



**University of Verona**  
**Department of Biotechnology**  
**Department of Computer Science**



***Ad hoc* improvement in Biotechnology thanks to  
*ad hoc* application of Computer Science**



**Elisa Salvetti**  
**Giovanna E. Felis**



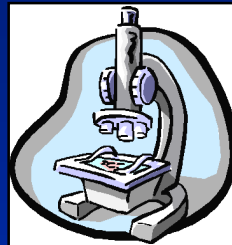
**Diego Dall'Alba**  
**Davide Quaglia**

**Verona, May 18<sup>th</sup>, 2010**

## Dry work in a wet world: computation in biotechnology

Organized merging of computation linked to experimental biology is the most important challenge in modern laboratories.

### BIOTECHNOLOGY



- experiments with many data points

- production of large amount of data from high-throughput technologies

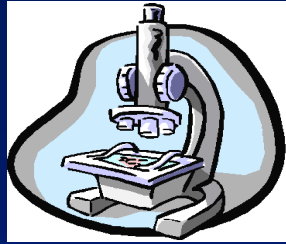
INTEGRATION  
AMONG  
DISCIPLINES



### COMPUTER SCIENCE

- development of specific softwares bundled with instruments

- databases creation for data storage and organization

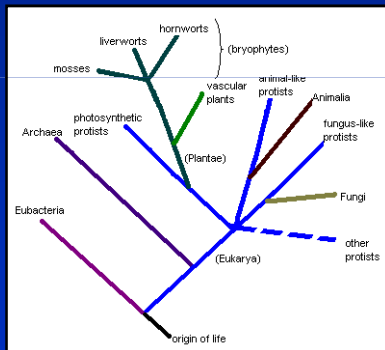


## BIOTECHNOLOGY

# Study case: food microbiology and bacterial taxonomy

## BACTERIAL TAXONOMY

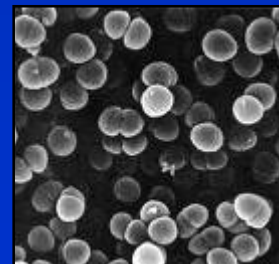
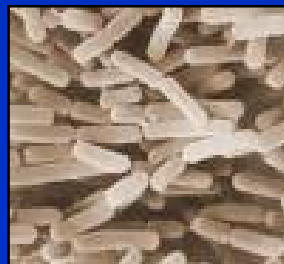
Phylogenetic study and evolution of bacteria

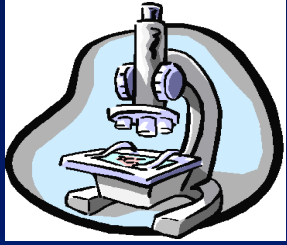


## FOOD MICROBIOLOGY

The study of microorganisms which inhabit, create or contaminate food

*Lactobacillus,*  
*Pediococcus,*  
*Paralactobacillus*  
genera

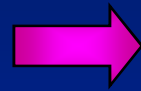




**BIOTECHNOLOGY**

## Study case: basic concepts in microbiology and bacterial taxonomy

**BACTERIAL TAXONOMY**



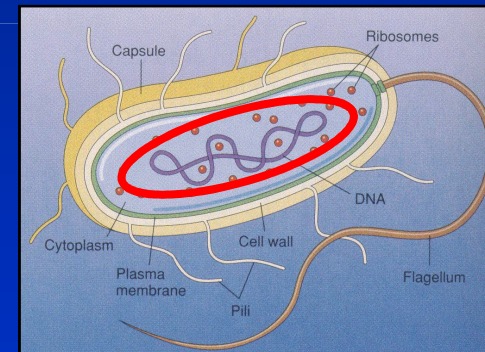
*Polyphasic approach*

- Genotypic information
- Phenotypic information

**GENOTYPE**

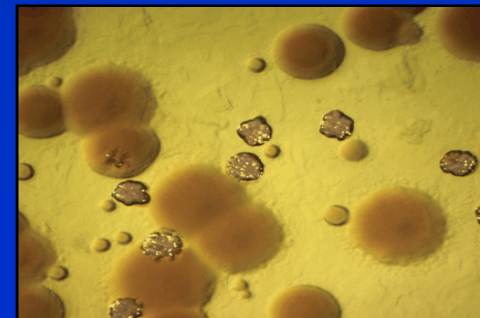
Genetic constitution of the microorganism

