國立清華大學
National Tsing Hua University

TBI

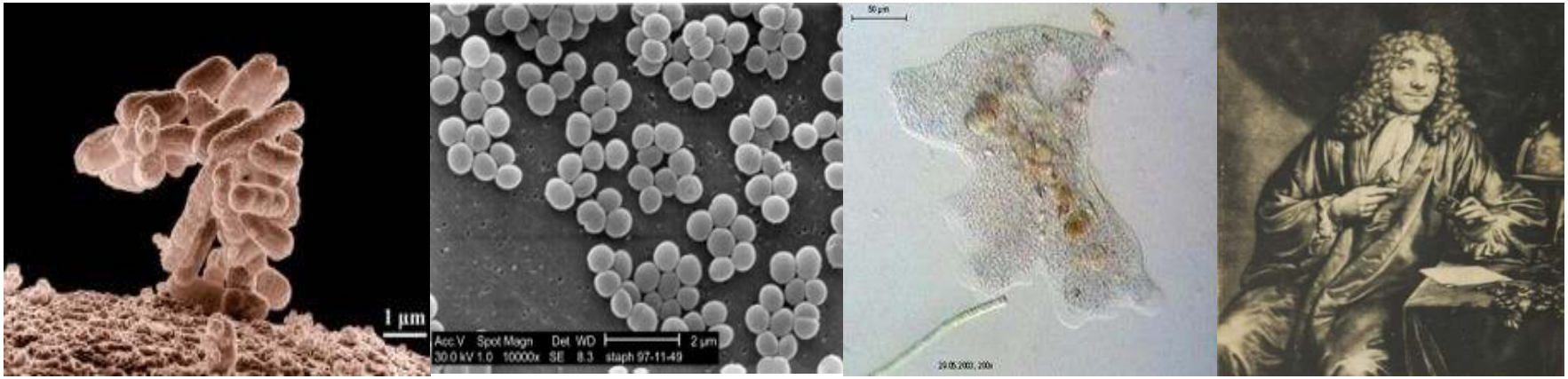
Proteome analysis tool for Microorganisms

GPDB (Genome Profile DataBase)

李季青

國立清華大學 資訊工程系

Introduction: Microorganisms



- Various morphological and environmental distributing properties
- Microbial identification and classification is more difficult than higher organisms.

Conserved molecular biomarkers

- Since 1980s
- 16S rRNA gene (1500 bp)
  - ⇒ Ubiquity
  - ⇒ Conserved region
  - ⇒ Variable region
- 16S rRNA + Other biomarkers → modern taxonomy

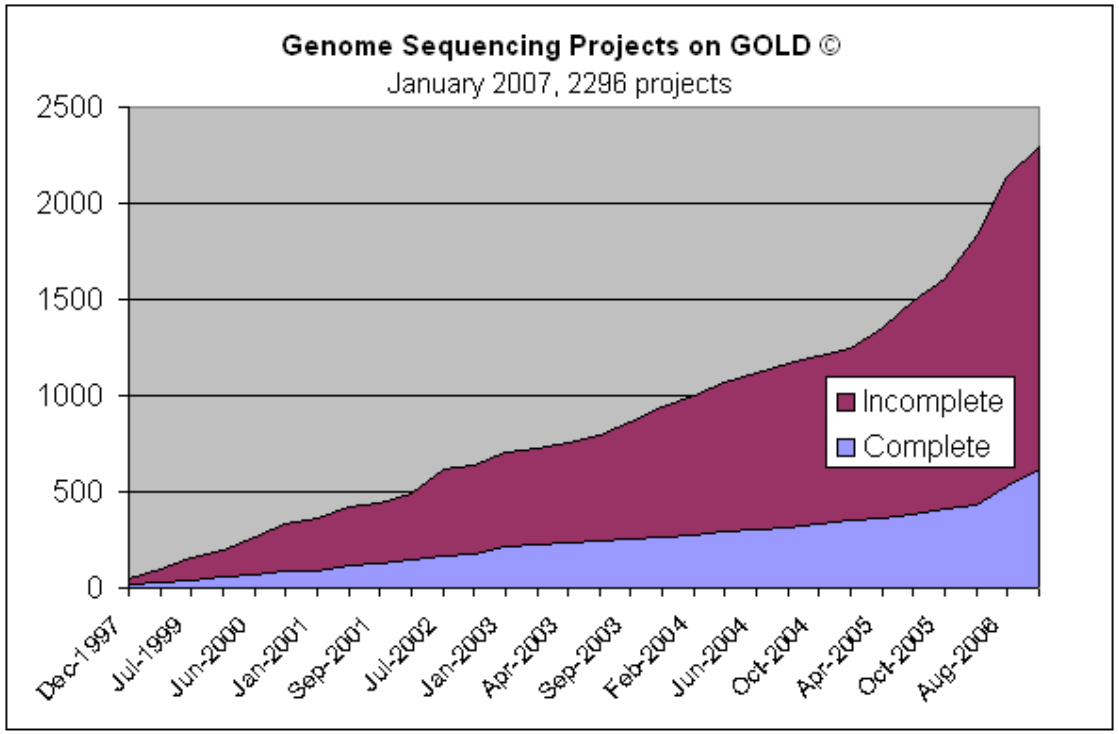
Biochemical typing
Metabolic tests
DNA-DNA hybridization
GC contents
Chemotaxonomic markers
...

⇨ May not universal of all organisms

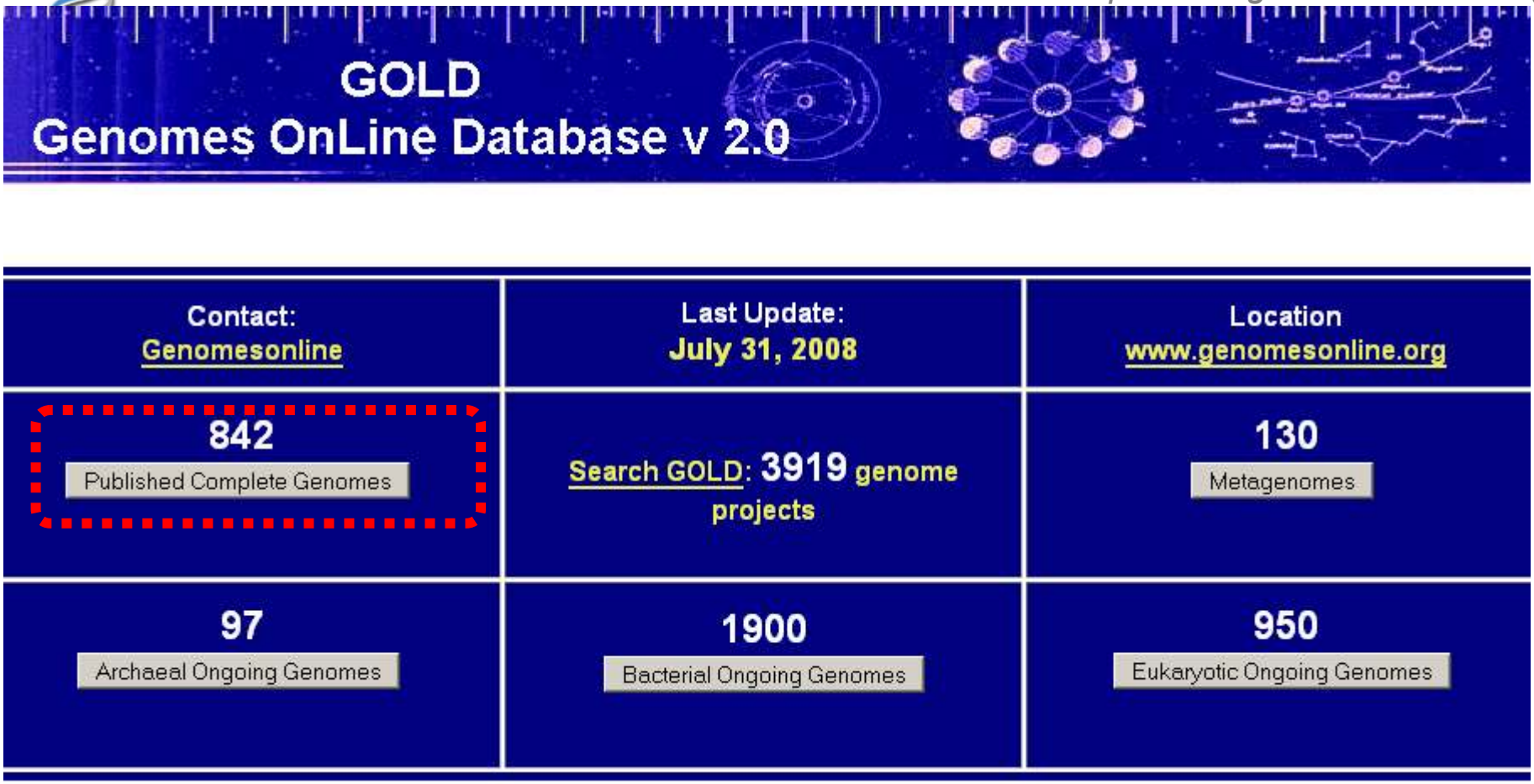
Genome sequences

How about biochemical / metabolic features, life cycle, GC-contents, type of host cell … ?

They are determined by entire genome.



Data from: http://www.genomesonline.org/

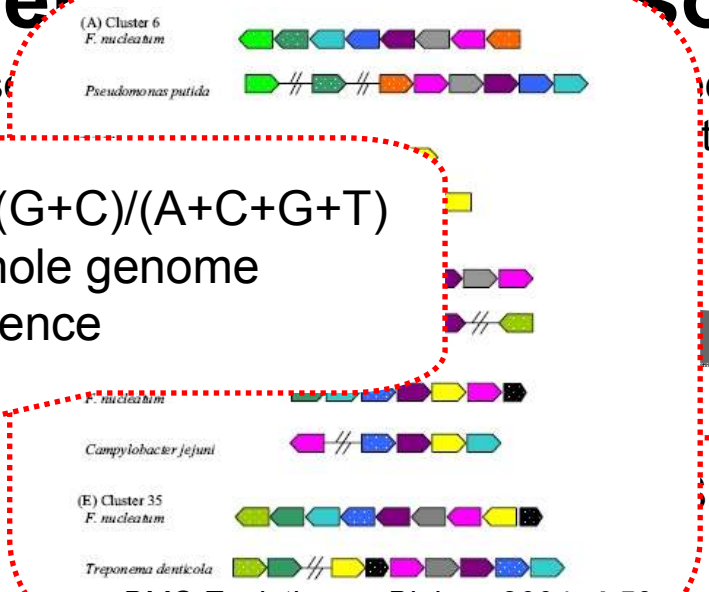


* Most of complete genomes are microbes

- Valuable sources for comparative genomics studies

Genome comparison methods

Long multiple-s… …cause evolution events (gene rearrangement… …t such high rates.



*BMC Evolutionary Biology 2004, **4:50***

100*(G+C)/(A+C+G+T) of whole genome sequence

GC content

…hared gene content

Whole genome sequence alignment

Alignment-free statistical method

Chromosome gene order

Resol…

Computing time

Quick and Dirty

How we study genomic data?

- Information derived from both nucleotide and protein sequence in a genome-wide scale.
- Provide and compare features of the fully sequenced organisms in a graphic and easy-reading way.
- On-line graphic browsing interface and use hierarchical clustering method to compare and view the difference between these organisms

The profile

Current Genome Profile

- Basic information -
  - ✓ Species name, taxonomy, # of chromosome/plamid , genome size, orf number…
- Nucleotide composition -
  - ✓ ATGC composition, GC/AT content, N-nucleotide frequency (n=2,3), Codon usage…
- Amino acid composition -
  - ✓ Amino acid group composition, N-peptide frequency distribution (n=1,2), Proteome length, Mw, pI distribution…

Basic information

GPDB Genome Profile DataBase

Home | Browse | Virtual 2D | Compare | Download | Status | Help | Comment

[Bacteria] - Helicobacter pylori 26695

Species Name :

Helicobacter pylori 26695 (Taxonomy id : 85962)

Genome list :

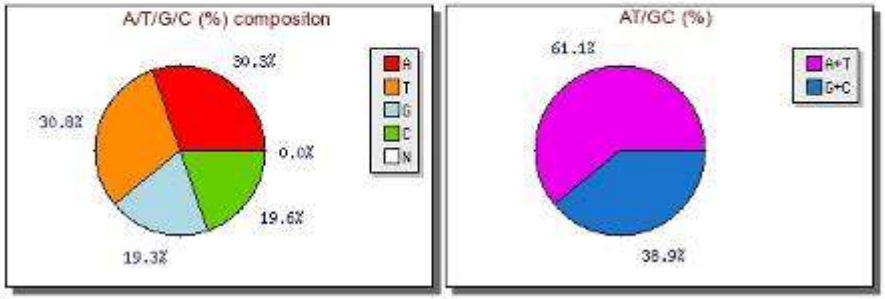
#	Accession	gi	Description	Size(bp)	Orfs
1	NC_000915	15644634	Helicobacter pylori 26695, complete genome	1667867	1576

Total Genome Size = 1.67 (Mb) , ORFs = 1576

Nucleotide composition :

A	T	G	C	N
30.3 % (505397 bp)	30.8 % (514075 bp)	19.3 % (321273 bp)	19.6 % (327080 bp)	0.0 % (42 bp)

A+T content = 61.12 %, G+C content = 38.87 %



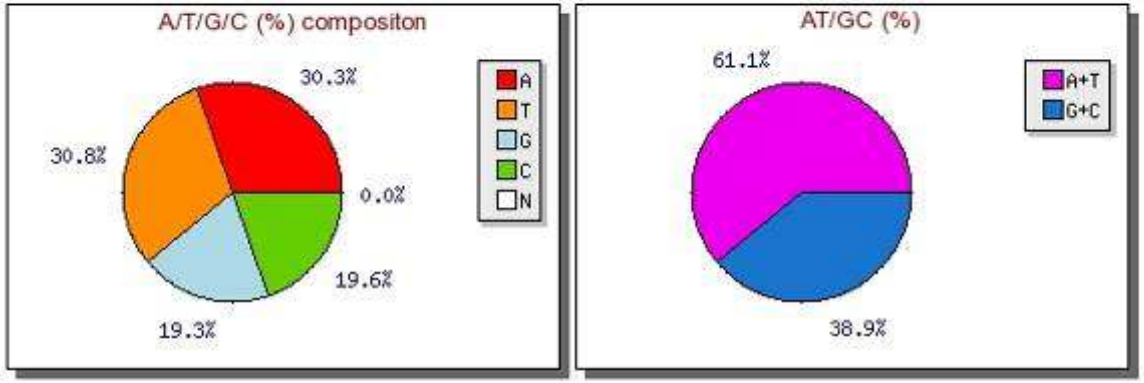
Nucleotide Composition

GC/AT content, GC/AT Skew

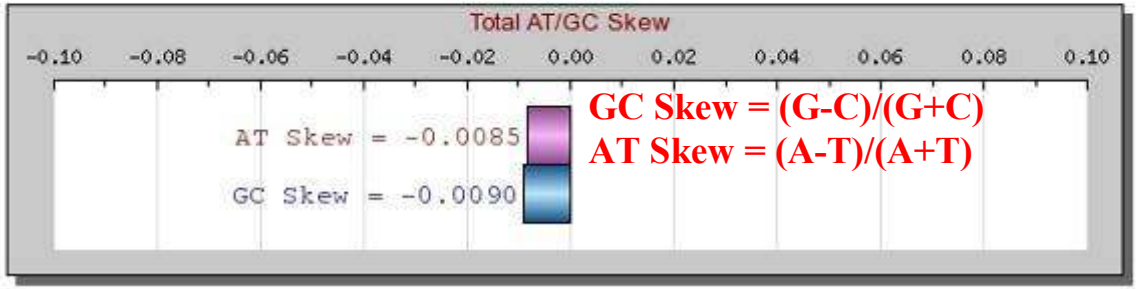
Nucleotide composition :

A	T	G	C	N
30.3 % (505397 bp)	30.8 % (514075 bp)	19.3 % (321273 bp)	19.6 % (327080 bp)	0.0 % (42 bp)

