Information Society and Media

Transcriptomics and Phylogenetics Applications in GRID

Outline

- Objectives
- Delivery name and description
- Material and methods
- Results
- Publications
- Conclusion
- Acknowledgment

Objectives

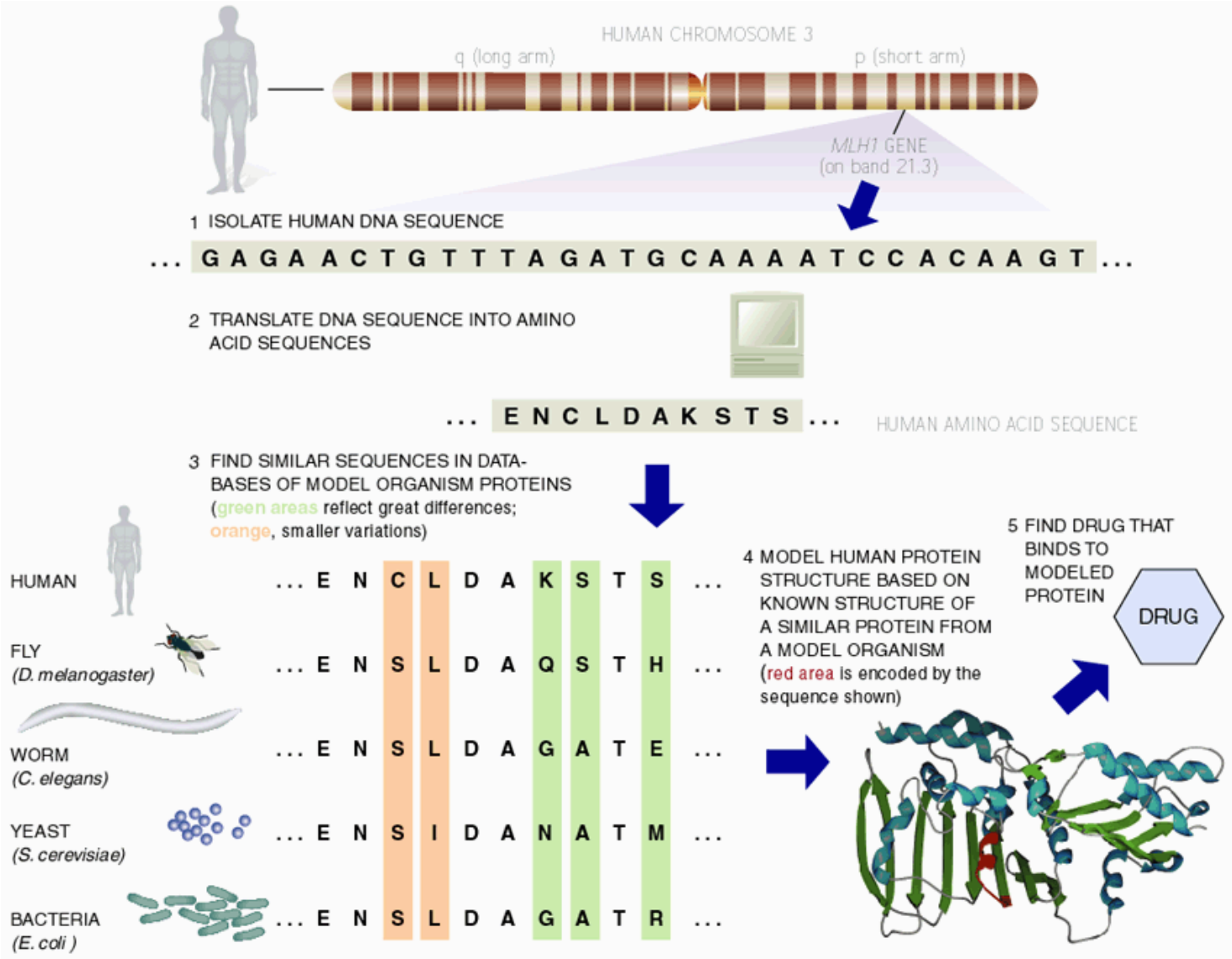
Aim of WP3 : use of computational GRIDs to analyse trascriptomics data and to perform application of Phylogenetic methods based on estimates trees.

- To perform algorithmic tools for gene expression data analysis in GRID: evaluate the computational tools for extracting biologically significant information from gene expression data.
- Algorithms will focus on clustering steady state and time series gene expression data, multiple testing and meta analysis of different microarray experiments from different groups, and identification of transcription sites.

Objectives

- **Objective 1**: Validation of a Grid infrastructure to perform applications of Phylogenetics
- The reconstruction of phylogenetic relationships between gene sequences is a crucial step towards the understanding of their evolution and is done by the building of phylogenetic trees that encapsulate the evolutionary information within its branches which provides a meaningful way to order large amounts of data