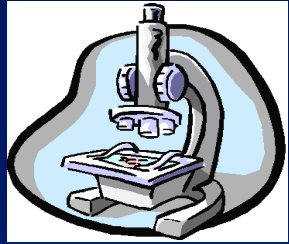
- **GENOME**
- **GENES**



**PHENOTYPE**

Observable characteristics of the microorganism

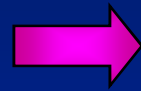




**BIOTECHNOLOGY**

## Study case: basic concepts in microbiology and bacterial taxonomy

**BACTERIAL TAXONOMY**



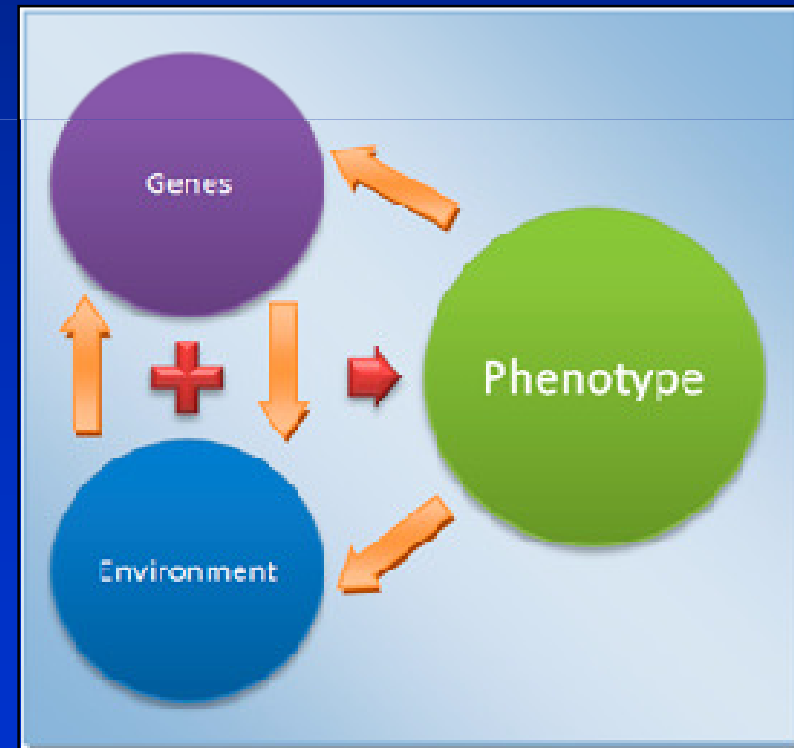
*Polyphasic approach*

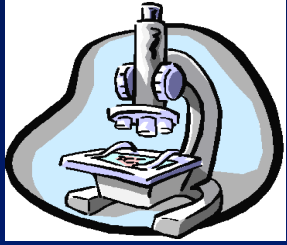
- Genotypic information
- Phenotypic information

**PHENOTYPE**



Result of the **expression of genes** and the **influence of environmental factors**





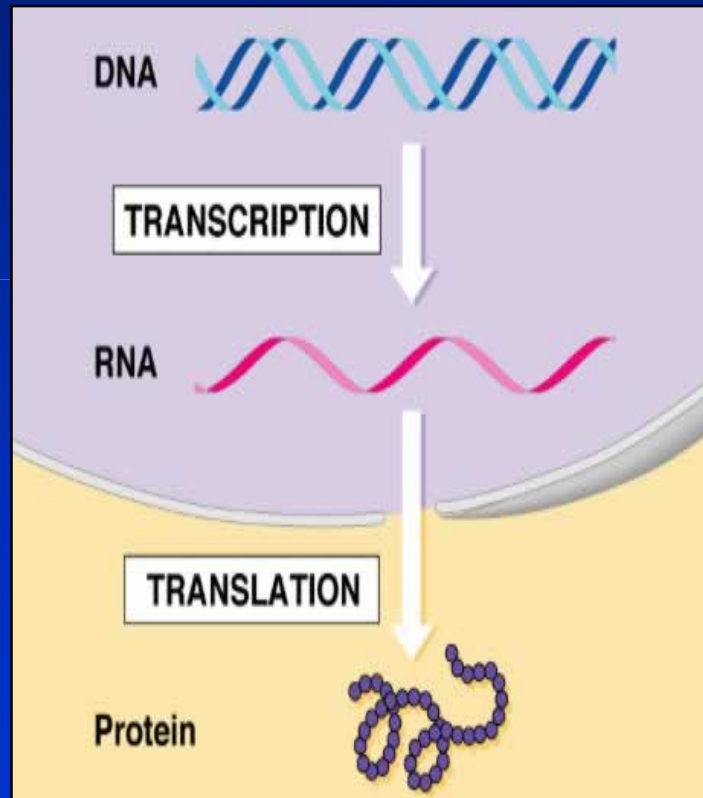
**BIOTECHNOLOGY**

## Study case: basic concepts in microbiology

**GENOTYPE**

### GENE EXPRESSION

- Interpretation of the genetic code
- Origin of organism's phenotype through the properties of expression products (proteins)