A+T content = 61.12 %, G+C content = 38.87 %



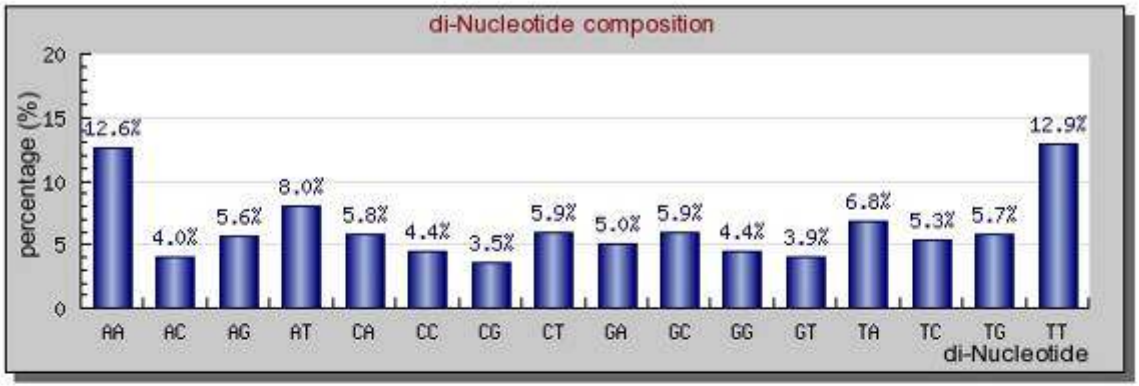
Total GC/AT Skew



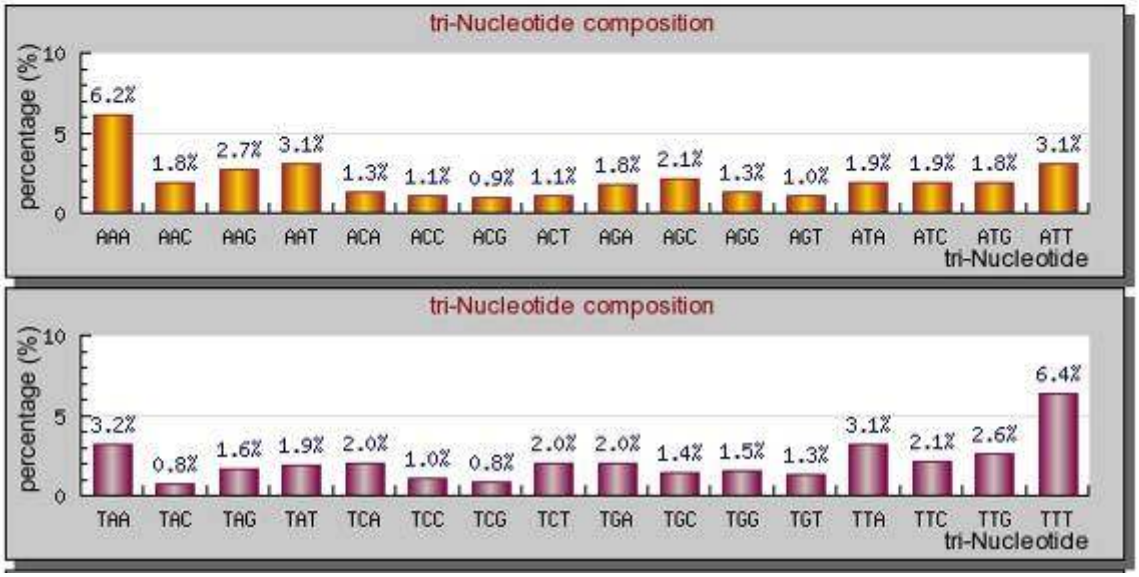
GC Skew = (G-C)/(G+C)
AT Skew = (A-T)/(A+T)

N-nucleotide Composition (N=2,3)

di-Nucleotide composition :



tri-Nucleotide composition :

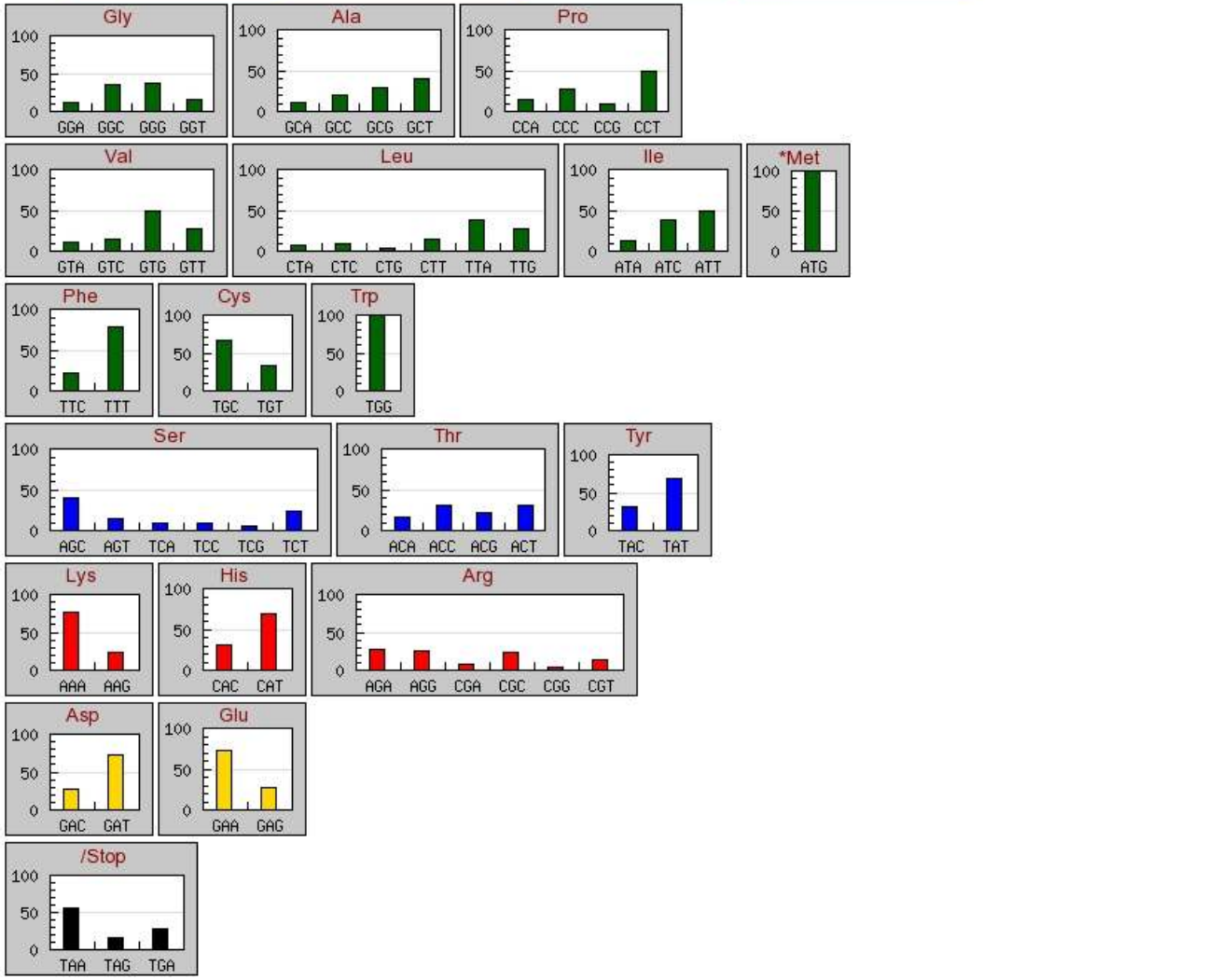


Codon Usage

Codon Usage :

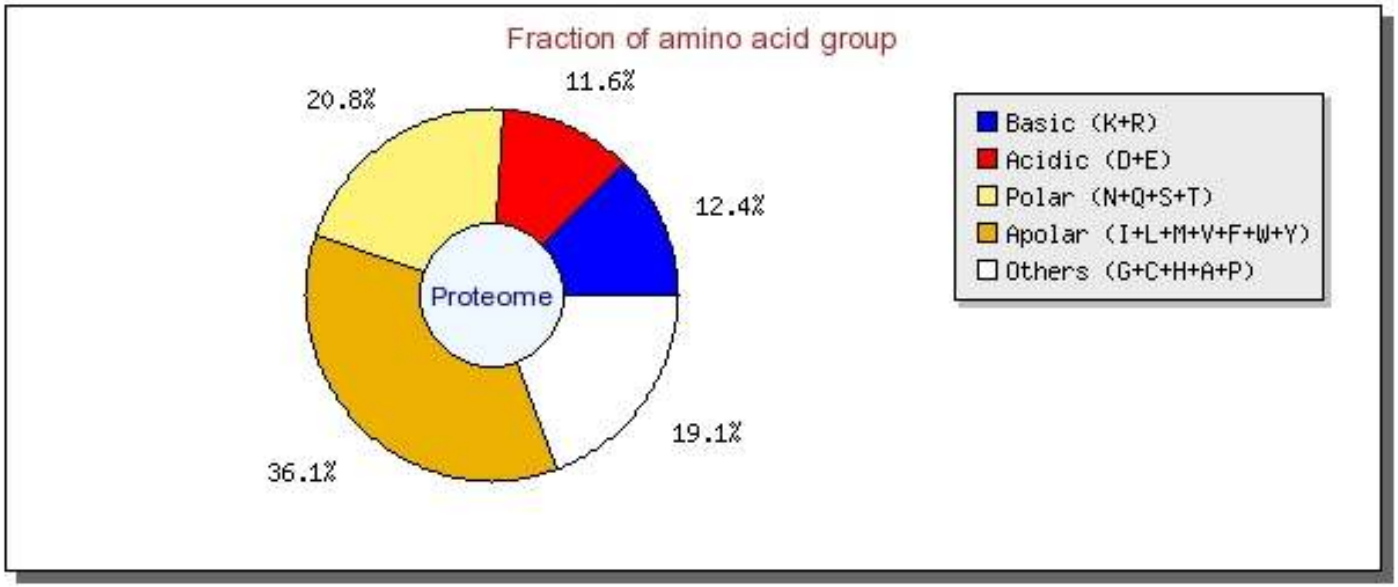
- **Total Codon Usage (%)**

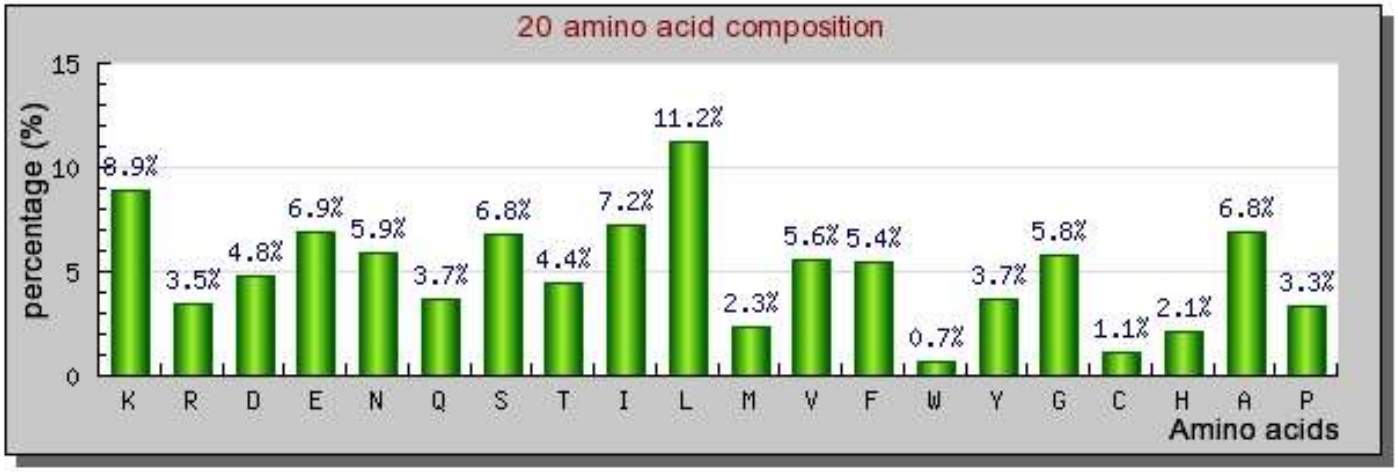
color: aploar amino-acid | polar amino acid | positive amino acid | negitive amino acid | stop codon



Amino Acid Composition

Amino acid composition :

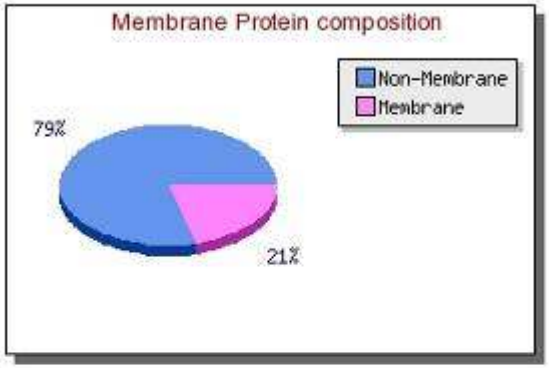




Proteome Distribution-
Non-membrane/Membrane protein & Charge distribution

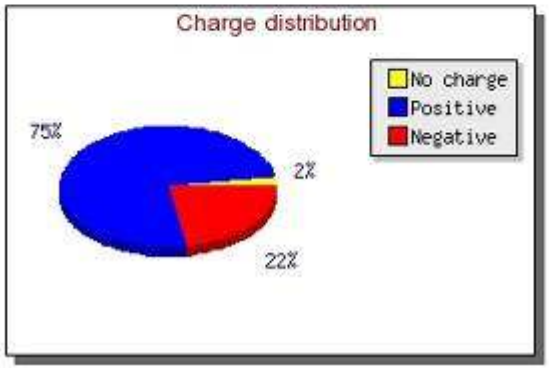
Proteome Distribution :

- **Non-Transmembrane/Transmembrane protein distribution**



TMHMM prediction

- **Charge distribution**



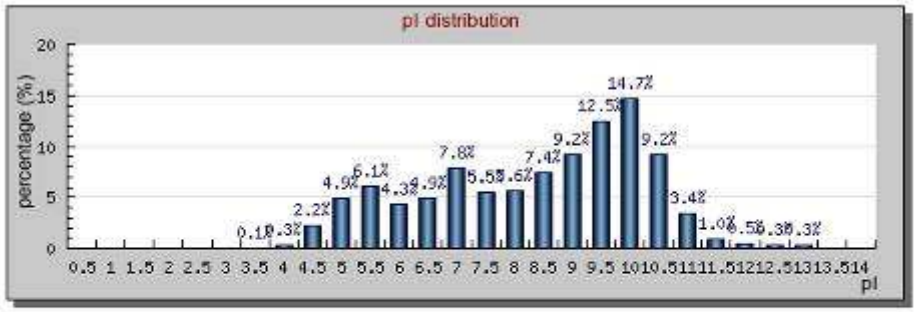
protein with

- **positive charge**
- **negitive charge**
- **no charge**

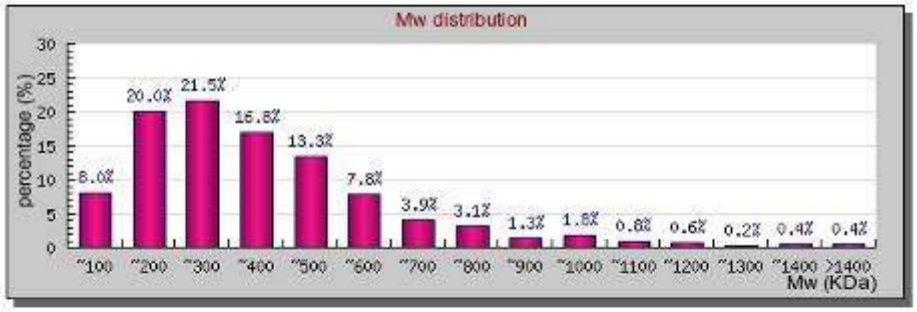
Proteome Distribution-

pI & Mw & Length

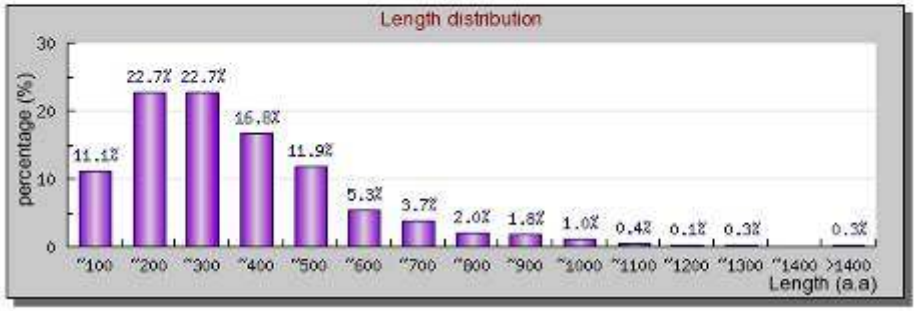
- **Isoelectric point(pI) distribution**



- **Molecular Weight distribution**



- **Length distribution**

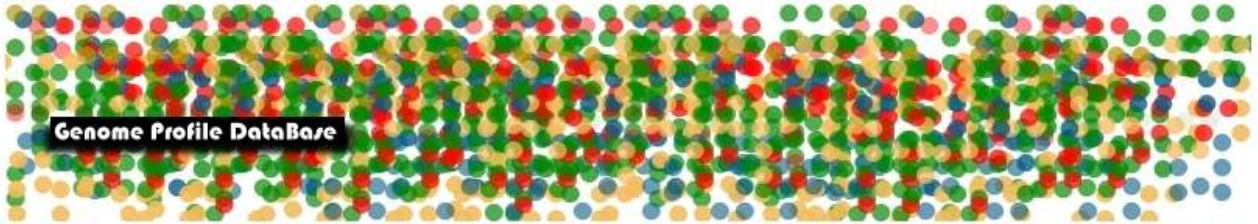


http://gpdb.life.nthu.edu.tw/

GPDB

Genome Profile DataBase

Home | Login | Free Registration | Comment



What's Genome Profile DataBase (GPDB) ?

GPDB has been developed to provide and compare features of the fully sequenced organisms in a graphic and easy-reading way. We focus on these prokaryotes including bacteria and archaea. In this post-genome era, we try to grab the information derived from both nucleotide and protein sequence in a genome-wide scale. We provide some "Genome Profile", develop on-line graphic browsing interface and use hierarchical clustering method to compare and view the difference between these organisms. The original sequence data and annotations are from NCBI GeneBank and RefSeq databases. The species name system is from Taxonomy database. We wrote the perl program called "Genome Profile Pipeline" which can automatically mirror data from NCBI ftp site and continue to parse, caculate, and analyse data.