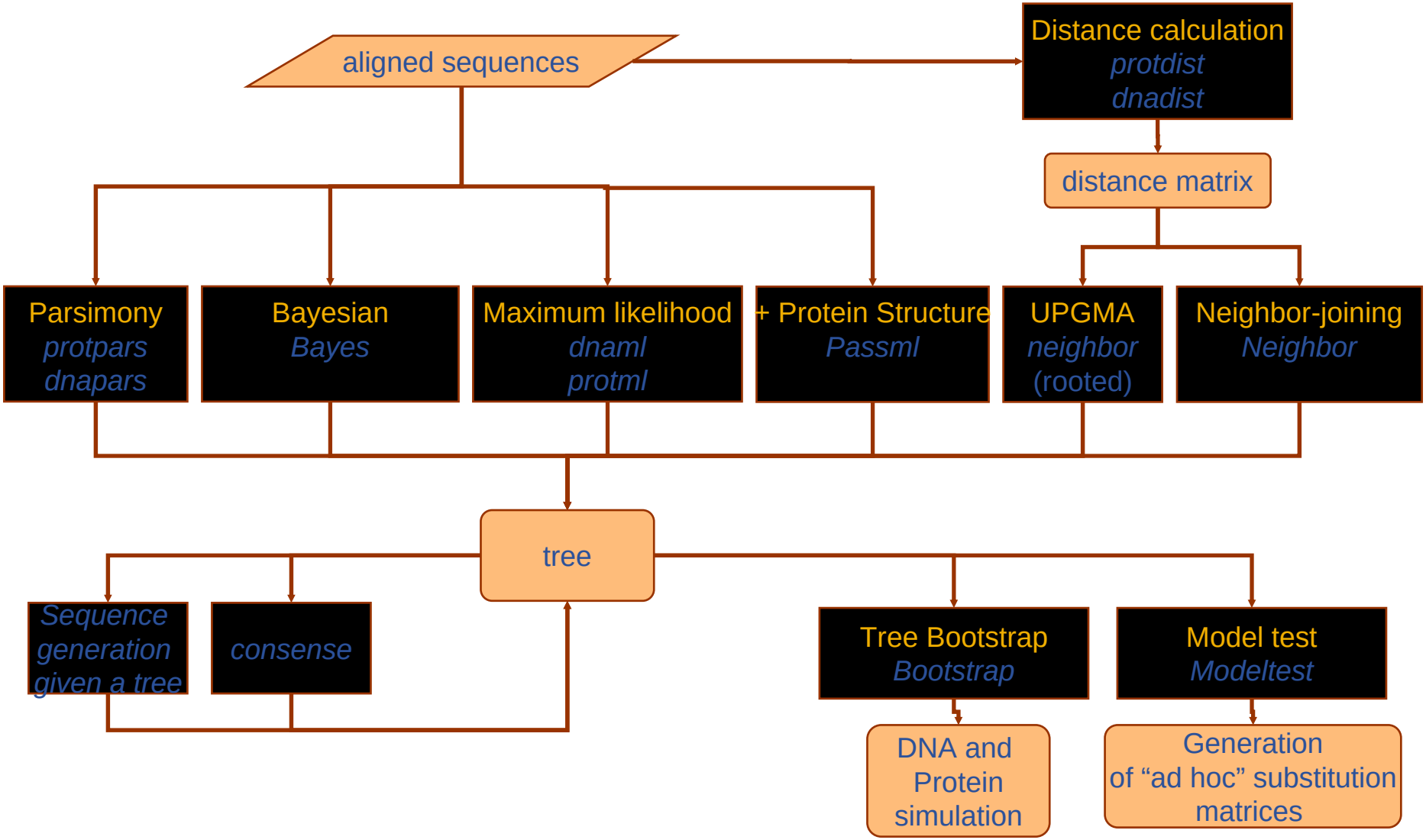
Phylogenetics : the reasoning



Comparison of Methods

Neighbour-joining	Maximum parsimony	Maximum likelihood
Uses only pairwise distances	Uses only shared derived characters	Uses all data
Minimizes distance between nearest neighbors	Minimizes total distance	Maximizes tree likelihood given specific parameter values
Very fast	Slow	Very slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, homoplasy rare)	Good for very small data sets and for testing trees built using other methods

Phylogenetics in GRID



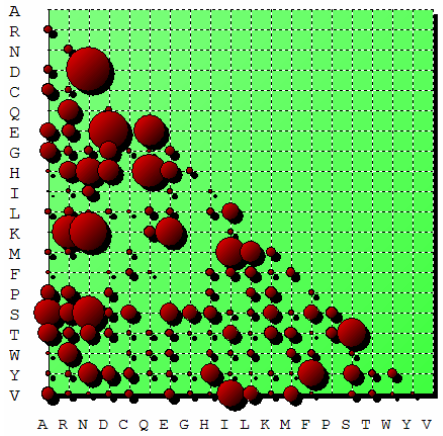
Maximum Likelihood

a) Alignment of protein sequences

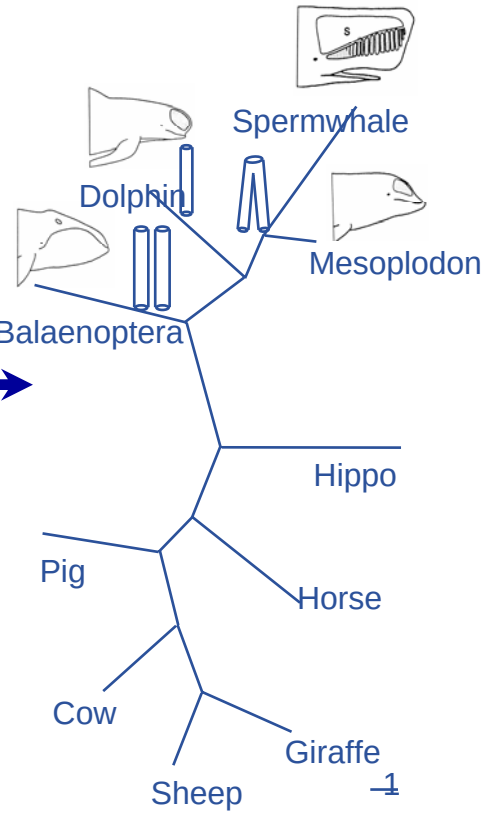
b) Model of evolution + Statistical framework

