



Biomed **GRID**school

Varenna, Italy, 14-19 May 2007



EGEE Bioinformatics Meeting #4
Common EU-EGEE
EU-BioinfoGRID meeting
Varenna, Italy, 19 May 2007

GRID approach for Bioinformatics and System Biology



Milanesi Luciano

National Research Council

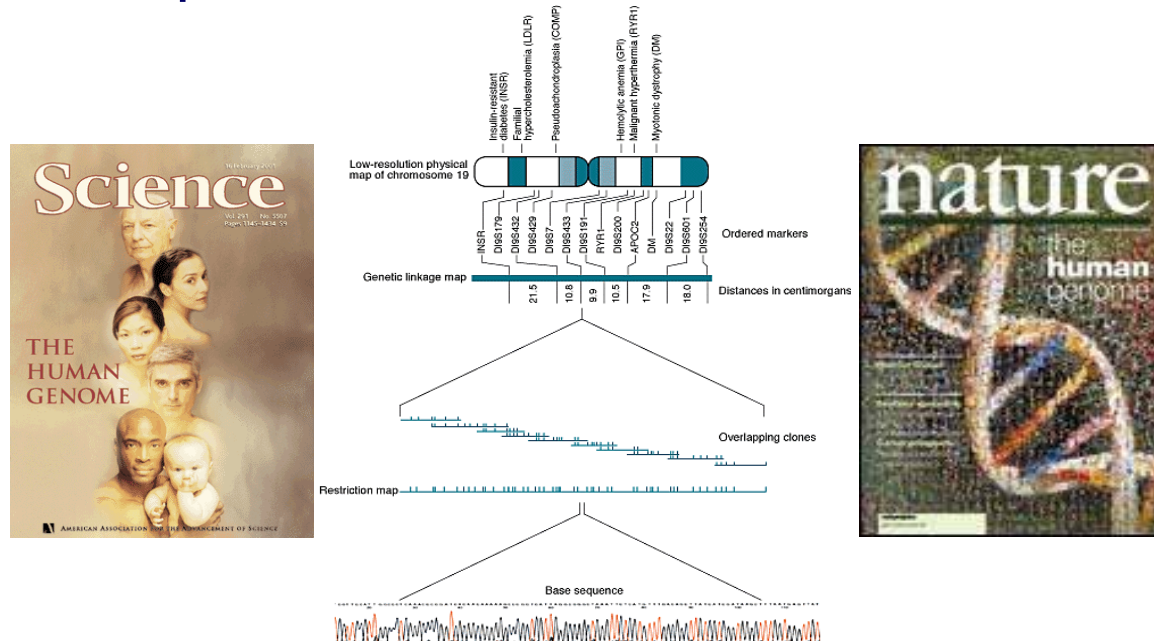
Institute of Biomedical Technologies, Milan, Italy

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 transcriptome, proteome, metabolome are the different topics developed in this class.



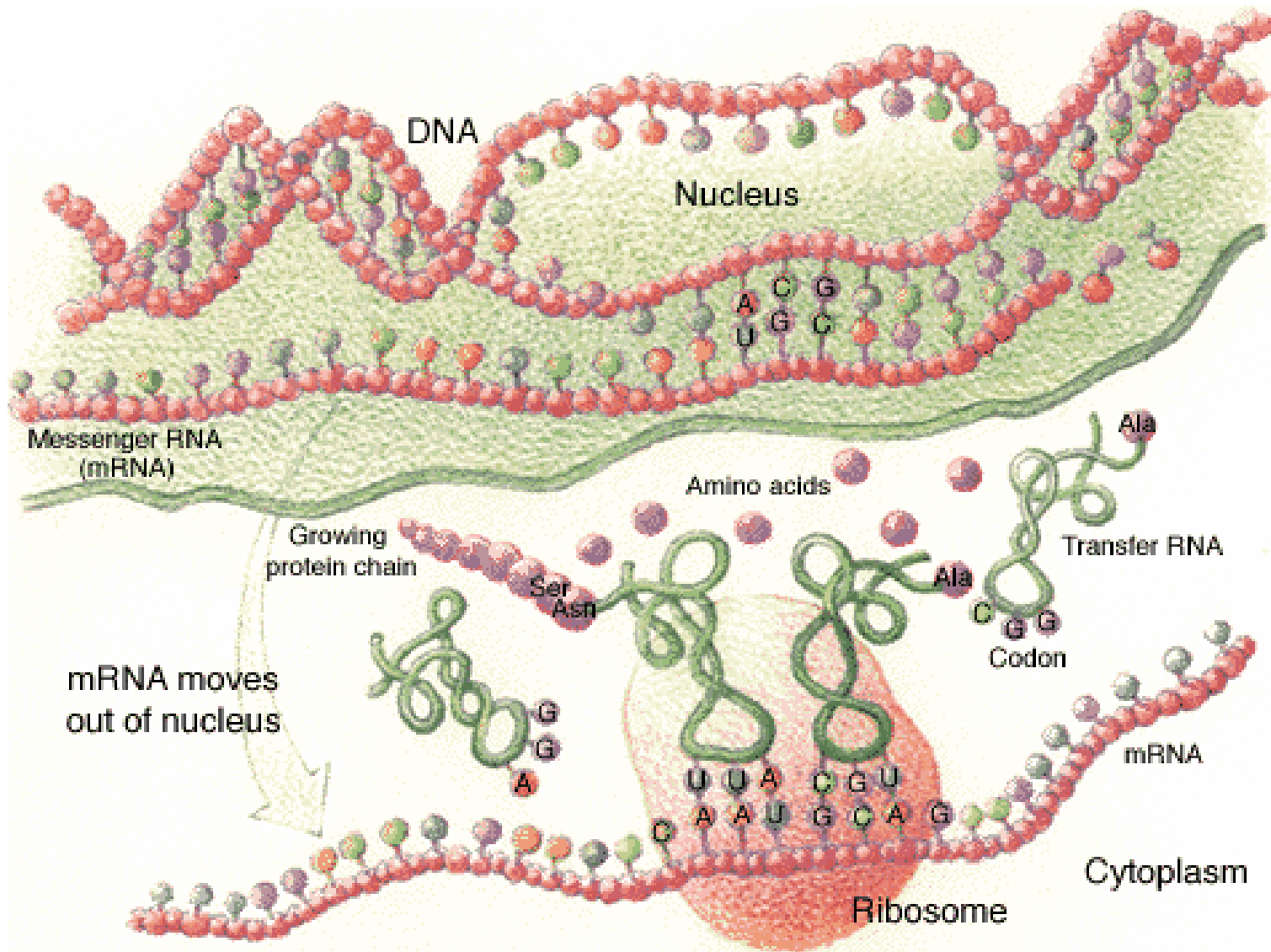


Genome-wide analysis

- 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



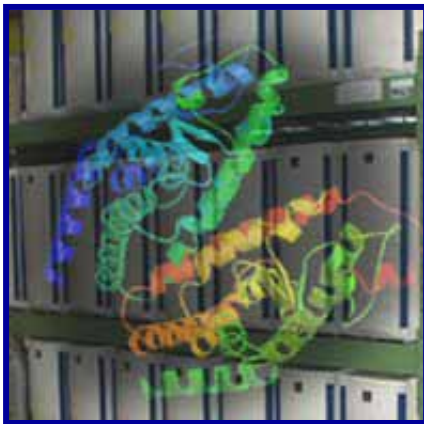
Gene to Protein



Networks of resources

BioinfoGRID

- The potential of **new biological and biomedical technological platforms** in connection with **HPC and GRID** technology will be particularly useful to deal with the increasing amount, complexity, and heterogeneity of biological and biomedical data.
- **Bioinformatics applications for eHealth** have become an ideal research area where computer scientists can apply and further develop new intelligent computation methods, in both experimental and theoretical cases.





