



Systems Biology Application in Grid

BioinfoGRID Workshop

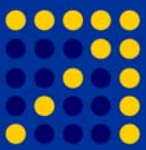
Healthgrid 2007 Conference

Geneva, April 24th 2007

Roberta Alfieri

Institute for Biomedical Technologies, CNR, Milan, Italy

CILEA, Milan, Italy



- **Introduction to Systems Biology and Cell Cycle Modelling**
- **Cell Cycle Model Simulation Machinery**
 - Database Model Section
 - Simulation Engine
 - User Interface
- **Parameter Estimation and Grid Technology**



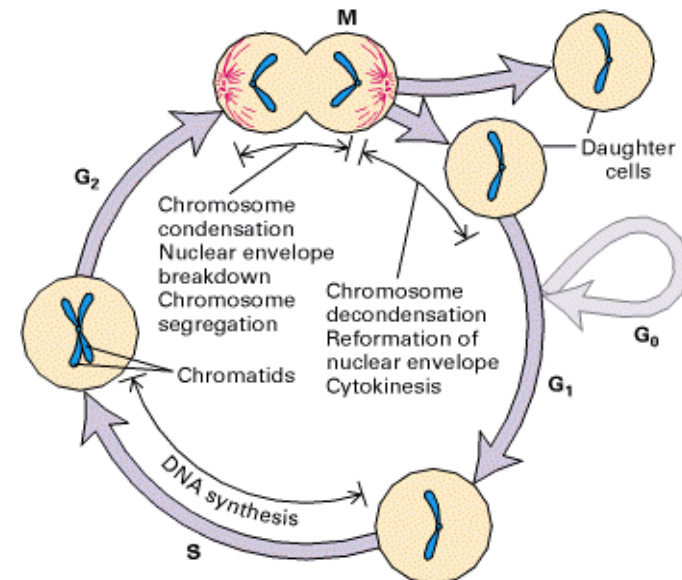
The Biological Problem: Cell Cycle

BioinfoGRID

Cell Cycle:

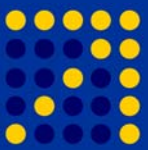
- repeated sequence of events which leads the division of a mother cell into daughter cells;
- Biological process frequently studied in correlation to **tumour disease**;
- It is considered a valuable target for **drug discovery** in the context of cancer and neurodegenerative disease.

The cell cycle is a frequently investigated process in systems biology, especially through **mathematical modelling**, to verify the impact differently regulated genes can have in normal and cancer cells.



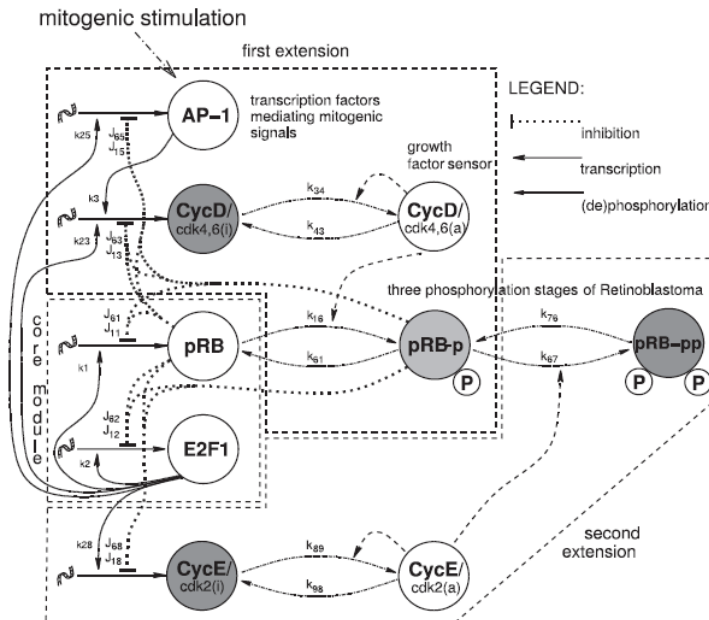


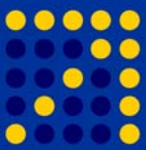
- Systems biology studies **how biological functions emerge** from the interactions between components of living systems;
- The **complexity** of this biological process lies in the high number of genes and networks of protein interactions involved;
- The **quantification** of the behaviour of each cell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