Current Status

Last update 2006/8/3

345 organisms (27 Archaea, 338 Bacteria)

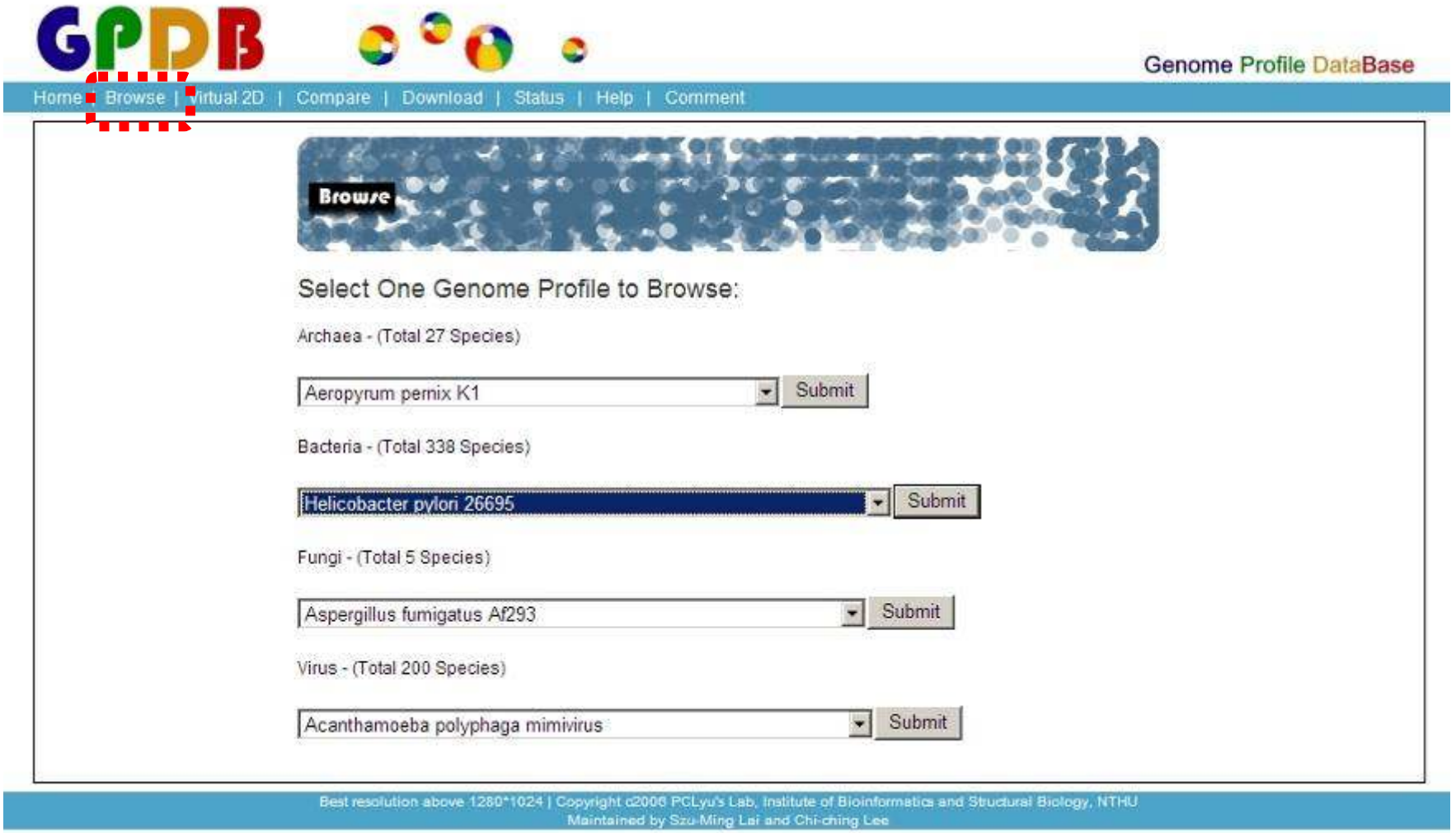
Development Platform

Powered by Apache+PHP+MySQL under RedHat Linux 9.0.

Version 2.5 Update Log

2006-10-06	Add registration function.
2006-08-24	Add sorting function.
2006-08-20	New style.
2006-08-12	New compare interface.
2006-08-03	Update database to 2006/08/03.
2006-07-	Fix automatically update scripts
2006-06-10	Java robot for collecting web resource.
2006-04-30	Automatically update scripts.
2006-02-06	Version 2.5 project engage.
2004-11-13	Add Virtual 2D menu to main page.
2004-04-04	Add Total Codon Usage Browse & Comparison.
2004-03-29	Free Registration Form added.
2004-03-25	Add Total GC/AT Skew and Transmembrane protein distribution
2004-03-24	Release Version 2.0
2004-03-23	Virtual 2D gel completed.
2004-03-19	Fix Mw/Length scale.
2004-02-26	The release version of GPDB is 1.0. Add four-type comparison options.
2003-12-15	Version 0.1. Platform builded including basic architecture ,database design, web interface.

Browse – *Helicobacter pylori* 26695



TBI

So many genome profile, and then ?

GPDB

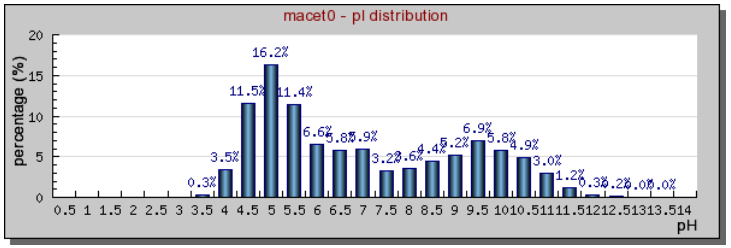
Genome Profile DataBase

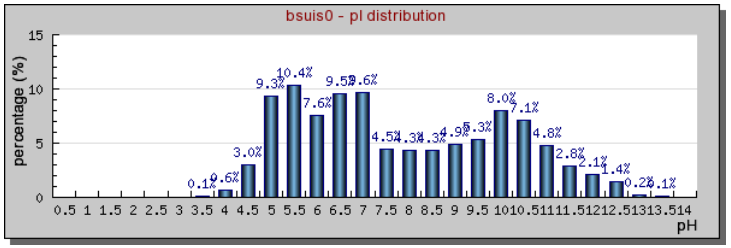
Home | Browse | Virtual 2D | Compare | Download | Status | Help | Comment

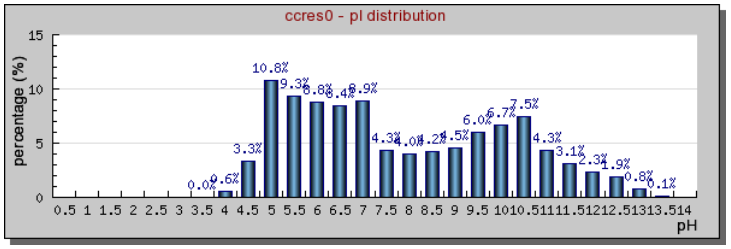
Does whole genome profile reflect phylogeny or environment or …… ?

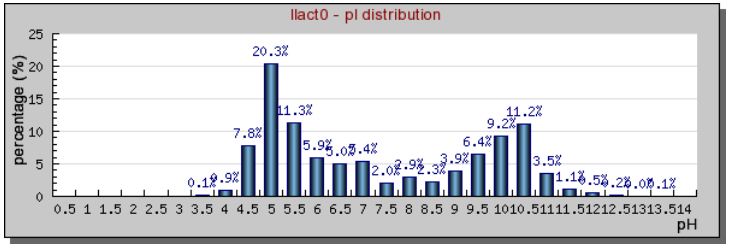
- Phylogenomics- whole genome scale.
- Is it possible to explore it, using different whole genome profile.
- We don't know the answer.
- Then, we try.

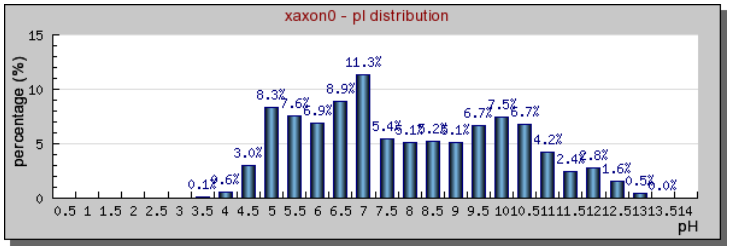
Ex: pI distribution

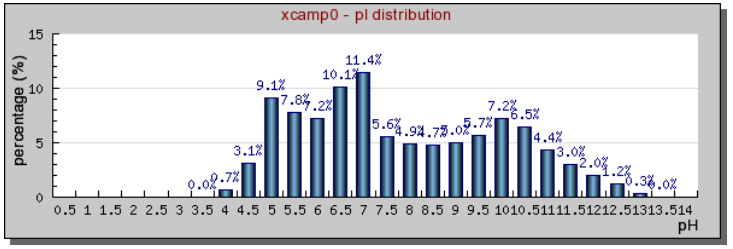


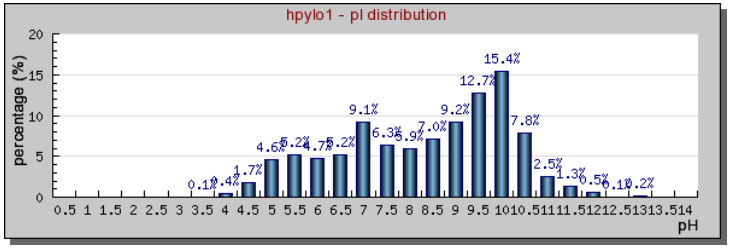


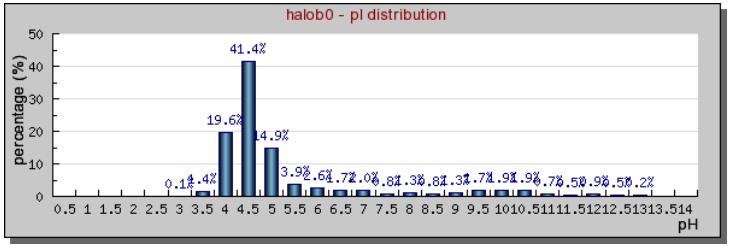




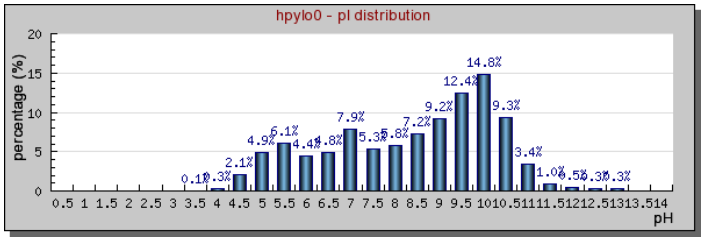


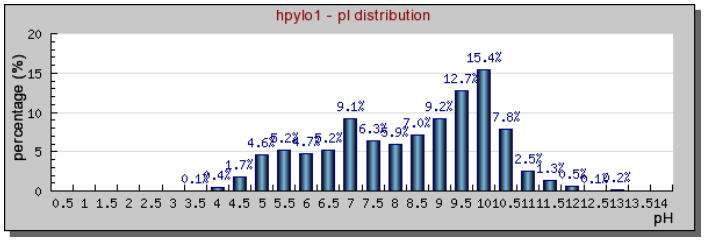




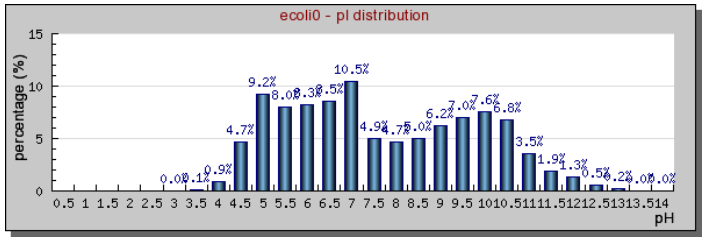


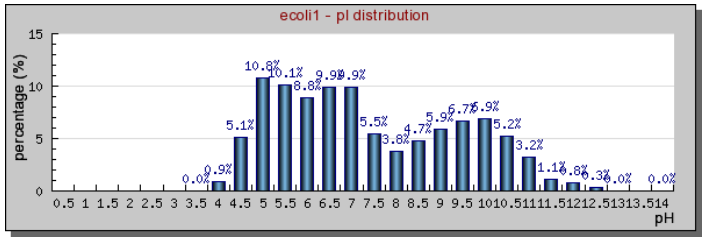
Similar ?