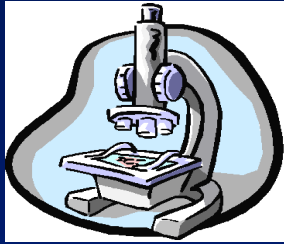
**GENE**



**PROTEIN**

organic compounds which are **essential for the microorganism**

**PHENOTYPE**

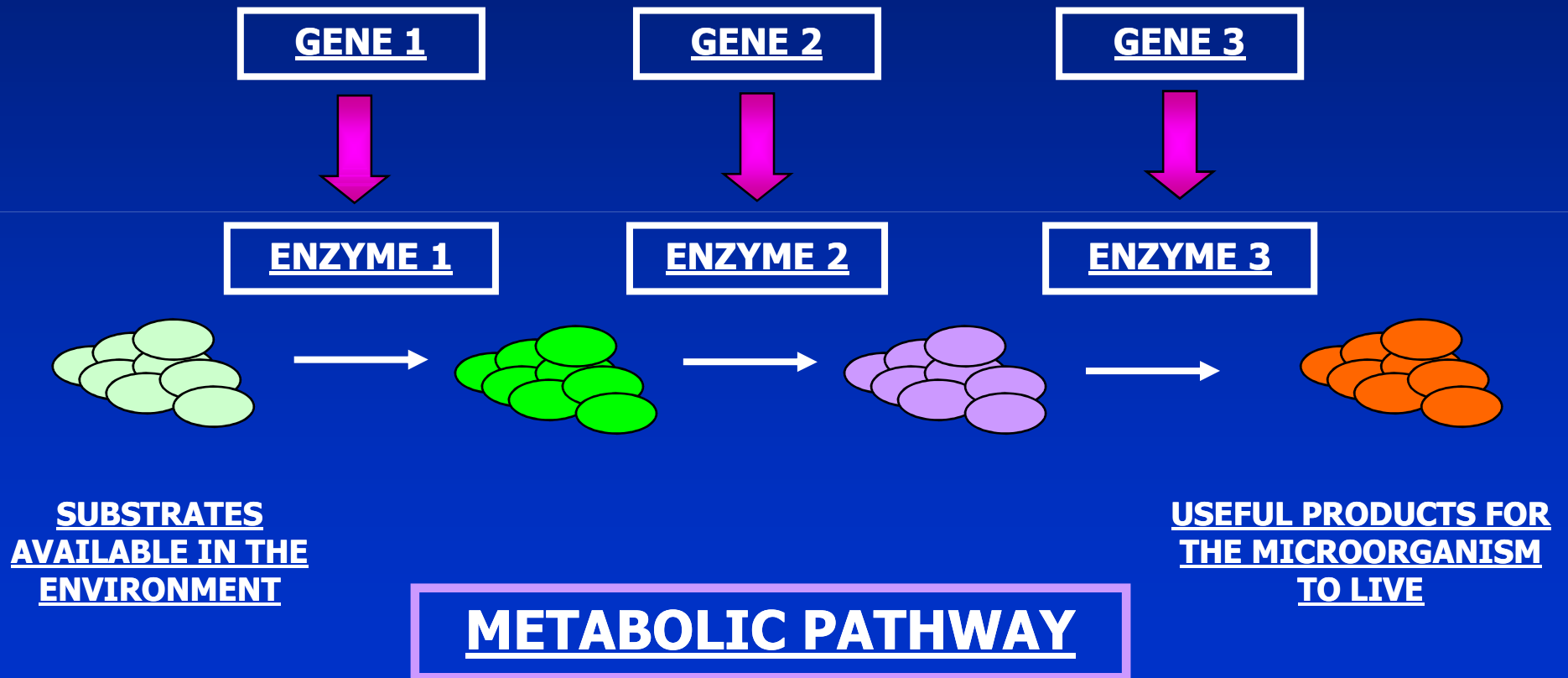


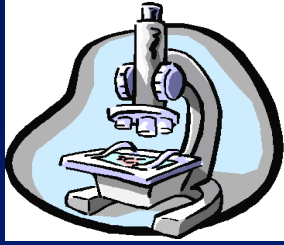
**BIOTECHNOLOGY**

## The study case: basic concepts in microbiology

**PROTEIN = ENZYME**

Catalysis of biochemical reactions  
inside the microorganism





## BIOTECHNOLOGY

# Study case: workflow of the analysis

1. Collection of phenotypic data of species inside a taxonomic group

2. Detection of the heterogeneous phenotypic characters

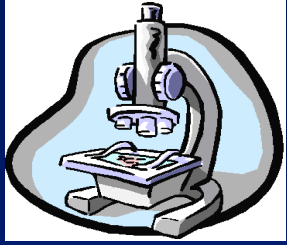
3. Detection of the metabolic pathways involved

4. Investigation of the metabolic pathways on the genomes available

5. Analysis of each gene and its localization on each genome

6. Comparative analysis of these traits between more genomes





## **BIOTECHNOLOGY**

# **Study case: workflow of the analysis**

**1. Collection of phenotypic data of species inside a taxonomic group**

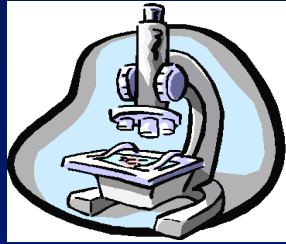
**2. Detection of the heterogeneous phenotypic characters**

**3. Detection of the metabolic pathways involved**

**4. Investigation of the metabolic pathways on the genomes available**

**5. Analysis of each gene and its localization on each genome**

**6. Comparative analysis of these traits between more genomes**



## BIOTECHNOLOGY

# Study case: workflow of the analysis

**4. Investigation of the metabolic pathways on the genomes available**



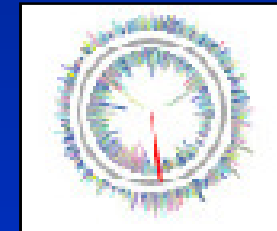
[www.genome.jp/Kegg/](http://www.genome.jp/Kegg/)  
[www.genome.jp/kegg/pathway](http://www.genome.jp/kegg/pathway)



**5. Analysis of each gene and its localization on each genome**



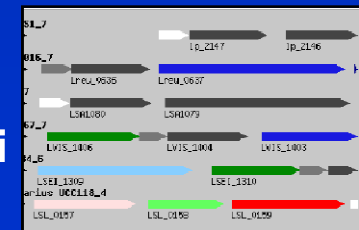
[www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/)  
[www.ncbi.nlm.nih.gov/sites/genome](http://www.ncbi.nlm.nih.gov/sites/genome)