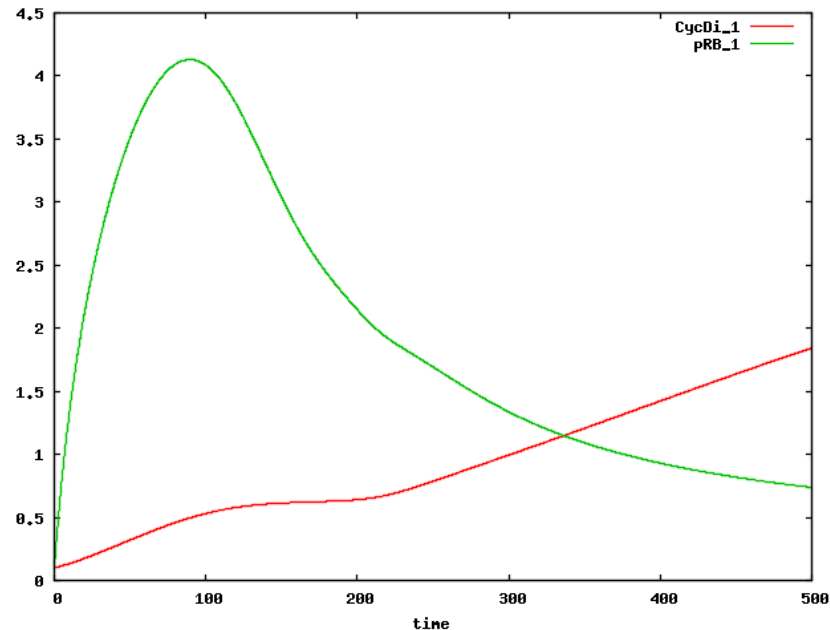
- **What is modelling?**

- The act of describing something in a schematic representation, usually on a smaller scale (general definition);
- Design and analysis of a mathematical representation of a biological system to outline unknown properties of that system, the emergent properties (systems biology definition).



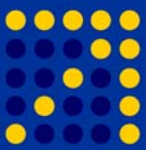


- **Mathematical representation of a biological process:**
 - Set of kinetic equations to define biochemical reactions
 - System of Ordinary Differential Equations to describe the dynamic behaviour of the model components
 - Initial parameters for kinetic equations
 - Initial concentration of the model species

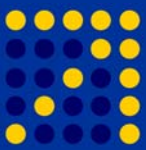




- Simulation of an **ODE system** is possible on a single workstation: the numerical integration of an ODE system is not very time consuming;
- **Parameter estimation**, the evaluation of the best set of parameters which define the model relating to a specific experimental dataset; **requires High Performance Computing techniques** since the computational load needed in finding the best model is very great;
- The estimation of the kinetic parameters *in silico* is performed **by computing a number of ODE systems** with different parameters and verifying the best solution.



- Introduction to Systems Biology and Cell Cycle Modelling
- **Cell Cycle Model Simulation Machinery**
 - **Database Model Section**
 - Simulation Engine
 - User Interface
- Parameter Estimation and Grid Technology



