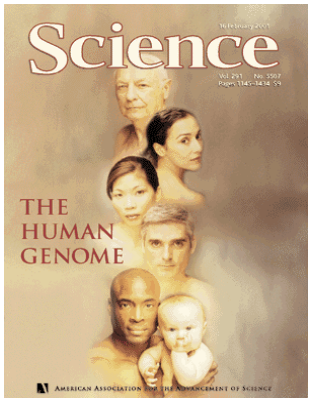
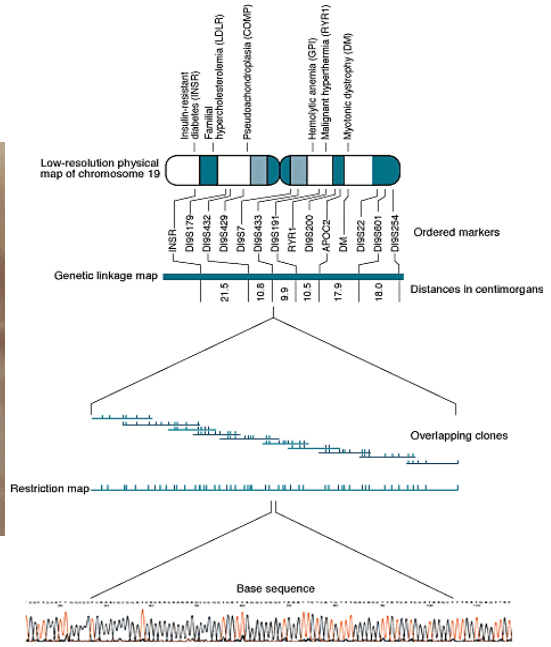
BioinfoGRID Project
Bioinformatics Grid Application for Life Science

Milanesi Luciano (Project coordinator)
National Research Council
Institute of Biomedical Technologies,
Milan, Italy luciano.milanesi@itb.cnr.it

Introduction: Post-genomic

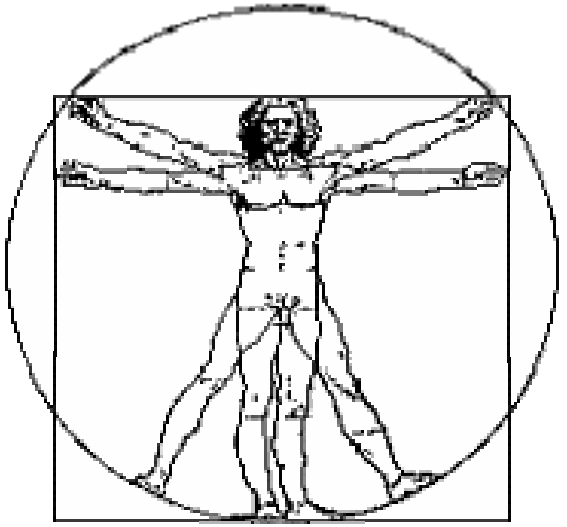
- “Post-genomic” focuses on the new tools and new methodologies emerging from the knowledge of genome sequences.
- Production and use of DNA micro arrays, analysis of transciptome, proteome, metabolome are the different topics developed in this class.

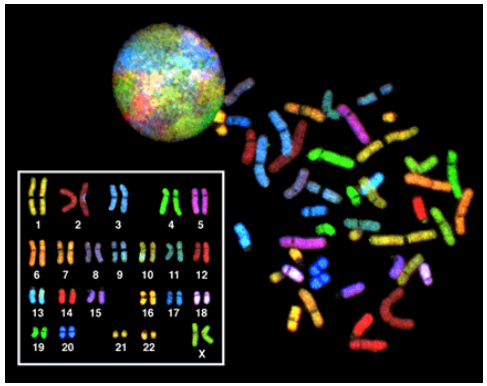




The human organism:

- ~ 3 billion nucleotides
- ~ 30,000 genes coding for
- ~ 100,000-300,000 transcripts
- ~ 1-2 million proteins
- ~ 60 trillion cells of
- ~ 300 cell types in
- ~14,000 distinguishable morphological structures





Genome-wide analysis

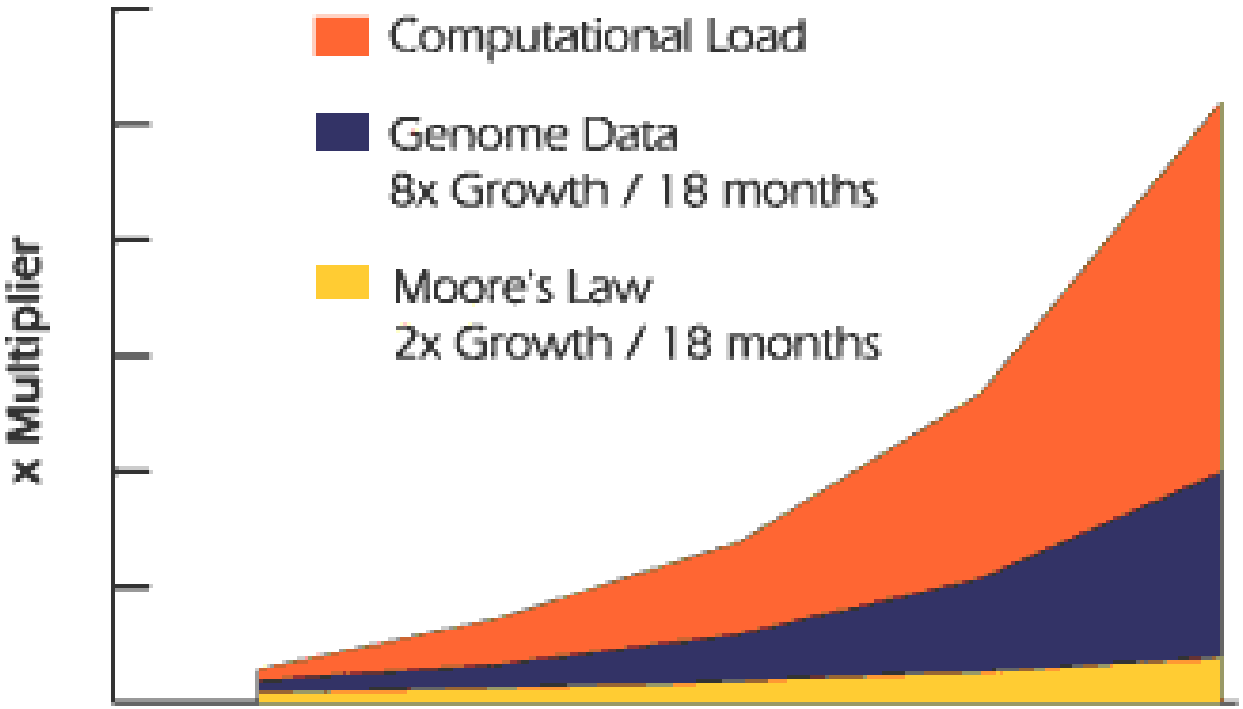
- Current interest in the genome-wide analysis of cells at the level of transcription ('**transcriptome**') and translation ('proteome'), the third level of analysis is the 'metabolome'.
- The term '**metabolome**' refers to the entire complement of all the small molecular weight metabolites inside a cell suspension of interest.
- A new level of experiments are required to obtain an overall picture of **when, where, and how gene are expressed**.
- The **functional genomics** includes:
- The analysis of gene expression profiles at the mRNA and protein levels
- The analysis of polymorphism or mutation patterns in the genome

Human Genetic Diversity

- Any two individuals differ in about **3×10^6 bases (0.1%).**
- The population is now about **5×10^9**.
- A catalog of all sequence differences would require **15×10^{15}** entries.
- This catalog may be needed to find the rarest or most complex disease genes.
- Less than **5%** of the **3×10^6 bases genome** encodes genes.
- A conservative estimate of the number of genes gives a value of about **25.000 genes**.
- The structural diversity within the proteins encoded by these genes is considerably greater than this small number of true genes.

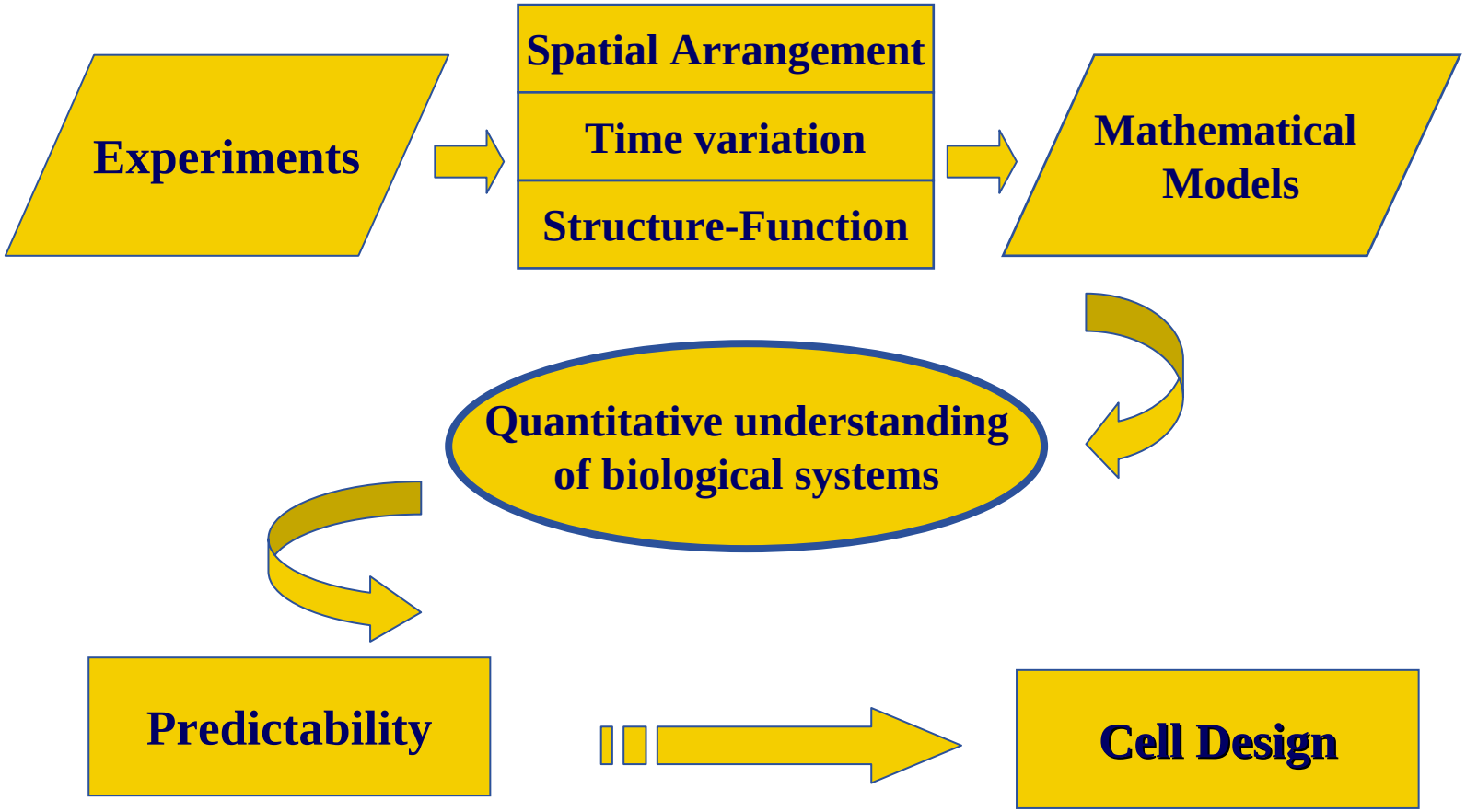
Database explosion

- In the very beginning of the genome sequencing era, Walter Gilbert and colleagues warned of database explosion, stemming from the exponentially increasing amount of incoming DNA sequence



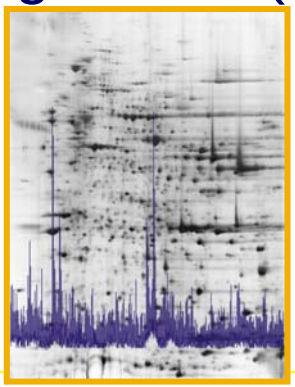
Quantitative Biology

In silico comparative analysis of genes can be used in cell cycle for structural motifs determination in the homologous protein families and for system biology

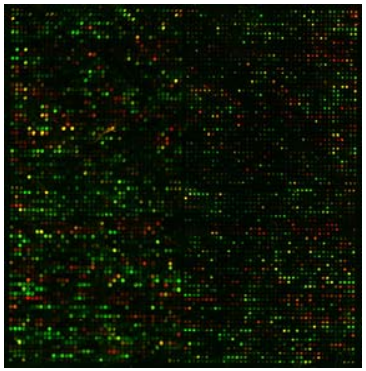


Functional Genomics

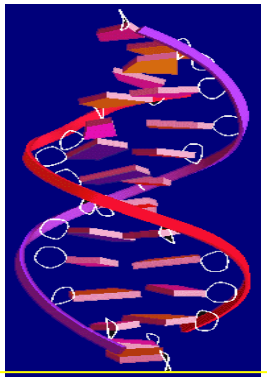
- A new level of systematics experiments is required to obtain an overall picture of When, Where, and How gene are expressed. The study of *functional genomics* includes:
- The analysis of gene expression profiles at the mRNA and protein levels (*Proteome*) and
- The analysis of polymorphism or mutation patterns in the genome (eg. *DNA chips*).