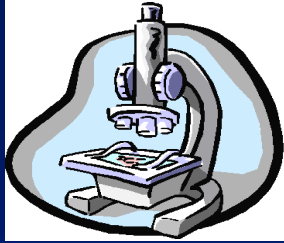


**6. Comparative analysis of these traits between more genomes**



<http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi>





## **BIOTECHNOLOGY**

# **Study case: workflow of the analysis**

- 2 metabolic pathways: 45 genes
- 19 genome sequences available

**4. Investigation of the metabolic pathways on the genomes available**



Control the metabolic pathways on each genome and annotation of the accession number of the genes (38 times)

**5. Analysis of each gene and its localization on each genome**

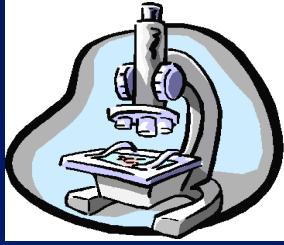


Submission of the accession number related to every gene to control its localization on each genome (855 times)

**6. Comparative analysis of these traits between more genomes**



Submission of the accession number related to every gene and to every genome to compare their position simultaneously (855 times)



## BIOTECHNOLOGY

# Study case: workflow of the analysis

- 2 metabolic pathways: 45 genes
- 19 genome sequences available

**4. Investigation of the metabolic pathways on the genomes available**



**5. Analysis of each gene and its localization on each genome**



**6. Comparative analysis of these traits between more genomes**

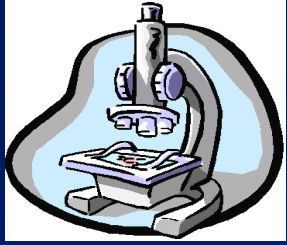


• NOT AUTOMATED STEPS

• TIME CONSUMING PROCESSES

• LIMITED DATA ANALYSIS

• TRICKY DATA VISUALIZATION

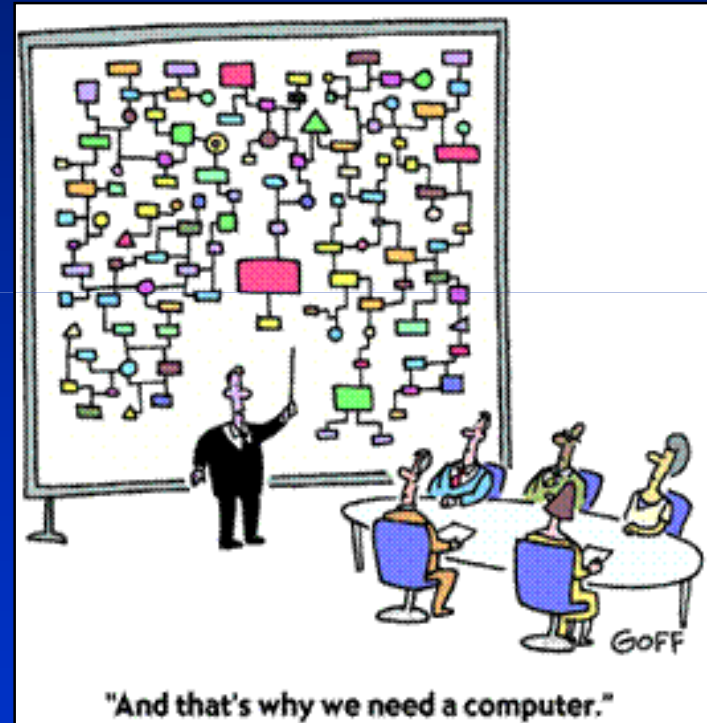


**BIOTECHNOLOGY**



# Study case: pitfalls of the workflow

- NOT AUTOMATED STEPS
- TIME CONSUMING PROCESSES
- TRICKY DATA VISUALIZATION
- LIMITED DATA ANALYSIS





**COMPUTER SCIENCE**

## **Computer science introduction**

**Starting from the workflow described we have developed a specific software, codename ByoGear:**

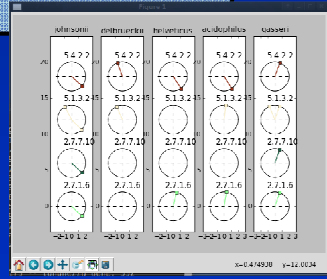
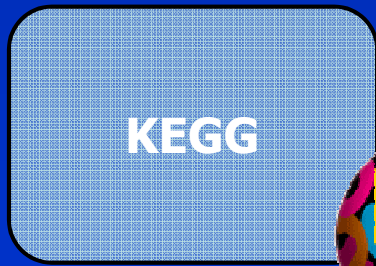
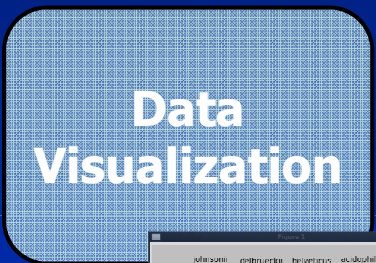
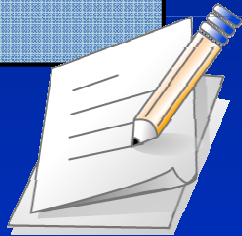
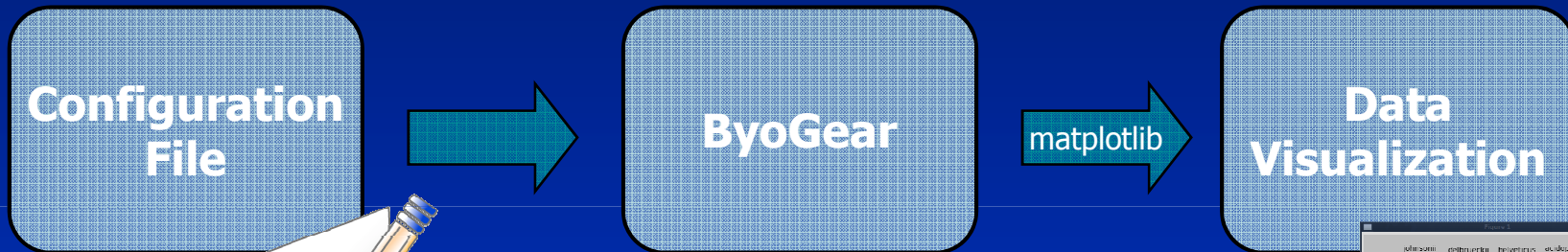
- **Easier and faster procedure execution**
- **Very simple to use (almost no learning needed)**
- **Automatic data visualization**
- **Improvement in the comparison between different genomes**