c) Tree

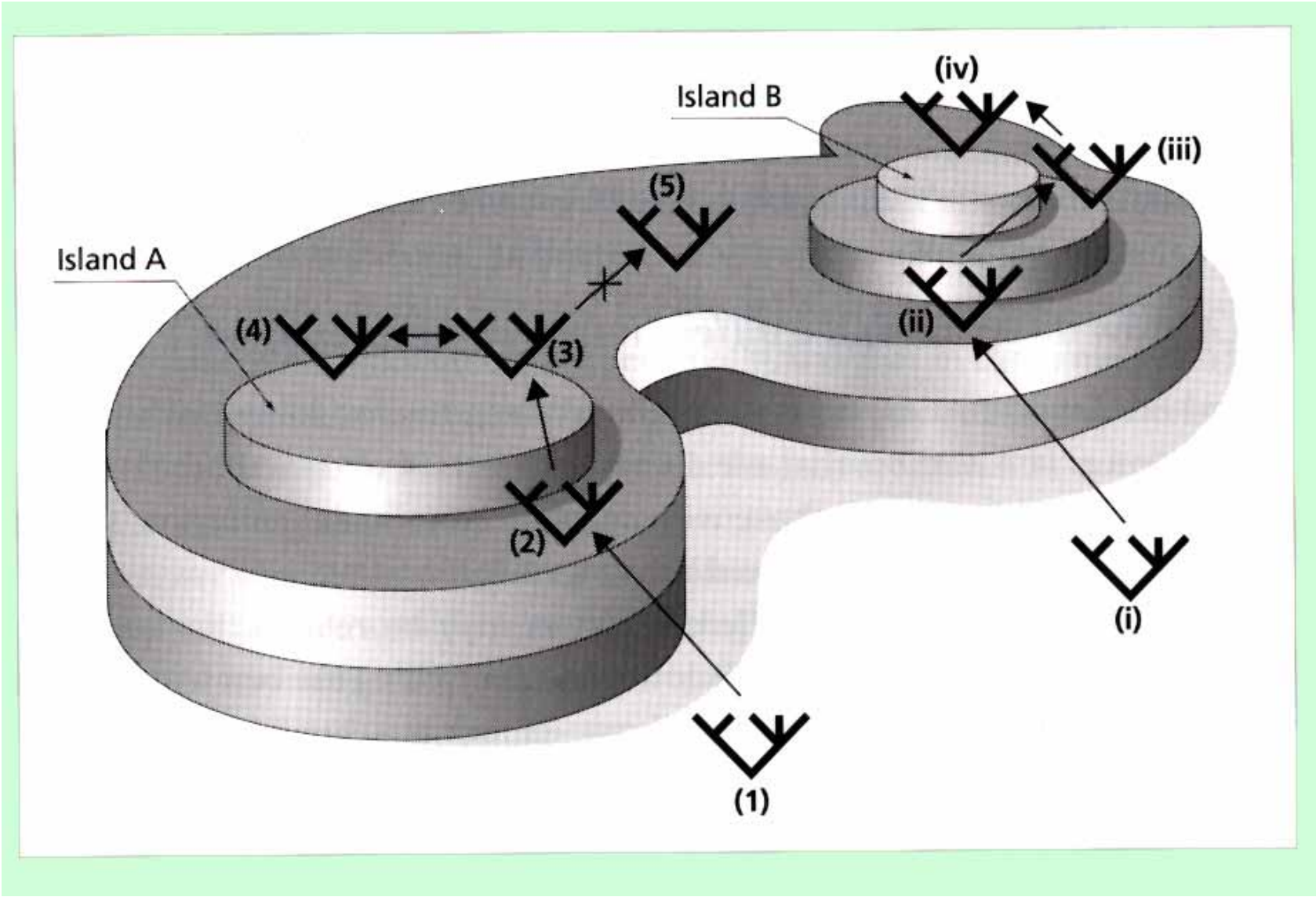
```
                1                 379
Balaenoptera    MTNIRKTHPLMKII..W
Cow             MTNIRKSHPLMKIV..W
Dolphin         MTNIRKTHPLMKIL..W
Giraffe         MINIRKSHPLMKIV..W
Hippo           MTNIRKSHPLMKII..W
Horse           MTNIRKSHPLIKII..W
Mesoplodon      MTNIRKTHPLMKIV..-
Pig             MTNIRKSHPLMKII..W
Sheep           MINIRKTHPLMKIV..W
Sperm whale     MTNIRKSHPLMKII..W
```