Proteins (Proteomics)

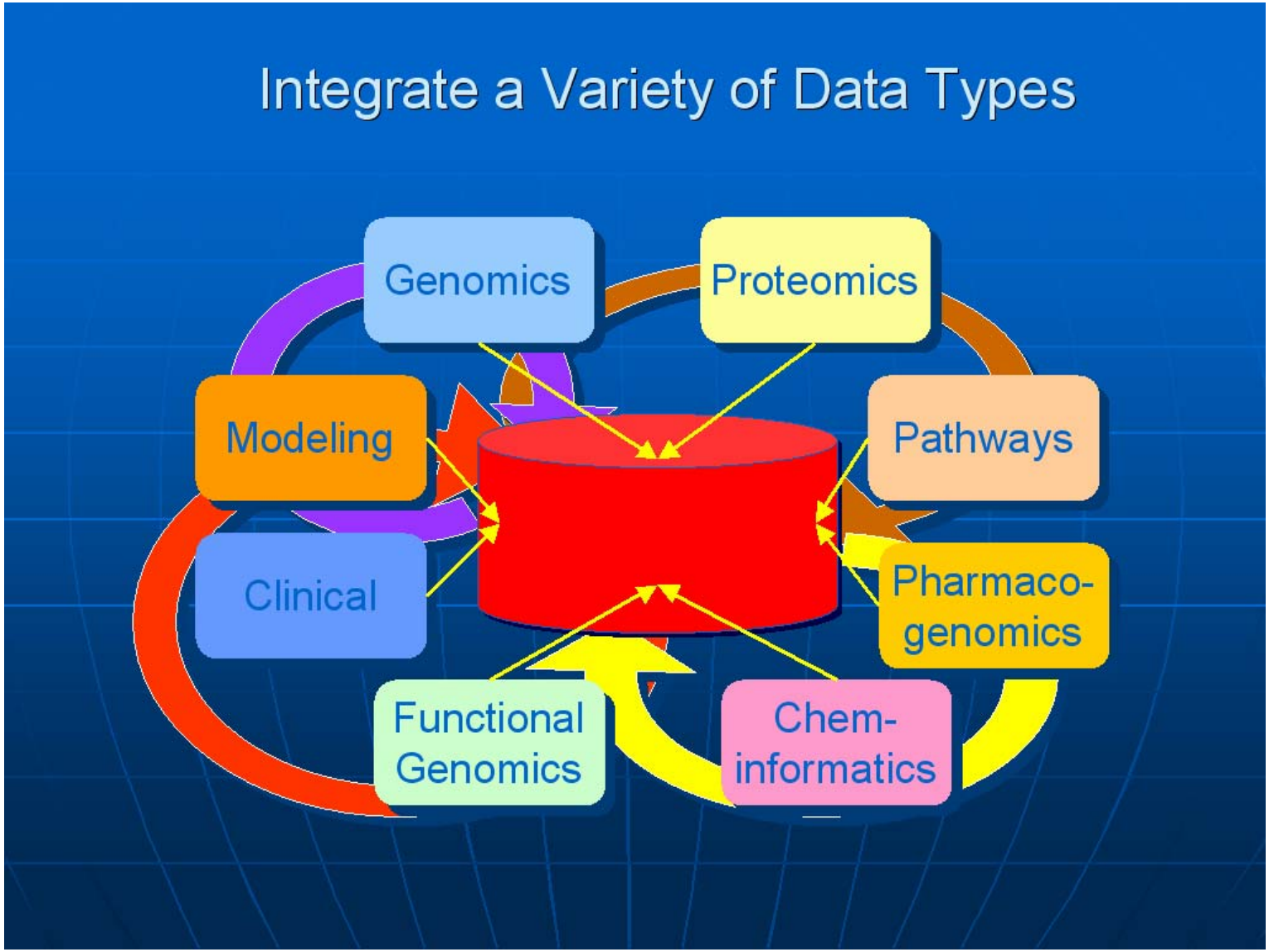


Microarray (Trascriptomics)



Gene & SNPs (Genomics)

System Biology for Health



Population Disease

Disease resistant population

Disease susceptible population

Genotype all individuals for thousands of SNPs

ATGATTATAG

geneX

ATGTTTATAG

Resistant people all have an ‘A’ at position 4 in *geneX*, while susceptible people have a ‘T’

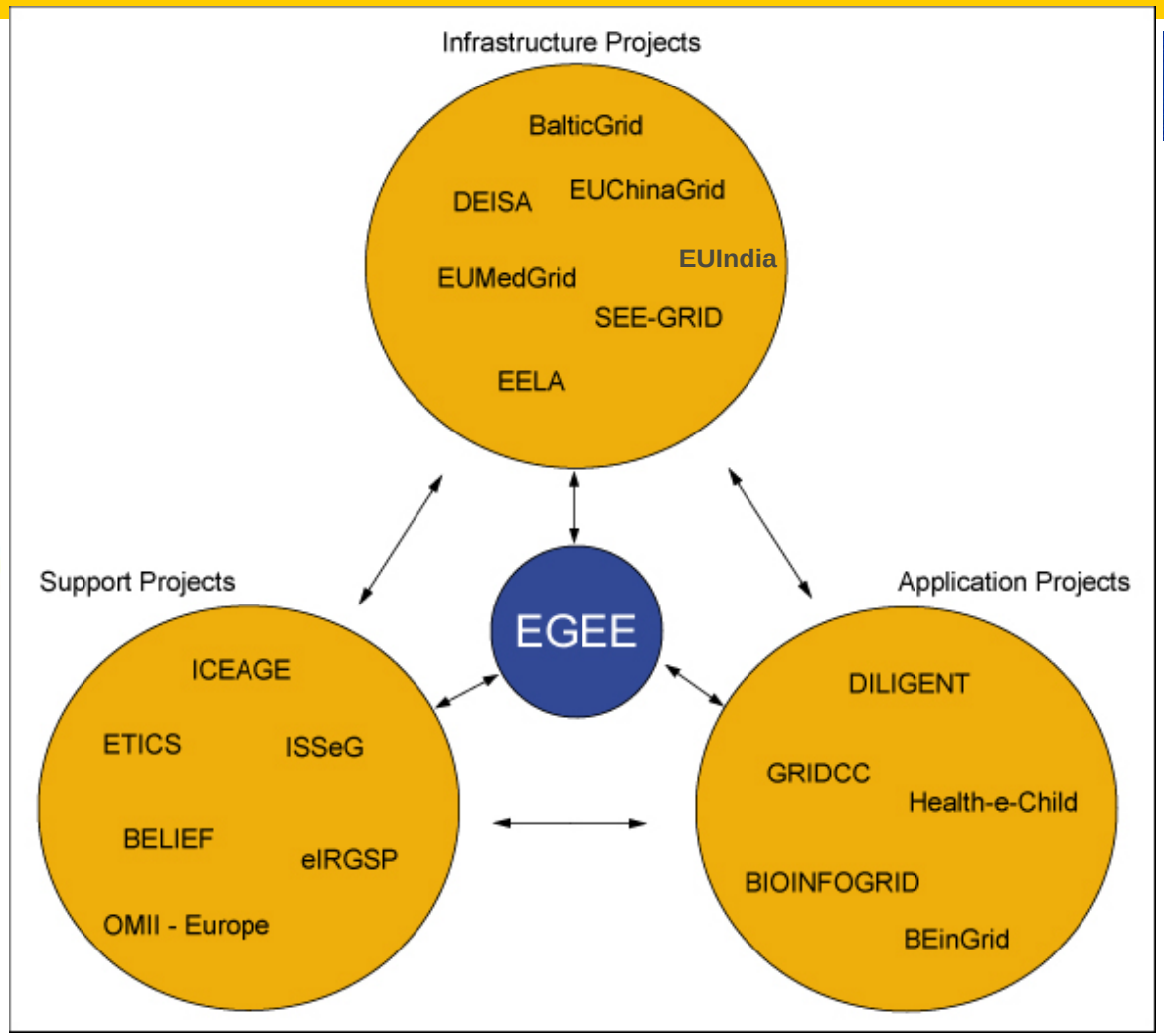
GRID and *omics

- A key development in the computational world has been the arrival of ***de novo* design algorithms** that use all available spatial information to be found within the target to **design novel drugs**.
- Coupling these algorithms to the rapidly growing body of information from structural genomics together with the new **ICT technology (eg. HPC, GRID, Web Services, ecc**.)
- provides a powerful new possibility for exploring design to a broad spectrum of genomics targets, including more challenging techniques such as:
- **protein–protein interactions, docking, molecular dynamics, system biology**, **gene network** ecc.

The grid application aspects.

