - Create an enzyme-genome table
  - Then use the table for various enzyme sequence analysis tasks

Screenshot:
Pathway Selection



- 11 categories
- 123 pathways
- Users can upload the previous Enzyme-Genome table datatype to continue analysis

Screenshot:
Genome Selection



ComPath : Comparative Pathway Analysis

- Pathway 00010=Glycolysis / Gluconeogenesis was selected.
- Select genomes from taxonomy tree or alphabetical genome list

Genome Tree by Taxonomical Order

Archaea Submit Reset

Crenarchaeota
Thermoprotei
- a.pernix — Aeropyrum pernix, complete genome.
- s.solfataricus — Sulfolobus solfataricus, complete genome.
- s.tokodaii — Sulfolobus tokodaii, complete genome.
- p.aerophilum — Pyrobaculum aerophilum, complete genome.

Euryarchaeota
Archaeoglobi
- a.fulgidus — Archaeoglobus fulgidus DSM 4304, complete genome.

Halobacteria
- halobacterium — Halobacterium sp. NRC-1 plasmid pNRC100, complete sequence.

Methanobacteria
- m.thermoautotrophicum — Methanobacterium thermoautotrophicum str. Delta H complete genome.

Methanococci
- m.jannaschii — Methanococcus jannaschii complete genome.
- m.maripaludis — Methanococcus maripaludis, complete genome.

Methanomicrobia
- m.acetivorans — Methanosarcina acetivorans str. C2A, complete genome.
- m.mazei — Methanosarcina mazei strain Goe1, complete genome.

Methanopyri
- m.kandleri — Methanopyrus kandleri AV19, complete genome.

Thermococci
- p.abyssi — Pyrococcus abyssi, complete genome.
- p.horikoshii — Pyrococcus horikoshii, complete genome.

- 250 genomes from KEGG database
- Users can select genomes by taxonomical and alphabetical order

Enzyme-Genome Table

- An enzyme-genome table allows for tests on whether a certain enzyme in a given pathway is present or missing using sequence analysis techniques.
- Information in this table can be easily saved, uploaded, transferred.
- Users also can upload their sequence set, e.g., an entire set of predicted proteins in a newly sequenced genome, and perform annotation of the sequences in terms of KEGG pathways.

Screenshot:
KEGG's Ortholog Table – STATIC!

Organism	E2.7.1.11	E4.1.2.13		E5.3.1.1	E1.2.1.12	E2.7.2.3	E5.4.2.1
	Phospho-fructokinase	Aldolase		Triose-phosphate isomerase	Glyceraldehyde-3P dehydrogenase	Phosphoglycerate kinase	Phosphoglycerate mutase
		class II	class I				
hsa [P | G | T]	5211(PFKL) 5213(PFKM) 5214(PFKP)		226(ALDOA) 229(ALDOB) 230(ALDOC)	7167(TPI1)	2597(GAPD) 26330(GAPDS)	5230(PGK1) 5232(PGK2)	441531 (LOC441531) 5223(PGAM1) 5224(PGAM2) 669(BPGM)
mmu [P | G | T]	18641(Pfkl) 18642(Pfkm) 56421(Pfkp)		11674(Aldoa) 11676(Aldoc) 230163(Aldob)	21991(Tpi1)	14433(LOC14433) 14447(Gapds) 407972(Gapd)	18655(Pgk1) 18663(Pgk2) 432633 (LOC432633)	12183(Bpgm) 18648(Pgam1) 56012(Pgam2)
rno [P | G | T]	25741(Pfkl) 65152(Pfkm)		24189(Aldoa) 24190(Aldob) 24191(Aldoc)	246267 (LOC246267) 24849(Tpi1)	24383(Gapd) 66020(Gapds)	24644(Pgk1)	24642(Pgam1) 24959(Pgam2)
gga [P | G | T]							
dre [P | G | T]			321664 369193	192309 192310	406367		
dme [P | G | T]	CG4001-PA CG4001-PB CG4001-PC		CG6058-PA CG6058-PB CG6058-PE CG6058-PF	CG2171-PA CG2171-PB	CG12055-PA CG8893-PA CG9010-PA	CG3127-PA CG9961-PA	CG1721-PA CG17645-PA CG7059-PA CG7059-PC CG7059-PD
cel [P | G | T]	C50F4.2 Y71H10A.1a Y71H10A.1b		T05D4.1	Y17G7B.7	F33H1.2 K10B3.7 K10B3.8 T09F3.3	T03F1.3	