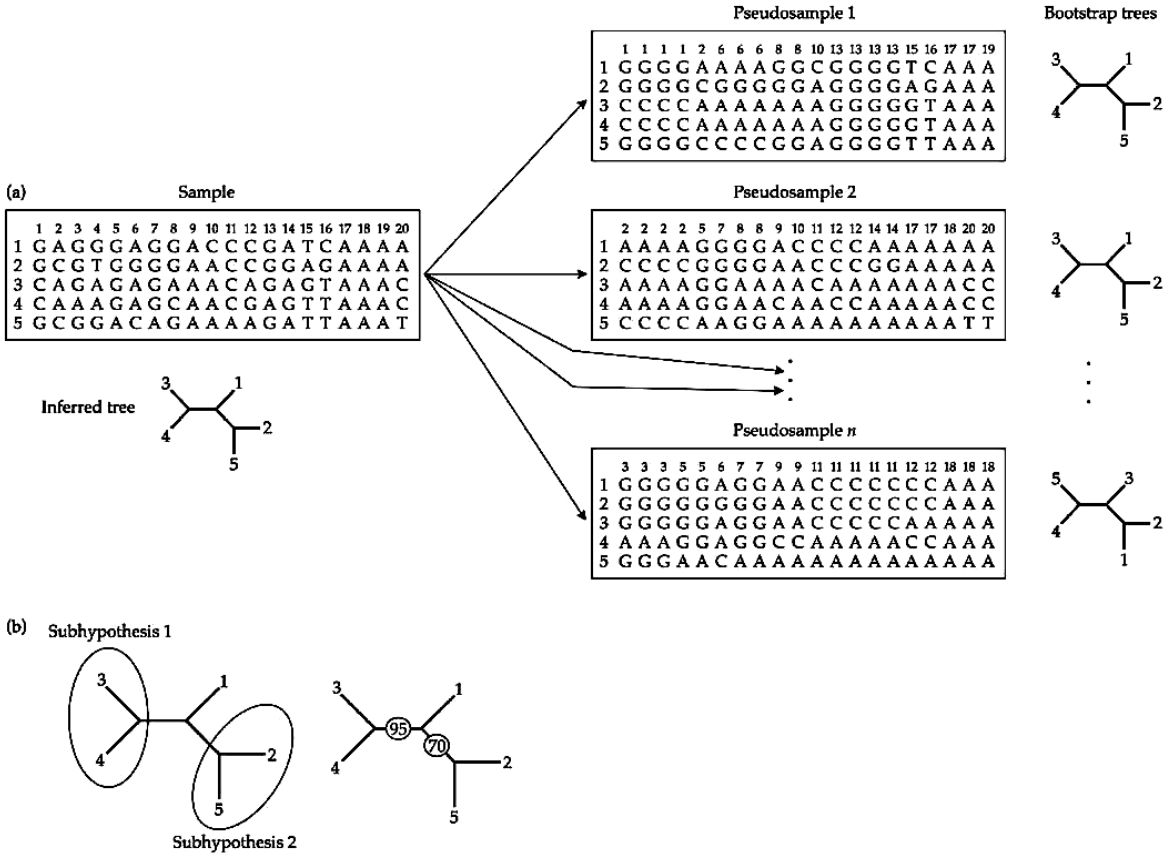
$$L = \Pr(D \mid Tree, \text{model})$$



Applications of Phylogenetics



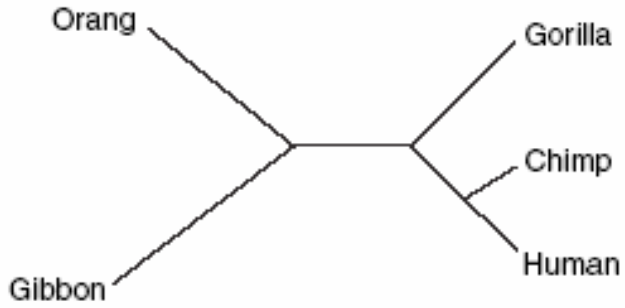
Bootstrap


(a)
Sample
Inferred tree
Pseudosample 1
Pseudosample 2
Pseudosample n
Bootstrap trees
(b)
Subhypothesis 1
Subhypothesis 2
95
70

Statistical test of models

(a)

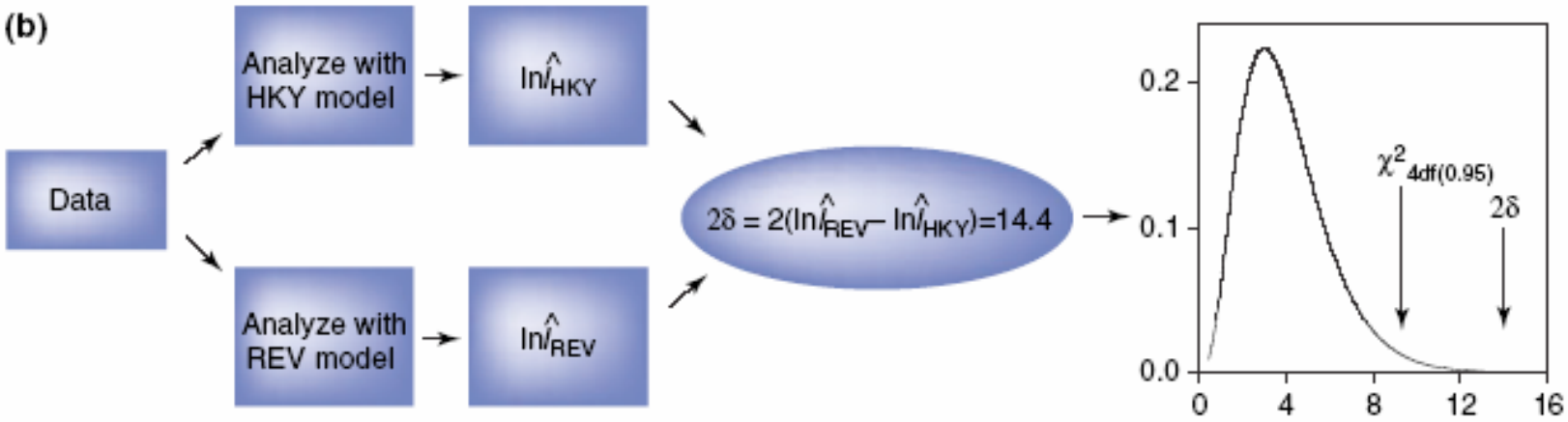
Human	AAGCTTCACC GGCGCAGTCA ...
Chimp	AAGCTTCACC GGCGCAATTA ...
Gorilla	AAGCTTCACC GGCGCAGTTG ...
Orangutan	AAGCTTCACC GGCGCAACCA ...
Gibbon	AAGCTTTACA GGTGCAACCG ...