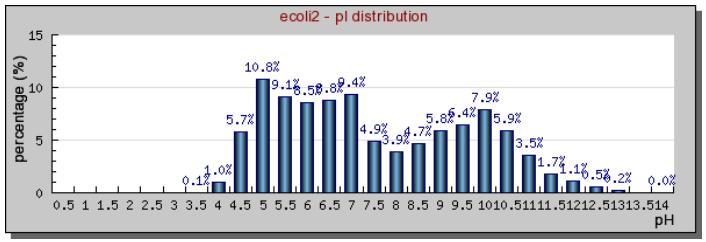


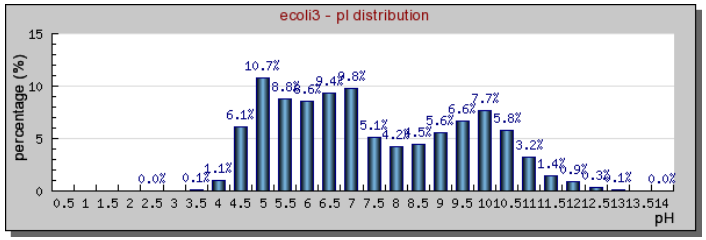


Helicobacter pylori



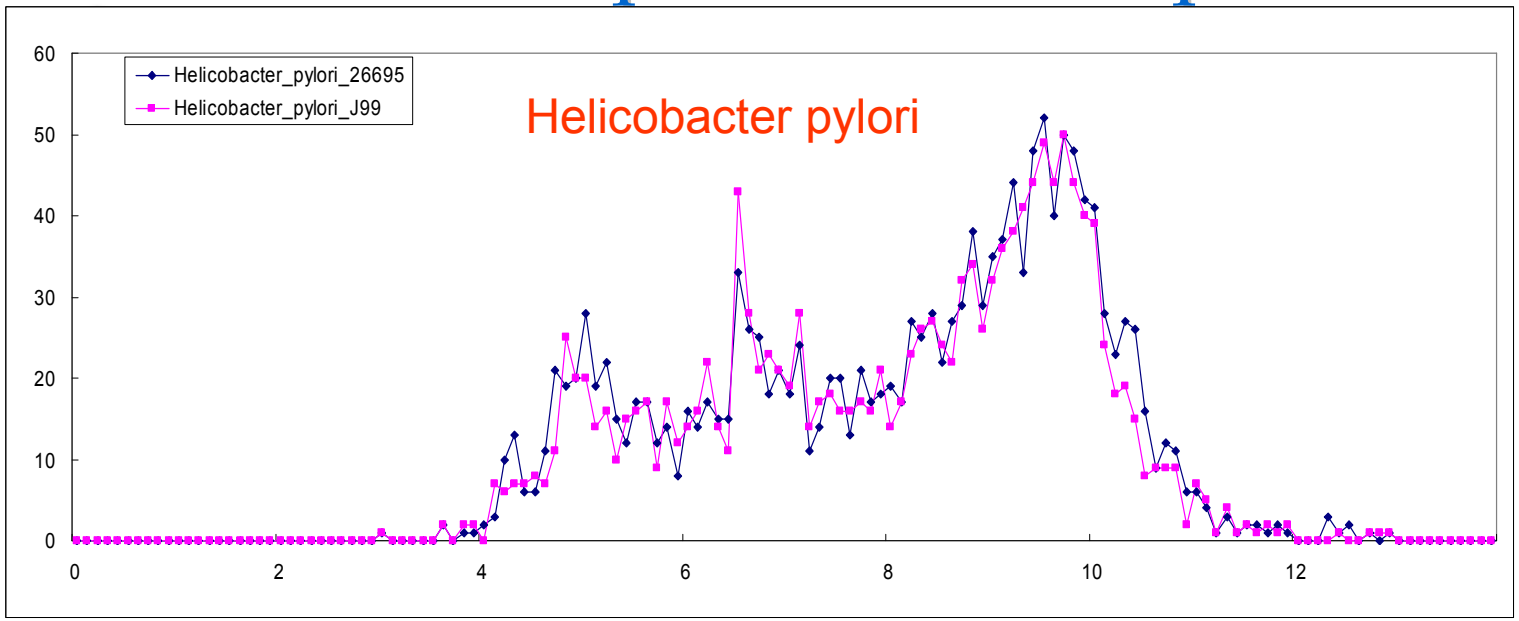


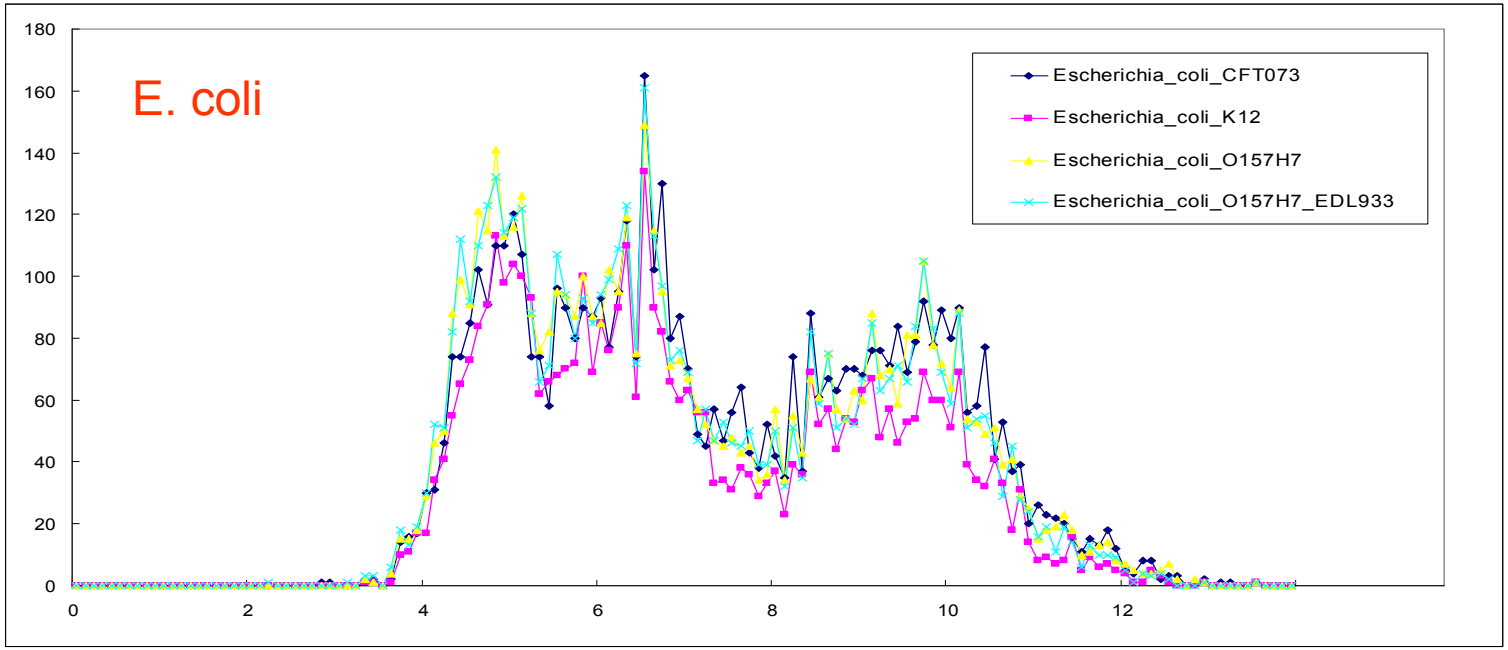




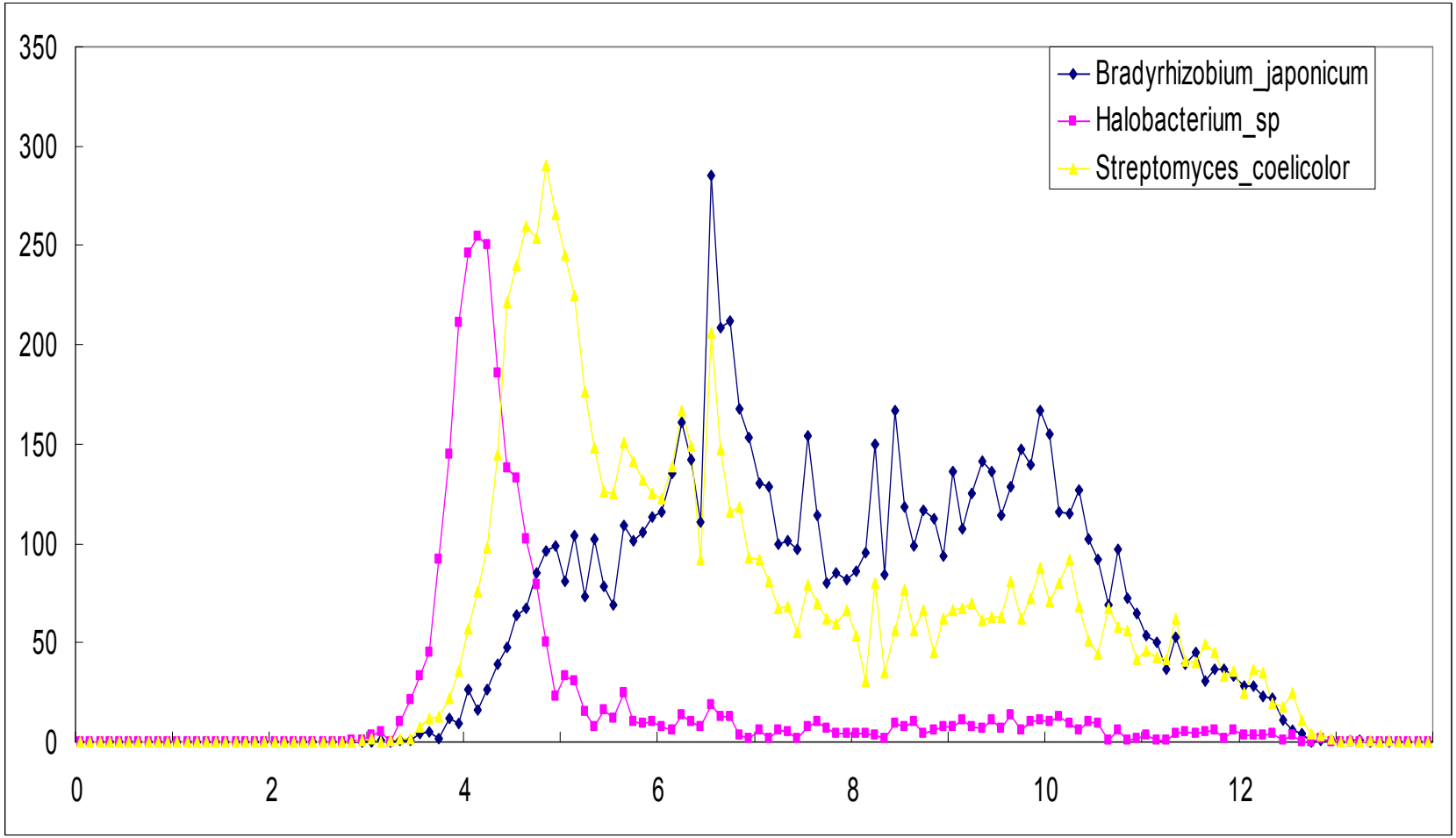
E. coli

Similar species, similar pI distribution

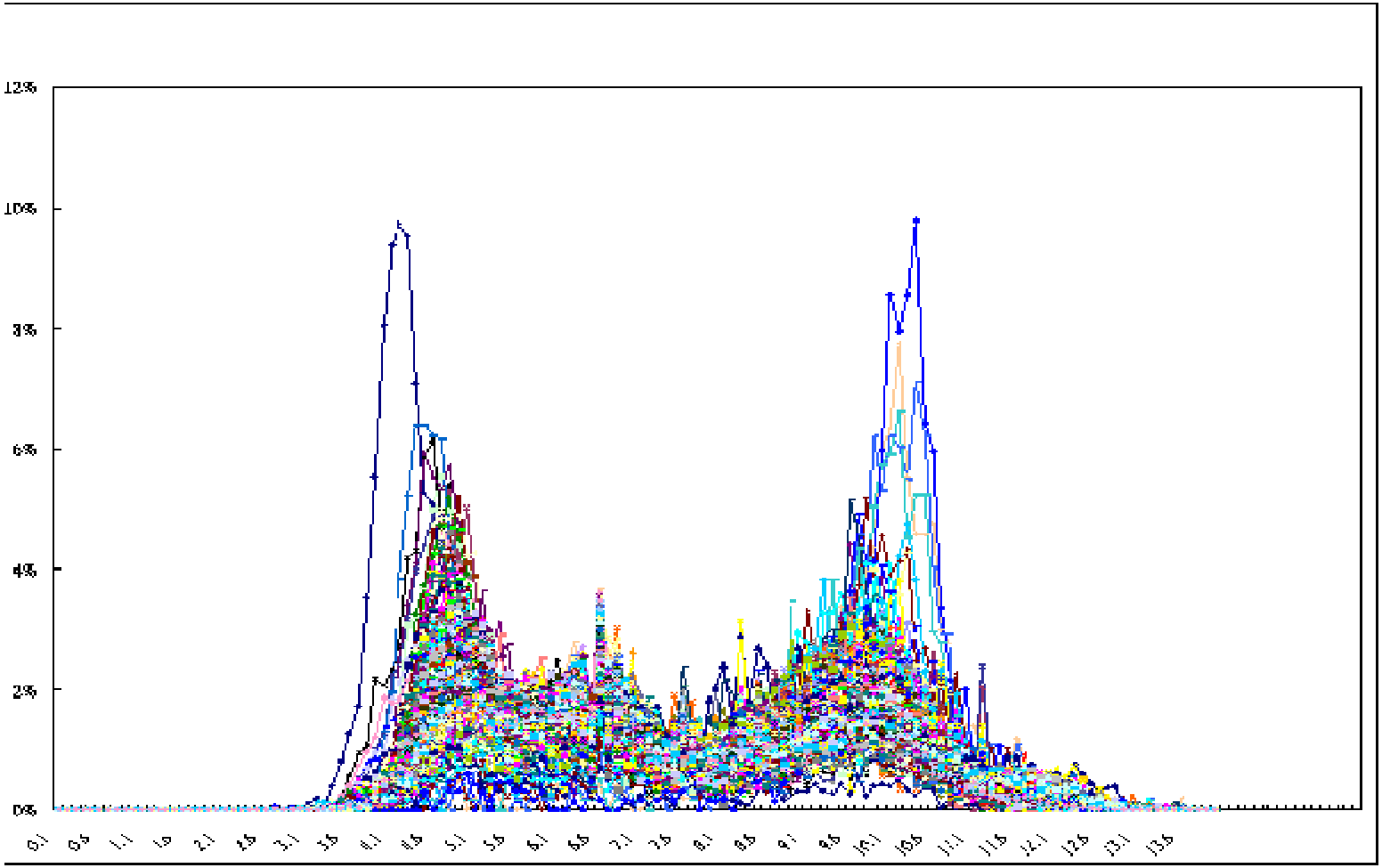




Different species, different pI distribution



pI distribution Over 100 Species

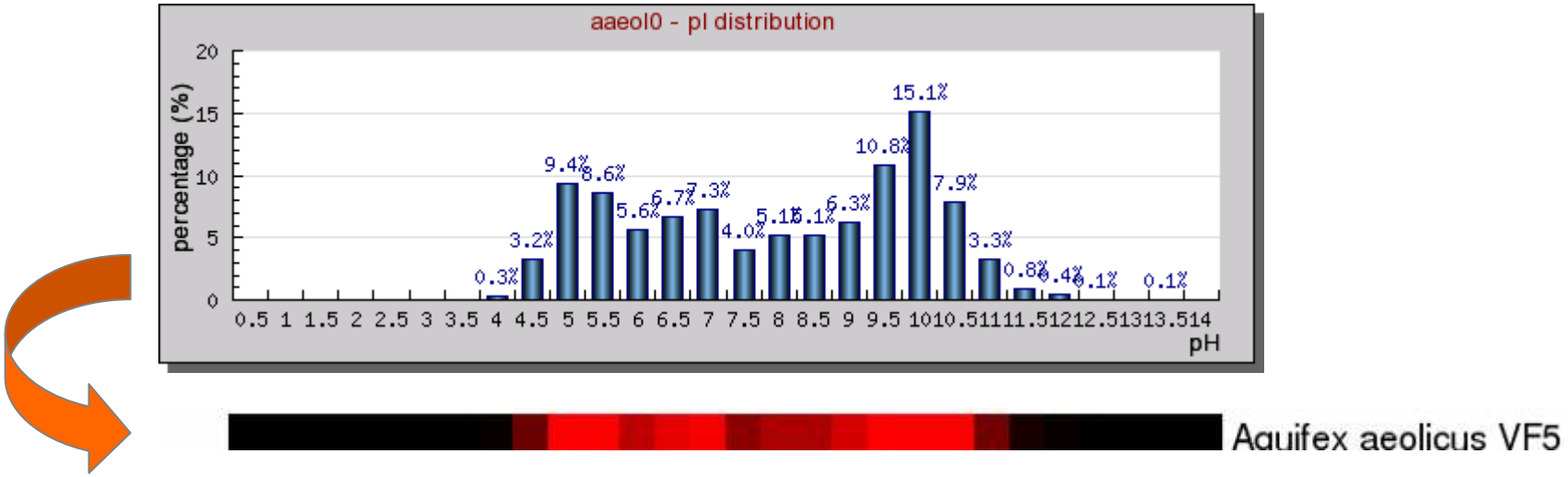


How to compare ?

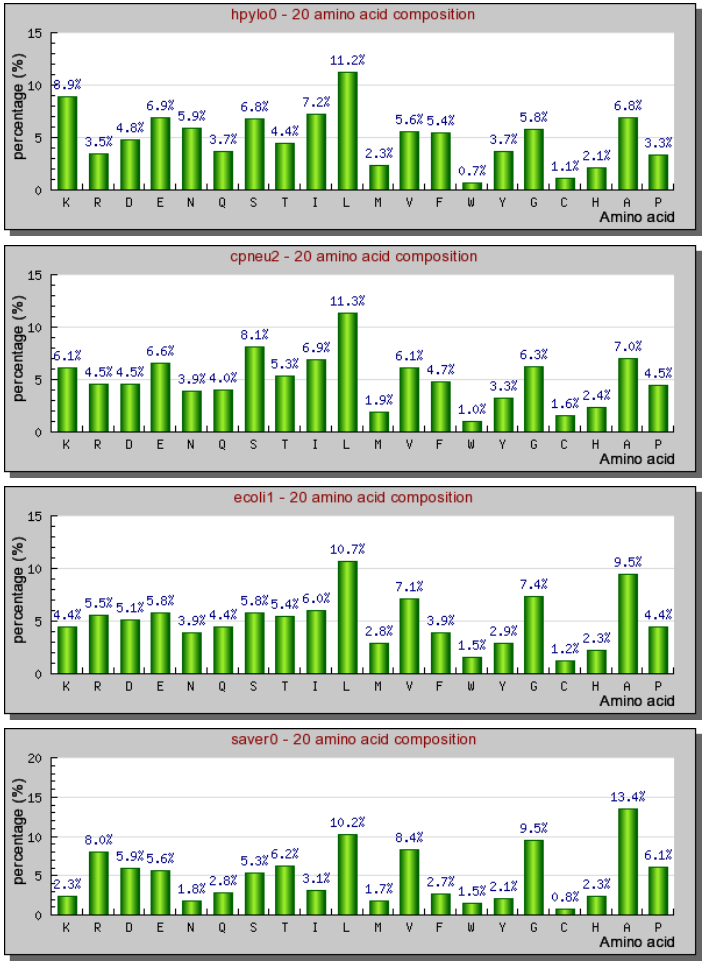
- Interactive on-line analysis to help us to explore the different combination.
- Easy-reading.

Transform

[Bacteria] - Aquifex aeolicus VF5



Transform




Escherichia coli K12


Chlamydophila pneumoniae J138


Helicobacter pylori 26695


Streptomyces avermitilis MA-4680

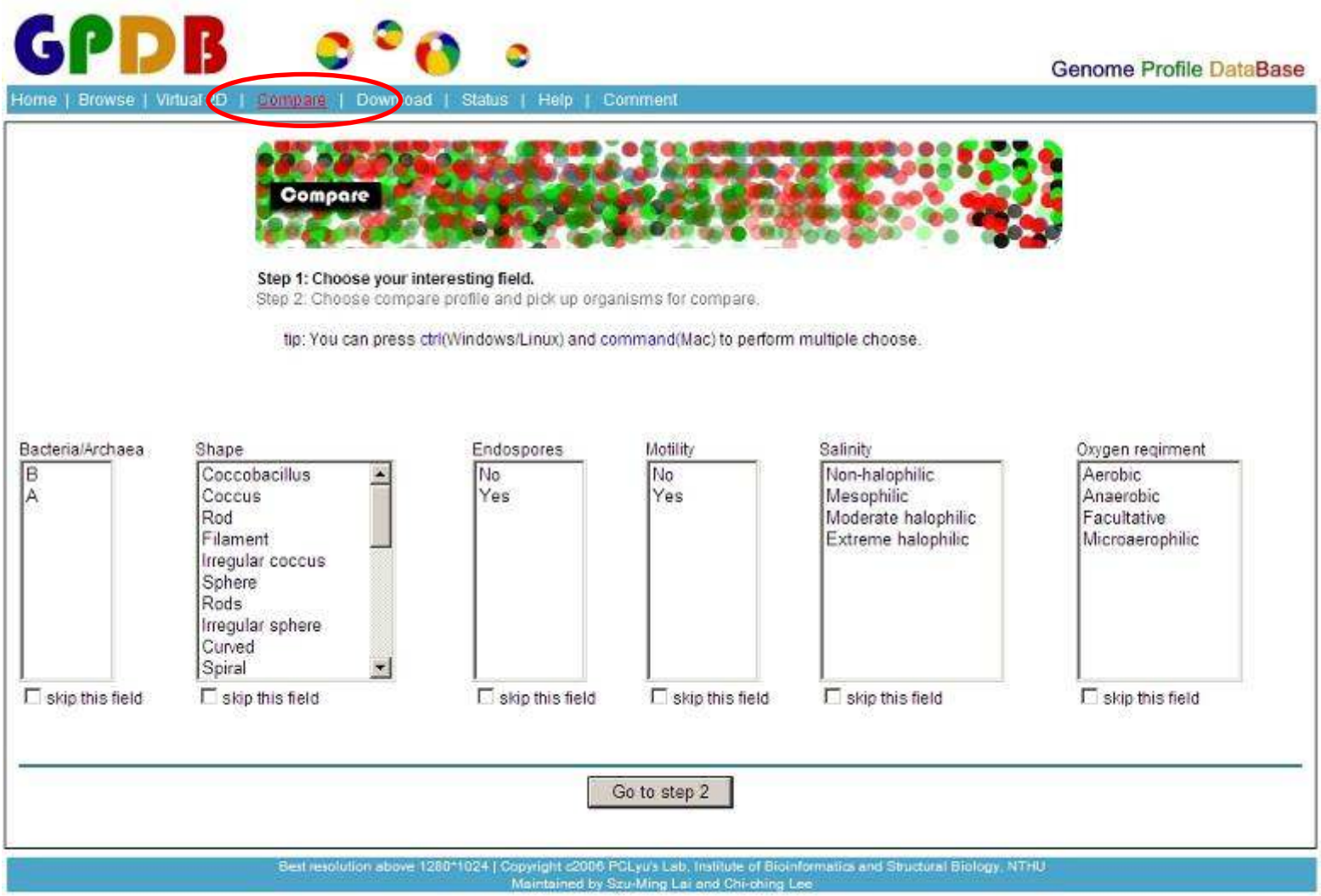
Clustering


Escherichia coli K12
Chlamydophila pneumoniae J138
Helicobacter pylori 26695
Streptomyces avermitilis MA-4680