Screenshot:

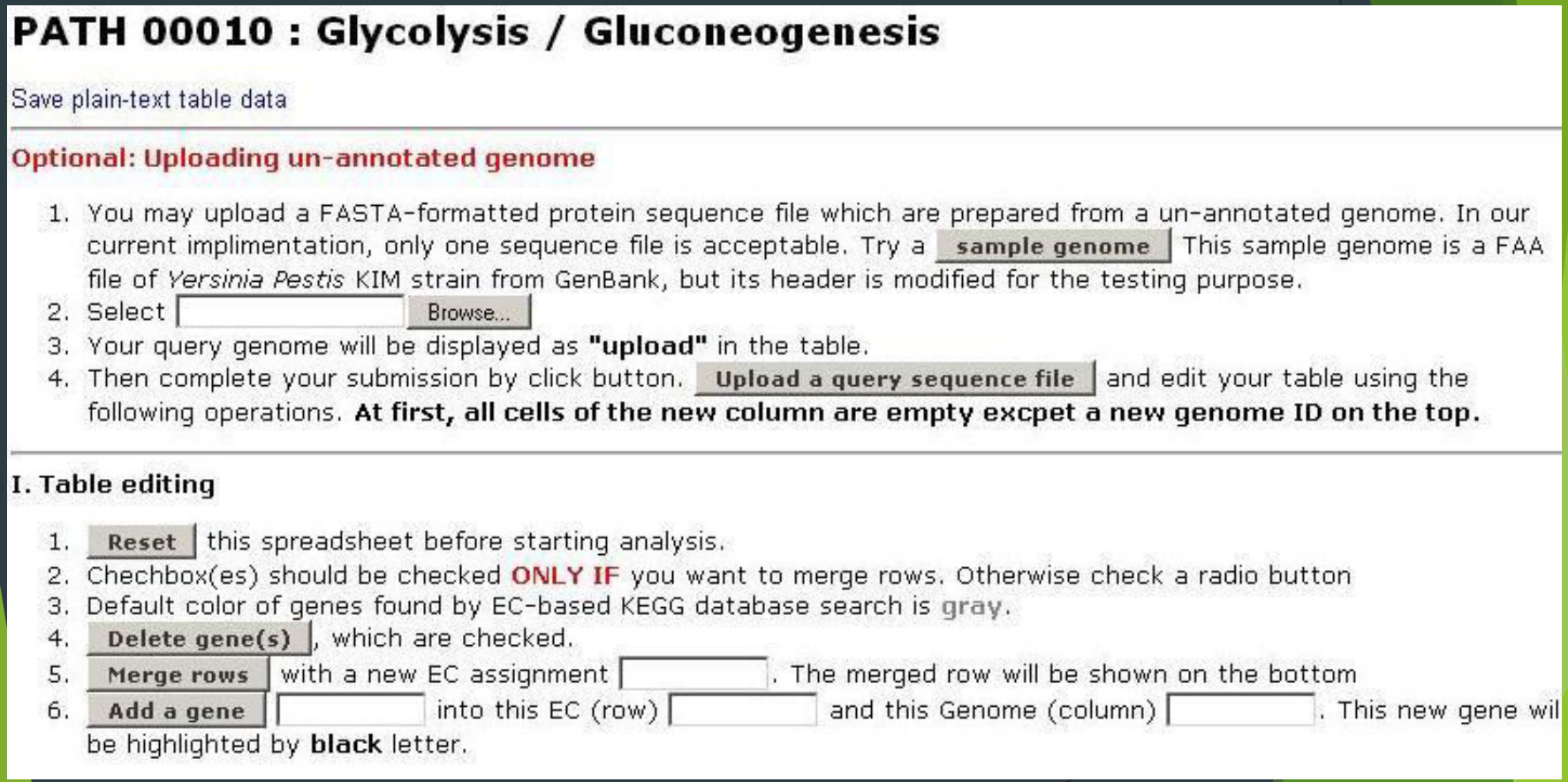
ComPath' Enzyme-Genome Table – INTERACTIVE!