$\ln\hat{l}_{HKY} = -2665.42$

$\ln\hat{l}_{REV} = -2658.22$

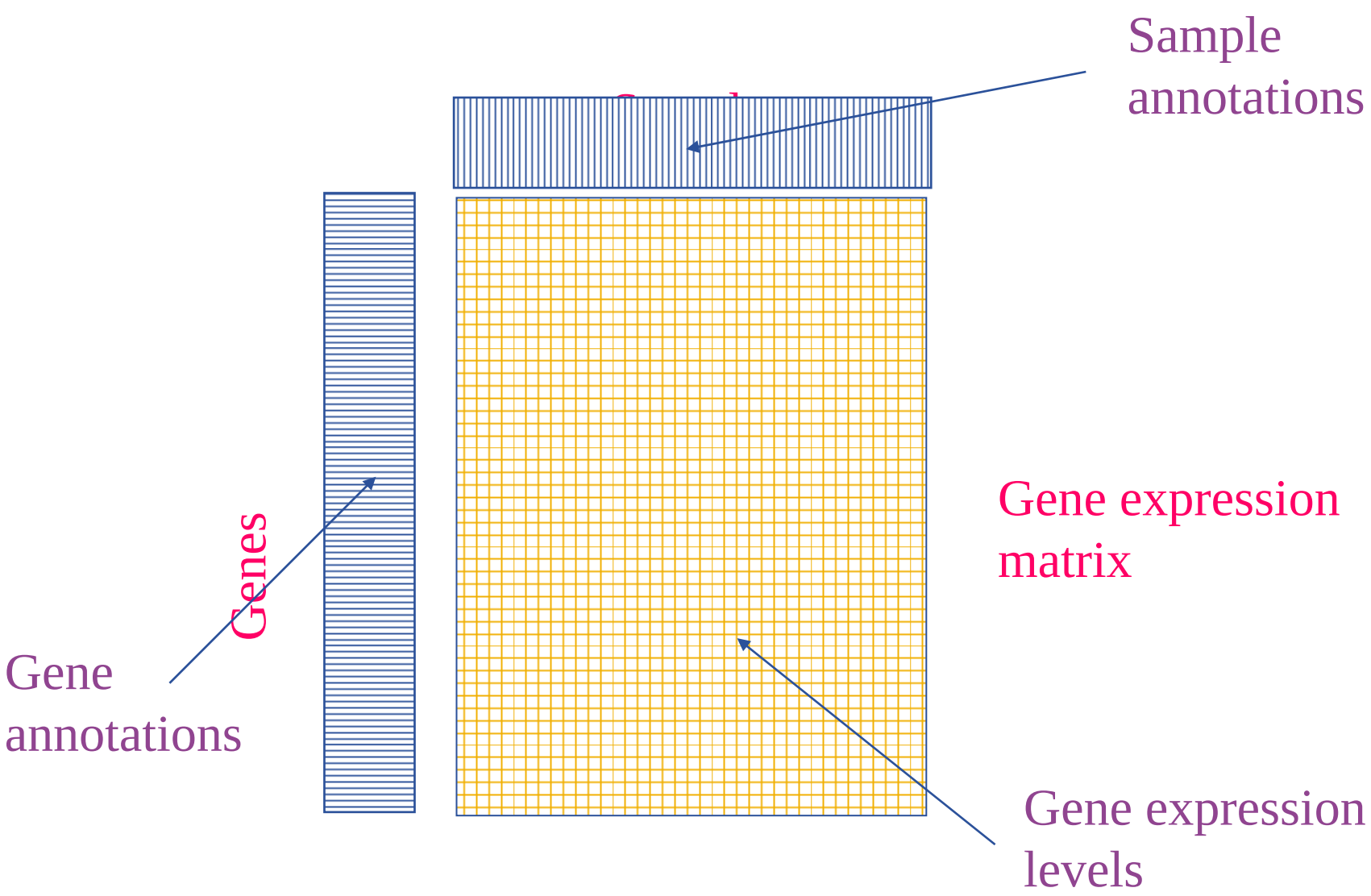
(b)



Data analysis for microarray in GRID

- **Objective 2**: Validation of data analysis for microarray in GRID.
- Transcriptomics encapsulates the analysis of global gene expression patterns and is a key area due to the fact that the adaptation to variable conditions of a cell or an organism as well as its development and differentiation is largely determined by the profile of gene expression.
- DNA array technology, allowing simultaneous large-scale expression measurements on great number of genes using minute samples, is increasingly becoming the method of choice for gene expression profiling. The applications of microarray are deployed in research fields such as pharmacology, diagnostics and environmental monitoring.