- The massive potential of Grid technology will be indispensable when dealing with both the complexity of models and the enormous quantity of data, for example, in searching the human genome or when carry out simulations of molecular dynamics for the study of new drugs.
- The BIOINFOGRID projects proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure created by EGEE

Related EU projects

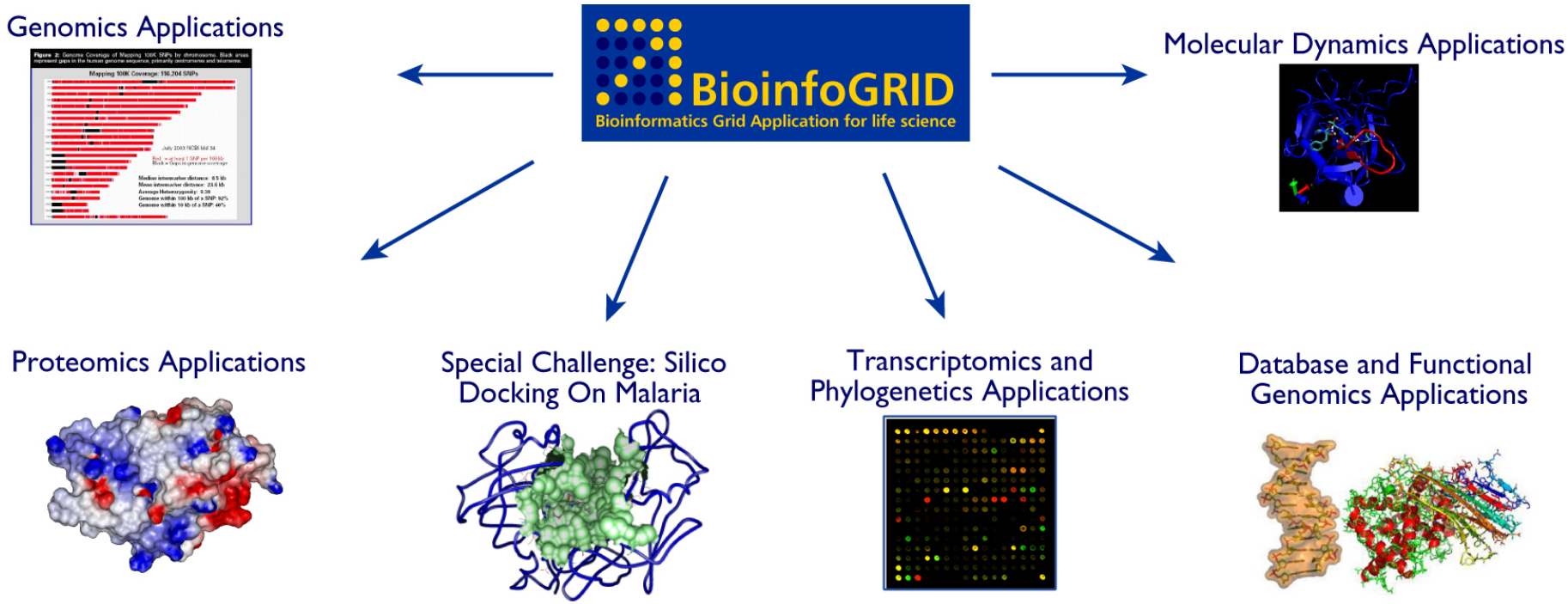


BioinfoGRID Project

- The **BIOINFOGRID project** proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure by EGEE and EGEEII projects.
- In the BIOINFOGRID initiative we plan **to evaluate genomics, transcriptomics, proteomics and molecular dynamics applications studies based on GRID technology**.
- The project start date: 1st January 2006
- The project finish date: 31 December 2007

BioinfoGRID Objective

- **Objective of the BioinfoGRID project**



BioinfoGRID Work Packages

Work-package No	Work Package title
WP1	Genomics Applications in GRID
WP2	Proteomics Applications in GRID
WP3	Transcriptomics Applications in GRID
WP4	Database and Functional Genomics Applications
WP5	Molecular Dynamics Applications
WP6	Coordination of technical aspects and relation with Grid infrastructure Projects, user training, application support and resources integration.
WP7	Dissemination and Outreach.
WP8	Project Management Office

Genomics applications

Aim : use of computational GRID to analyse molecular biological data at the genomic scale

Description

- **the GRID Portal system**: unification of larger groups of bioinformatics tools into single analytical steps and their optimization for GRID
- **GRID analysis of cDNA data:** computer- aided functional annotation of cDNAs in order to optimize sensitivity and specificity

Genomics applications in GRID

- **GRID analysis of genomic databases**: integration of precomputed data, gene identification, differentiation of pseudogenes, comparative genome analysis, etc.

- **Multiple alignments:** testing of new algorithms for computationally very demanding alignment procedures, optimization for GRID.

```
PRTC          TWFLVGLVSWG-EGCGLLHNYGVYTKVSRYLDWIHGHIRDKEAPQKSWAP----------
FA10          TYFVTGIVSWG-EGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAKSHAPEVITSSPLK
FA7           TWYLTGIVSWG-QGCATVGHFGVYTRVSQYIEWLQKLMRSE-------PRPGVLLRAPFP
THROMBIN      RWYQMGIVSWG-EGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE-----------------
FA9           TSFLTGIISWG-EECAMKGKYGIYTKVSRYVNWIKEKTKLT-------------------
KALLIKREIN    MWRLVGITSWG-EGCARREQPGVYTKVAEYMDWILEKTQSSDGKAQMQSPA---------
FA11          VWHLVGITSWG-EGCAQRERPGVYTNVVEYVDWILEKTQAV-------------------
TRYB1         TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-------------------
TRYB2         TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-------------------
TRYA          TWLQAGVVSWD-EGCAQPNRPGIYTRVTYYLDWIHHYVPKKP-------------------
KLKE          --QLQGLVSWGMERCALPGYPGVYTNLCKYRSWIEETMRDK-------------------
CTRL          TWVLIGIVSWG-TKNCNVRAPAVYTRVSKFSTWINQVIAYN-------------------
```

Proteomics Applications in GRID

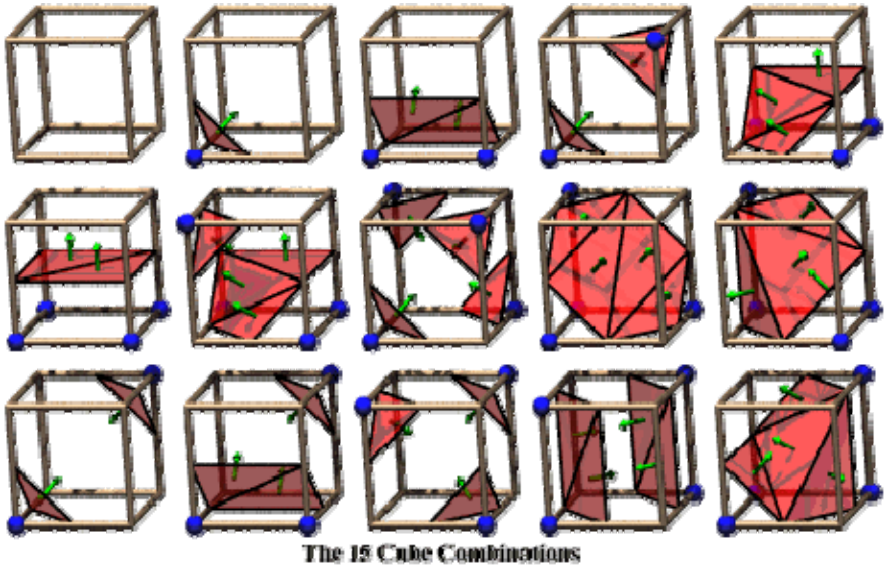
Aim : use of computational GRIDs to analysis molecular biological data in proteomics

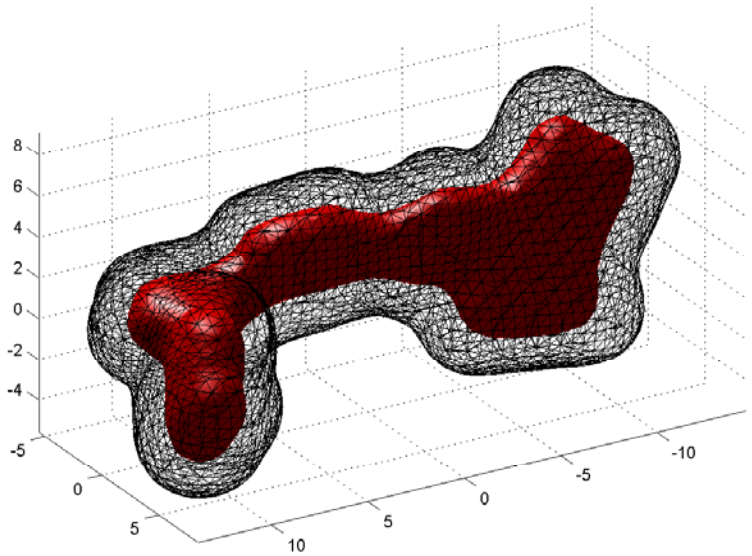
Description

- **Perform functional protein analysis in GRID** by using the functional protein domain annotations on large protein families using GRID and related databases.

Proteomics Applications

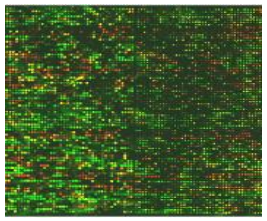
- **Protein surface calculation in GRID. :** the grid will be used to elaborate the volumetric description of the protein obtaining a precise representation of the corresponding surface.





Transcriptomics applications

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



Description

- **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.

Transcriptomics applications

Organism environment: -age, sex, energy sources (food composition), time of sampling, conditions of sampling, treatment

Location Environment: organ, substructure(lobes),tissue

Cellular Environment-histology/pathology, chemistry, physical

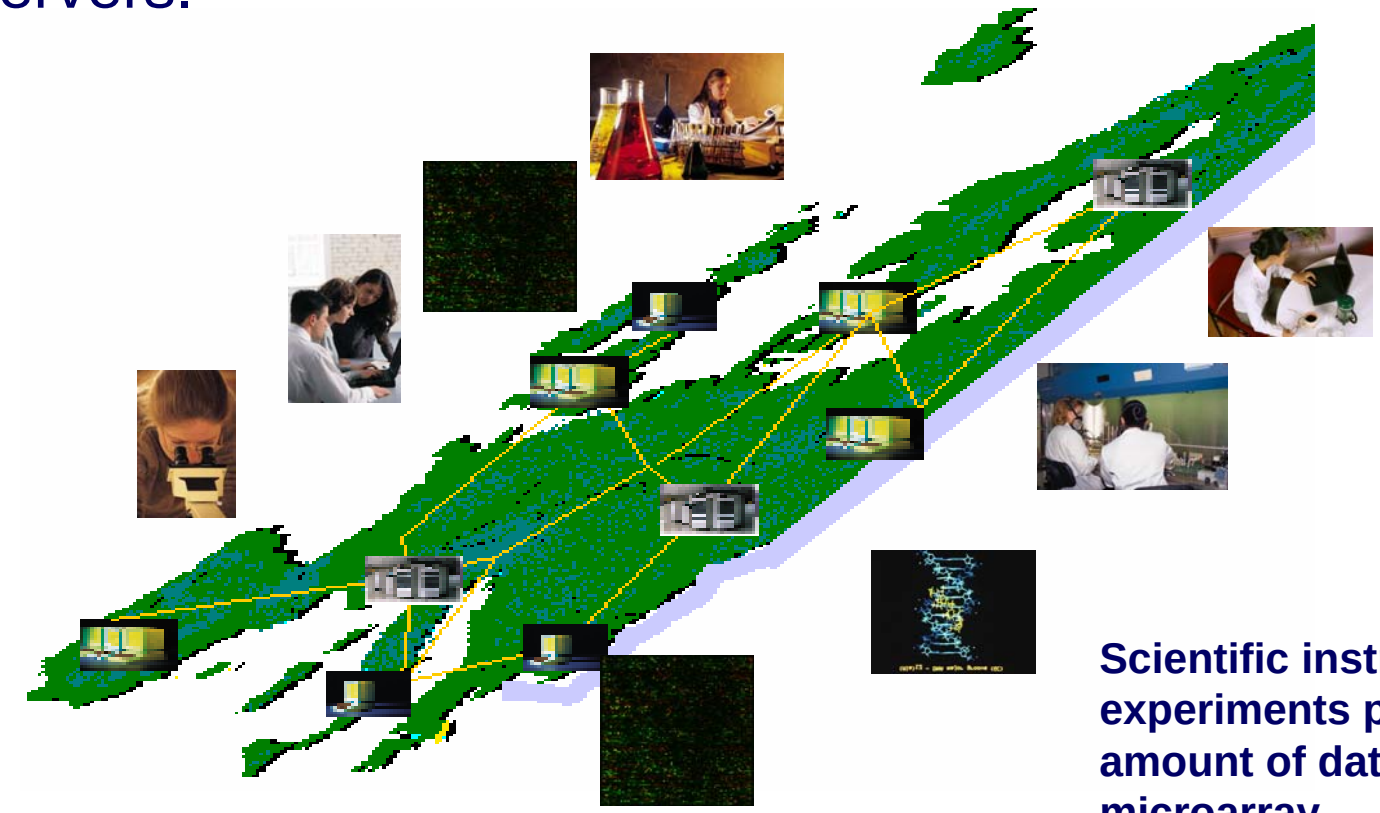
Subcellular environment-organelles,ultrastructural pathology

Molecular environment

Gene expression

Transcriptomics applications

Data analysis specific for bioinformatics allow the GRID user to store and search genetics data, with direct access to the data files stored on Data Storage element on GRID servers.