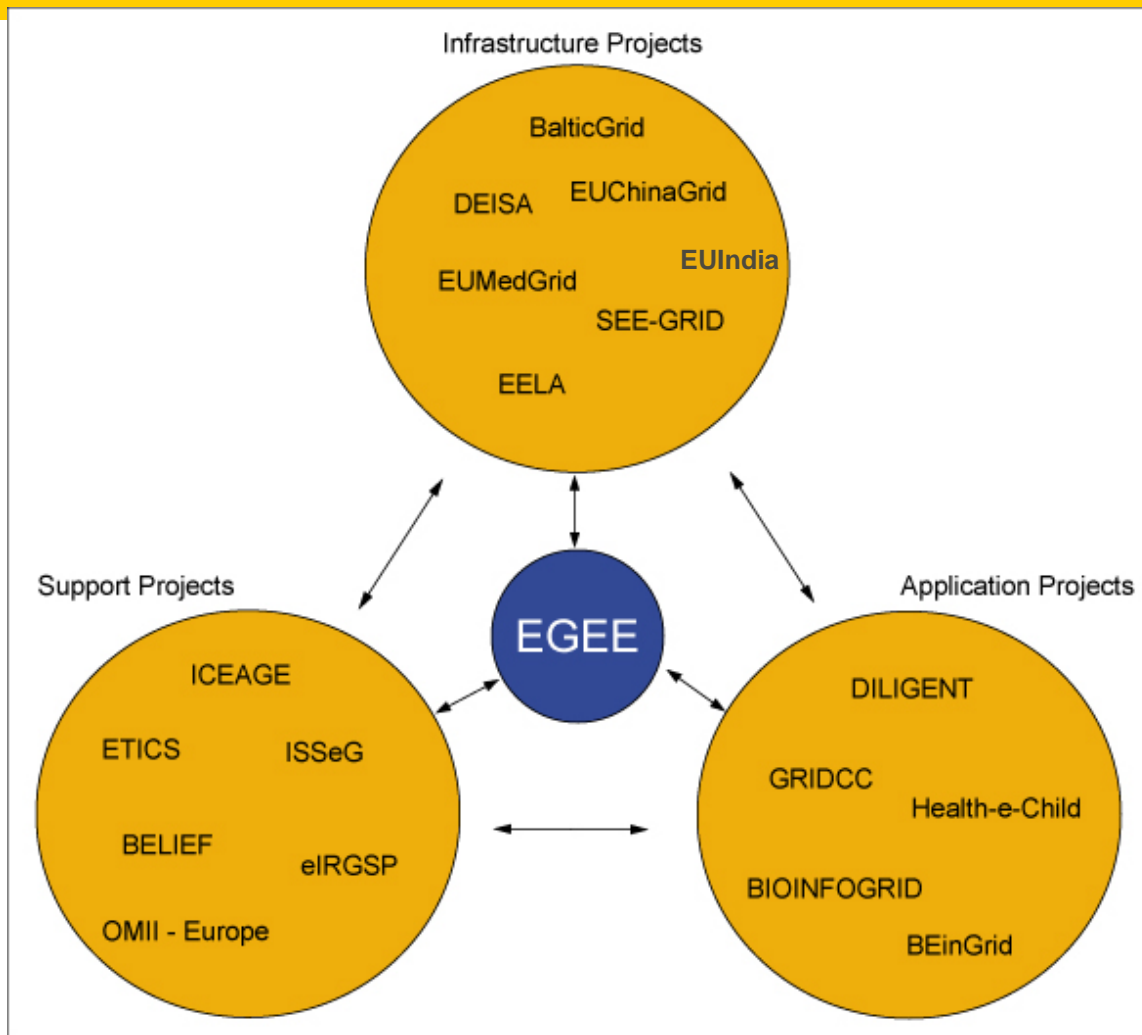
- A key development in the computational world has been the arrival of ***de novo* design algorithms** that use all available spatial information to be found within the target to **design novel drugs**.
- Coupling these algorithms to the rapidly growing body of information from structural genomics together with the new **ICT technology (eg. HPC, GRID, Web Services, Bioinspired networks 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.**

Related EU projects

BioinfoGRID



ISSeG



e-IRGSP





- 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



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

Description

- **GRID Bioinformatics User Tools**: 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