Gene expression data

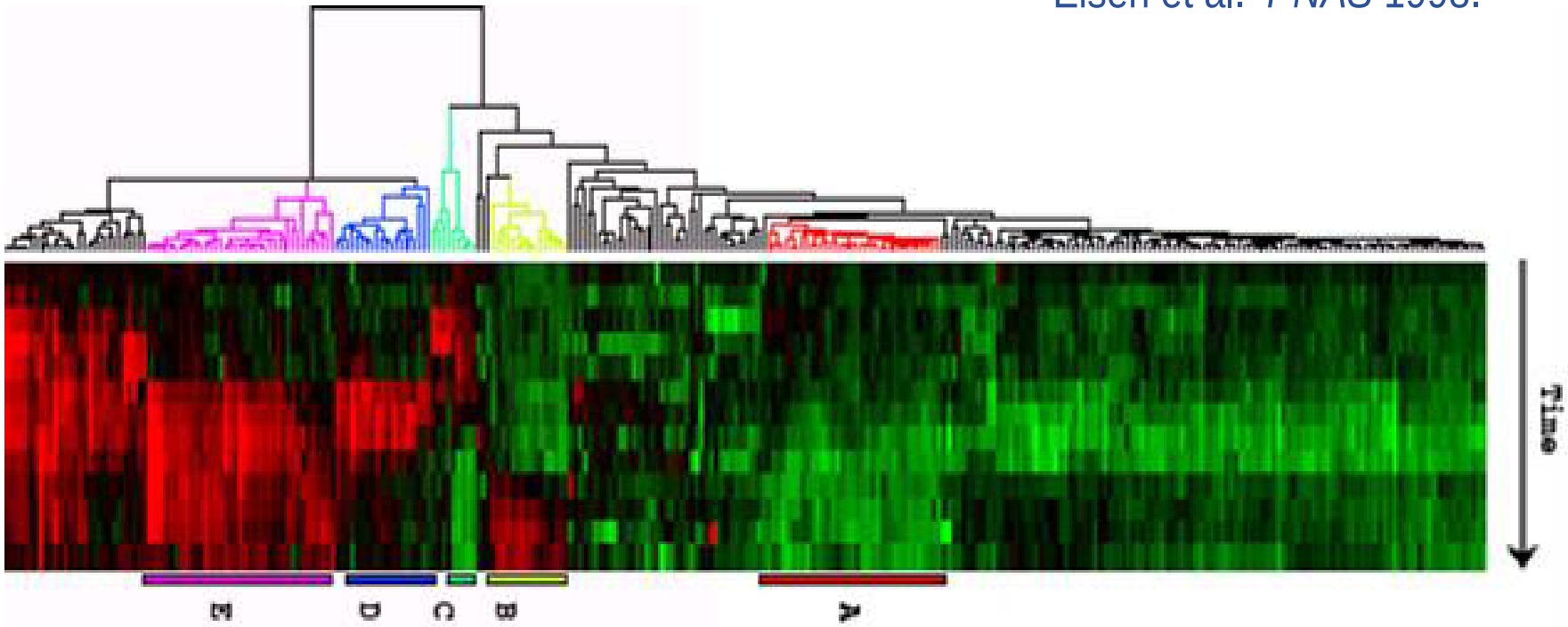


Gene expression data Clustering

- Clustering
  - **Idea:** Groups of genes that share similar function have similar expression patterns
    - Hierarchical clustering
    - k-means / fuzzy k-means
    - Bayesian approaches
    - Projection techniques
      - Principal Component Analysis
      - Independent Component Analysis
- Classification
  - **Idea:** A cell can be in one of several states
    - (Diseased vs. Healthy, Cancer X vs. Cancer Y vs. Normal)
  - Can we train an algorithm to use the gene expression patterns to determine which state a cell is in?
    - Support Vector Machines
    - Decision Trees
    - Neural Networks
    - K-Nearest Neighbors

Gene expression data Clustering

Eisen et al. *PNAS* 1998.

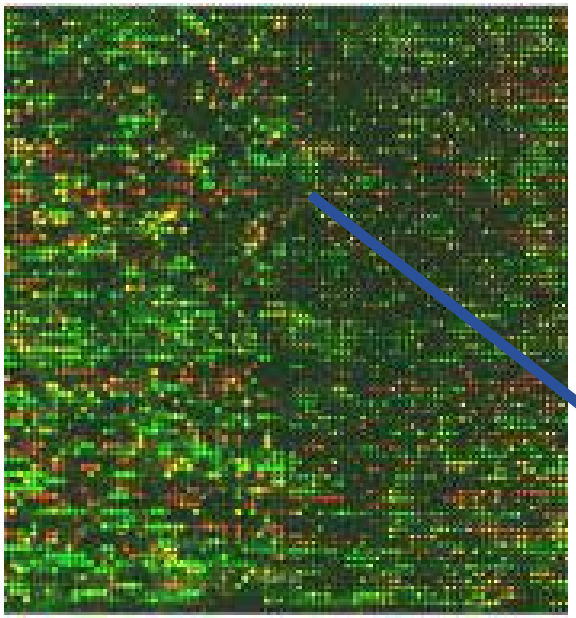


Green = Expression level low with respect to reference sample.
Red = Expression level high with respect to reference sample.
Black = Expression level comparable to reference sample.

The columns are ordered such that similar expression profiles neighbor each other.