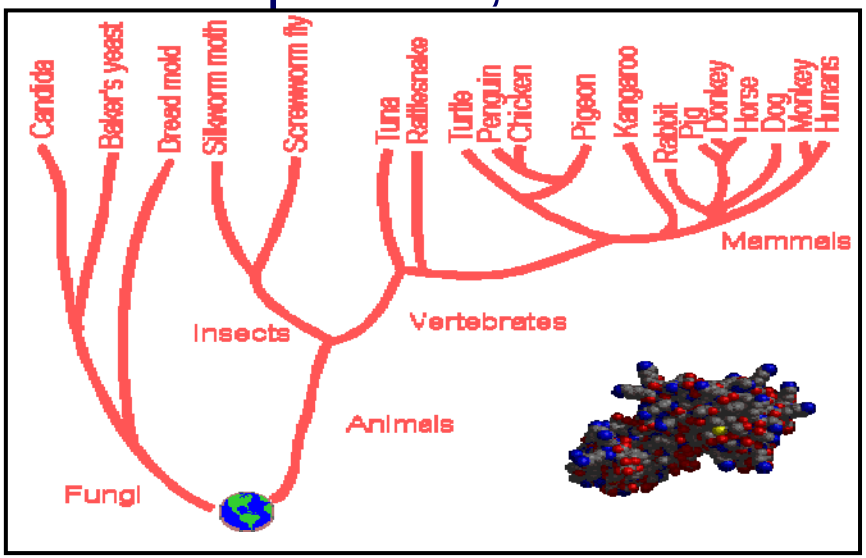


Researchers perform their activities regardless geographical location, interact with colleagues, share and access data

Scientific instruments and experiments provide huge amount of data from microarray

Phylogenetic application

- **Phylogenetics :** Reconstructing the evolutionary history of a group of taxa is major research thrust in computational biology and a standard part of exploratory sequence analysis. An evolutionary history not only gives relationships among taxa, but also an important tool for inferring the universal **tree of life**, inferring structural, physiological, and biochemical properties of sequences from other similar sequences, and reconstruction of tissue evolution.

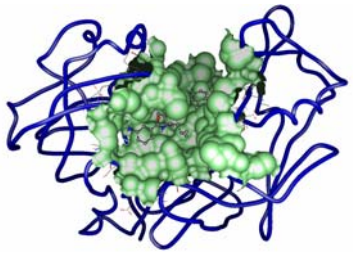


Molecular applications

Aim : The objective is to docking and Molecular Dynamics simulations, which usually take a very long time to complete the analysis.

Description

- **Wide *In Silico* Docking On Malaria initiative WISDOM-II:** This project perform the docking and molecular dynamics simulation on the GRID platform for *discovery new targets for neglected diseases* . Analysis can be performed notably using the data generated by the WISDOM application on the EGEE infrastructure.



Database Applications

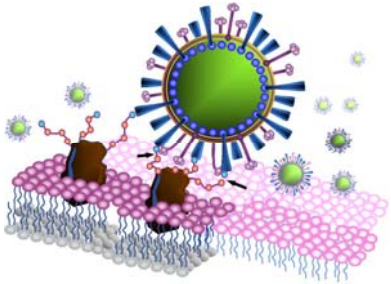
Aim : To mange the biological database, by using the GRID EGEE infrastructure.

Description

- **Biological database on GRID:** these databases will be complemented by others that are publicly available in Internet, by using GRID and web services where appropriate.
- **Functional Analogous Finder:** By using the GO terms and the associations to gene products it is possible to compare the total associated GO terms and their ascending parents to validate the functional analogy between two gene products

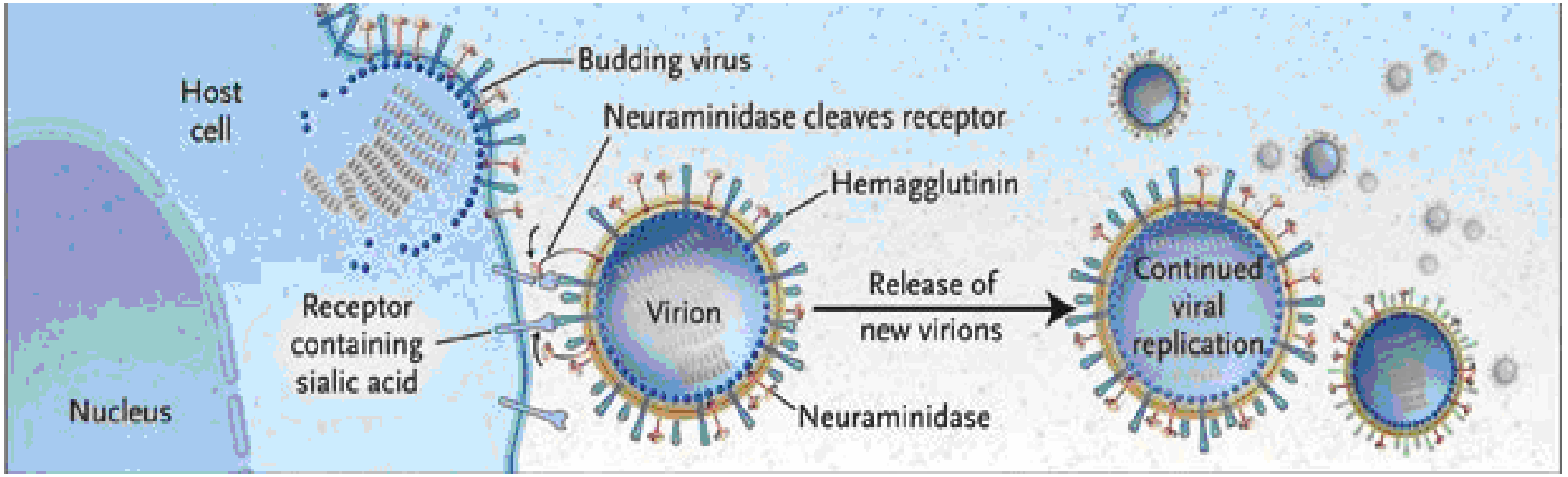
In Silico Docking On Avian Flu

A collaborative EGEE project led by Academia Sinica in Taiwan, CNRS-IN2P3 in France and the European SSA BioinfoGRID project, was set up to identify new drugs for the potential variants of the Influenza A virus analysing 300,000 possible drug components against the avian flu virus H5N1 using the EGEE Grid infrastructure.



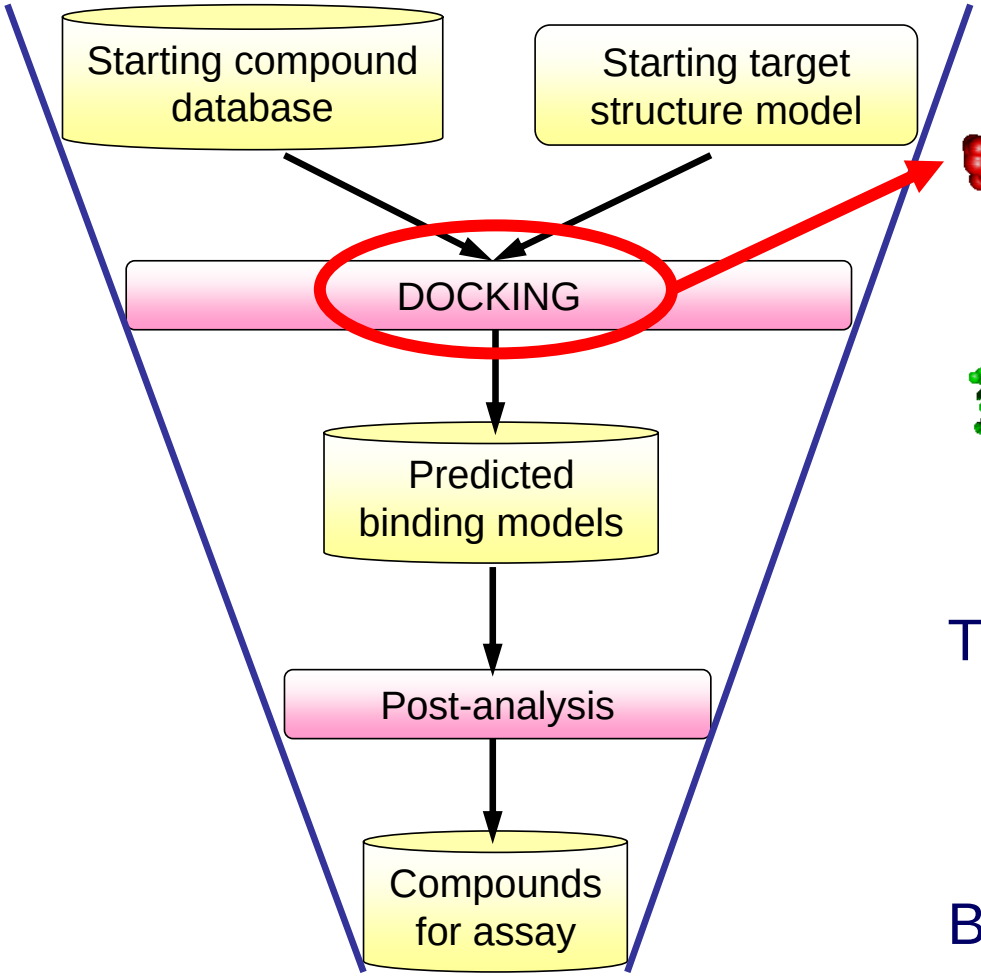
In silico virtual screening requires intensive computing, of the order of a few TFlops during one day to compute 1 million docking probabilities or for the molecular modelling of 1000 ligands on one target protein.