BioinfoGRID

- **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-EGCARKGKYGIIYTKVTAFLKWI DRSMKTRGLPKAKSHAPEVITSSPLK
FA7      TWYLTGIVSWG-QGCATVGHFGVYTRVSQYIEWLQKLMRSE-----PRPGVLLRAPFP
THROMBIN RWYQMGIVSWG-EGCDRDGKYGFYTHVFRLLKKWIIQKVIDQFGE-----
FA9      TSFLTGIISWG-EECAMKGKYGIIYTKVSRYVMWIKKTKLT-----
KALLIKREIN MURLVGITSWG-EGCARREQPGVYTKVAEYMDWILEKTQS SDGKAQMOSP A-----
FA11     VVHLYGITSWG-EGCAQRE R PGVYTNVVEYVDWILEKTQAV-----
TRYB1    TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIH HYVPKKP-----
TRYB2    TWLQAGVVSWG-EGCAQPNRPGIYTRVTYYLDWIH HYVPKKP-----
TRYA     TWLQAGVVSWD-EGCAQPNRPGIYTRVTYYLDWIH HYVPKKP-----
KLKE     --QLQGLVSWGME RCALPGYPGVYTNLCKYRSWIEETMRDK-----
CTRL     TWVLI GIVSWG-TKNCNVRA PAVYTRVSKFSTWINQVIAYN-----

```



Proteomics Applications in GRID

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

Description

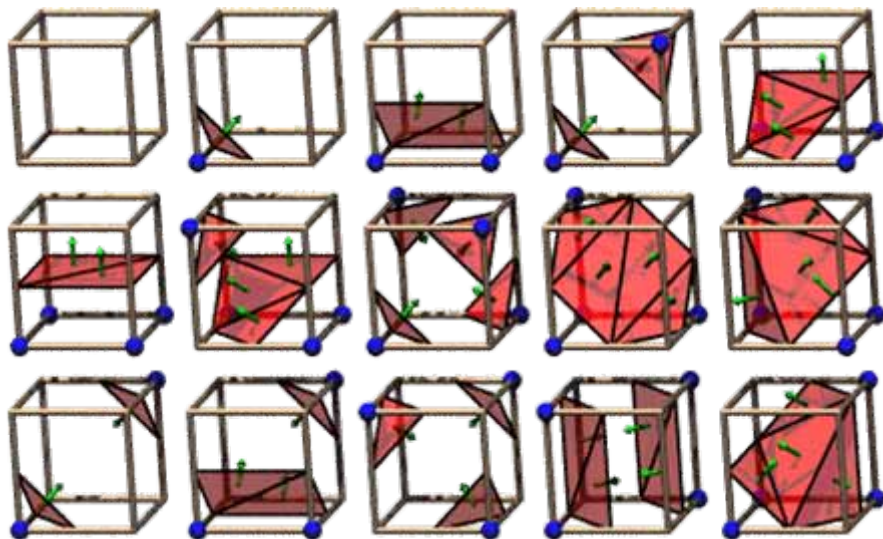
- **Perform functional protein analysis in GRID** by using the functional protein domain annotations on large protein families using GRID and related databases.



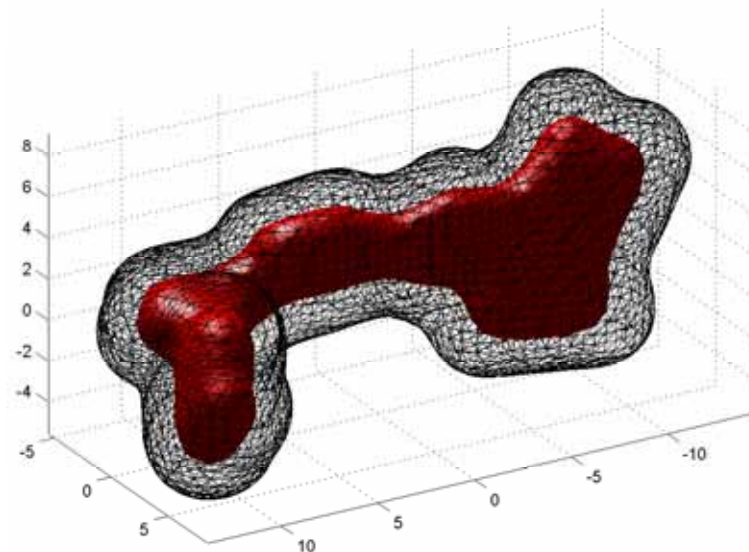


Proteomics Applications in GRID

- **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.



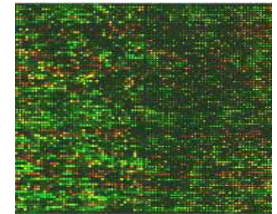
The 15 Cube Combinations





Transcriptomics applications

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



Description

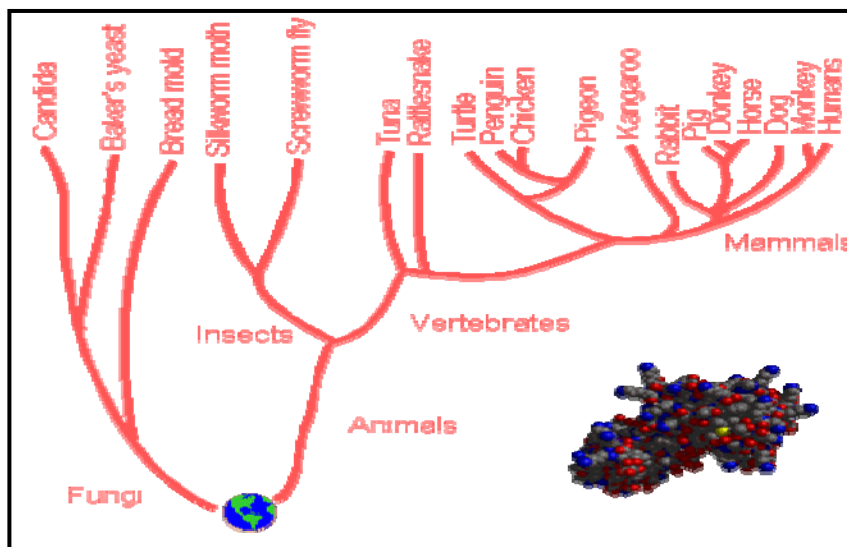
- **To perform algorithmic tools for gene expression data analysis in GRID:** evaluate the computational tools for extracting biologically significant information from gene expression data.