IV. Genome-EC table

1. Genomes to be searched can be limited by checking checkbox(es) on the top row. If not, ComPath uses all genomes in this table.
2. Please make sure that an EC number is set from the third column.

Merge	All/None	Select One	mtu	bsu	bha	hin	eco	aae	bsu	hpy	mge
☐	○ ○	○ 1.1.1.1	☐ Rv0162c ☐ Rv0761c ☐ Rv1530 ☐ Rv1862	☐ BG11902 ☐ BG11941 ☐ BG13553	☐ BH1829	☐ HI0185	☐ b0356 ☐ b1241 ☐ b1478 ☐ b3589	☐ aq_1240 ☐ aq_1362	☐ BG11902 ☐ BG11941 ☐ BG13553		
☐	○ ○	○ 1.1.1.2	☐ Rv3045	☐ BG12562					☐ BG12562	☐ HP1104	
☐	○ ○	○ 1.1.1.27		☐ BG19003	☐ BH3937				☐ BG19003		☐ MG460
☐	○ ○	○ 1.1.1.37	☐ Rv1240	☐ BG11386 ☐ BG13206	☐ BH3158	☐ HI1031 ☐ HI1210	☐ b0801 ☐ b3236 ☐ b3575	☐ aq_1665 ☐ aq_1782	☐ BG11386 ☐ BG13206		
☐	○ ○	○ 1.1.99.8									
☐	○ ○	○ 1.2.1.1	☐ Rv0761c	☐ BG11902	☐ BH1829	☐ HI0185	☐ b0356		☐ BG11902		
☐	○ ○	○ 1.2.1.10	☐ Rv3535c				☐ b0351 ☐ b1241				
☐	○ ○	○ 1.2.1.12	☐ Rv1436	☐ BG10827 ☐ BG12592	☐ BH3149 ☐ BH3560	☐ HI0001	☐ b1779	☐ aq_1065	☐ BG10827 ☐ BG12592	☐ HP0921 ☐ HP1346	☐ MG301
☐	○ ○	○ 1.2.1.19									
Merge	All/None	Select One	mtu	bsu	bha	hin	eco	aae	bsu	hpy	mge
☐	○ ○	○ 1.2.1.3	☐ Rv0147 ☐ Rv0223c	☐ BG11355 ☐ BG11903	☐ BH0539 ☐ BH0681				☐ BG11355 ☐ BG11903		

Screenshot:
Upload Query Genome and Table Editing Functions



PATH 00010 : Glycolysis / Gluconeogenesis

Save plain-text table data

Optional: Uploading un-annotated genome

1. You may upload a FASTA-formatted protein sequence file which are prepared from a un-annotated genome. In our current implimentation, only one sequence file is acceptable. Try a [sample genome] This sample genome is a FAA file of *Yersinia Pestis* KIM strain from GenBank, but its header is modified for the testing purpose.
2. Select [ ] [Browse...]
3. Your query genome will be displayed as **"upload"** in the table.
4. Then complete your submission by click button. [Upload a query sequence file] and edit your table using the following operations. **At first, all cells of the new column are empty excpet a new genome ID on the top.**