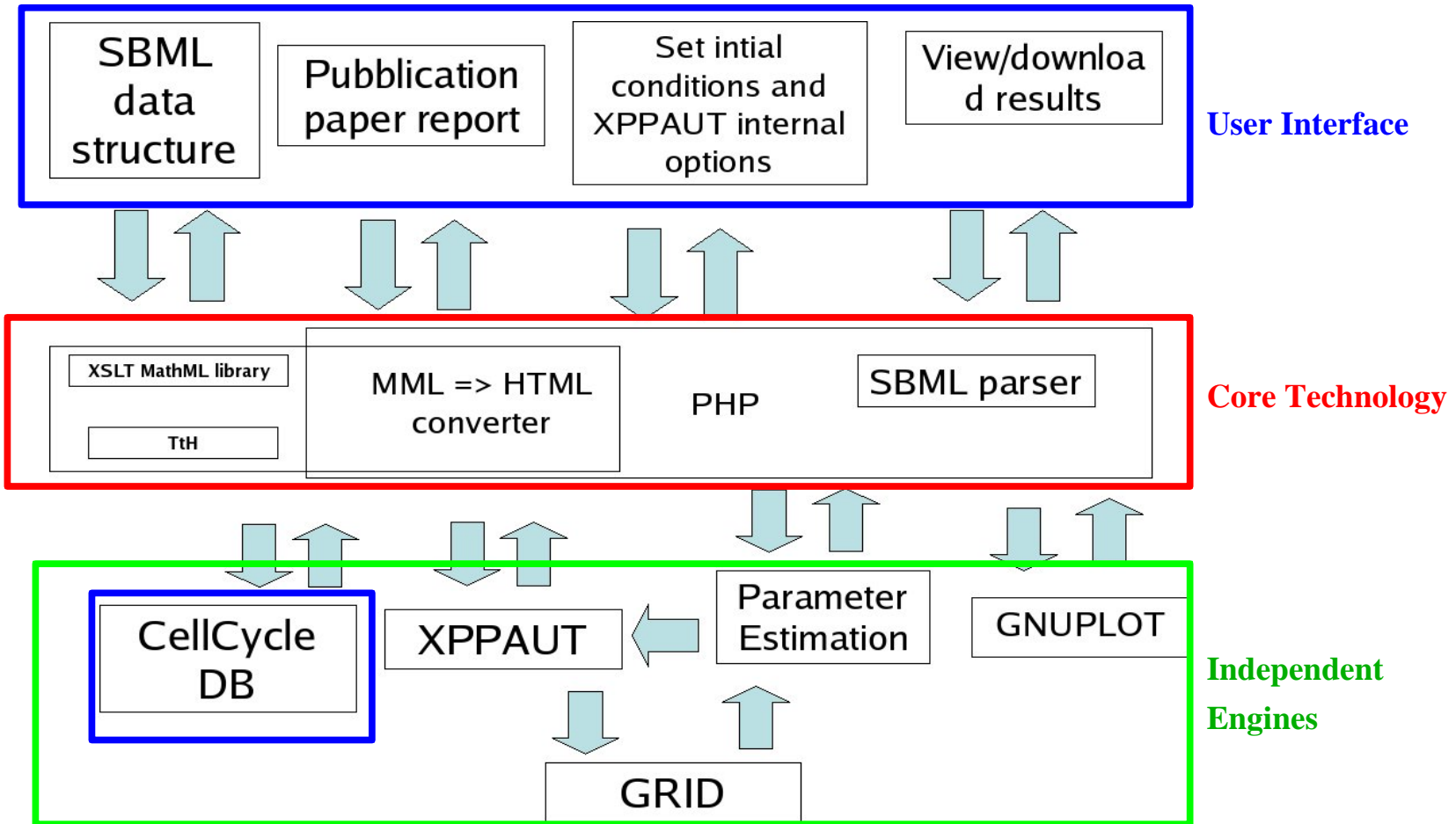
- **Cell Cycle Database:** relational database which integrates information about genes and proteins involved in yeast and mammalian cell cycle process;
- **Database section dedicated to cell cycle mathematical models**
 - **Model publication data** (information on the published models, such as the detailed publication data, authors, PubMed ID, abstract, journal information, diagram of the model, protein involved in the model, XML file, where available);
 - **SBML data structure** (SBML components of the model, including its mathematical expressions);
 - **Simulation section** (model simulation using XPPAUT, direct results retrieval in graphical formats).

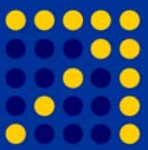


- Introduction to Systems Biology and Cell Cycle Modelling
- **Cell Cycle Model Simulation Machinery**
 - Database Model Section
 - **Simulation Engine**
 - User Interface
- Parameter Estimation and Grid Technology



Simulation Engine Workflow





- The **pipeline** is composed of a series of **PHP scripts** and allows the visualization and the computation of SBML models through:
 - Data retrieval from Cell Cycle Database;
 - SBML parser;
 - MathML to HTML converter: pipeline for the translation of the SBML mathematical expression for their visualization on web interface;
 - XPPAUT which is the simulation software chosen:
 - allows the solution of differential equations using many different options for the numerical algorithm;
 - widely used for the modelling of different biological pathways;
 - requires simply formatted input file.