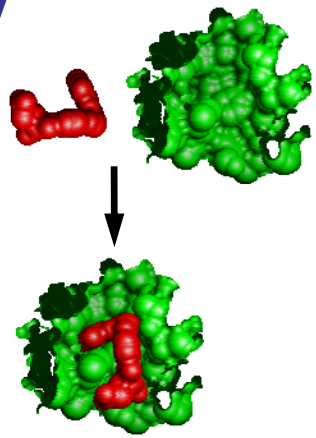
Neuraminidase Target



The neuraminidase viruses is considered a valid target for antiviral drugs

Virtual Screening





Docking: predict how small molecules bind to a receptor of known 3D structure

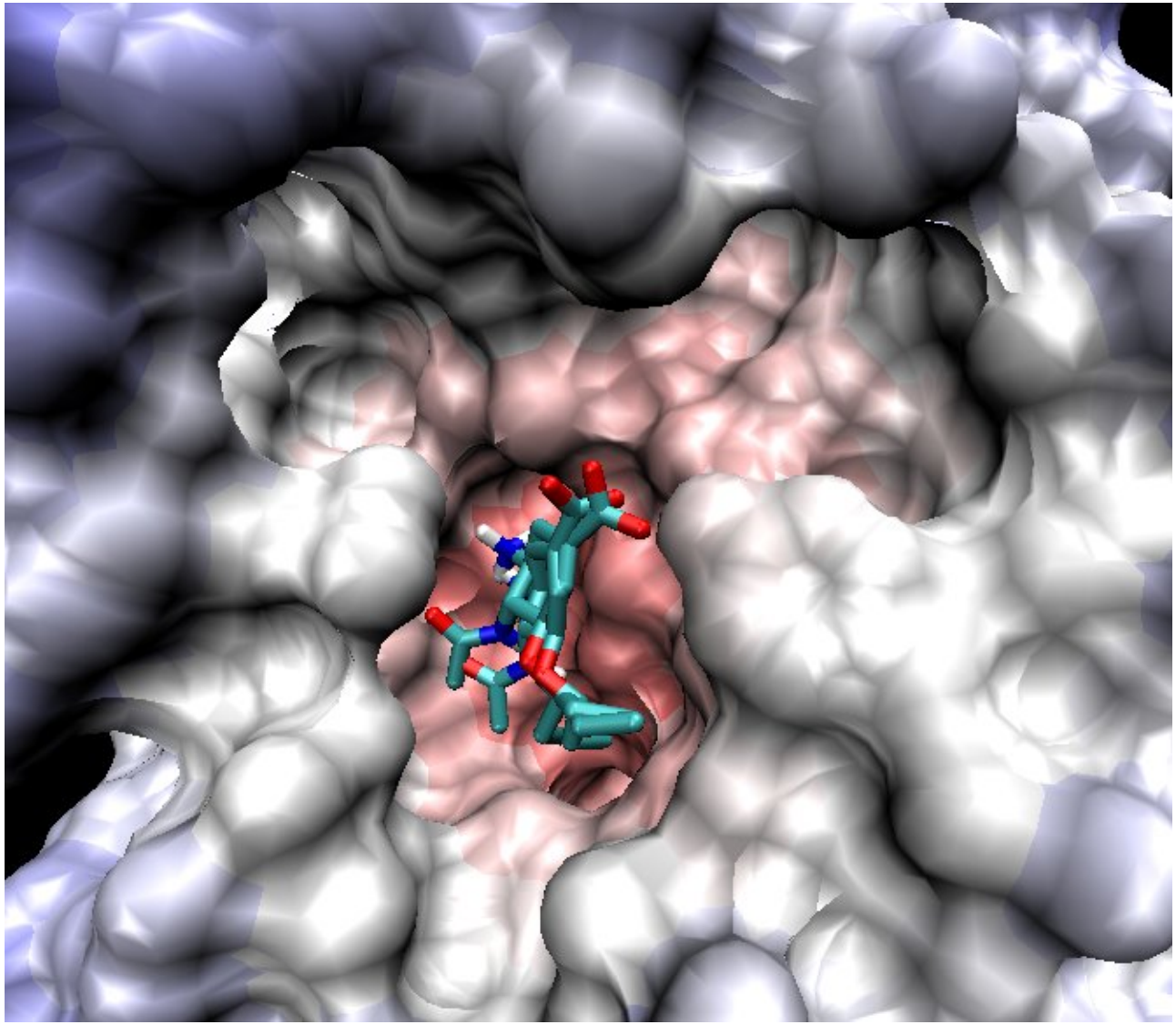
There are successful examples
- rapid,
- cost effective…

But there are limitations
- CPU and storage needed

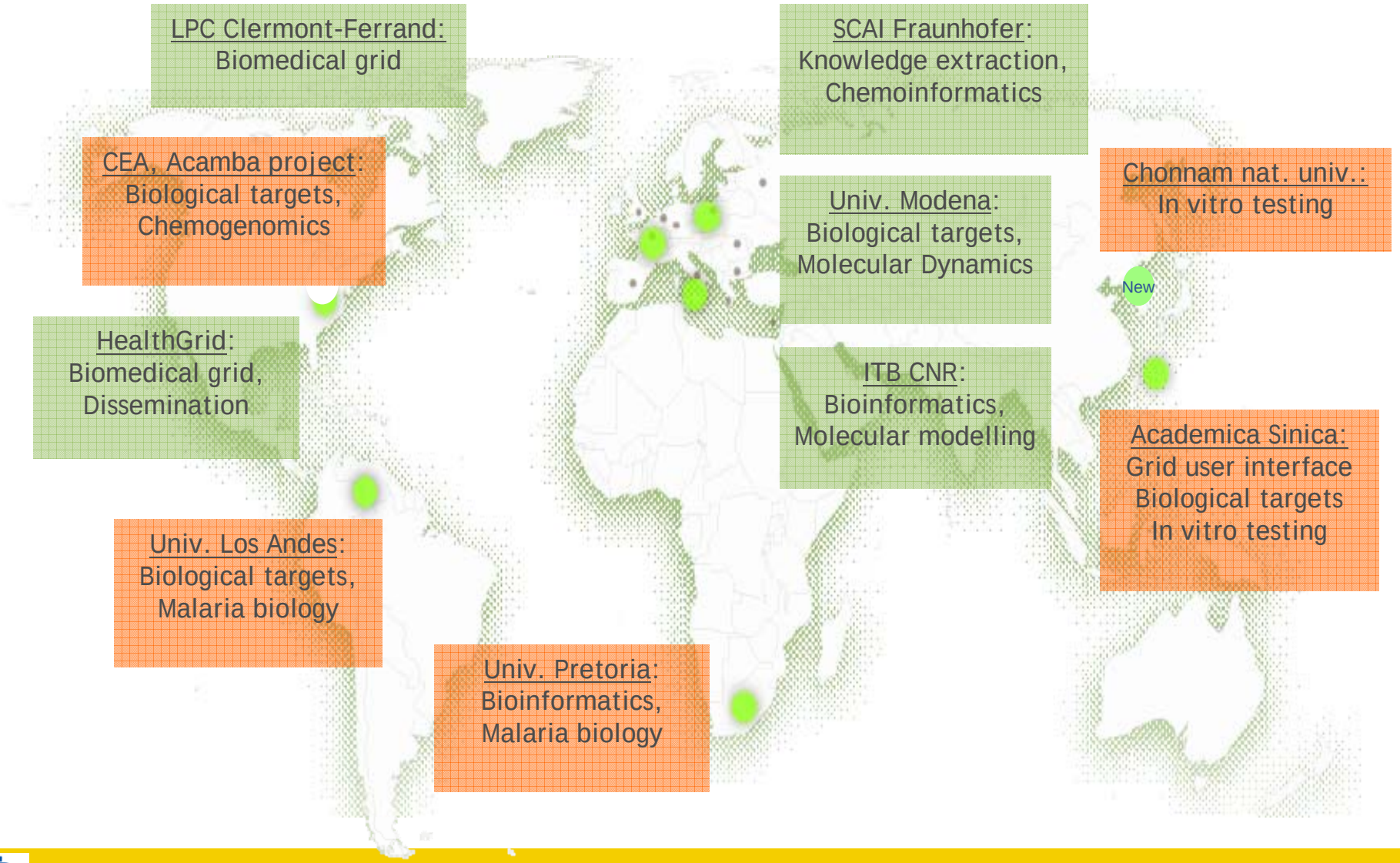
Influenza A Neuraminidase

- Grid-enabled High-throughput *in-silico* Screening against Influenza A Neuraminidase
- Encouraged by the success of the first EGEE biomedical data challenge against malaria (WISDOM), the second data challenge battling avian flu was kicked off in April 2006 to identify new drugs for the potential variants of the Influenza A virus. Mobilizing thousands of CPUs on the Grid, the 6-weeks high-throughput screening activity has fulfilled over 100 CPU years of computing power.
- In this project, the impact of a world-wide Grid infrastructure to efficiently deploy large scale virtual screening to speed up the drug design process has been demonstrated.