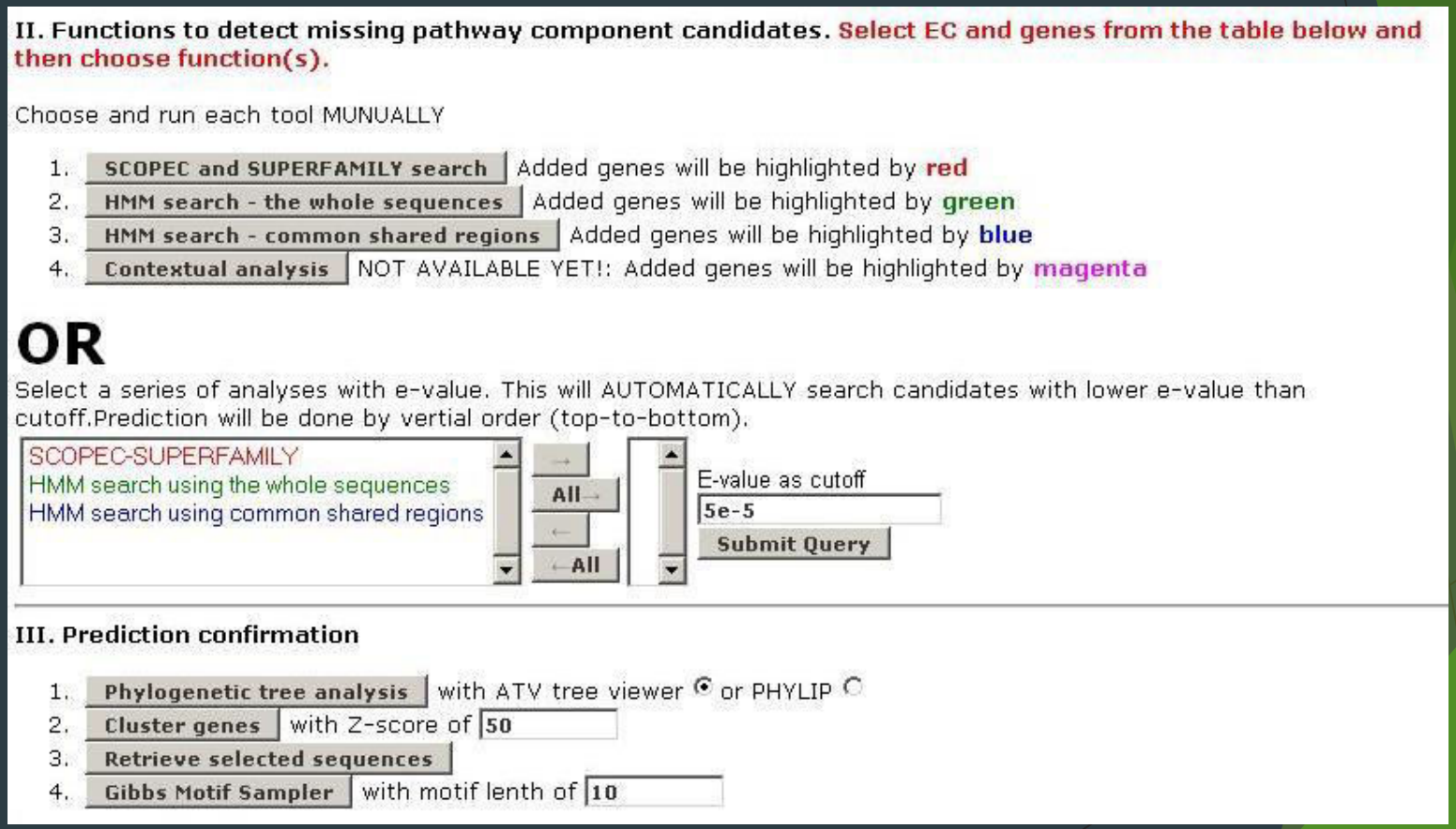
I. Table editing

1. [Reset] this spreadsheet before starting analysis.
2. Chechbox(es) should be checked **ONLY IF** you want to merge rows. Otherwise check a radio button
3. Default color of genes found by EC-based KEGG database search is gray.
4. [Delete gene(s)], which are checked.
5. [Merge rows] with a new EC assignment [ ]. The merged row will be shown on the bottom
6. [Add a gene] [ ] into this EC (row) [ ] and this Genome (column) [ ]. This new gene wil be highlighted by **black** letter.

Sequence Analyses

- Missing enzyme search
  - Pairwise (FASTA) and multiple sequence alignment (CLUSTALW),
  - Domain search using SCOPEC/SUPERFAMILY and PDB domains
  - Domain-based analysis using hidden markov models (HMM),
  - Contextual sequence analysis (currently not available)
- Sequence analysis for further investigation
  - Phylogenetic analysis of enzymes in selected genomes,
  - Gibbs motif sampler.
  - BAG clustering
  - Contextual sequence analysis (currently not available)

Screenshot:
Sequence Analysis Functions



Conclusion

The Kyoto Encyclopedia of Genes and Genomes is a vast library of information gathered from fully sequenced genomes, genes, proteins, pathways, and chemical compounds pertaining to over a hundred different species of both prokaryotes and eukaryotes

ComPath is one of the tools that could aid the data extraction from KEGG

References

- http://bioinformatics.bc.edu/~clotelab/misc/PPT/keggSeo.ppt
- bio.informatics.indiana.edu/sunkim/talks/
- http://www.docstoc.com/docs/512126/ComPath-Comparative-Metabolic-Pathway-Analysis-Tool