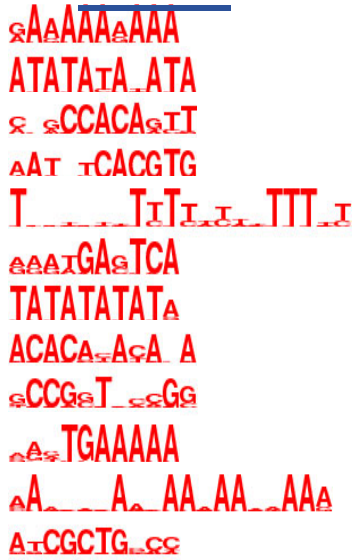
Gene expression data Analysis

Gene expression

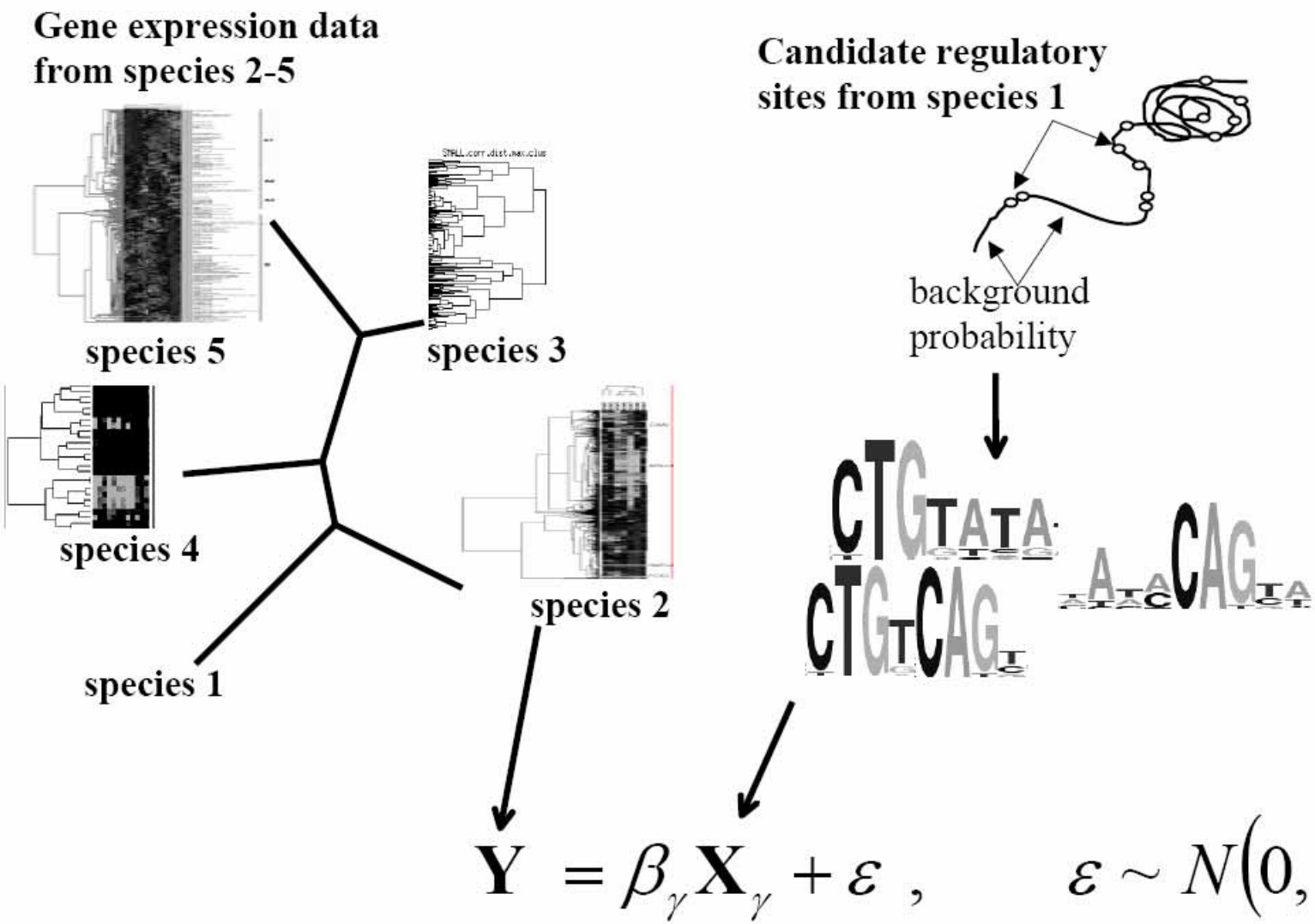


$$Y = \beta_\gamma X_\gamma + \varepsilon\ , \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