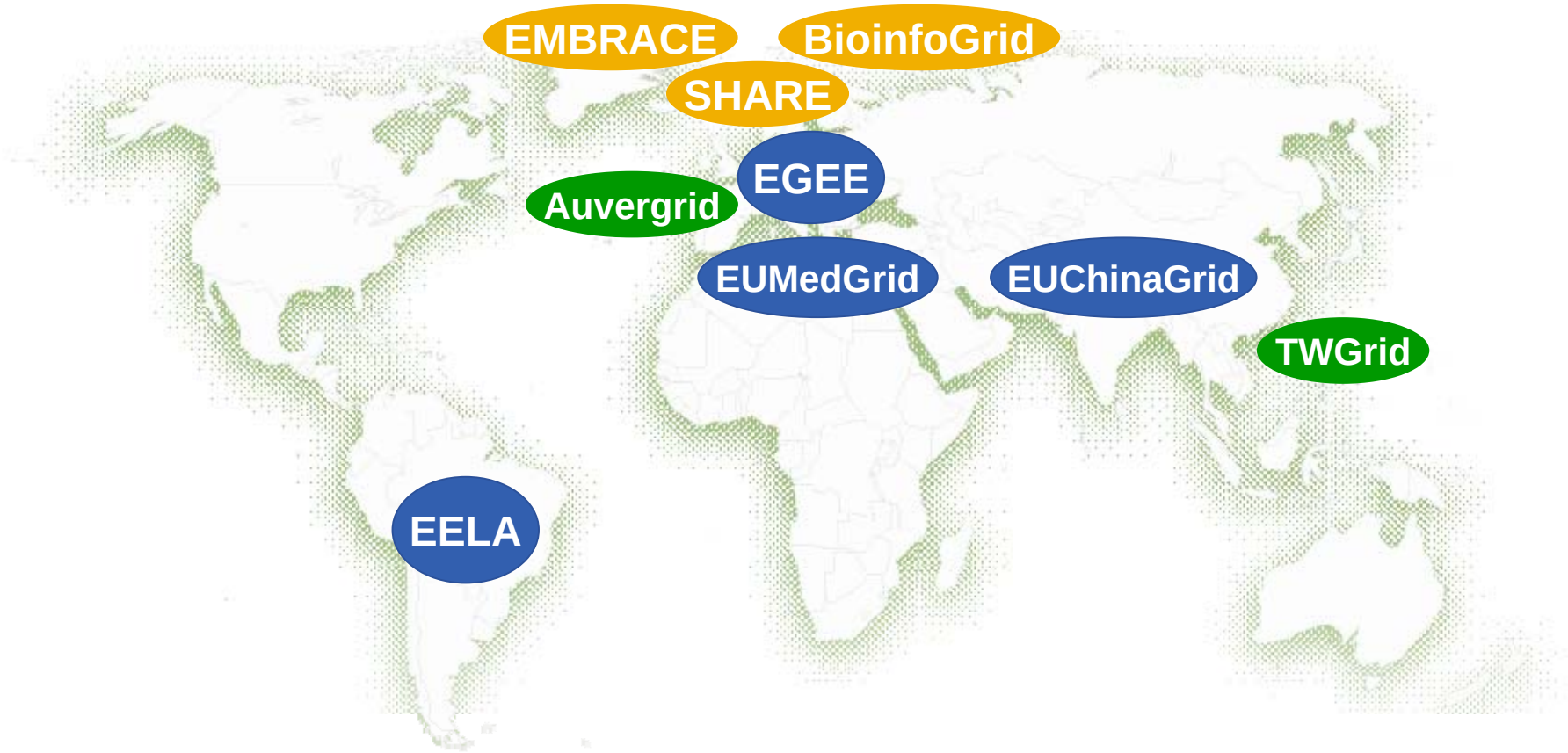
Docking



A grid for Malaria and Avian Influenza



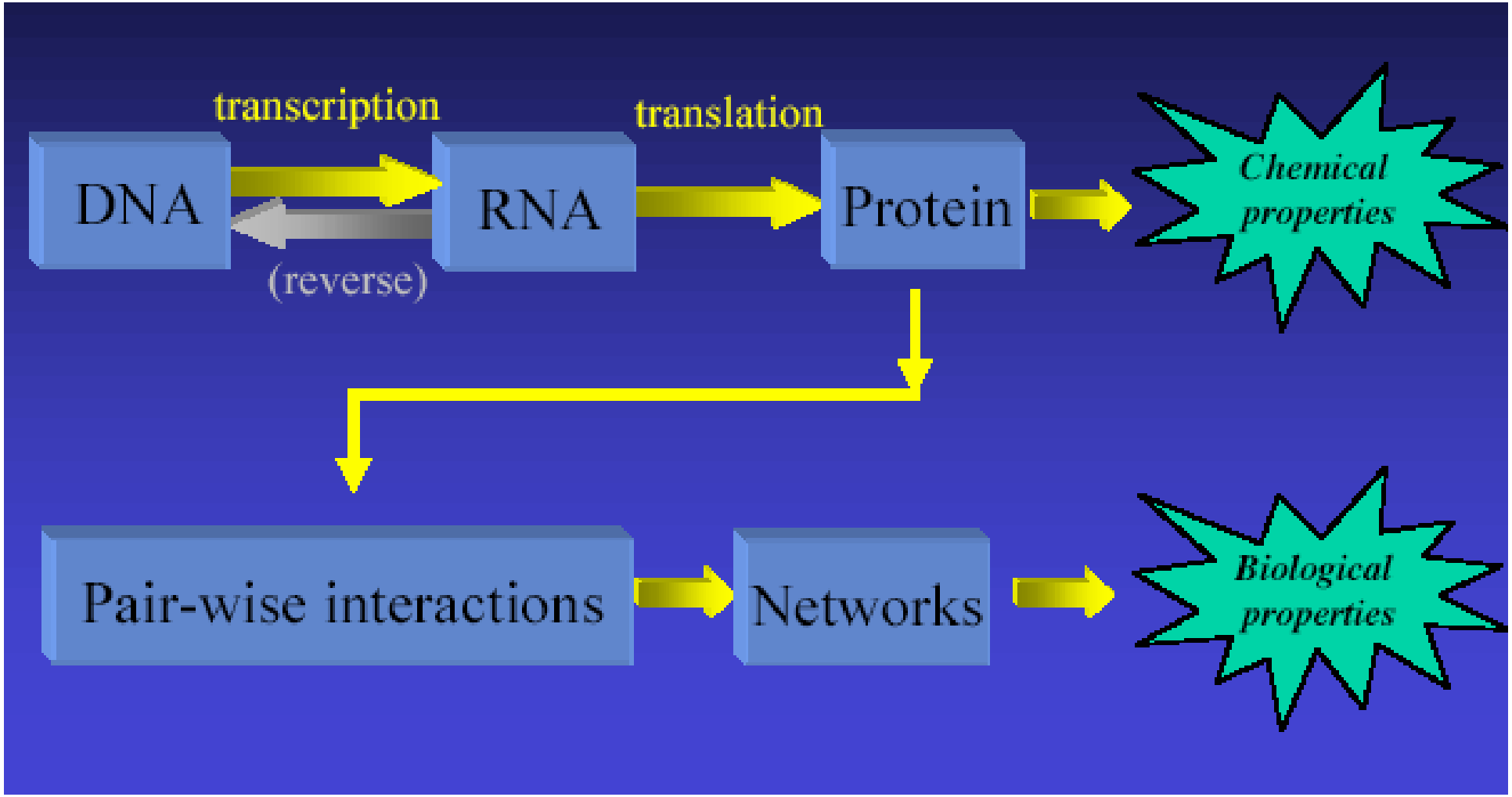
Networks of people



: European grid infrastructure : European grid project

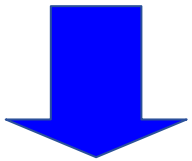
: Regional/national grid infrastructure

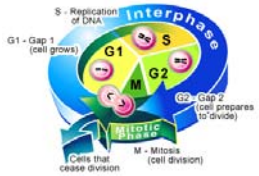
Systems Biology



Cell Cycle: Motivation

- Cell cycle is a complex biological process that implies the interaction of a large number of genes
- Disease studies on tumour proliferation are related with the de-regulation of cell cycle
- It will be useful finding as quickly as possible information related to all the genes involved in this cellular process





- We implement a new resource which collects useful information about the human cell cycle to support studies on genetic diseases related to this crucial biological process

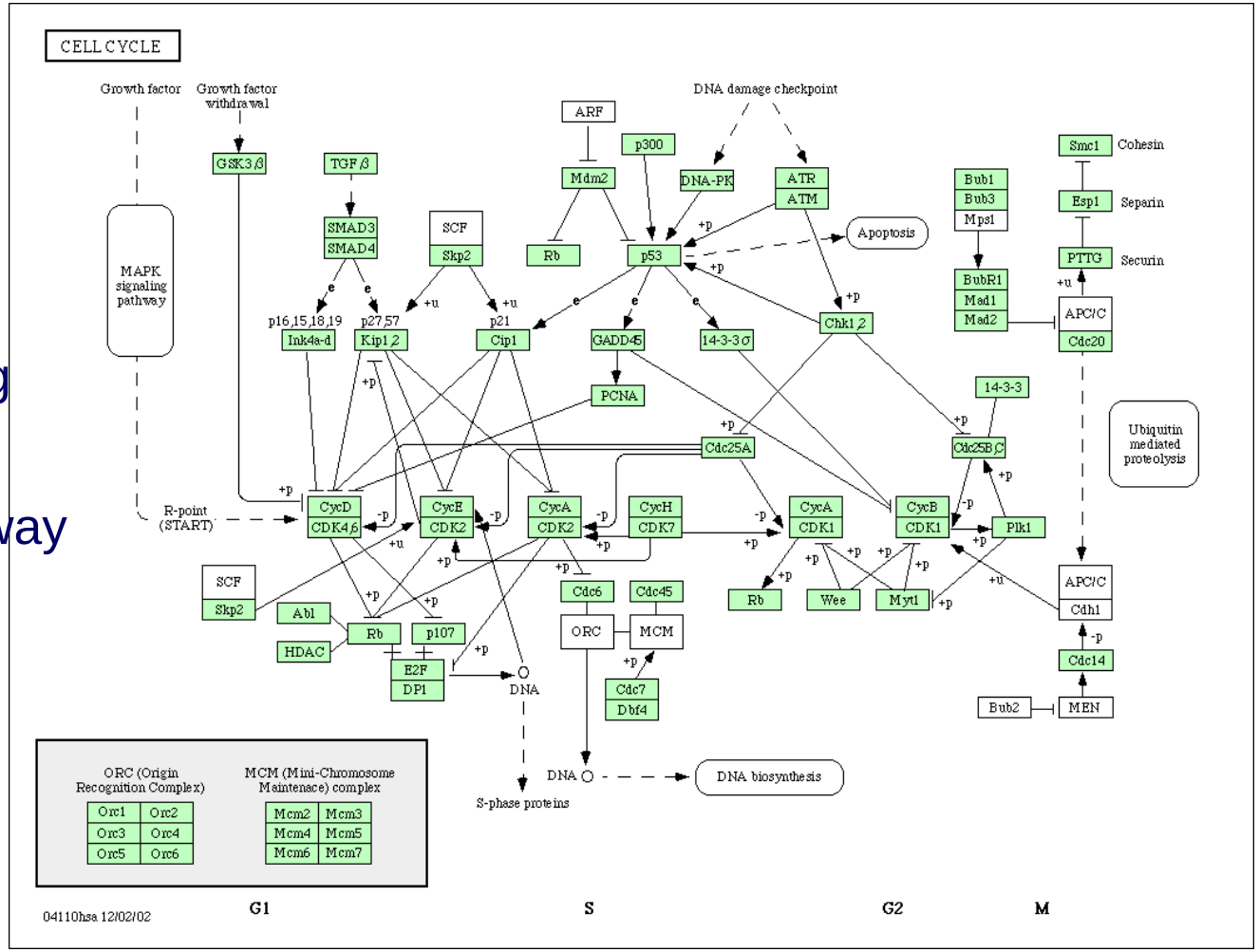
HCCDB:Human Cell Cycle DB

Development of a “Human Cell Cycle Database” HCCDB

It contains genes involved in

- Cell cycle pathway
  - MAPK signalling pathway
  - Apoptosis pathway from KEGG
  - Checkpoints
  - Mitosis

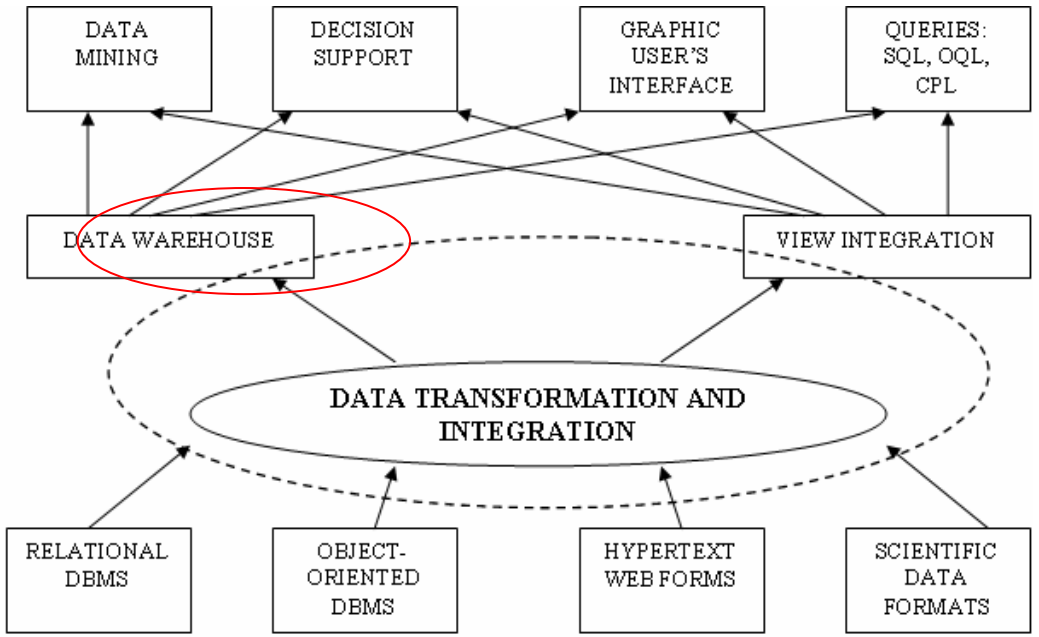
from REACTOME



Data Integration

Data integration system from many biological resources:

NCBI,
Ensemble,
Kegg,
Reactome,
dbSNP,
MGC,
DBTSS,
Unigene,
QPPD,
TRANSFAC
UniProt,
InterPro,
PDB,
TRANSPATH,
BIND,
MINT,
IntAct