- Introduction to Systems Biology and Cell Cycle Modelling
- **Cell Cycle Model Simulation Machinery**
 - Database Model Section
 - Simulation Engine
 - **User Interface**
- Parameter Estimation and Grid Technology

Cell Cycle Database Model Section

BioinfoGRID



CCDB Cell Cycle Database

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

[Publication paper](#)
[SBML components - formulas](#)
[Simulate this model](#)

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

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

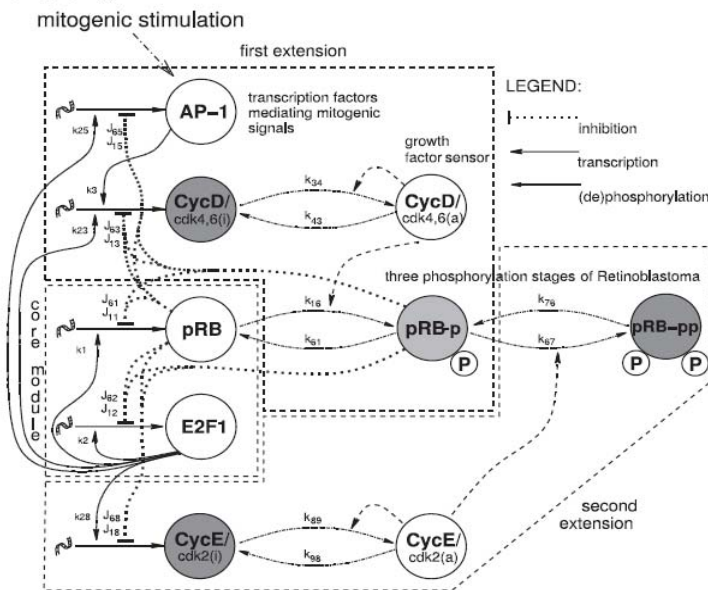
Abstract:

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



Species*

Species id	Value	Species id	Value	Species id	Value	Species id	Value	Species id	Value
pRB_1	0.1 mole	pRBp_1	0.1 mole	E2F1_1	0.1 mole	CycDi_1	0.1 mole	CycDa_1	0.1 mole
AP1_1	0.1 mole	pRBpp_1	0.1 mole	CycEi_1	0.1 mole	CycEa_1	0.1 mole		

Parameters

Parameter id	Value	Parameter id	Value	Parameter id	Value	Parameter id	Value	Parameter id	Value
k1_1	1.0	Km1_1	0.5	J11_1	0.5	J61_1	5.0	k16_1	0.4
k61_1	0.3	phi_pRB_1	0.0050	kp_1	0.05	k2_1	1.6	a_1	0.04
Km2_1	4.0	J12_1	5.0	J62_1	8.0	phi_E2F1_1	0.1	k3_1	0.05
k23_1	0.3	J13_1	0.0020	J63_1	2.0	k34_1	0.04	Km4_1	0.3
phi_CycDi_1	0.023	k43_1	0.01	phi_CycDa_1	0.03	Fm_1	0.0050	k25_1	0.9
J15_1	0.0010	J65_1	6.0	phi_AP1_1	0.01	k67_1	0.7	k76_1	0.1
phi_pRBpp_1	0.04	phi_pRBp_1	0.06	k28_1	0.06	J18_1	0.6	J68_1	7.0
k89_1	0.07	Km9_1	0.0050	k98_1	0.0010	phi_CycEi_1	0.06	phi_CycEa_1	0.05

XPPAUT internal options setting

total	dt	method	tolerance	minimum step	maximum step	delay
500	0.1	stiff	0.001	1e-12	1	0

Users can:

➤ Change parameter values and protein

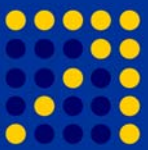
initial concentration

➤ Select the ODE solver

* when not provided in the original paper we suggest possible values

The simulation software is XPPAUT - Bard Ermentrout, 2002

- ♦ dt = sets step size used by the fixed step integrators and the output step for Gear and Stiff
- ♦ total = sets total time
- ♦ tolerance = sets the error tolerance for adaptive methods
- ♦ sets the minimum allowable time step for adaptive methods
- ♦ sets the maximum allowable time step for adaptive methods
- ♦ delay = sets the upper bound for the maximum delay, for use with Delay Differential Equation