- Algorithms will focus on clustering steady state and time series gene expression data, multiple testing and meta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

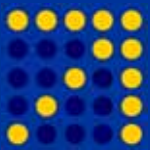


Phylogenetic application in GRID

BioinfoGRID

- **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 structural, physiological, and biochemical properties of sequences from other similar sequences, and reconstruction of tissue evolution.





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

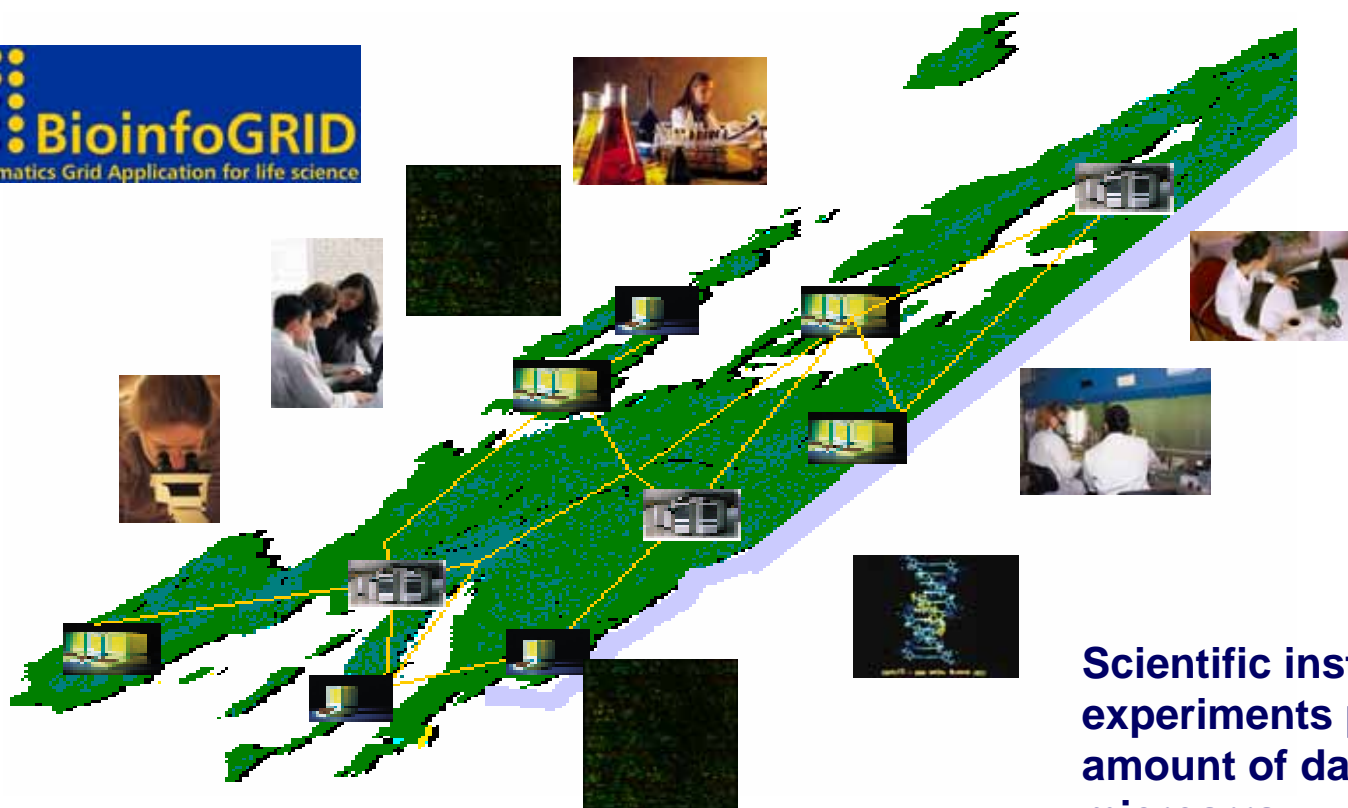
Description

- **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

Networking resources

BioinfoGRID

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

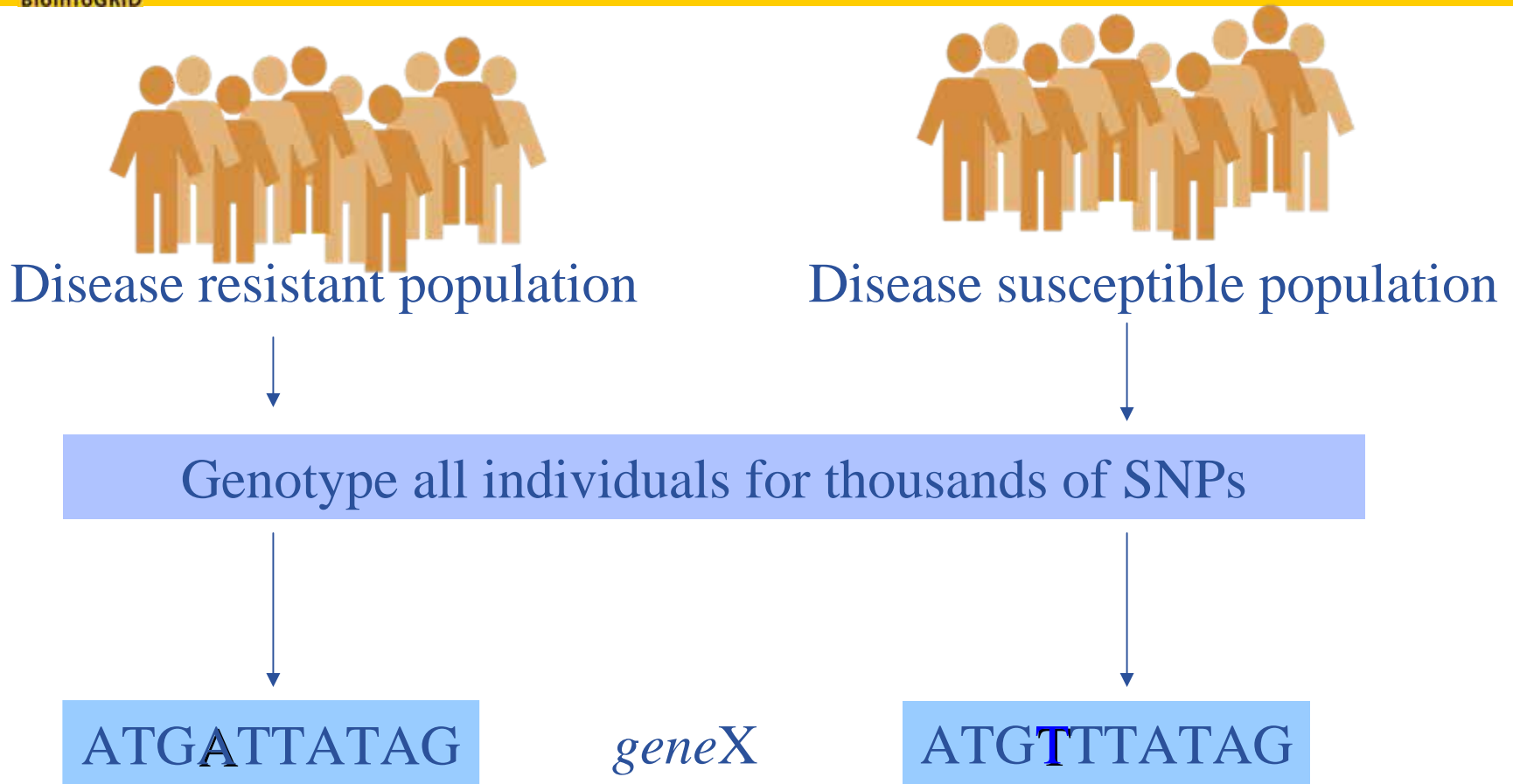


Researchers perform their activities regardless geographical location, interact with colleagues, share and access data

Scientific instruments and experiments provide huge amount of data from microarray



Disease Network



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

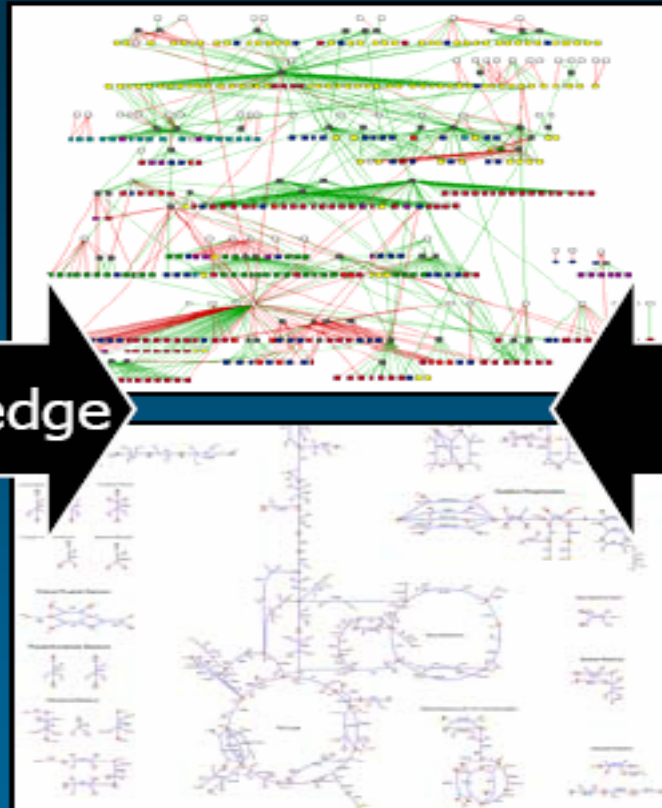


RininfoGRID

Regulatory Network Reconstruction

Network Reconstruction

Regulatory network



Knowledge

Data

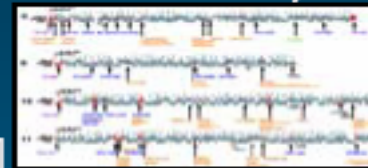
Metabolic network



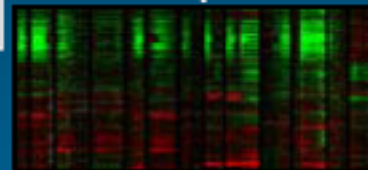
Promoter sequence

| | |
|--------------|-----------|
| GGTGGCAAAA | Rpn4 |
| AAAAGAATCA | Con4 |
| GAA TTCA GAA | HSE |
| AG GGGAA | Nrg1 |
| AAA CACGTTG | Cbf1 |
| AG ACTGG AG | RPS genes |
| TGATTGG | Hsp23.4 |

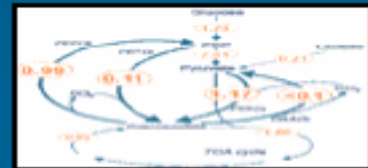
Location analysis



Gene expression



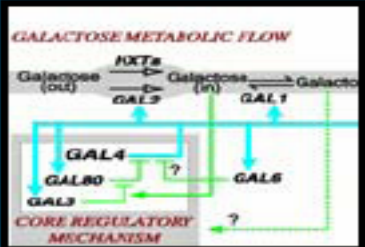
Metabolic fluxes



Genome annotation




Literature