COMPUTER SCIENCE

# Software Structure

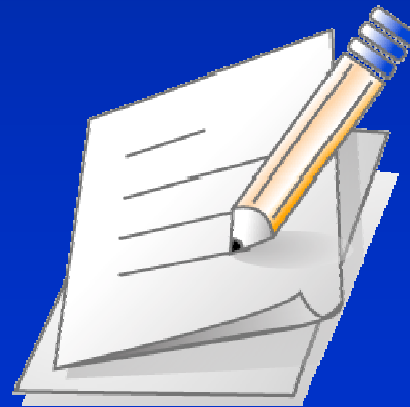




## COMPUTER SCIENCE

# Configuration File

- Very simple
- Useful to keep track of performed test
- Very easy to read (for human users and computer)
- Could be handled with any text editor



```
[Version=0.1]
```

```
[Species]
```

```
L. johsonii  
L. delbrueckii  
L. helveticus  
L. acidophilus  
L. gasserii
```

```
[Enzymes]
```

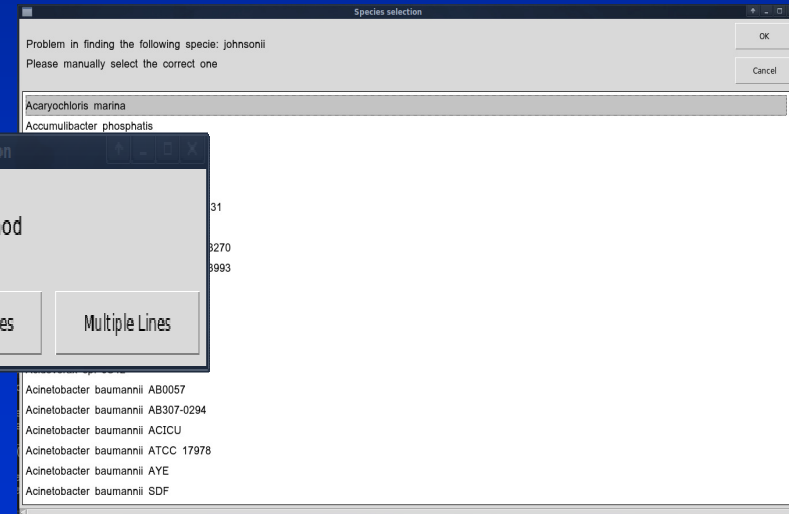
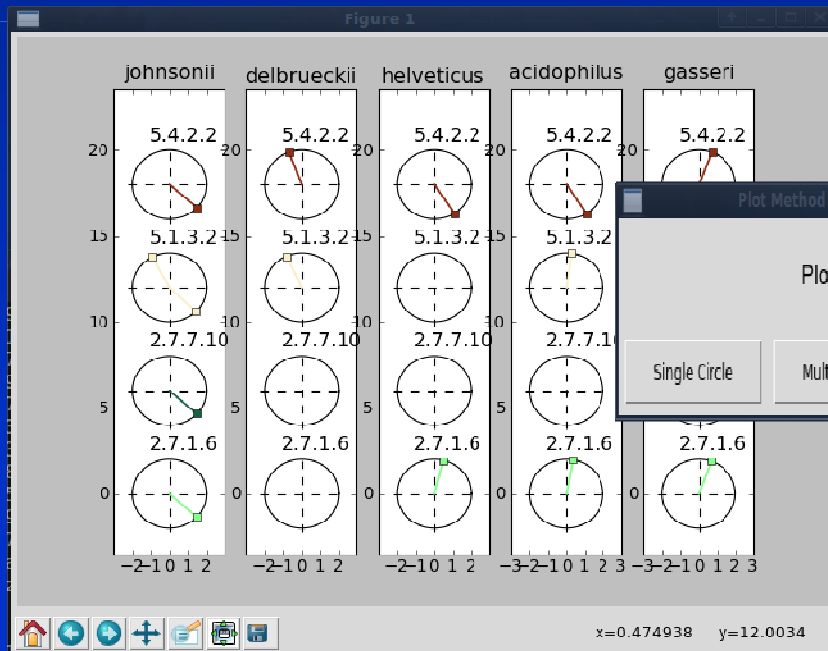
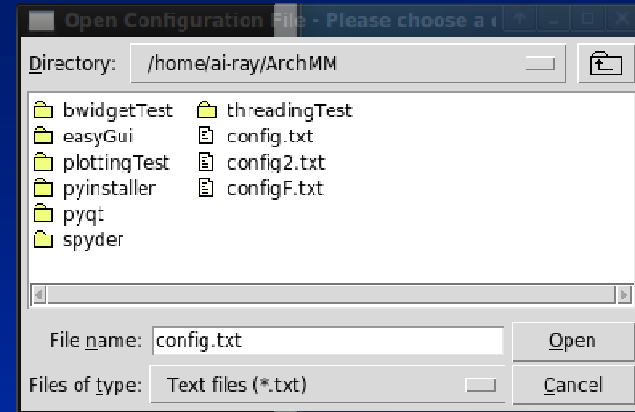
```
2.7.1.6  
2.7.7.10  
5.1.3.2  
5.4.2.2
```