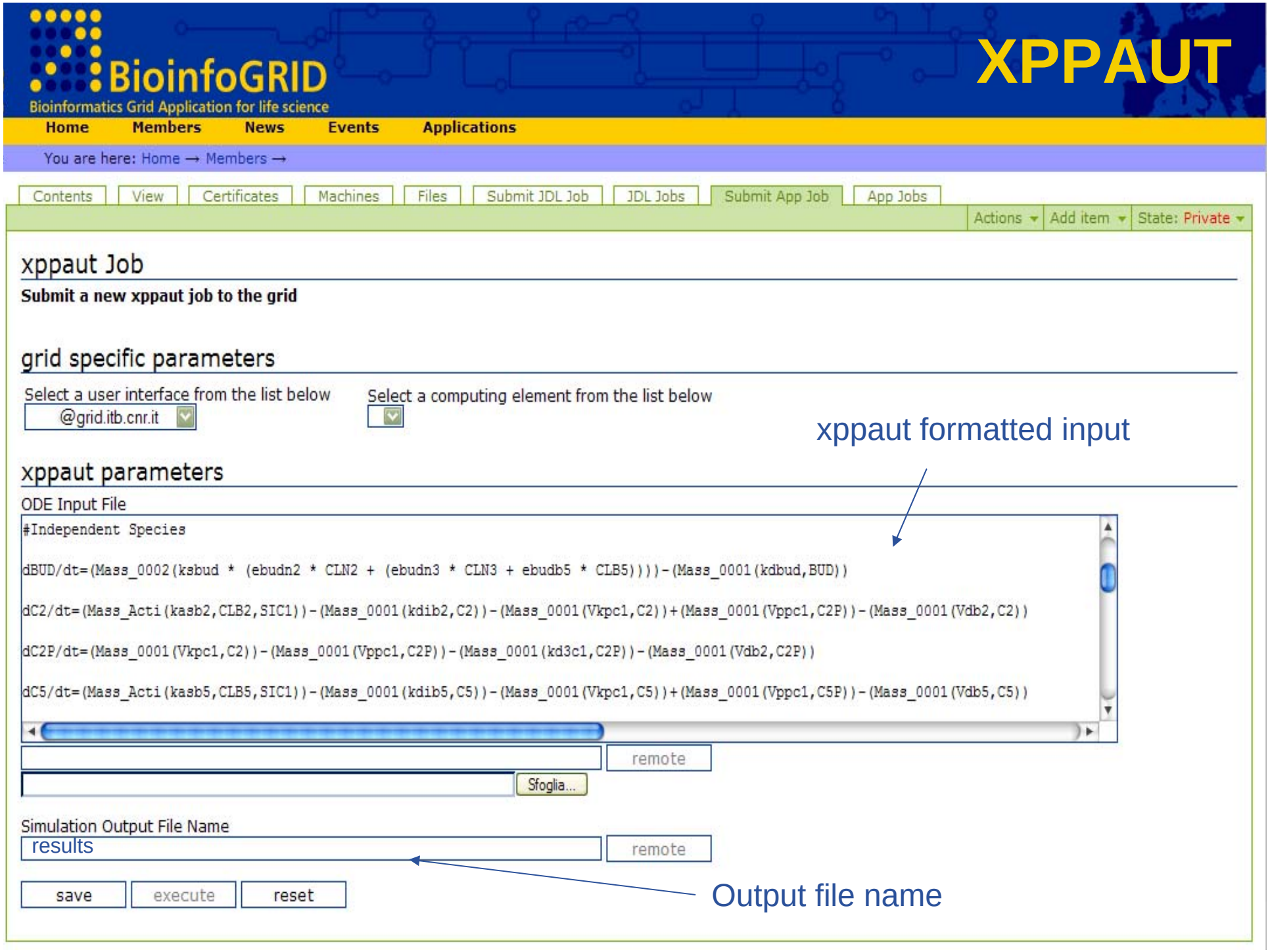
- Data Warehouse Approach

XPPAUT

Tool for simulate dynamical systems

can handle:

- Differential equations
- Delay equations
- Volterra integral equations
- Stochastic equations
- Discrete dynamical systems
- Markov processes
- Bifurcation


BioinfoGRID
Bioinformatics Grid Application for life science
XPPAUT
Home
Members
News
Events
Applications
You are here: Home → Members →
Contents
View
Certificates
Machines
Files
Submit JDL Job
JDL Jobs
Submit App Job
App Jobs
Actions
Add item
State: Private
xppaut Job
Submit a new xppaut job to the grid
grid specific parameters
Select a user interface from the list below
@grid.itb.cnr.it
Select a computing element from the list below
xppaut formatted input
xppaut parameters
ODE Input File
#Independent Species
dBUD/dt=(Mass_0002(ksbud * (ebudn2 * CLN2 + (ebudn3 * CLN3 + ebudb5 * CLB5))))-(Mass_0001(kdbud,BUD))
dC2/dt=(Mass_Acti(kasb2,CLB2,SIC1))-(Mass_0001(kdib2,C2))-(Mass_0001(Vkpc1,C2))+(Mass_0001(Vppc1,C2P))-(Mass_0001(Vdb2,C2))
dC2P/dt=(Mass_0001(Vkpc1,C2))-(Mass_0001(Vppc1,C2P))-(Mass_0001(kd3c1,C2P))-(Mass_0001(Vdb2,C2P))
dC5/dt=(Mass_Acti(kasb5,CLB5,SIC1))-(Mass_0001(kdib5,C5))-(Mass_0001(Vkpc1,C5))+(Mass_0001(Vppc1,C5P))-(Mass_0001(Vdb5,C5))
remote
Sfoglia...
Simulation Output File Name
results
remote
save
execute
reset
Output file name

Simulation Section

BioinfoGRID

Simulation results

Swat M, Kel A, Herzel H: Bifurcation analysis of the regulatory modules of the mammalian G1/S transition. - 2004