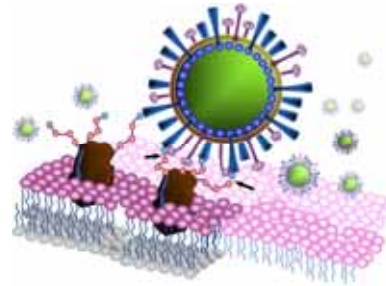


Curated databases





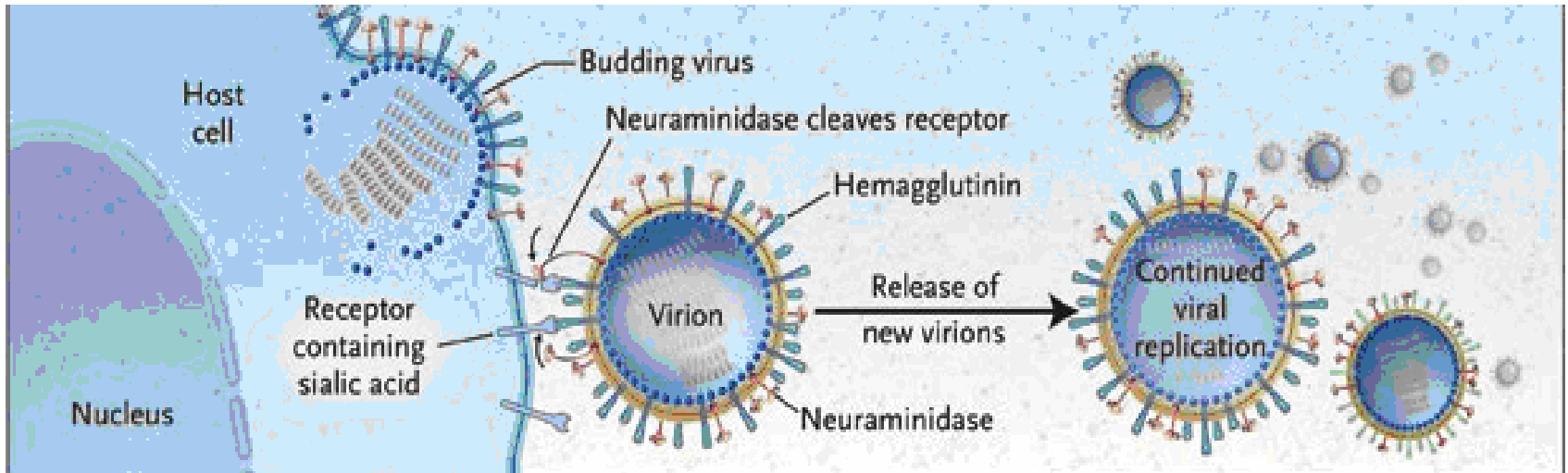
A collaborative EGEE project led by Academia Sinica in Taiwan, CNRS-IN2P3 in France and the European SSA BioinfoGRID project, was set up to identify new drugs for the potential variants of the Influenza A virus analysing 300,000 possible drug components against the avian flu virus H5N1 using the EGEE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



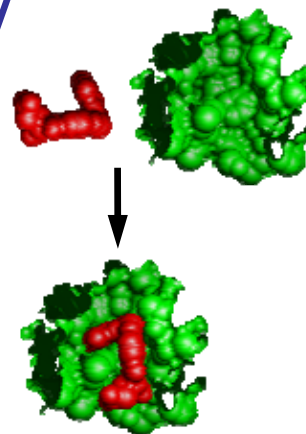
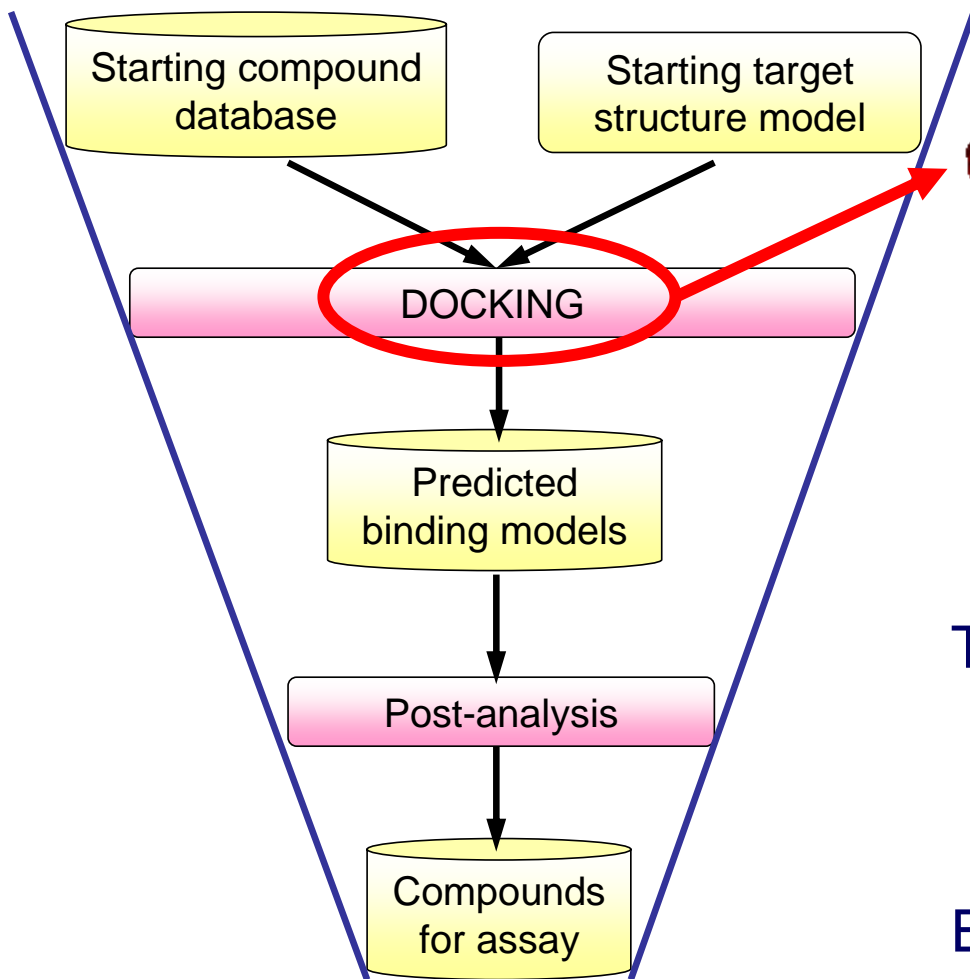
In silico virtual screening requires intensive computing, of the order of a few TFlops during one day to compute 1 million docking probabilities or for the molecular modelling of 1000 ligands on one target protein.

Neuraminidase Target

BioinfoGRID



The neuraminidase viruses is considered a valid target for antiviral drugs



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

There are successful examples

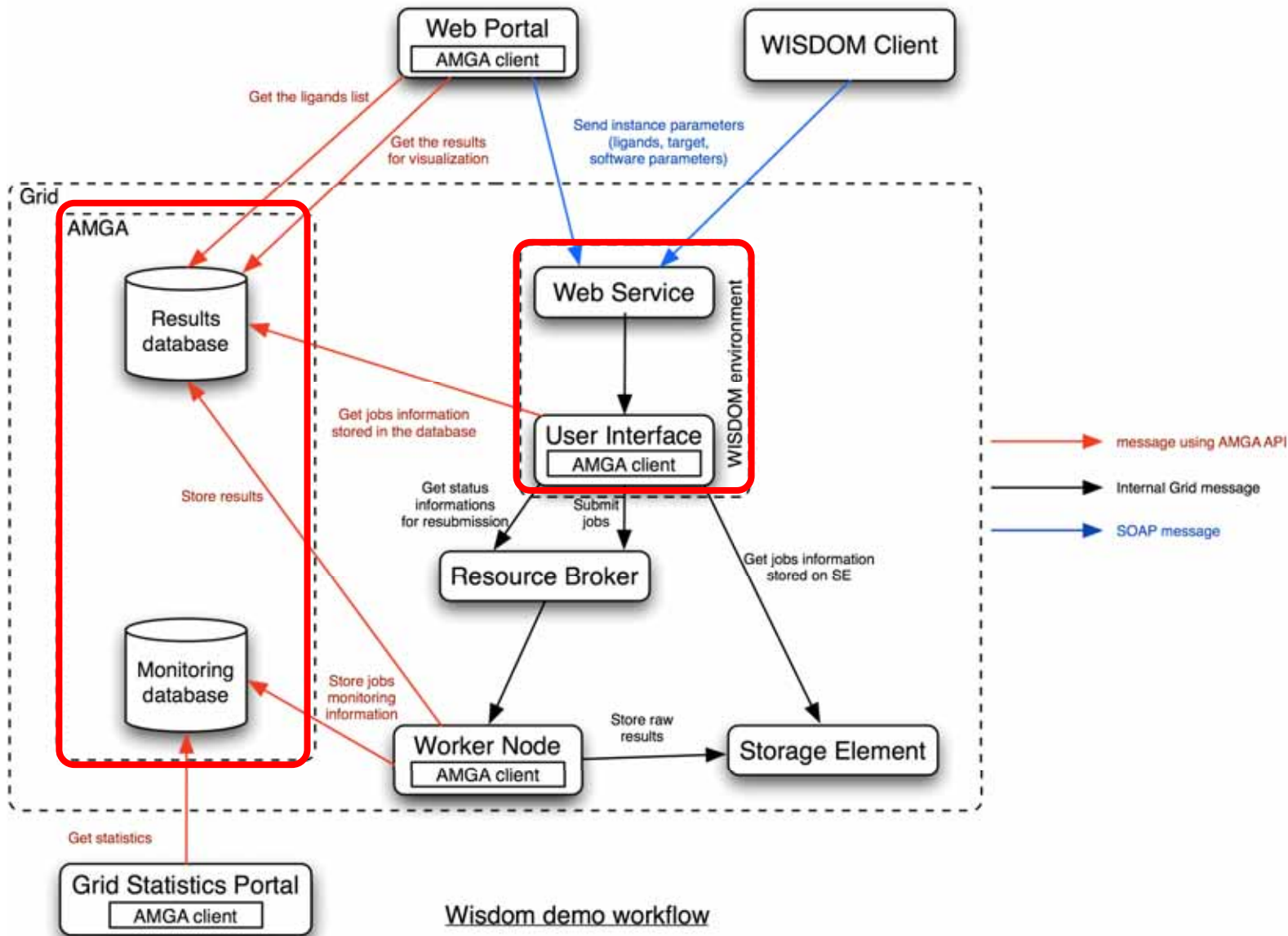
- rapid,
- cost effective...

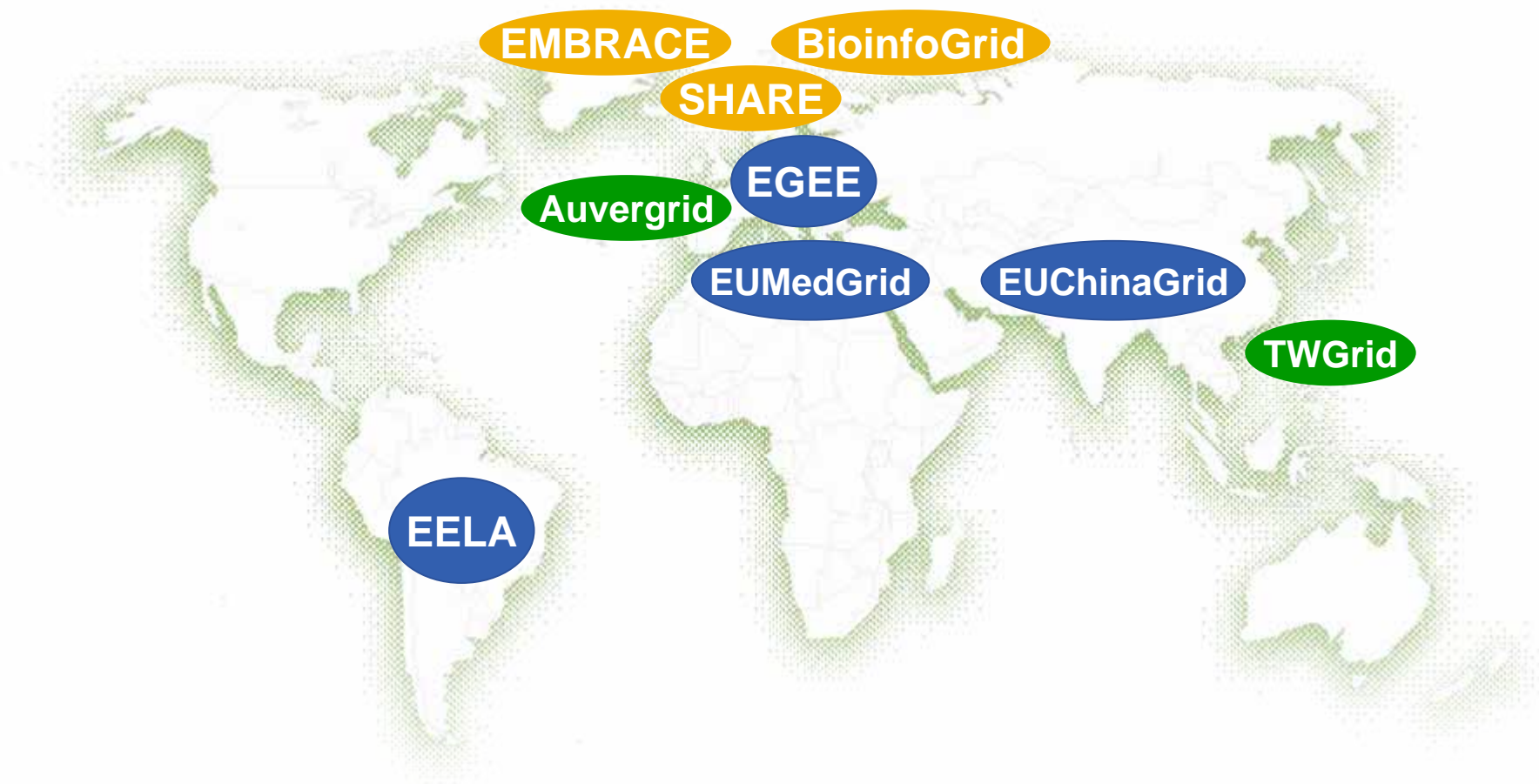
But there are limitations




- CPU and storage needed



The WISDOM Production System

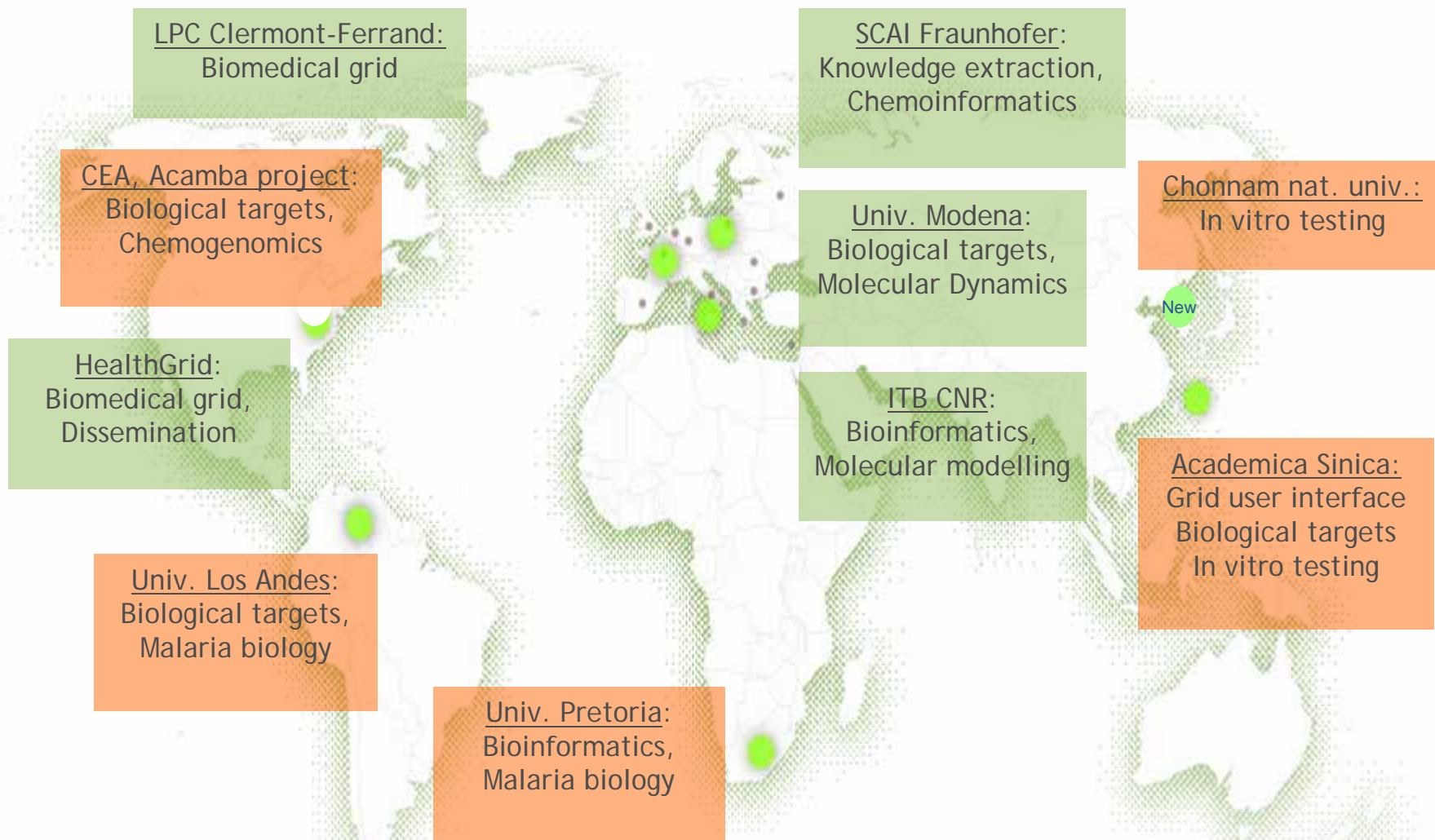




-  : European grid infrastructure
-  : European grid project
-  : Regional/national grid infrastructure



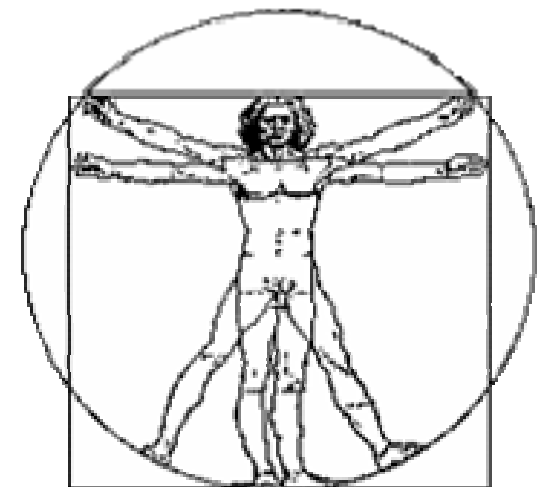
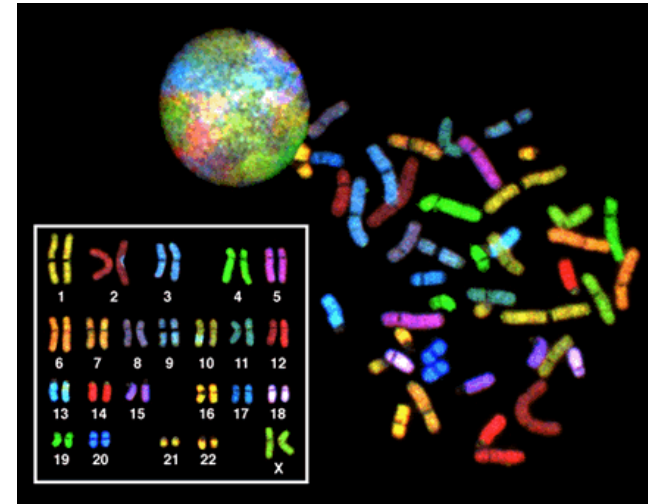
A grid for Malaria and Avian Influenza



The human organism:

BioinfoGRID

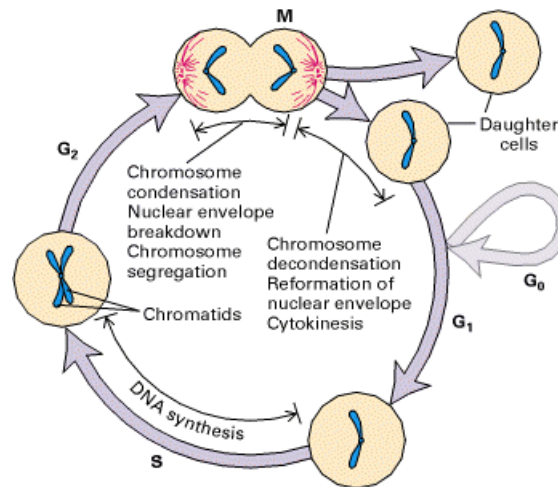
- ~ 3 billion nucleotides
- ~ 30,000 genes coding for
- ~ 100,000-300,000 transcripts
- ~ 1-2 million proteins
- ~ 60 trillion cells of