COMPUTER SCIENCE

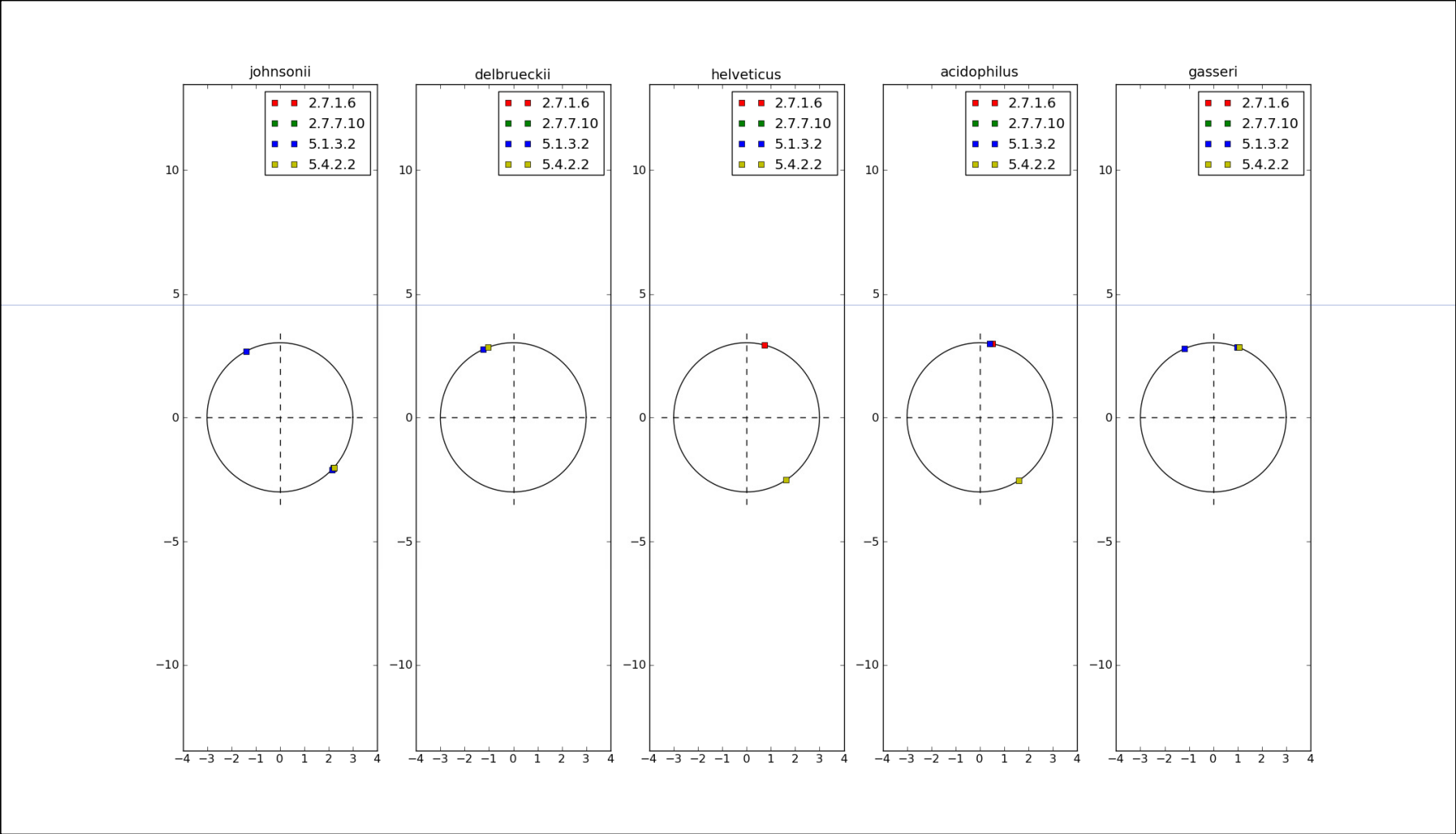
# ByoGear in action





# Data Visualization (1)

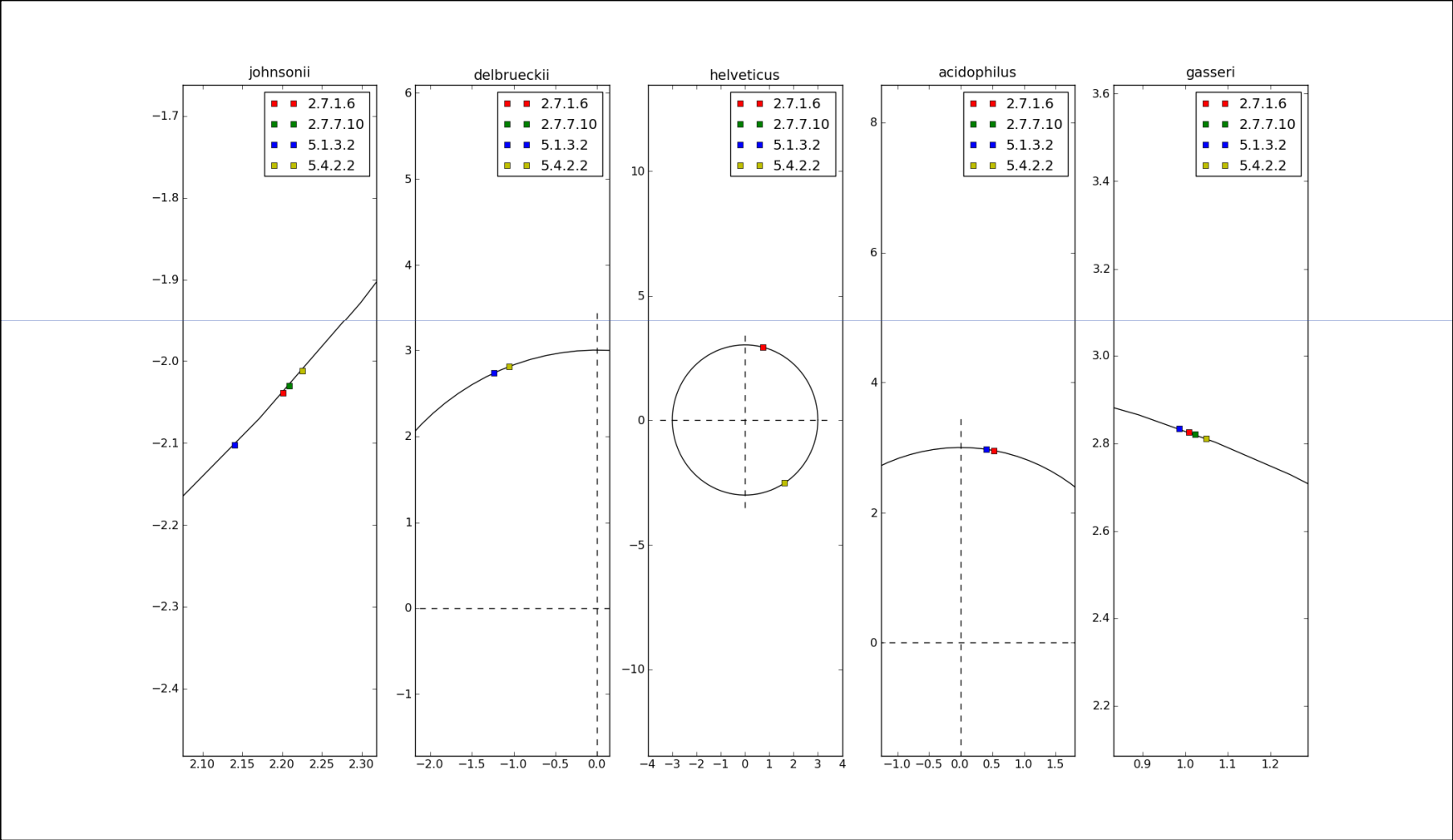
## COMPUTER SCIENCE





# Data Visualization (2)

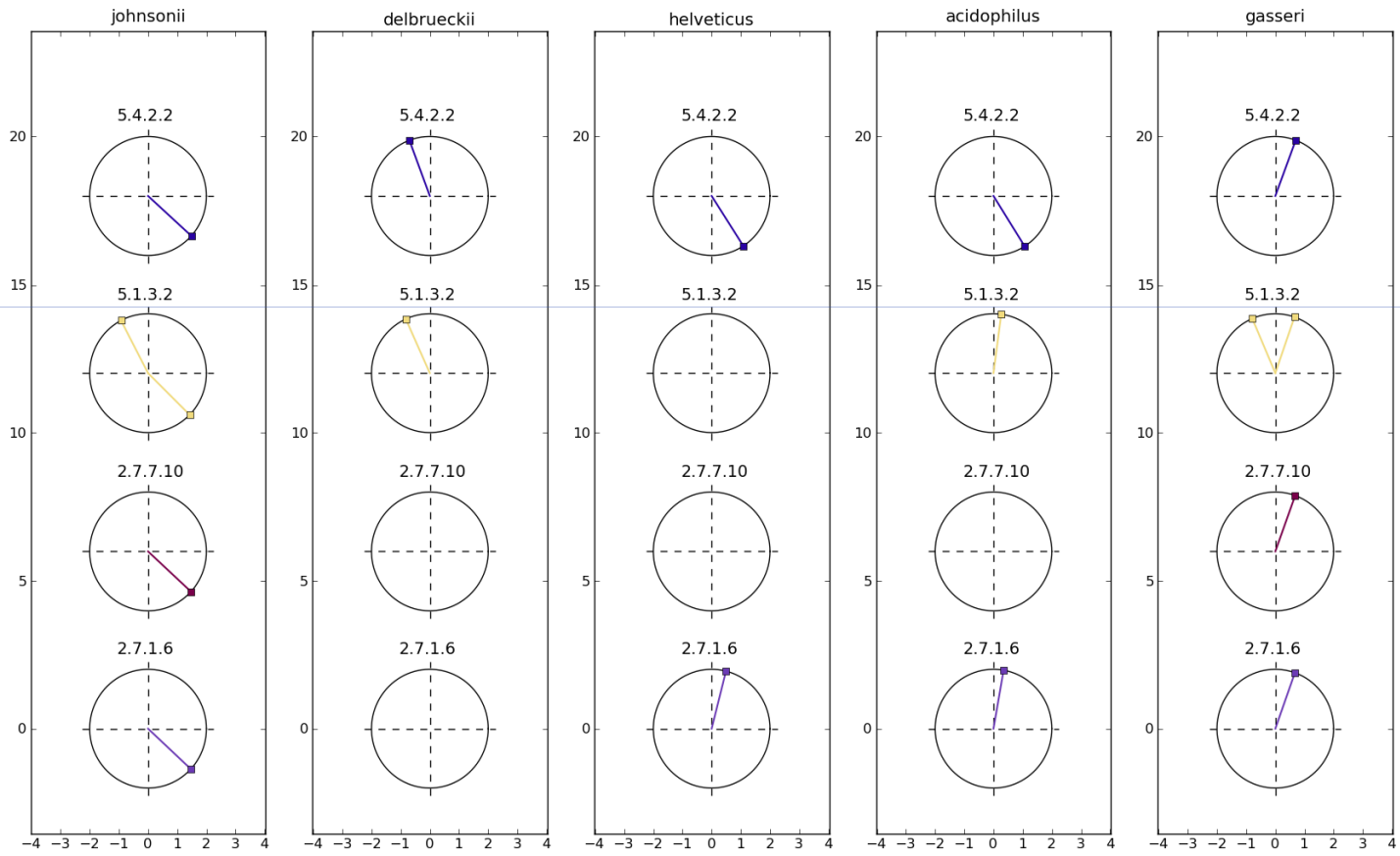
## COMPUTER SCIENCE





# Data Visualization (3)

## COMPUTER SCIENCE





COMPUTER SCIENCE

## Performance analysis

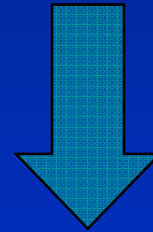
### Simple test case

#### 5 Species

- *Lactobacillus johnsonii*
- *Lactobacillus delbrueckii*
- *Lactobacillus helveticus*
- *Lactobacillus acidophilus*
- *Lactobacillus gasseri*

#### 4 Enzymes

EC 2.7.1.6  
EC 2.7.7.10  
EC 5.1.3.2  
EC 5.4.2.2



Mean computational time over 10 tests:

4 minutes and 30 seconds

Just the time for a (short :-)) coffee break

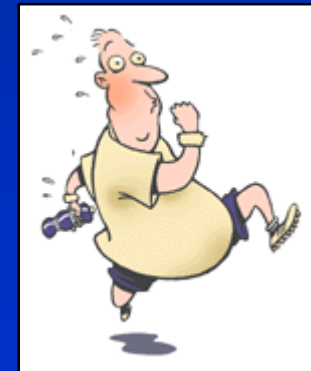




## COMPUTER SCIENCE

# Performance comments

- Great improvement in the execution time...
  - We can further improve