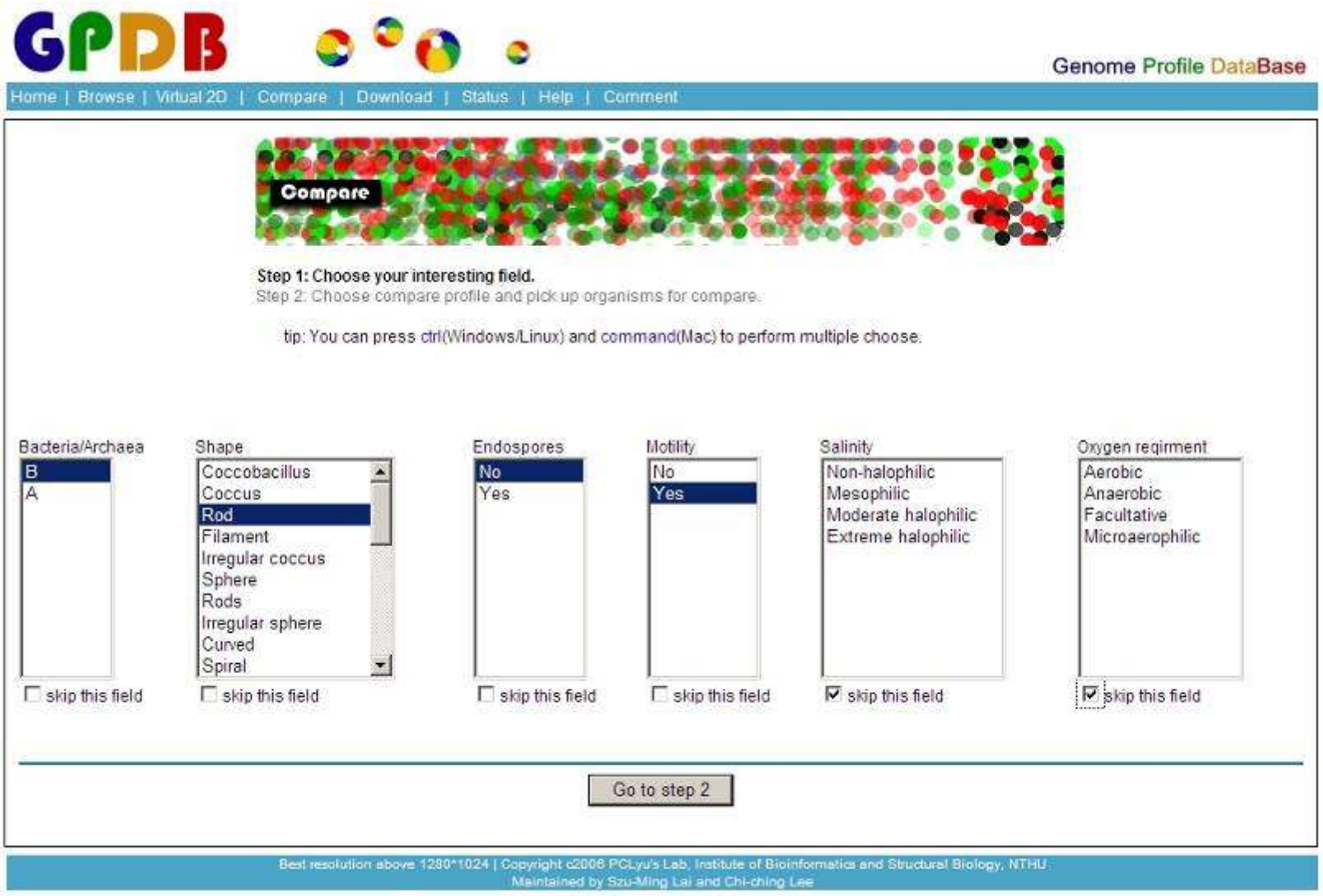
Clustering

- There are many clustering methods.
- We use Euclidean distances for Hierarchical Clustering.
- It just a easy way to read, not the only solution!

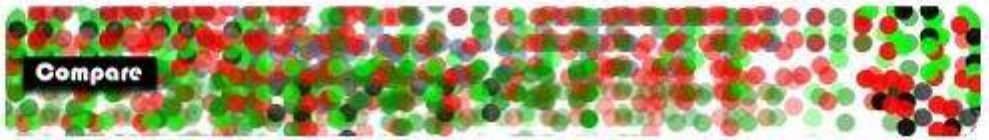
On-line compare



Choose your interesting field



Choose Profile



Step 1: Choose your interesting field.
Step 2: Choose compare profile and pick up organisms for compare.

Which profile do you want to compare?

- AT & GC Content
- Nucleotide composition
- di-nucleotide composition
- tri-Nucleotide composition
- Amino acid composition
- di-peptide composition
- Length distribution
- Molecular Weight distribution
- Isoelectric point distribution
- Total codon usage comparison

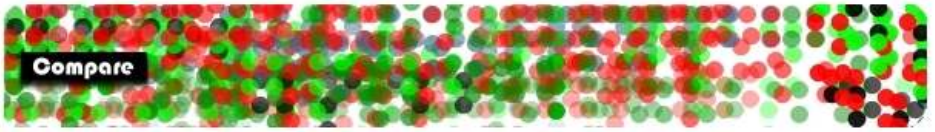
Which organisms do you want to compare?

tip: You can click title of the table to sort the organism list.

Submit Reset

	RefSeq ID	Organism	King	Shape	Endospores	Motility	Salinity	OxygenReq
☐	NC_007963	Chromohalobacter salexigens DSM 3043	B	Rod	No	Yes	Moderate halophilic	Aerobic
☐	NC_004557	Clostridium tetani E88	B	Rod	No	Yes	Non-halophilic	Anaerobic
☐	NC_007298	Dechloromonas aromatica RCB	B	Rod	No	Yes		Facultative
☐	NC_007722	Erythrobacter litoralis HTCC2594	B	Rod	No	Yes		Aerobic
☐	NC_008228	Pseudoalteromonas atlantica T6c	B	Rod	No	Yes		Aerobic
☐	NC_007613	Shigella boydii Sb227	B	Rod	No	Yes	Non-halophilic	Facultative
☐	NC_007606	Shigella dysenteriae Sd197	B	Rod	No	Yes	Non-halophilic	Facultative
☐	NC_007384	Shigella sonnei Ss046	B	Rod	No	Yes	Non-halophilic	Facultative
☐	NC_000853	Thermotoga maritima MSB8	B	Rod	No	Yes		Anaerobic
☐	NC_007404	Thiobacillus denitrificans ATCC 25259	B	Rod	No	Yes		Facultative

Choose Organisms



Step 1: Choose your interesting field.
Step 2: Choose compare profile and pick up organisms for compare.

Which profile do you want to compare?

- ○ AT & GC Content
- ○ Nucleotide composition ○ di-nucleotide composition ○ tri-Nucleotide composition
- ○ Amino acid composition ○ di-peptide composition
- ○ Length distribution ○ Molecular Weight distribution ◉ Isoelectric point distribution
- ○ Total codon usage comparison

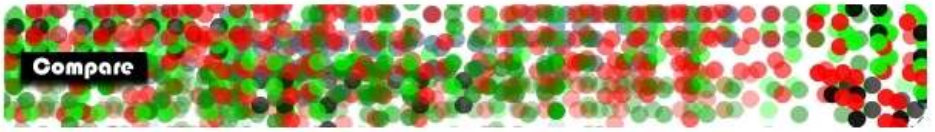
Which organisms do you want to compare?

tip: You can click title of the table to sort the organism list.

Submit Reset

	RefSeq ID	Organism	King	Shape	Endospores	Motility	Salinity	OxygenReq
☑	NC_007963	Chromohalobacter salexigens DSM 3043	B	Rod	No	Yes	Moderate halophilic	Aerobic
☑	NC_004557	Clostridium tetani E88	B	Rod	No	Yes	Non-halophilic	Anaerobic
☑	NC_007298	Dechloromonas aromatica RCB	B	Rod	No	Yes		Facultative
☑	NC_007722	Erythrobacter litoralis HTCC2594	B	Rod	No	Yes		Aerobic
☐	NC_008228	Pseudoalteromonas atlantica T6c	B	Rod	No	Yes		Aerobic
☑	NC_007613	Shigella boydii Sb227	B	Rod	No	Yes	Non-halophilic	Facultative
☑	NC_007606	Shigella dysenteriae Sd197	B	Rod	No	Yes	Non-halophilic	Facultative
☐	NC_007384	Shigella sonnei Ss046	B	Rod	No	Yes	Non-halophilic	Facultative
☑	NC_000853	Thermotoga maritima MSB8	B	Rod	No	Yes		Anaerobic
☑	NC_007404	Thiobacillus denitrificans ATCC 25259	B	Rod	No	Yes		Facultative

Submit



Step 1: Choose your interesting field.
Step 2: Choose compare profile and pick up organisms for compare.

Which profile do you want to compare?

- AT & GC Content
- Nucleotide composition
- di-nucleotide composition
- tri-Nucleotide composition
- Amino acid composition
- di-peptide composition
- Length distribution
- Molecular Weight distribution
- Isoelectric point distribution (selected)
- Total codon usage comparison

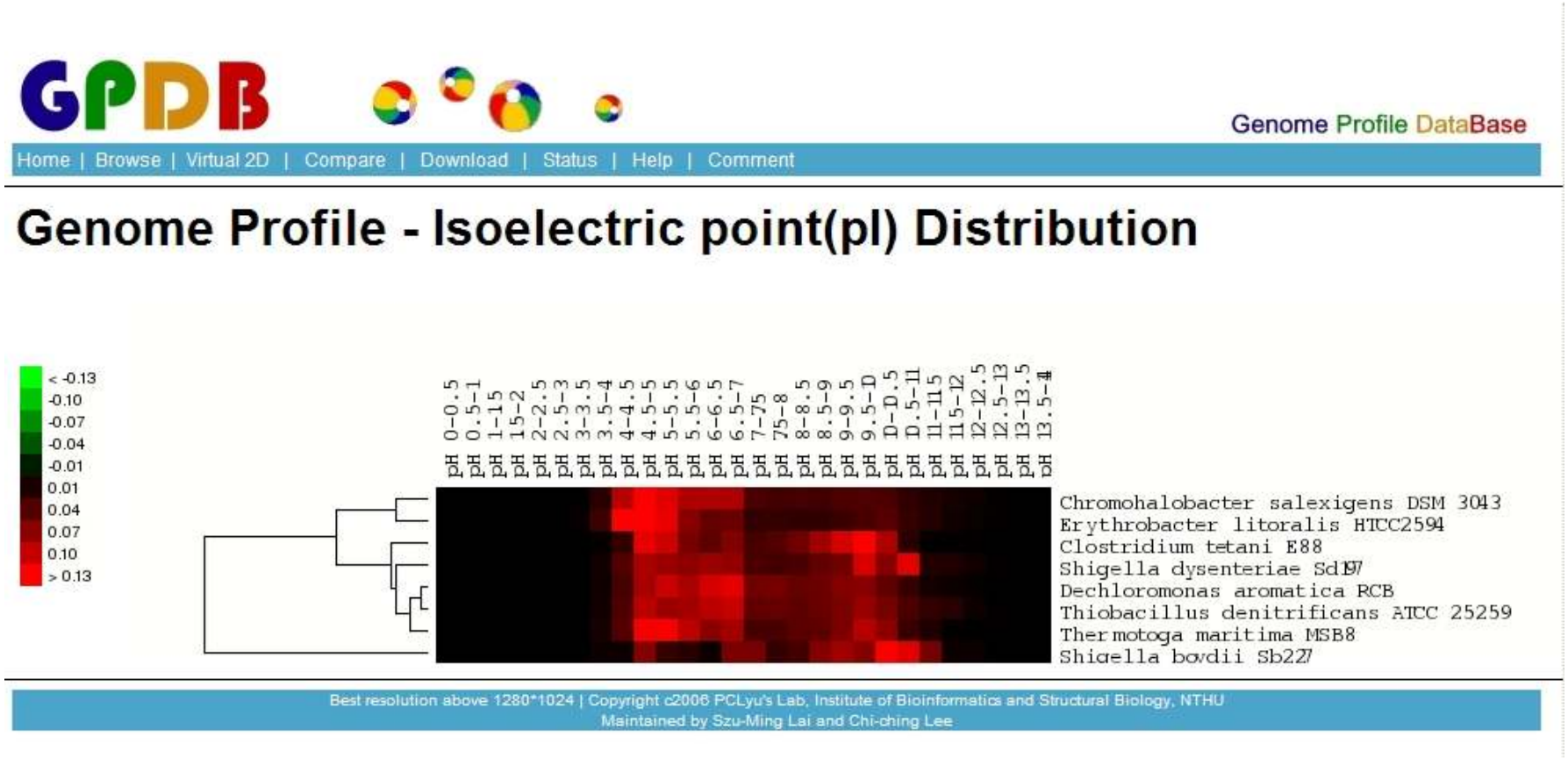
Which organisms do you want to compare?

tip: You can click title of the table to sort the organism list.

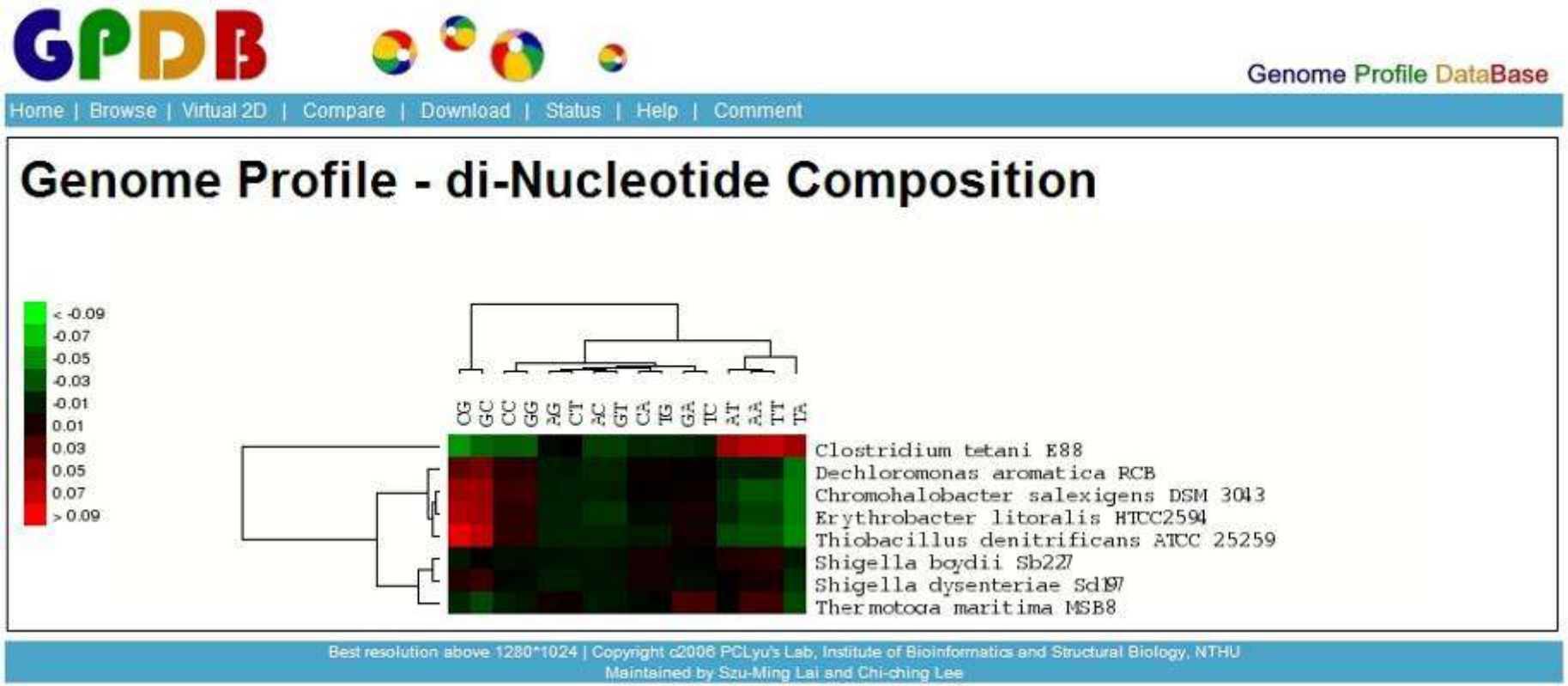
Submit Reset

	RefSeq ID	Organism	King	Shape	Endospores	Motility	Salinity	OxygenReq
☑	NC_007963	Chromohalobacter salexigens DSM 3043	B	Rod	No	Yes	Moderate halophilic	Aerobic
☑	NC_004557	Clostridium tetani E88	B	Rod	No	Yes	Non-halophilic	Anaerobic
☑	NC_007298	Dechloromonas aromatica RCB	B	Rod	No	Yes		Facultative
☑	NC_007722	Erythrobacter litoralis HTCC2594	B	Rod	No	Yes		Aerobic
☐	NC_008228	Pseudoalteromonas atlantica T6c	B	Rod	No	Yes		Aerobic
☑	NC_007613	Shigella boydii Sb227	B	Rod	No	Yes	Non-halophilic	Facultative
☑	NC_007606	Shigella dysenteriae Sd197	B	Rod	No	Yes	Non-halophilic	Facultative
☐	NC_007384	Shigella sonnei Ss046	B	Rod	No	Yes	Non-halophilic	Facultative
☑	NC_000853	Thermotoga maritima MSB8	B	Rod	No	Yes		Anaerobic
☑	NC_007404	Thiobacillus denitrificans ATCC 25259	B	Rod	No	Yes		Facultative