- ~ 300 cell types in
- ~14,000 distinguishable morphological structures





The Cell Cycle

BioinfoGRID



- **Cell Cycle:**

- repeated sequence of events which leads the division of a mother cell into daughter cells
- Biological process frequently studied in correlation to **tumour disease**
- It is considered a valuable target for **drug discovery** in the context of cancer and neurodegenerative disease



- Systems biology studies **how biological functions emerge** from the protein-protein interactions in the living systems;
- The **complexity** of this biological process relies in the high number of genes and networks of protein interactions involved in;
- The **quantification** of the behavior of each cell cycle components has a crucial role in the understanding the complex mechanism of cell cycle regulation.



System Biology: Cell Cycle

BioinfoGRID



CCDB Cell Cycle Database

Publication paper | SBML components - formulas | Simulate this model

- Home page
- Gene search
- Protein search
- Text search
- BLAST search
- Medata
- Links
- Acknowledgements

Bifurcation analysis of the regulatory modules of the mammalian G1/S transition.

(*) Swat M, Kel A, Herzel H - 2004 - *Bioinformatics*

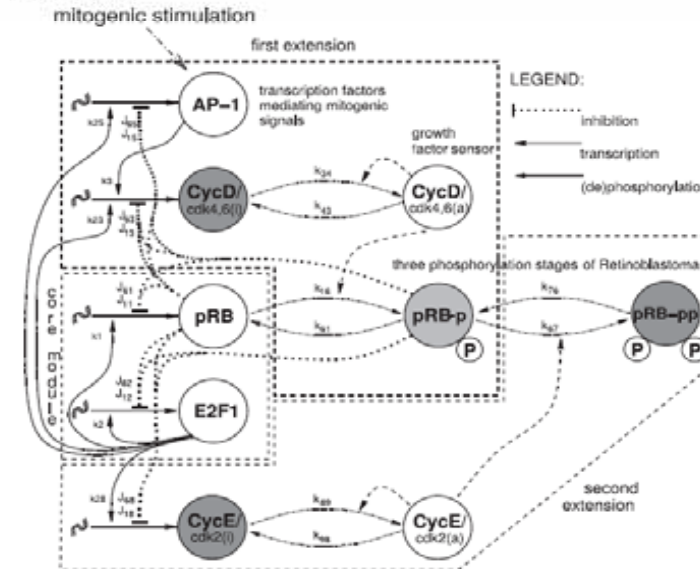
Abstract:

MOTIVATION: Mathematical models of the cell cycle can contribute to an understanding of its basic mechanisms. Modern simulation tools make the analysis of key components and their interactions very effective. This paper focuses on the role of small modules and feedbacks in the gene-protein network governing the G1/S transition in mammalian cells. Mutations in this