  - Modern computer are multi core, why do not exploit parallel execution?
  - We do not have to wait for sequential service
- 
- Let's take a look at the performance of improved version of ByoGear (multi-threaded version)





COMPUTER SCIENCE

## Performance of multi – thread version

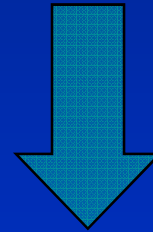
### Same test case

#### 5 Species

- *Lactobacillus johnsonii*
- *Lactobacillus delbrueckii*
- *Lactobacillus helveticus*
- *Lactobacillus acidophilus*
- *Lactobacillus gasseri*

#### 4 Enzymes

E.C. 2.7.1.6  
E.C. 2.7.7.10  
E.C. 5.1.3.2  
E.C. 5.4.2.2



Mean computational time over 10 tests is now:

Less than 40 seconds

Less coffee break for us :-)



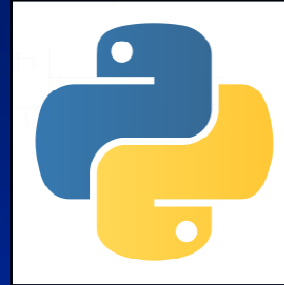


COMPUTER SCIENCE

## Distribution issue

### ByoGear is based on:

- Python
- Easygui
- Scipy
- Numpy
- Matplotlib
- SOAPpy
- Hidden ones...



Very complex to distribute, we ideally want a single executable file...

PyInstaller helps us in making this job straightforward







COMPUTER SCIENCE

## Final comments

### Pros

- Very Fast (compared to Human operator)
- Graphical User Interface makes user interaction easier
- Simple and clear data visualization

### Cons

- Few cases tested
- Still some HCI problems
- Not well organized source code (difficult to make changes)



COMPUTER SCIENCE

## Future work

**Improve software structure (re-engineering with Object Oriented approach)**

**Improve efficiency by redesigning data structure and inter-modules communication**

**Implement a better User Interface (based on pyQT)**

**More tests**





**University of Verona**  
**Department of Biotechnology**  
**Department of Computer Science**



# Thank you very much!



**Elisa Salvetti**  
**Giovanna E. Felis**



**Diego Dall'Alba**  
**Davide Quaglia**

**Verona, February 2<sup>nd</sup>, 2010**