Download XPPAUT input file

Download results file*

Select species to show on 2D plot

x

time

y series

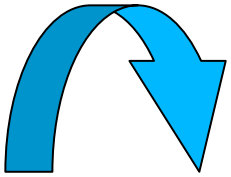
-

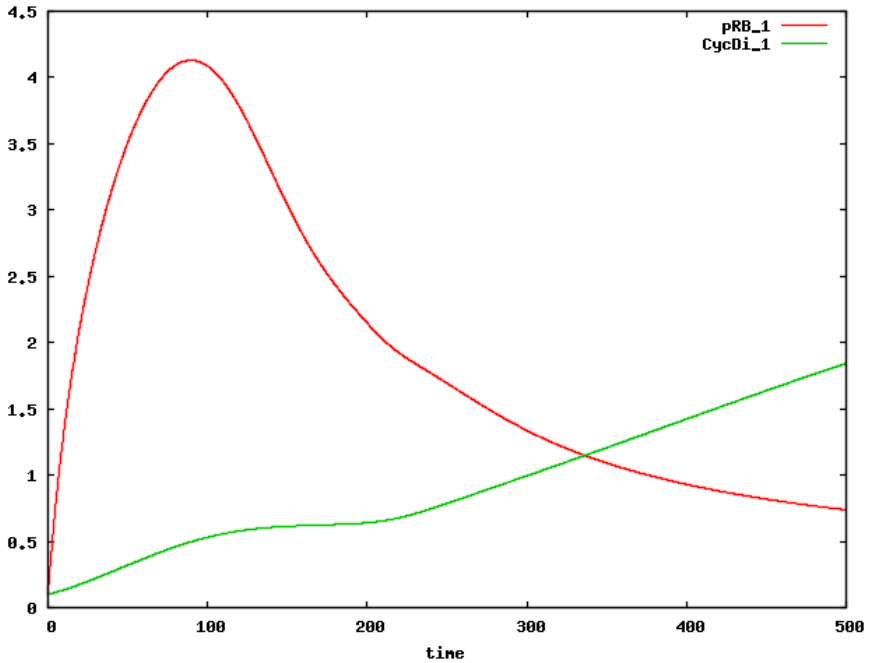
-

time
pRB_1
pRBp_1
E2F1_1
CycDi_1
CycDa_1
AP1_1
pRBpp_1
CycEi_1
CycEa_1

with GNUPLOT

cts order of variables in the input interface; first is time

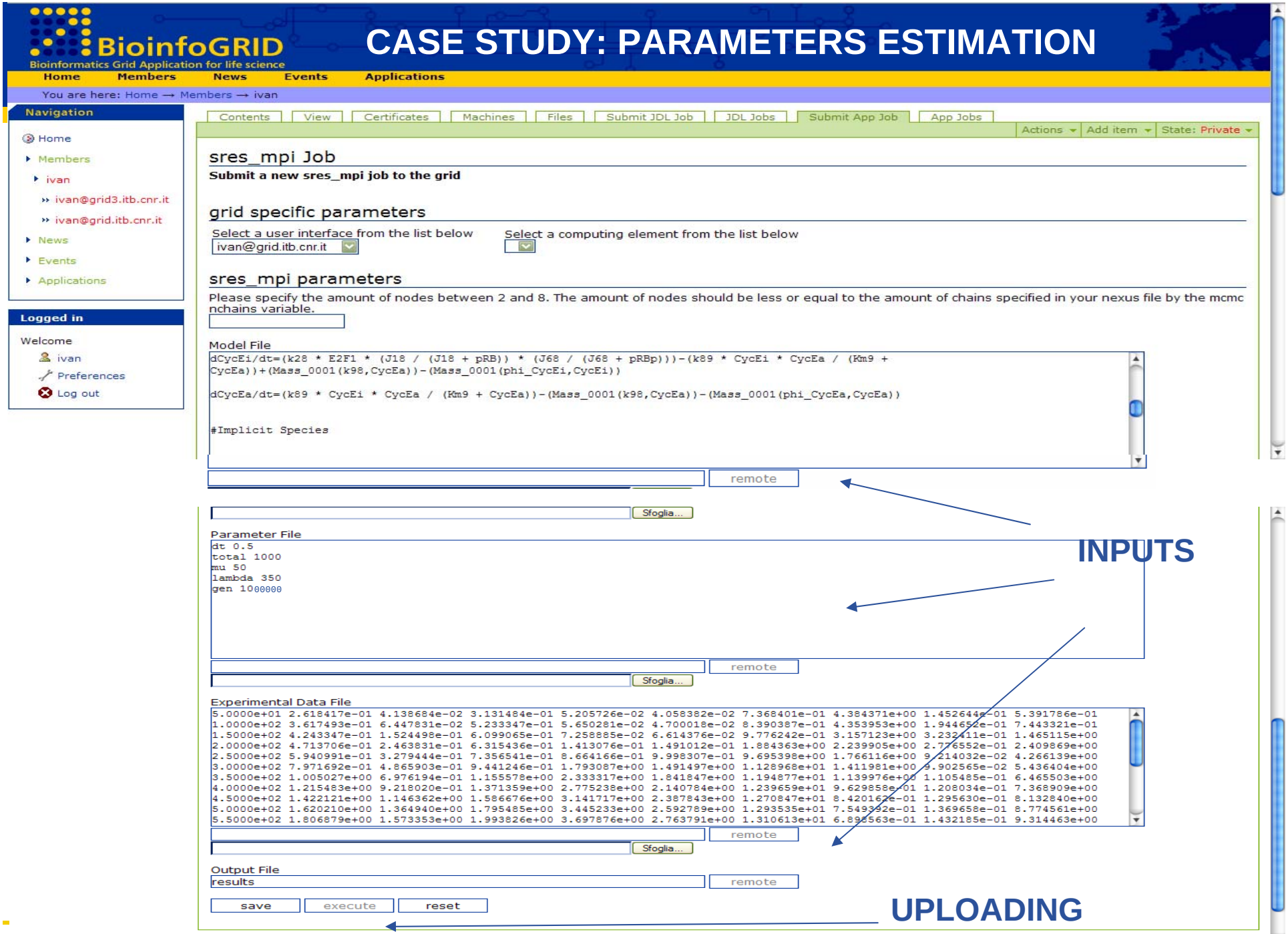




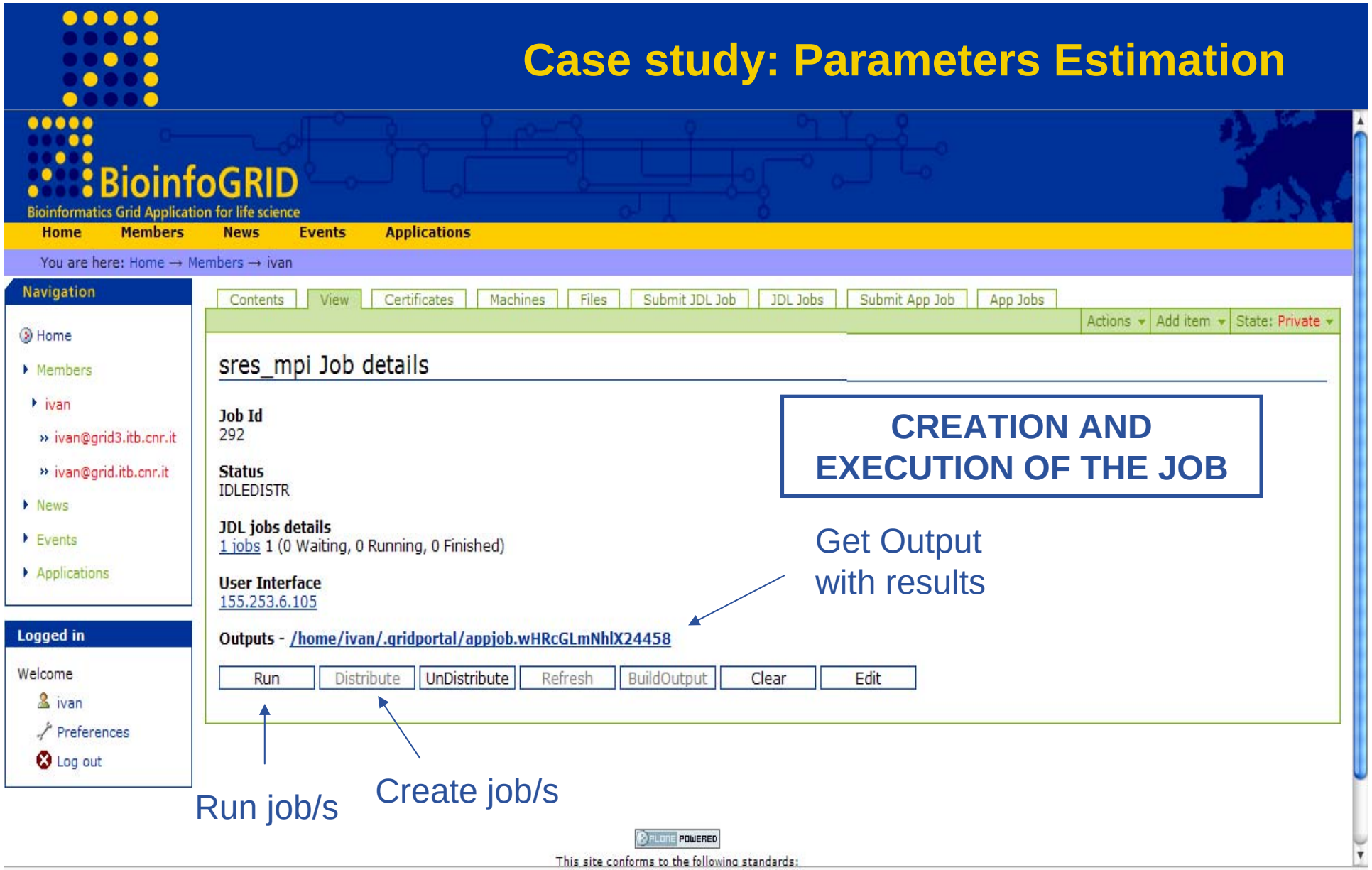
2D plot: image exported in png using GnuPlot

The simulation of a single ODE system describing a cell cycle model

CASE STUDY: PARAMETERS ESTIMATION



Case study: Parameters Estimation



Tissue Microarray in GRID

Genetic Diseases

↓

High throughput techniques (i.e. DNA microarray)

to screen the whole genome

↓

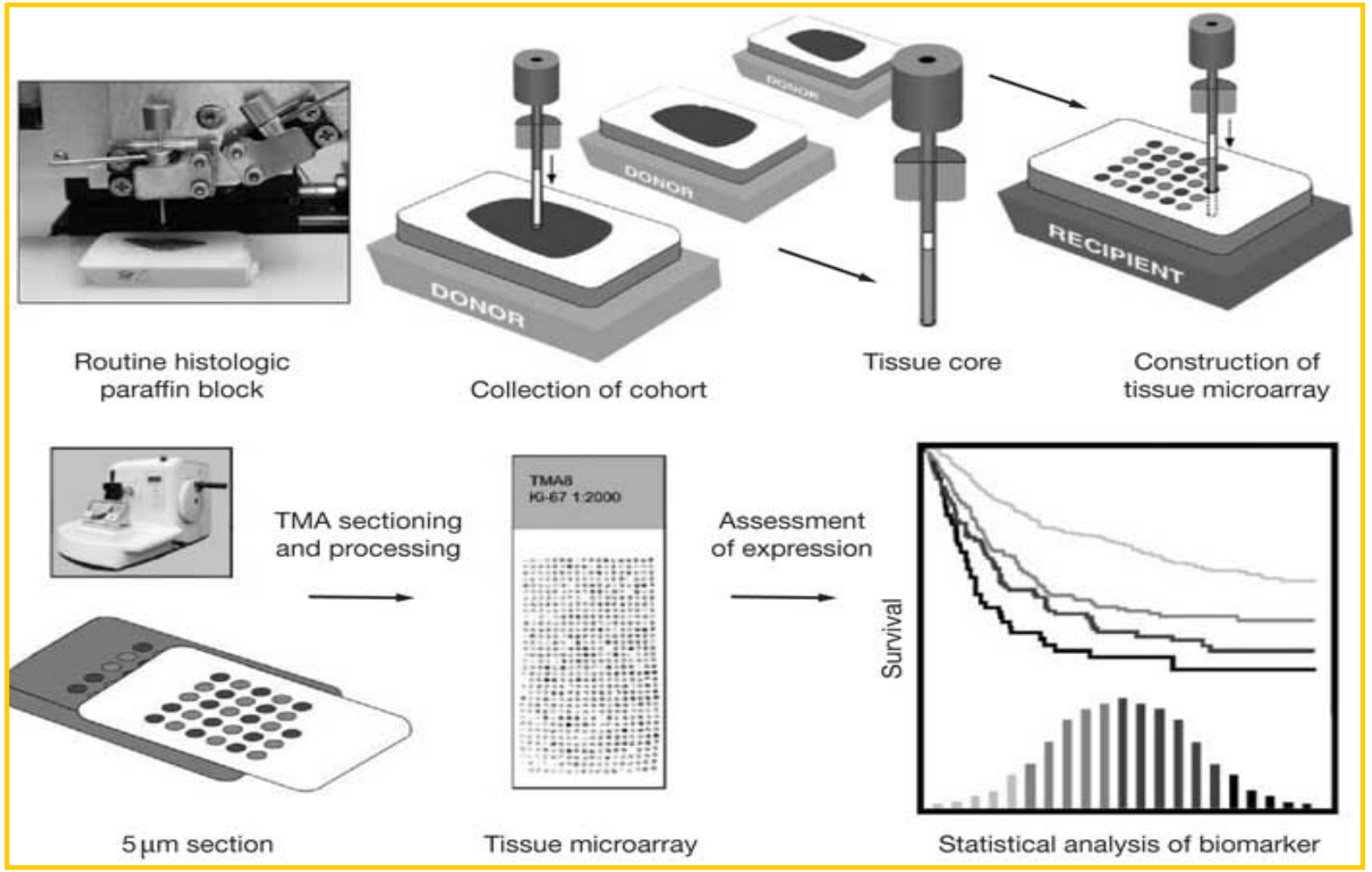
Low reliability

↓

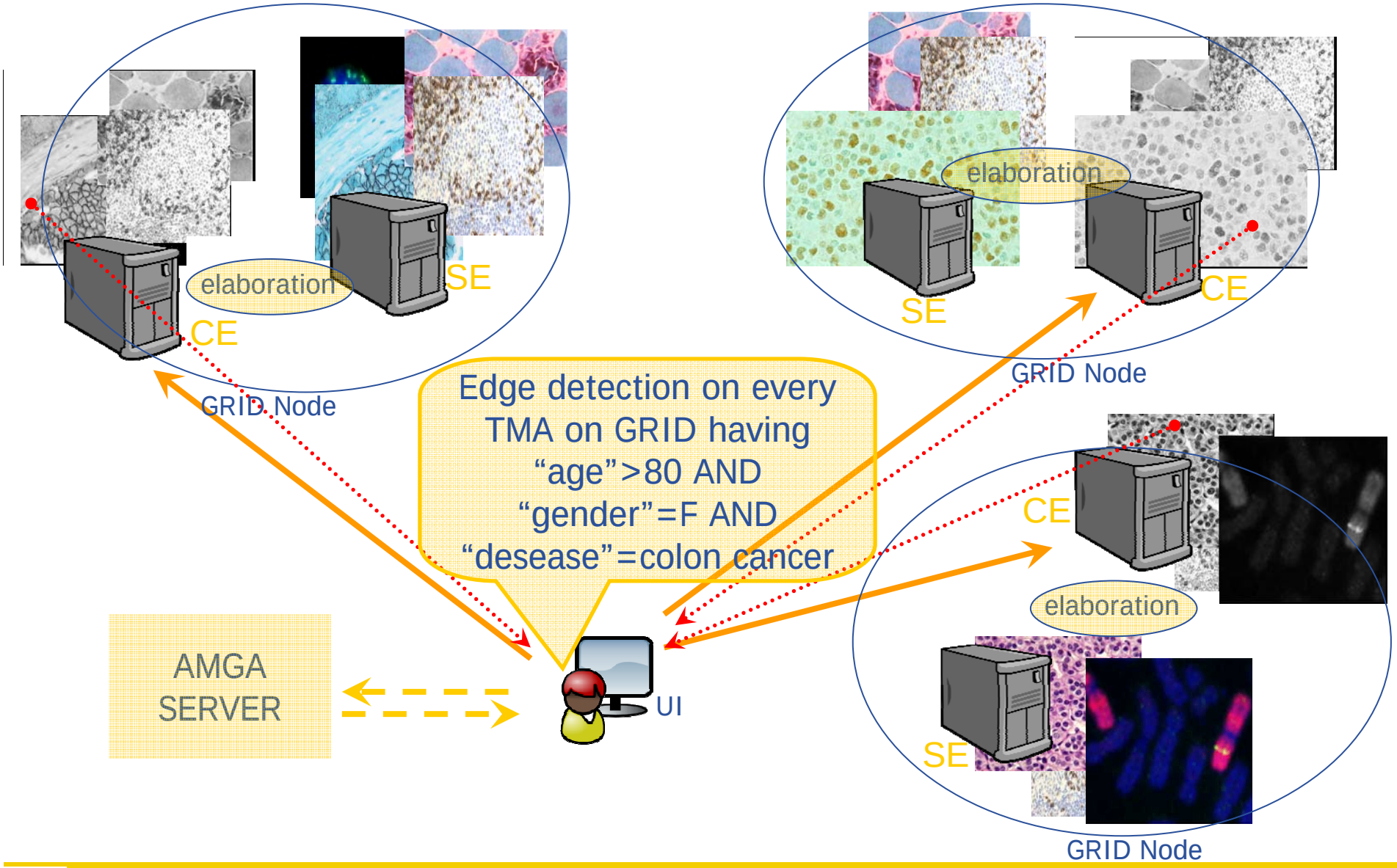
Validation through **Tissue Microarray**

Tissue Microarray technology

Genes and proteins detection



Tissue Microarray on GRID



Conclusion

- New technologies have been introduced to automate the analysis, and annotation of genomic, proteomic and Systems Biology data (eg. **Web services, Workflow, Data Mining, Agent, GRID, Ontology, Semantic Web**).

- A new generation of **algorithms and data mining** needs to be developed in order to be capable of connecting the biological information of genes, proteins and metabolic pathways with the patients' disease.

- The dedicated **HPC and GRID infrastructure** will be in a position to tackle the important role of developing new strategies for production and analysis of data in the fields of biotechnology and biomedicine.

- The **massive potential of HPC and Grid technology** will be indispensable when dealing with both the complexity of models and the enormous quantity of data.

Interaction with other projects

At present the BioinfoGRID project has established co-operations with the following projects:

- **EGEE**
- **BELIEF**
- **EMBRACE**
- **EUCHINAGRID**
- **EUMEDGRID**
- **EELA**
- **DILIGENT**
- **ICEAGE**
- **LITBIO**
- **LIBI**
- **HEALTHGRID**
- **SYMBIOMATICS**

Project dissemination

- Participation to conferences and events, specific to the various application fields, to present the results of the present project and disseminate the usefulness of the adoption of the Grid paradigm.

- **> 30 international conferences**

- **2 International BIOMED GRID School**

- **> 30 publications**
- **participation to the Malaria and Avian Flow challenges.**

Acknowledgments

- This work was supported by the:
- Italian FIRB-MIUR LITBIO: Laboratory for Interdisciplinary Technologies in Bioinformatics *http://www.litbio.org*,
- *BIOINFOGRID* *http://www.bioinfogrid.eu*
- EGEE Enabling Grid for E-science project
- Italian Bioinformatics Network ITLABIONET IRB-MIUR project.

Acknowledgements