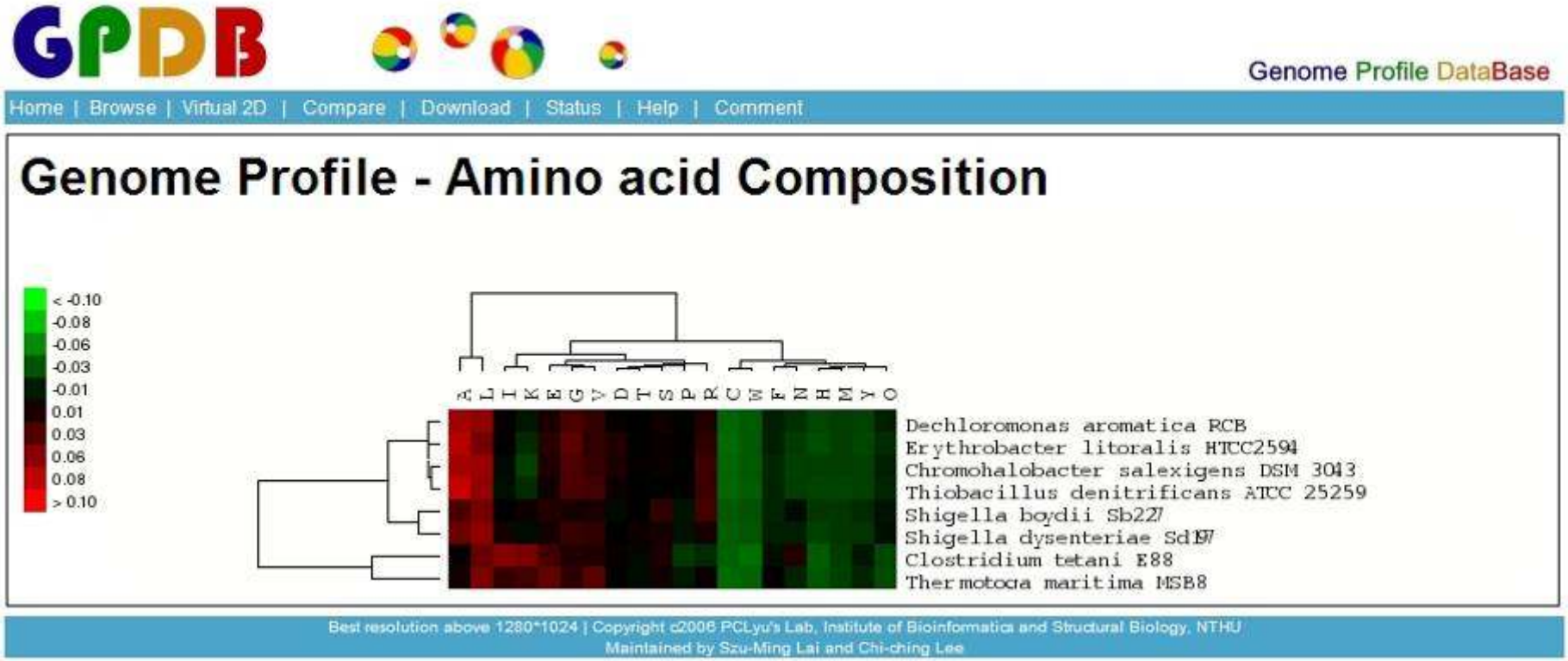
On-line Clustering



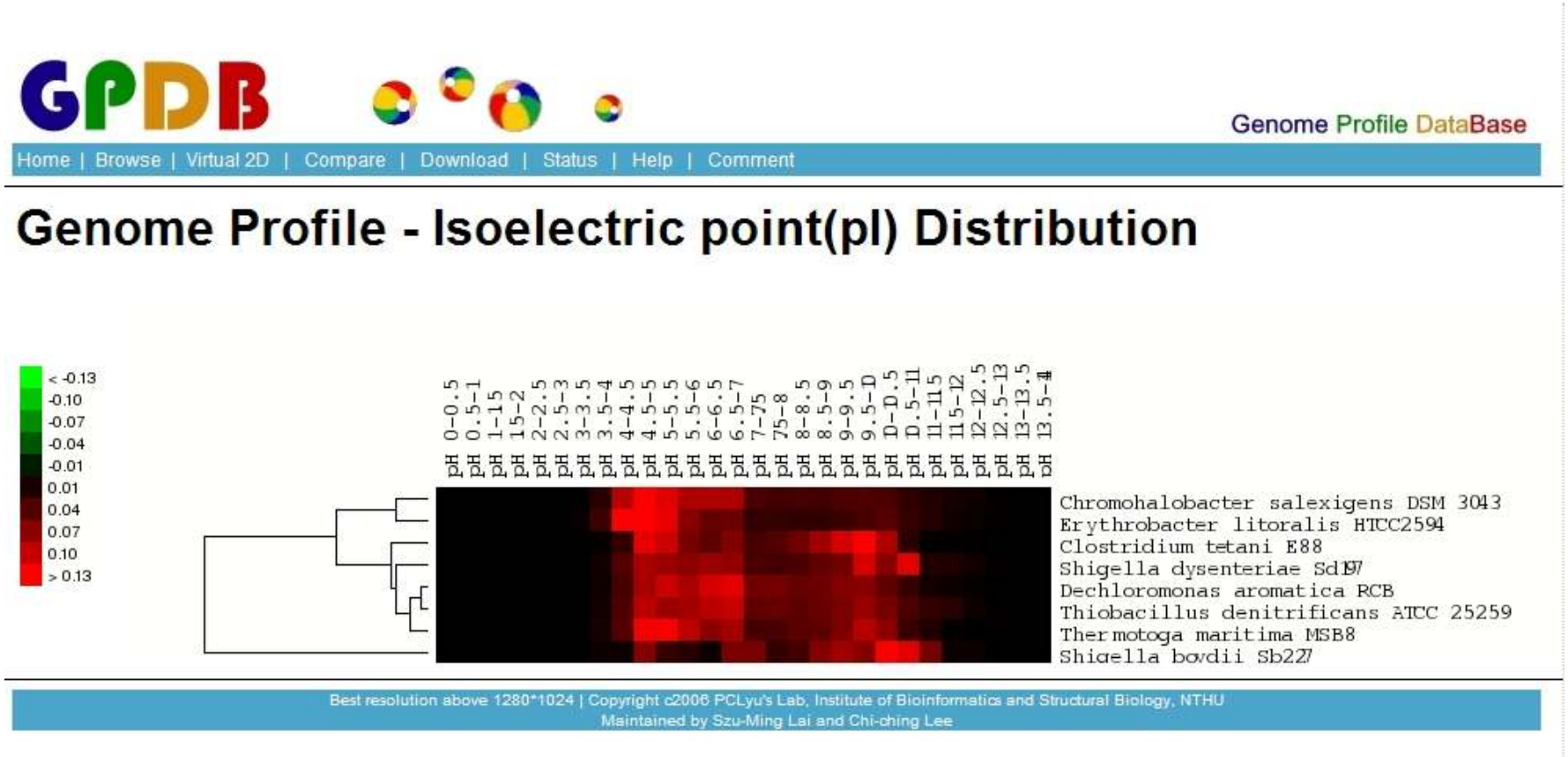
Di-nucleotide composition



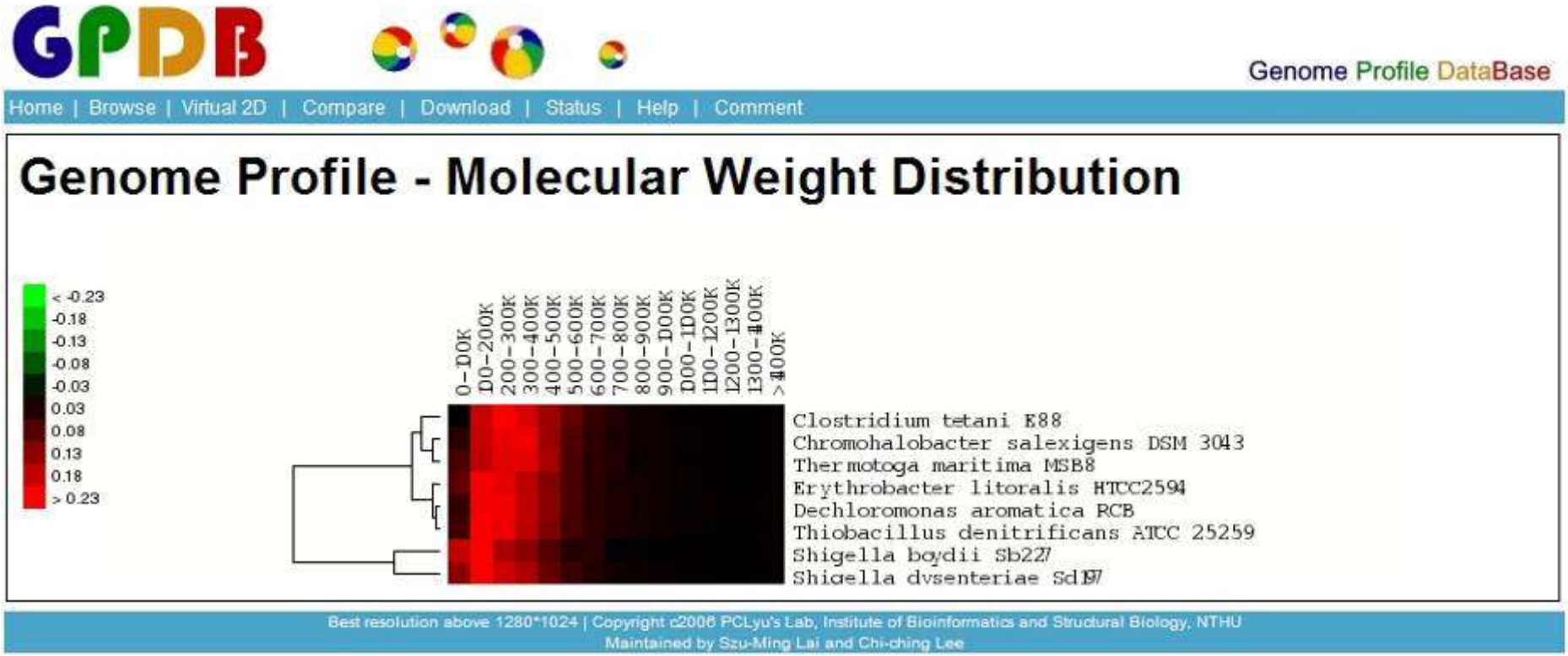
Amino acid composition



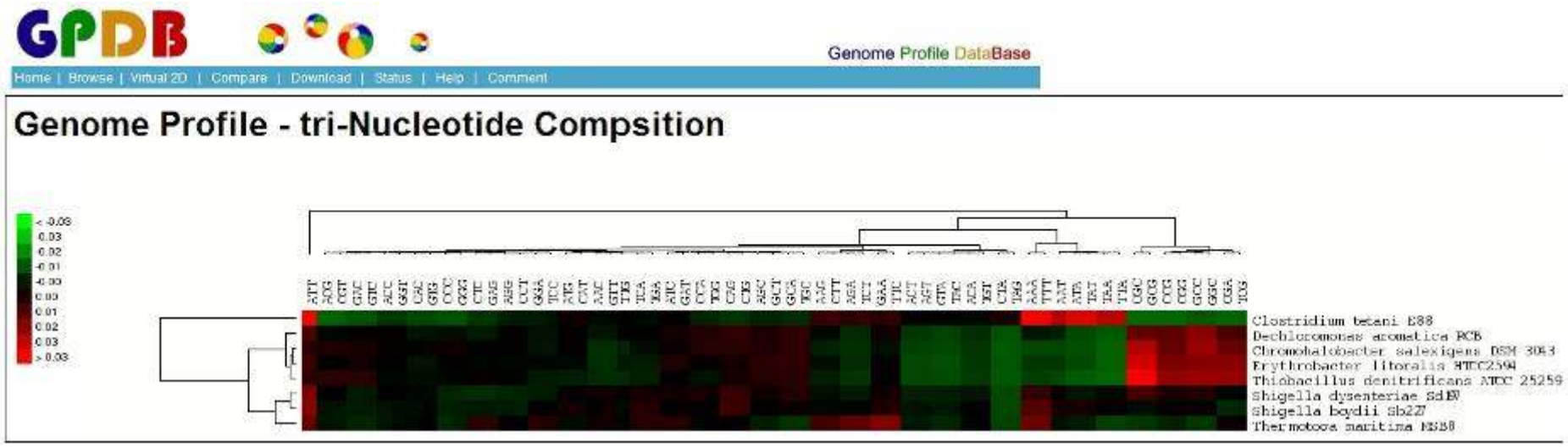
Isoelectric point distribution



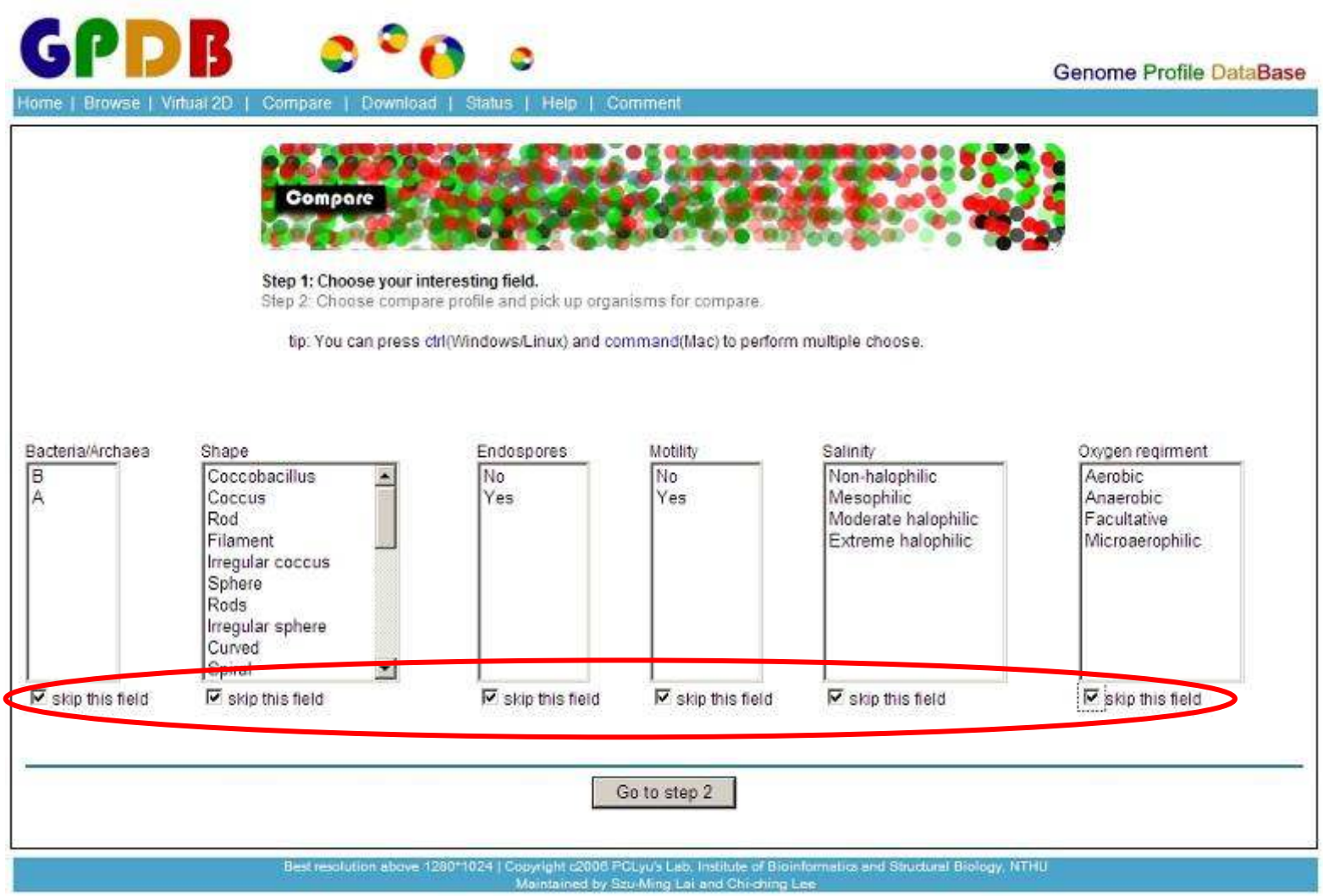
Molecular Weight distribution



Tri-nucleotide composition

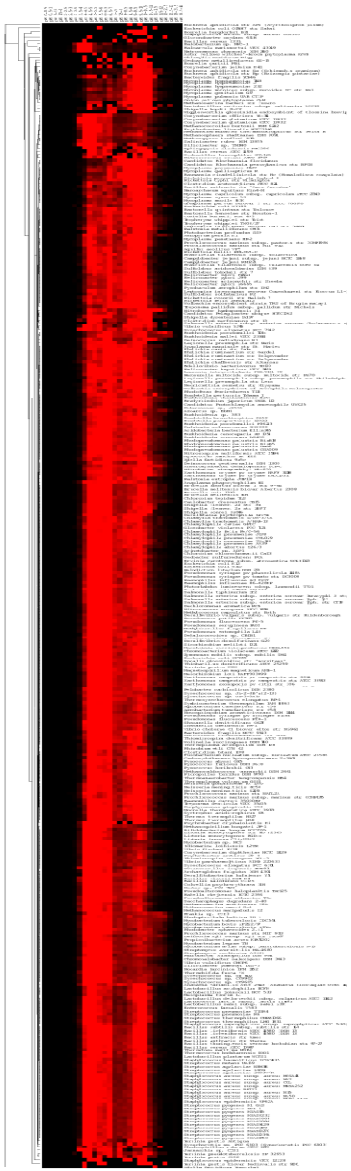


You can compare on profile in different combination or on the whole.