network may lead to uncontrolled cell proliferation. Bifurcation analysis helps to identify the key components of this extremely complex interaction network. **RESULTS:** We identify various positive and negative feedback loops in the network controlling the G1/S transition. It is shown that the positive feedback regulation of E2F1 and a double activator-inhibitor module can lead to bistability. Extensions of the core module preserve the essential features such as bistability. The complete model exhibits a transcritical bifurcation in addition to bistability. We relate these bifurcations to the cell cycle checkpoint and the G1/S phase transition point. Thus, core modules can explain major features of the complex G1/S network and have a robust decision taking function.

Organism:

Mammalian

Paper graph (*):

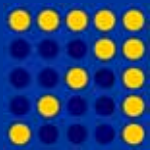


Model proteins:

- CCND1
- RB
- E2F1
- CDK4
- CDK6
- CDK2
- CCNE2
- CCNE1
- JUN_HUMAN
- CCND2
- TDP1
- TDP2

Links

- PubMed entry



Simulation Section

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

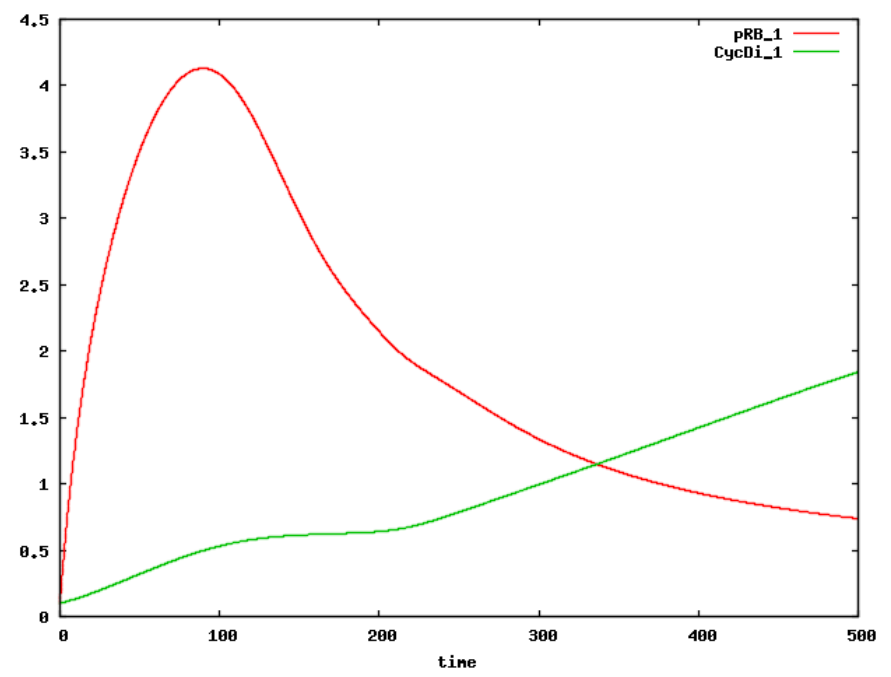
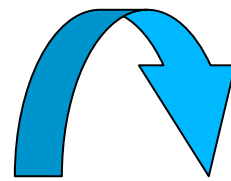
Download results file*

Select species to show on 2D plot

| | | | | |
|----------|---|---|---|---|
| x | | | | |
| time | | | | |
| y series | | | | |
| - | - | - | - | - |
| - | - | - | - | - |

time
pRB_1
pRBpp_1
E2F1_1
CycDi_1
CycDa_1
AP1_1
pRBpp_1
CycEi_1
CycEa_1

with GNUPLOT
its order of variables in the input interface; first is time



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

2D plot: image exported in png using GnuPlot



Genetic Diseases