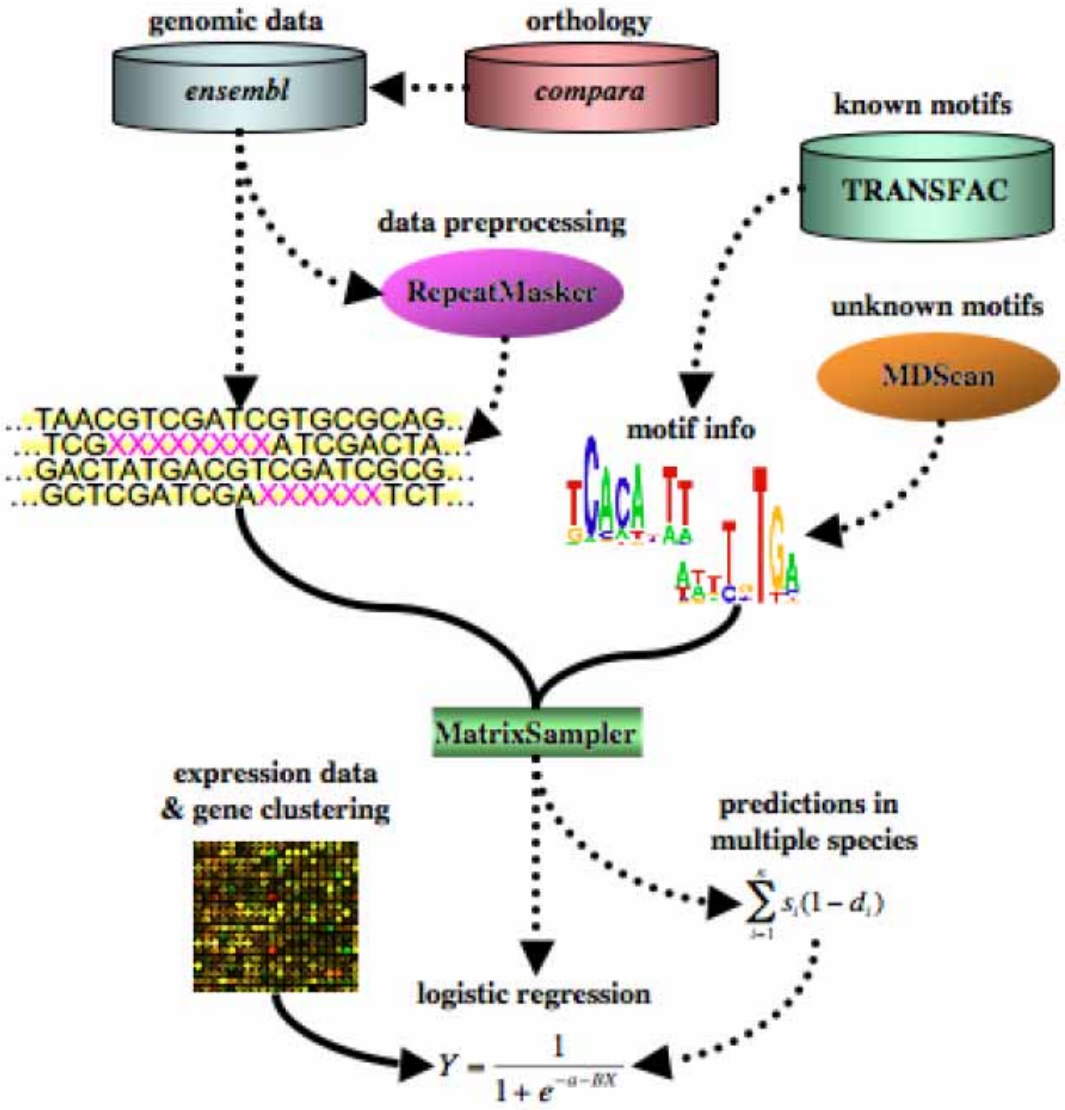
Motif



Gene expression data Analysis



Gene expression data Analysis



Publications

- Elizabeth van der Wath, Loukas Moutsianas, Richard van der Wath, Alet Visagie, Luciano Milanesi and Pietro Lio` Grid methodology for identifying co-regulated genes and transcription factor binding sites TNB Special Issue-06 (in Press)
- Angelini C, Cutillo L, De Feis I, van der Wath R, Lio' P, Identifying regulatory sites using neighborhood species. EVOBIO 2007 5th European Conference On Evolutionary Computation Machine Learning And Data Mining In Bioinformatics, Valencia, Spain, 11-13 April 2007 (in press)
- M. Brilli, R. Fani, P. Lio' MotifScorer: using a compendium of microarrays to identify regulatory motifs Bioinformatics 2006; doi: 10.1093/bioinformatics/btl607
- Pietro Lio' Wavelet algorithms for time series analysis and synchronization. Bioinformatics (submitted)
- Bianchi L. and Pietro Lio' Bioinformatics and forensic DNA evidences Brief. Bioinformatics

Conclusions

- We have selected a key clustering algorithms as methods for gene expression data analysis.
- Our current selection comprises of k-Means, Hierarchical, PCA (Principal Component Analysis) and SVD (Singular Value Decomposition).
- At present we are testing existing programs to reach optimal execution levels in a grid environment.
- Microarray services are offered at more than one research lab centre. Our aim is to rationalise and simplify the connection between them by using GRID enabling software.
- In order to evaluate the possibility to share Transcriptomics data, we will interconnect a few different laboratories in grid in order to share data and analysis tools. So far we have implemented software built by Cheng Li, named dChip.

Conclusions

- The final application interface will be integrated with the other work packages this procedure will be planned during the second year of the project activity.
- The test of the evaluations on cluster algorithms for microarray have been initiated.
- A complete data analysis chain tools has also been identified.
- It also became apparent that some jobs executes at record times where others may wait in a queue for long periods or simply abort.
- To overcome this problem it would be beneficial if the work flow can be constructed in such a way that it will continue if a certain percentage of jobs for a specific goal have finished if they are not interdependent.
- All the development was done from an INFN portal.

Acknowledgments

- Elizabeth Van der Wath
- Loukas Moutsianas
- Richard Van der Wath
- Hemant Kumar Choudhary
- Giorgio Maggi
- Roberto Barbera
- Ivan Porro
- Luciano Milanesi