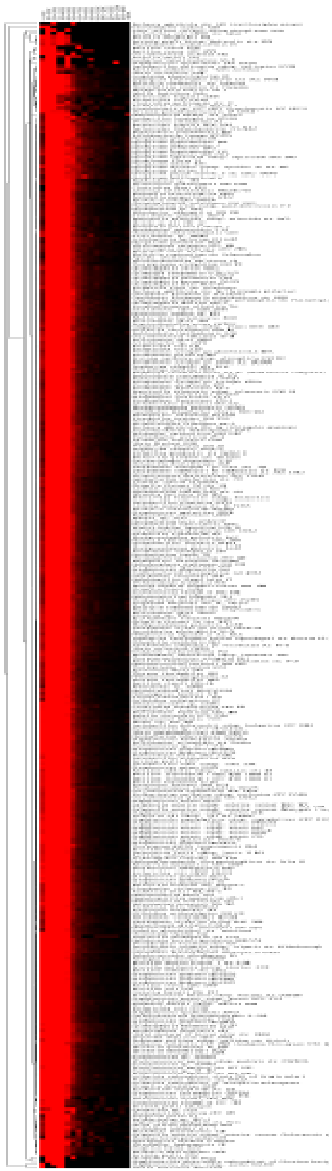


Compare 345 organisms (27 Archaea, 338 Bacteria)

Isoelectric point(pI) Distribution

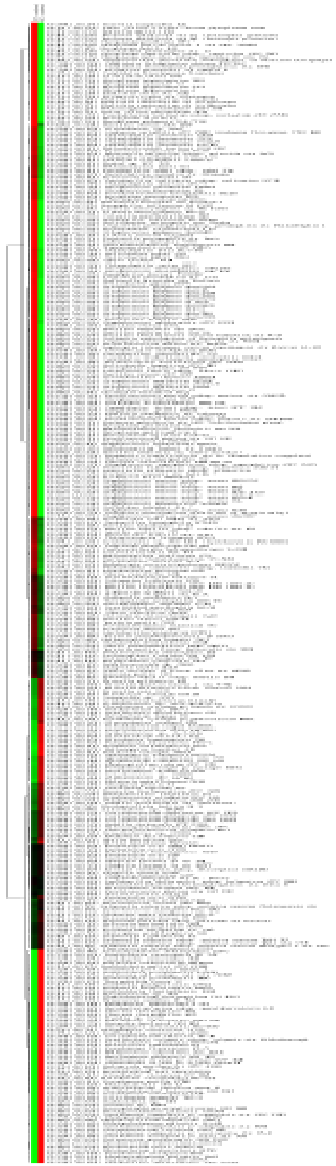


Molecular Weight Distribution



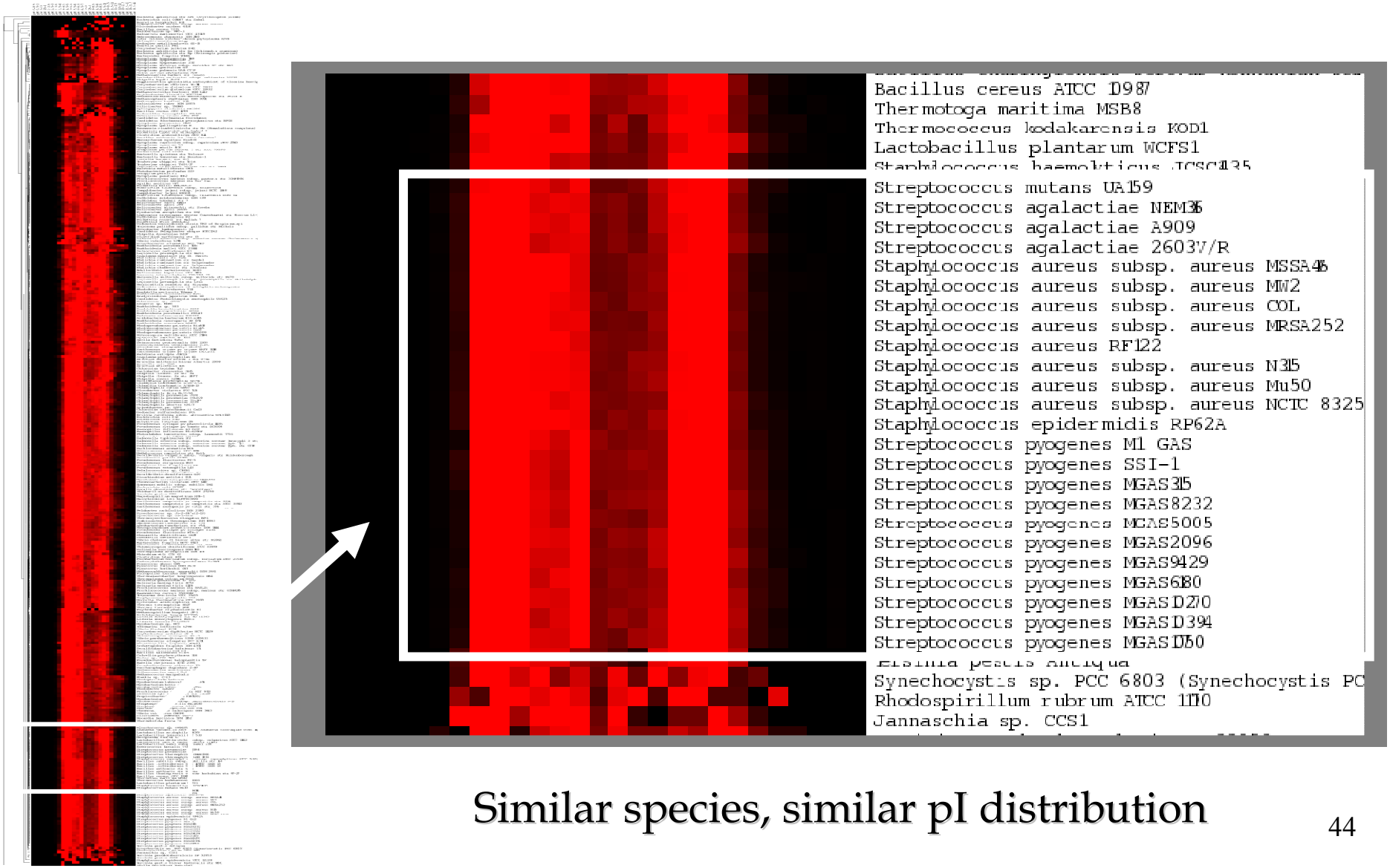
Nucleotide Composition

AT & GC Content



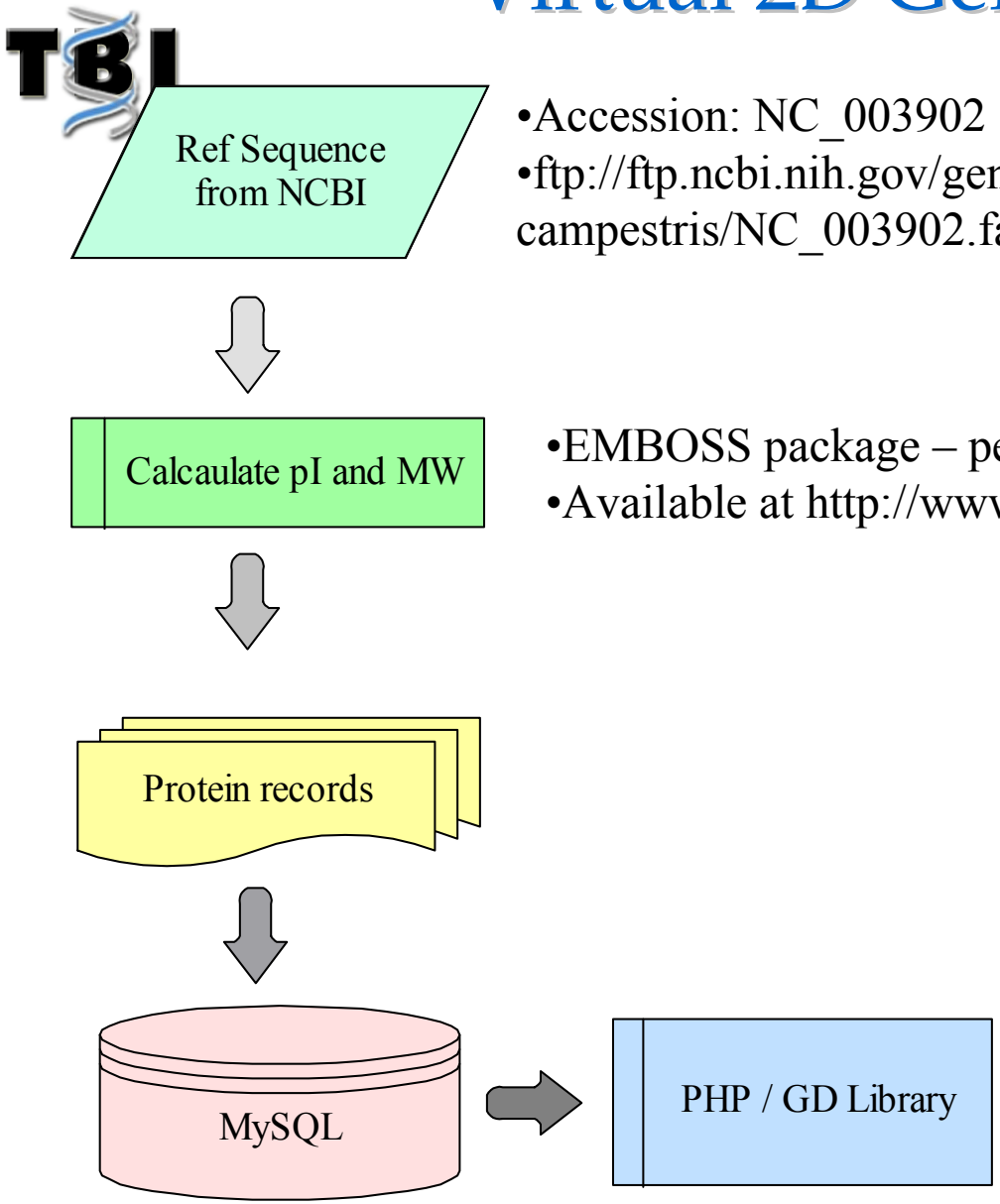
Compare 345 organisms (27 Archaea, 338 Bacteria)

Isoelectric point(pI) Distribution



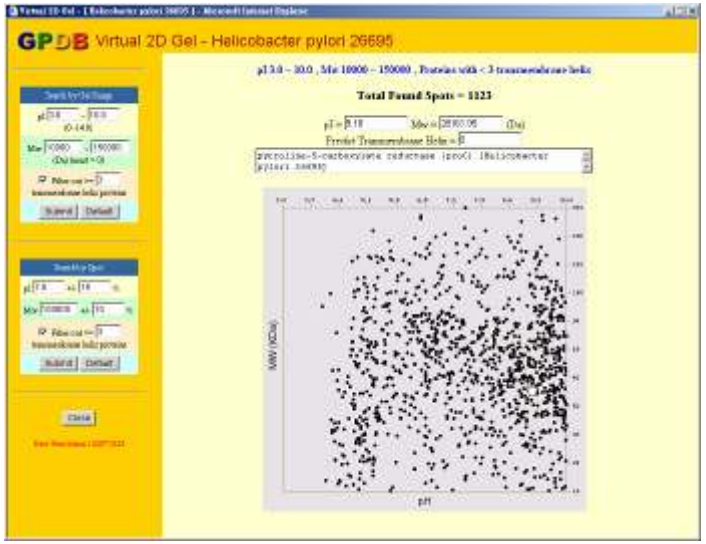
Similar species, similar pI distribution

Virtual 2D Gel Flowchart



- Accession: NC_003902
- ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Xanthomonas_campestris/NC_003902.faa

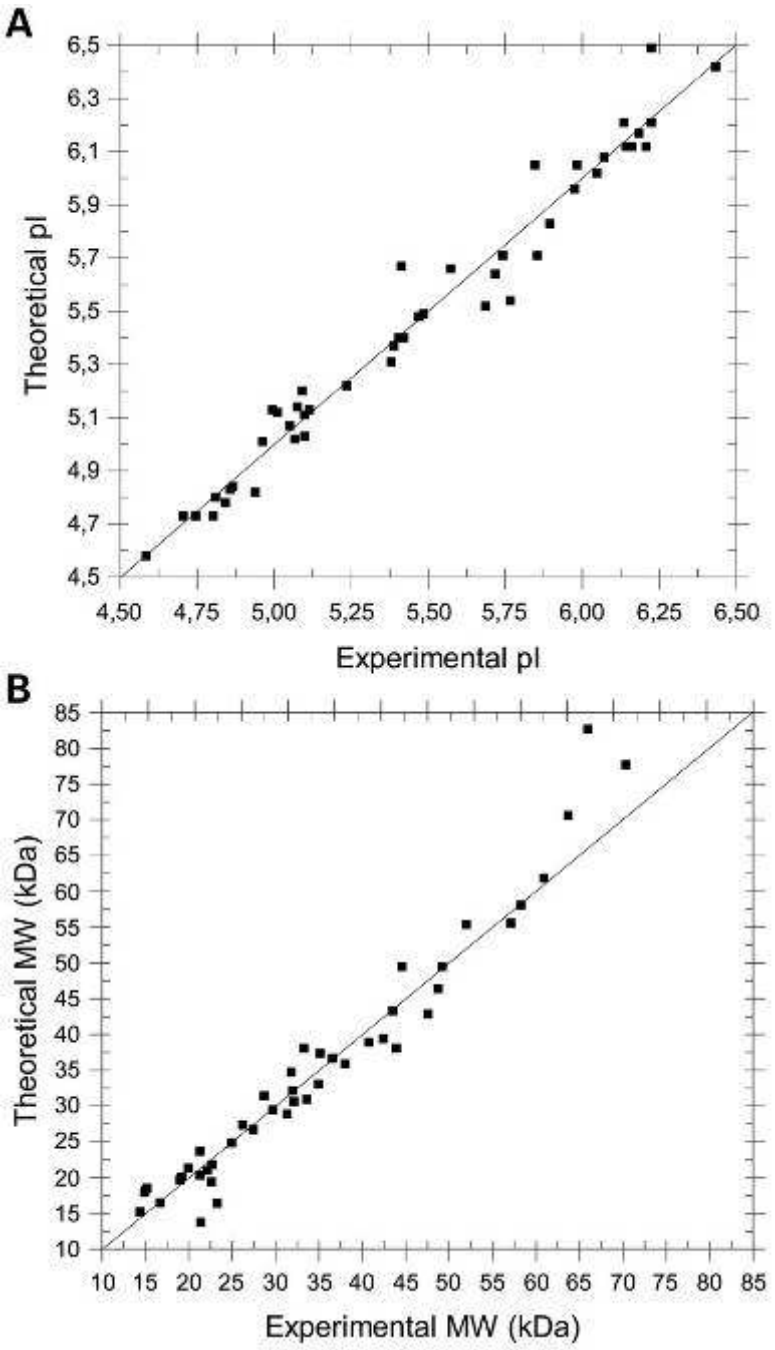
- EMBOSS package – pepstat
- Available at http://www.hgmp.mrc.ac.uk/Software/EMBOSS/



pI and MW
Theoritical vs. Experimental

Comparison between calculated and experimentally obtained pI and MW values for 47 randomly selected proteins from *Pseudomonas aeruginosa*.