High throughput techniques (i.e. DNA microarray)
to screen the whole genome

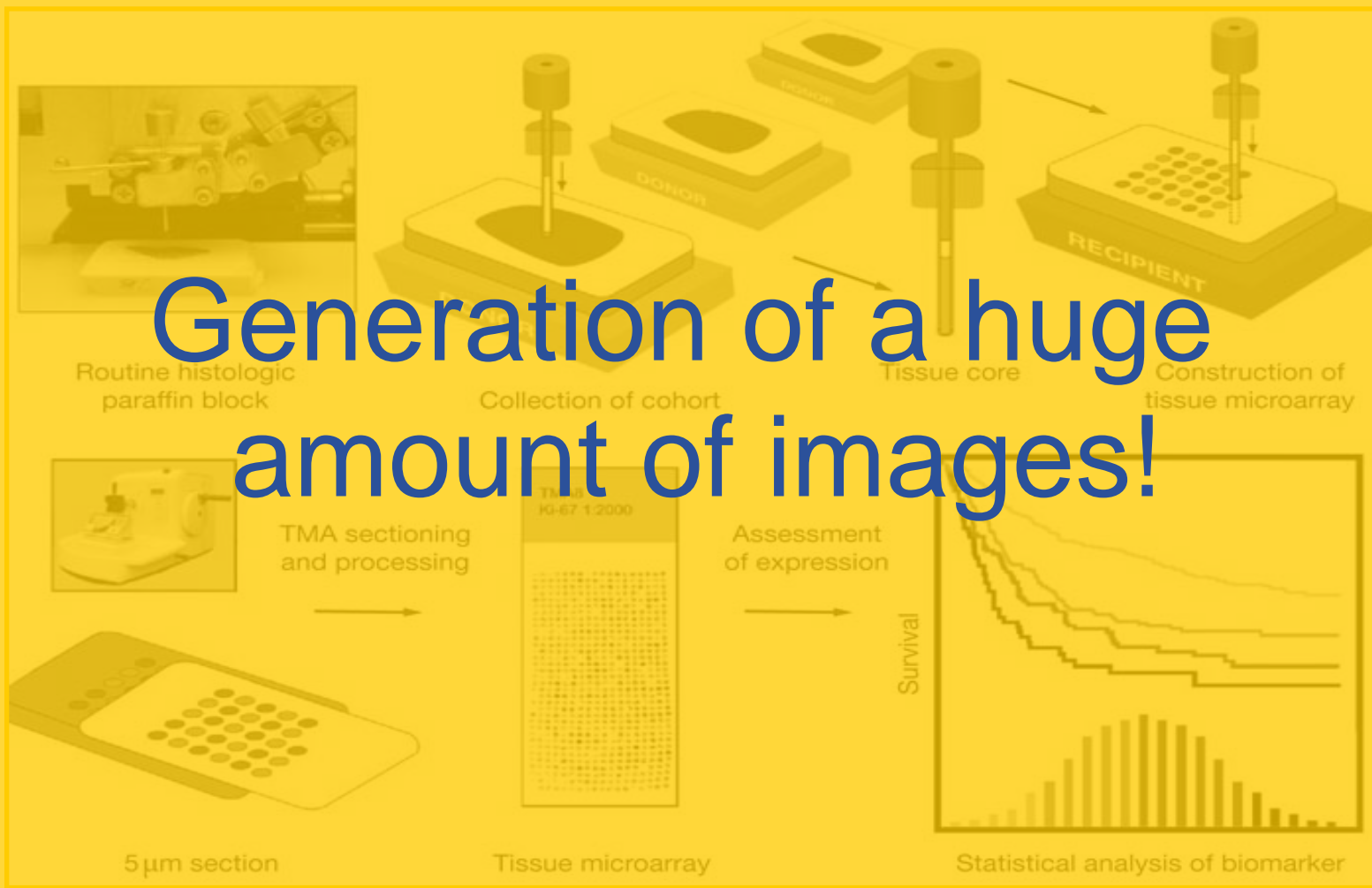
Low reliability

Validation through TMA

Tissue Microarray technology

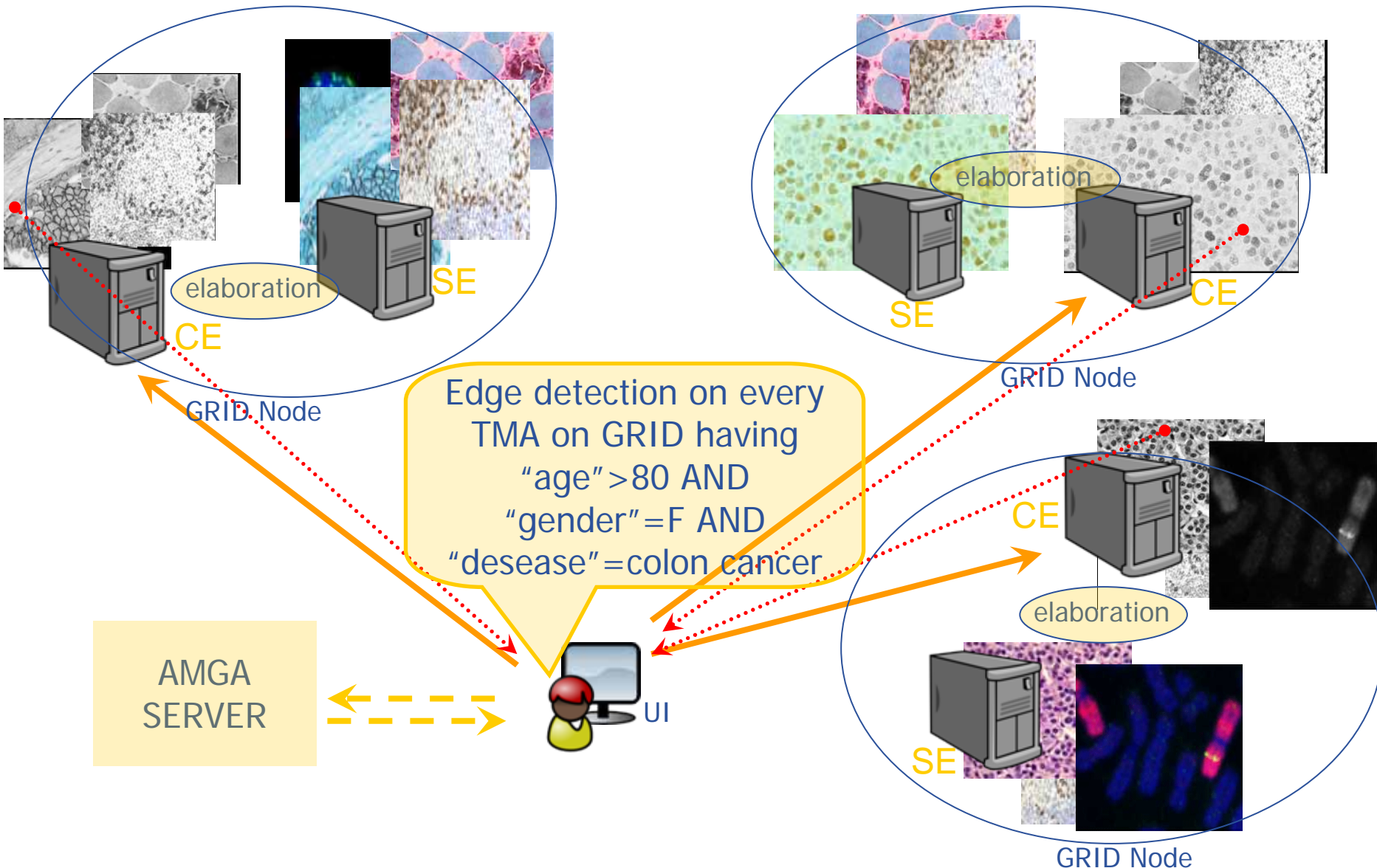
BioinfoGRID

Genes and proteins detection





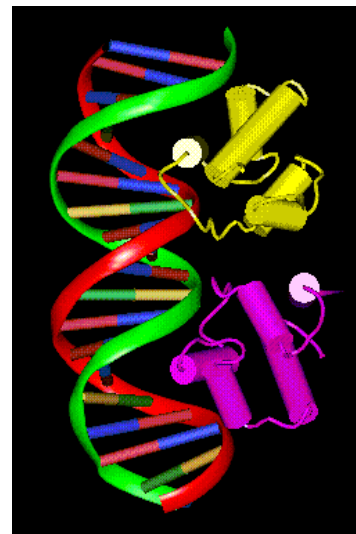
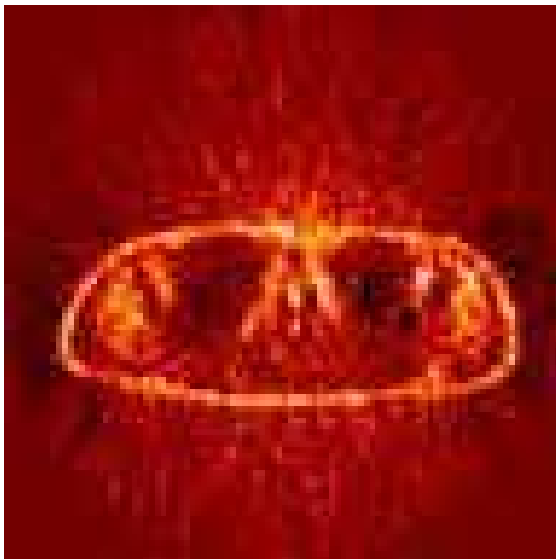
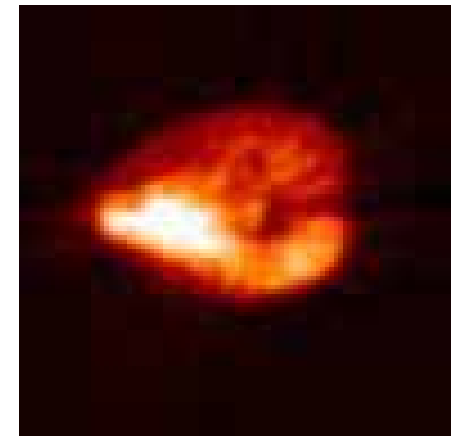
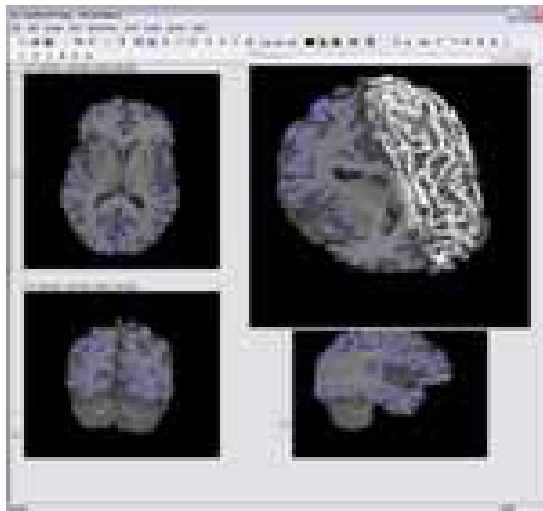
Tissue Microarray on GRID

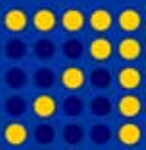




System Biology for Health

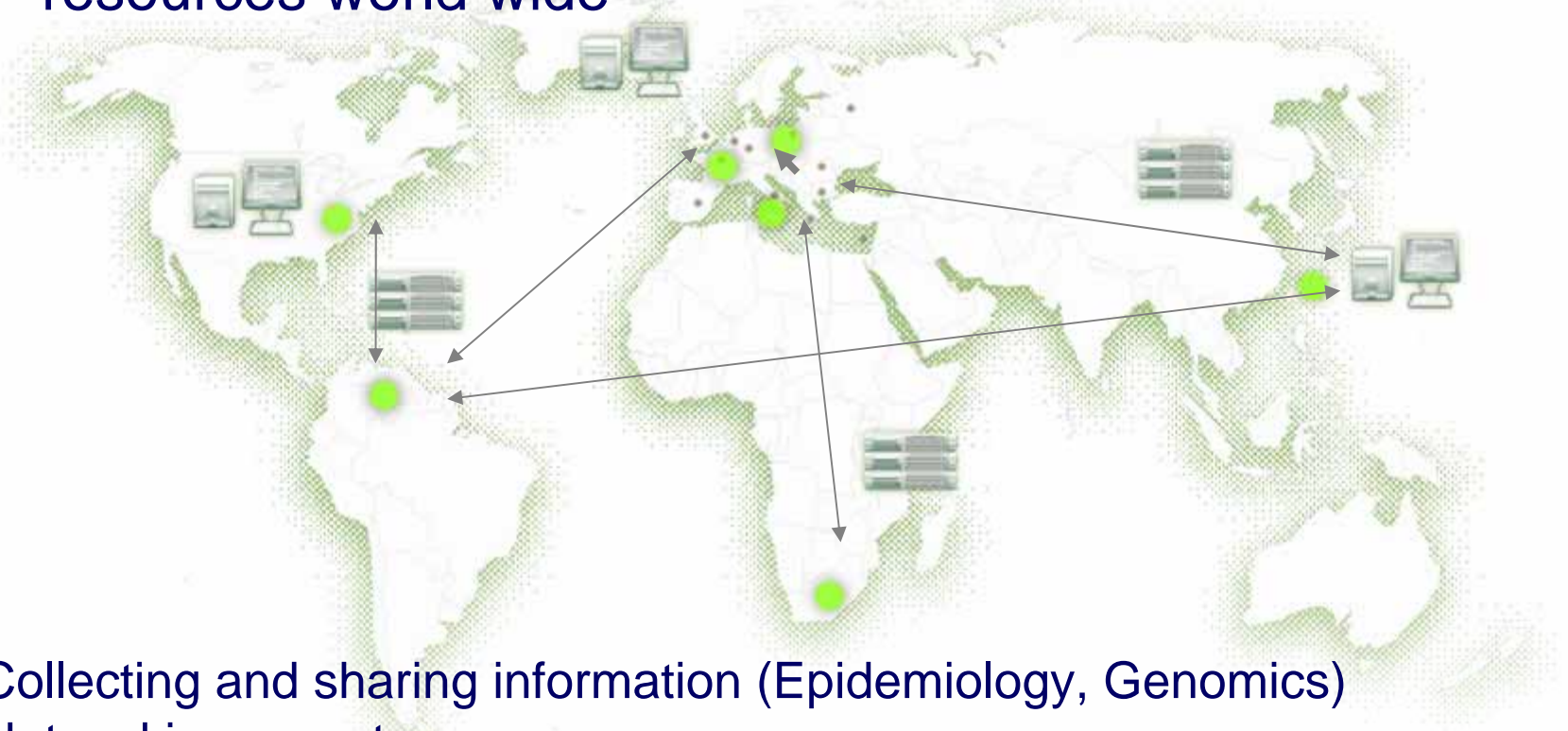
BioinfoGRID





Conclusion

- **Bioinformatics Challenges in Life Science** offer unprecedented opportunities for sharing information and resources world wide



- Collecting and sharing information (Epidemiology, Genomics)
- Networking experts
- Mobilizing resources routinely or in emergency (eg. vaccine, **drug discovery**)



BioinfoGRID

Acknowledgments



- BioinfoGRID *project*
<http://www.bioinfoGRID.eu>



- EGEE: Enabling Grid for E-science project
<http://www.eu.egEE.org>



- FIRB-MIUR LITBIO:
Laboratory for Interdisciplinary
Technologies in Bioinformatics
<http://www.litbio.org>,



Acknowledgments

- Wisdom
- <http://wisdom-demo.healthgrid.org>

IN2P3

INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE
ET DE PHYSIQUE DES PARTICULES



Academia Sinica
Genomics Research Center





BioinfoGRID

Acknowledgments

BIOMED GRID 2007

1st International Biomed GRID Summer School 2007

14-9 May Varenna – Italy



Information Society
and Media





BioinfoGRID

Welcome

BIOMED GRID 2008

2st International Biomed GRID Summer School 2008
May Varenna – Italy