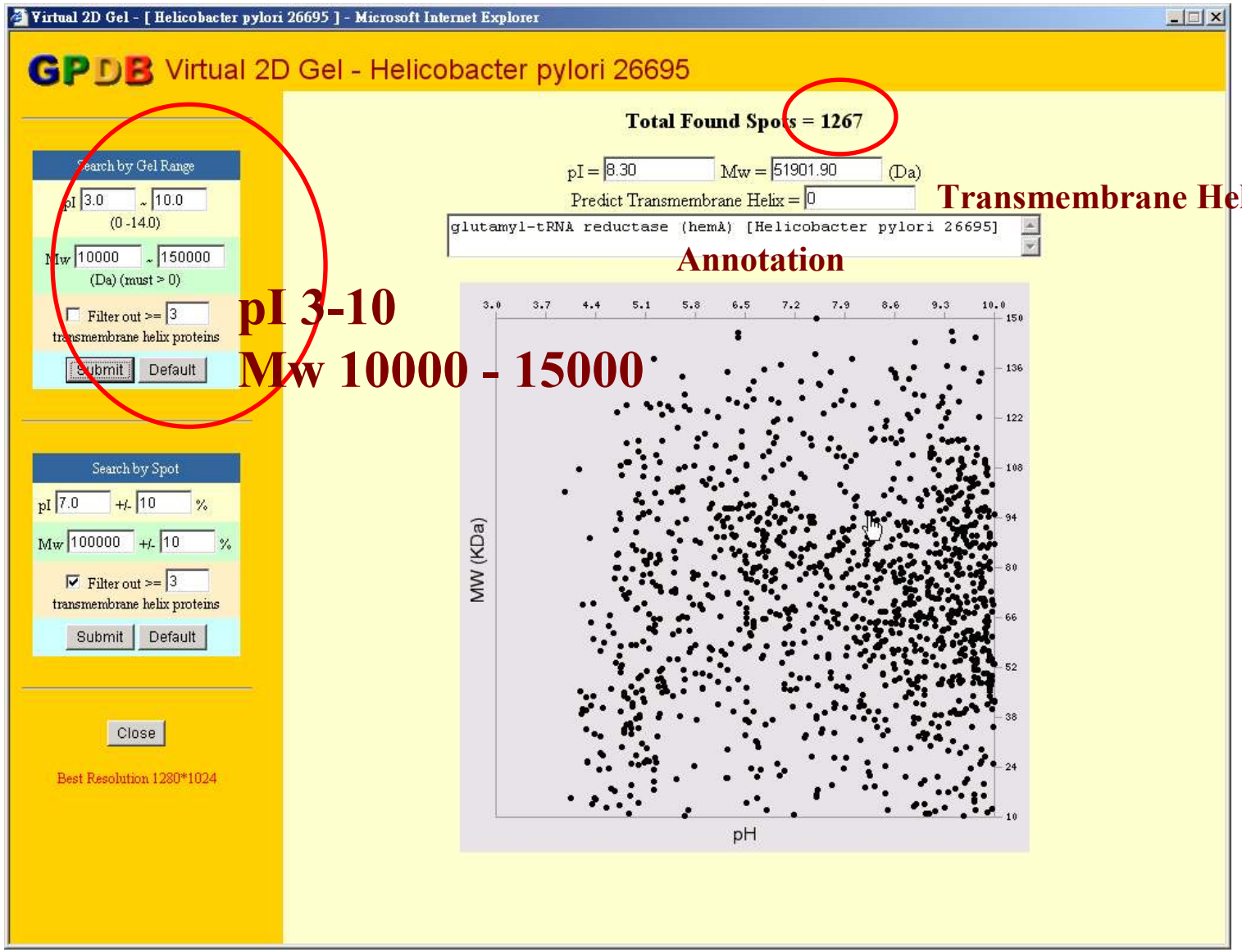
Nucleic Acids Research, 2003, Vol. 31, No. 13 3862-3865



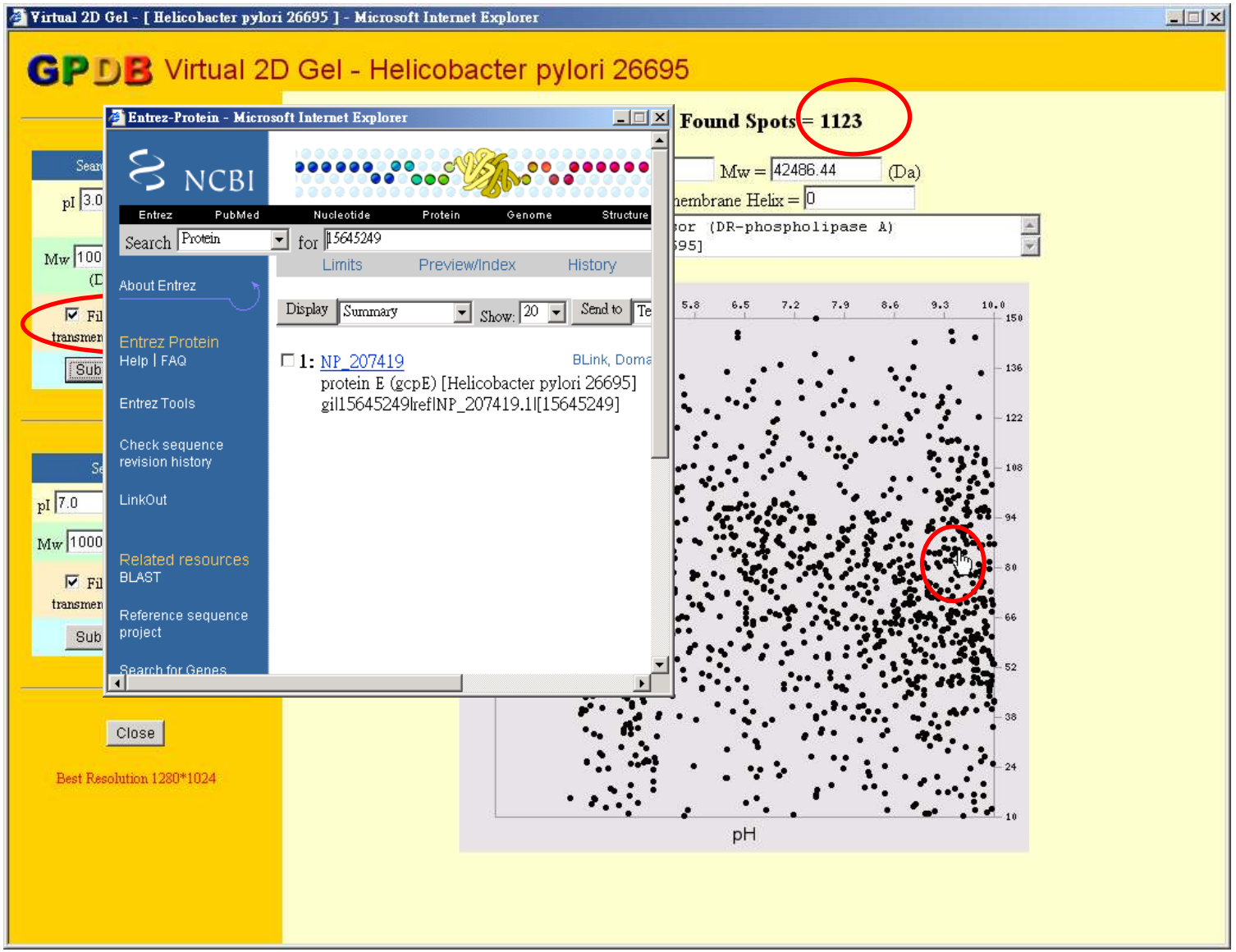
Virtual 2D Gel

- Search by different pH & mw range to simulate the real 2D gel.
- Search by spot range to guess the possible spot, ex:
  - pI = 5.7 +/- 5%
  - Mw = 100K +/- 5%
- Filter out transmembrane proteins.
  - TMHMM program
    - E. L.L. Sonnhammer, G. von Heijne, and A. Krogh.*In J. Glasgow et al., eds., Proc. Sixth Int. Conf. on Intelligent Systems for Molecular Biology , 175-182. AAAI Press, 1998.*
    - A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. *Journal of Molecular Biology, 305(3):567-580, January 2001.*
    - Moller S, Croning MD, Apweiler R. *Bioinformatics 2002 Jan;18(1):218*

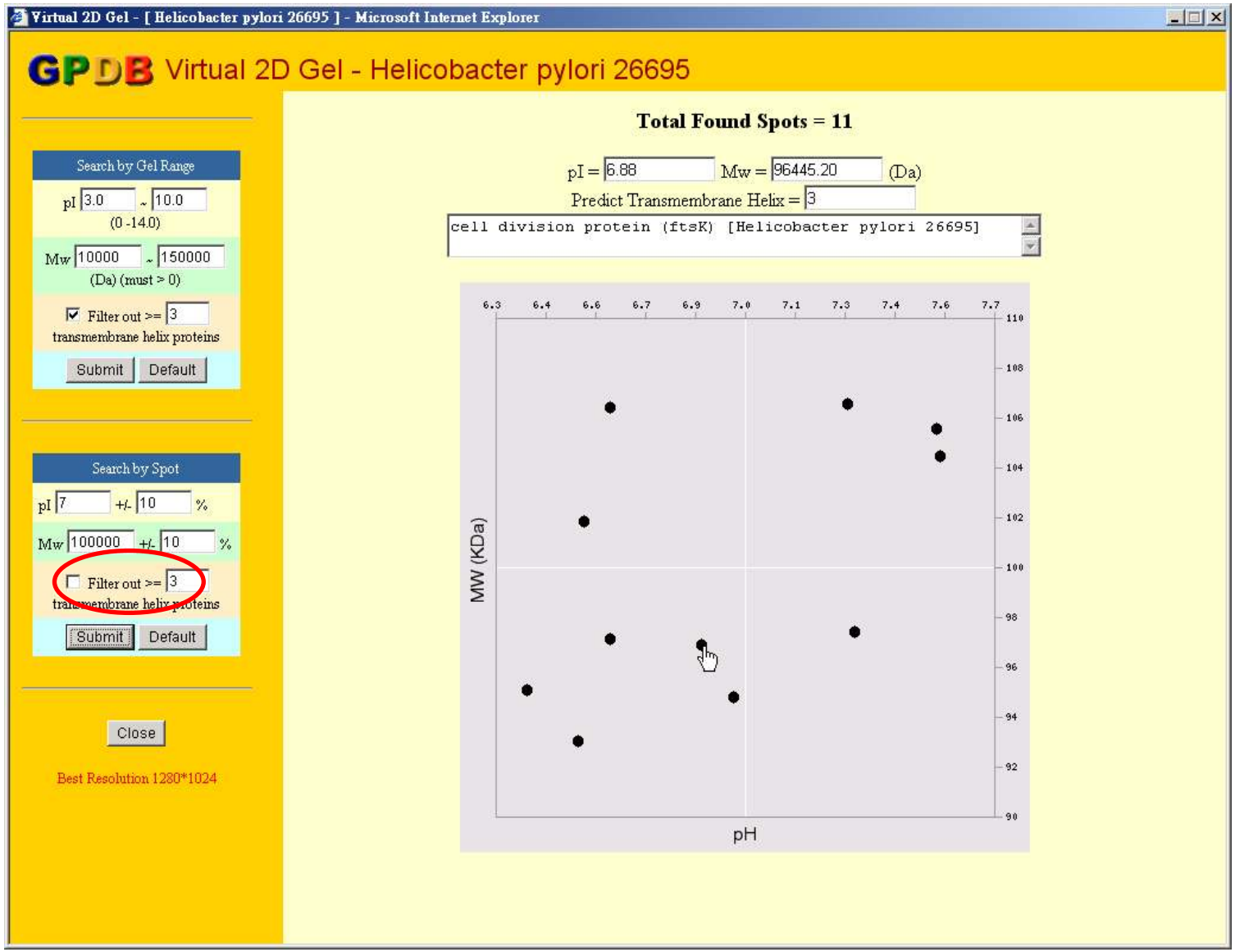
Search by Gel Range without TMHMM Filter



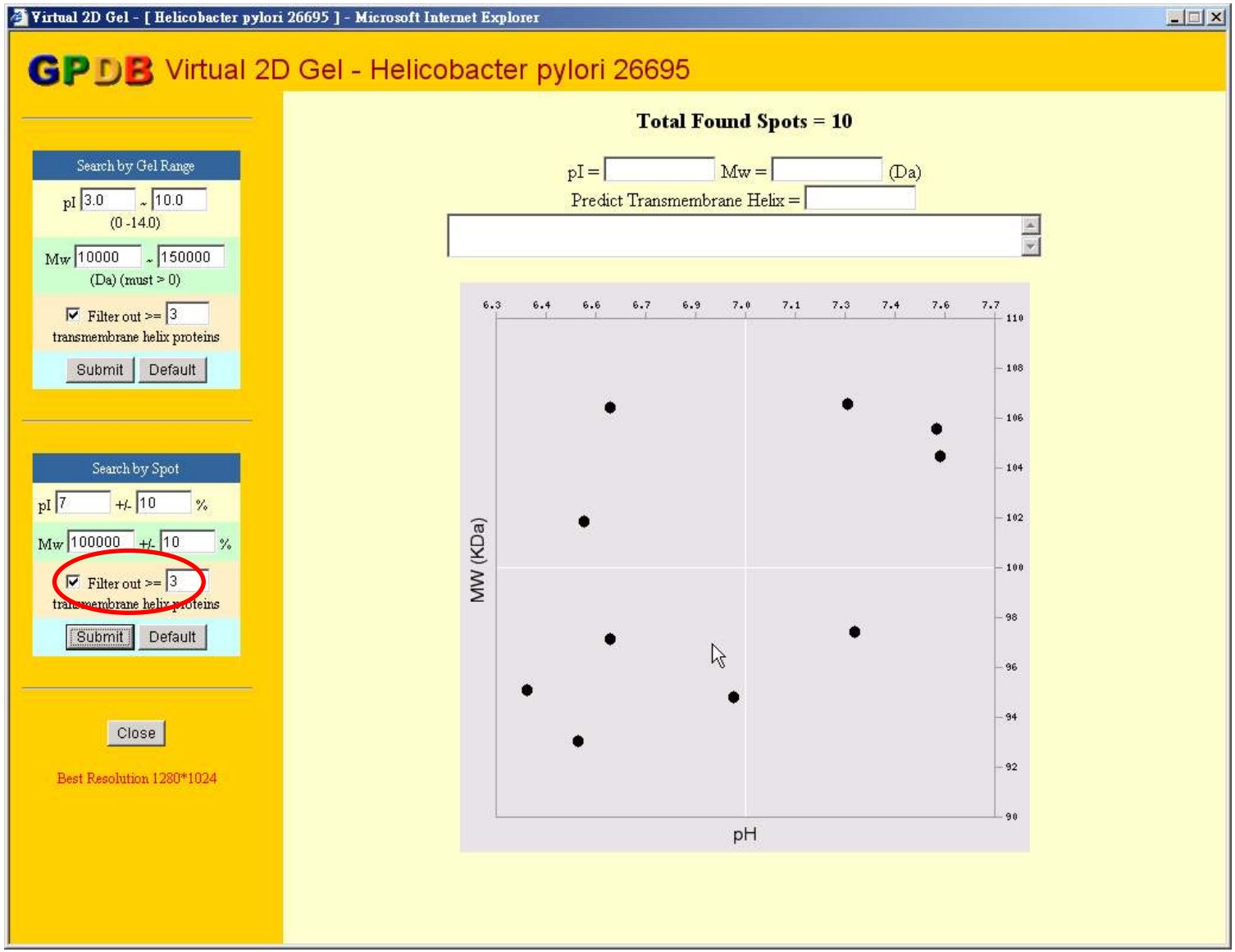
Search by Gel Range with TMHMM Filter



Search by Spot Range without TMHMM Filter



Search by Spot Range with TMHMM Filter



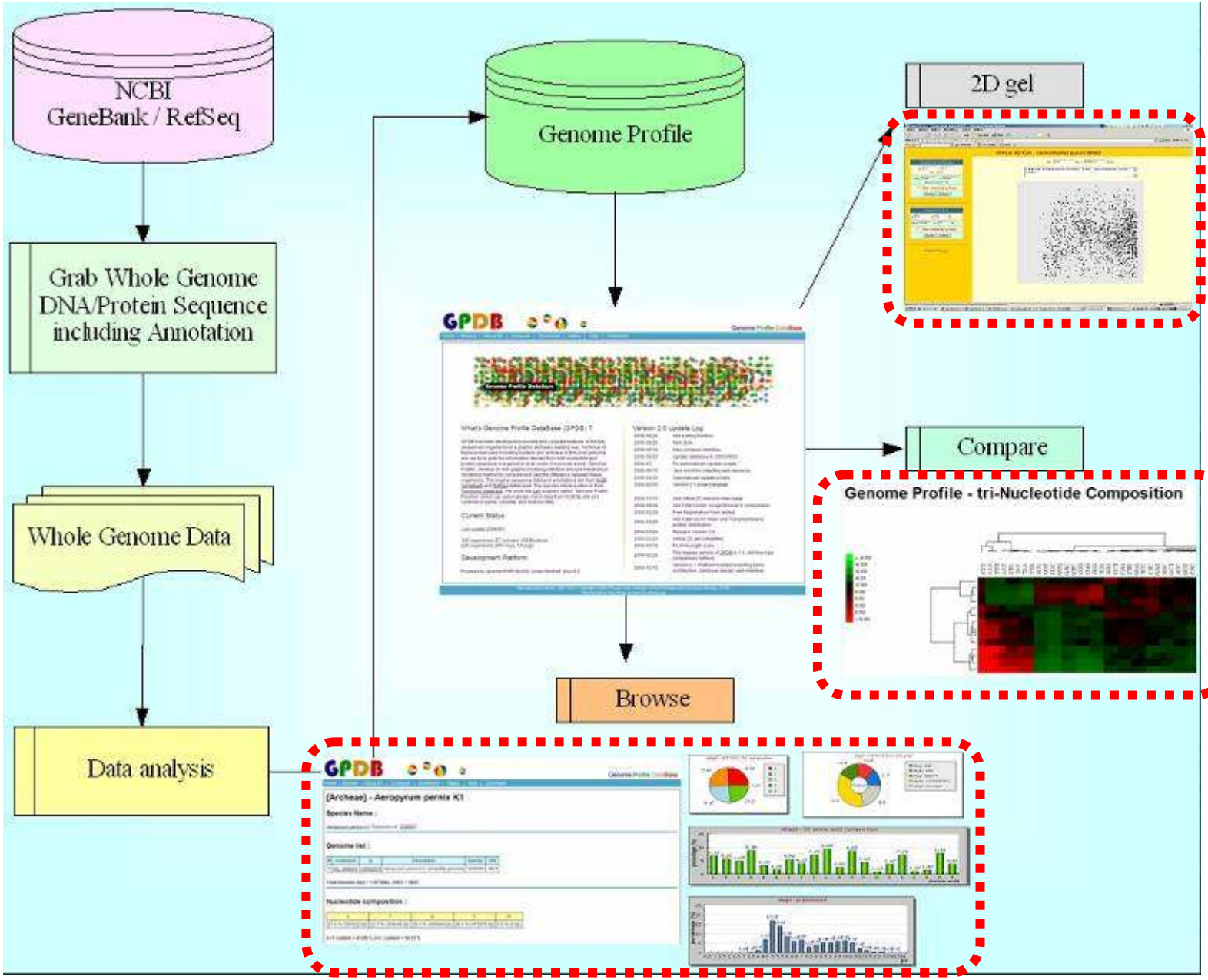
GPDB Current Status

- **550 organisms**
  (27 Archaea, 338 Bacteria, 200 Virus, 5 Fungi)
- **880 complete sequence**
- **Total - 1,176,013 protein (ORFs)**

Conclusions

- We construct the database (GPDB), which provides many whole-genome scale features.
- GPDB can let you explore the relationship between different species using genome-wide profiles.
- It can only compare one genome profile per time now.
- We hope it can compare multiple genome profiles per time in the future.
- We hope this can help to figure out some rules.

Summary


NCBI
GeneBank / RefSeq
Genome Profile
2D gel
Grab Whole Genome
DNA/Protein Sequence
including Annotation
GPDB
Whole Genome Data
Compare
Genome Profile - tri-Nucleotide Composition
Browse
Data analysis
GPDB