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

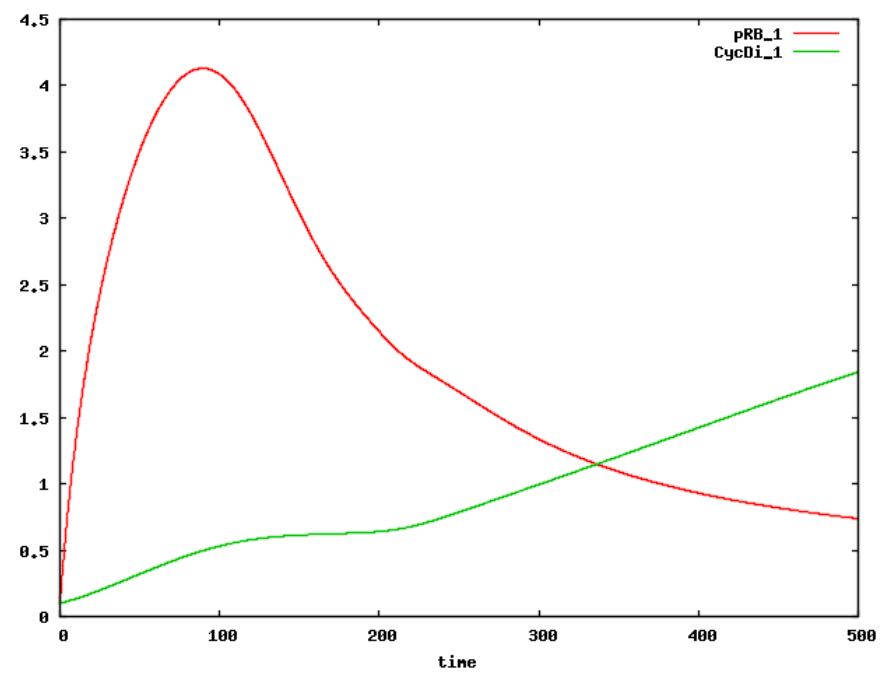
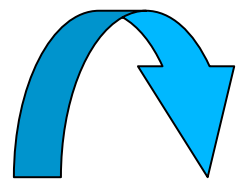
Download results file*

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

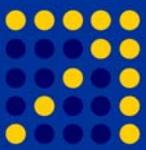


The simulation of a single ODE system describing a cell cycle model is possible

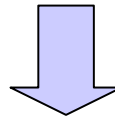
2D plot: image exported in png using GnuPlot



- Introduction to Systems Biology and Cell Cycle Modelling
- Cell Cycle Model Simulation Machinery
 - Database Model Section
 - Simulation Engine
 - User Interface
- **Parameter Estimation and Grid Technology**



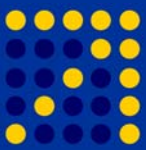
Estimate the model which fits with real biological data the best



Find the best parameter set which describes the real biological system

Possible approaches to parameter estimation:

- Deterministic mathematical methods
- Stochastic mathematical methods



● Deterministic approach

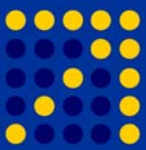
Algorithm of parameter estimation using **ODRPACK**: an estimation for the rate constant is given by minimizing the weighted orthogonal distance between the experimental data set and the calculated model;

- Zwolak et al, Parameter estimation for a mathematical model of the cell cycle in frog eggs, *J. Comput Biol.* 2005

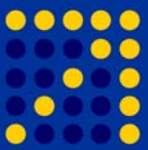
● Stochastic approach

Parameter estimation through an **Adaptive Swarm Algorithm** based on simulation of social behaviour in a flock of birds. This algorithm is highly suitable for constrained multi-objective optimization problems. The models are simulated over the grid through GridX meta scheduler and Globus.

- P.K. Dhar et al., Grid Cellware: the first grid-enabled tool for modelling and simulating cellular processes, *Bioinformatics*, 2005



- **Evolutionary algorithms:** population-based stochastic methods relying on the idea of biological evolution:
 - Iterative creation of new **generations of individuals (relying on the recombination of the best individuals of the previous generation)** in numerical forms to find solutions close to optimum (experimental data);
- Three groups of evolutionary methods:
 - Genetic Algorithm
 - Evolutionary Programming
 - **Evolutionary Strategies:** the most efficient and robust especially for continuous problems, like ODE systems resolution (N. Saravanan, 1995).



Simulation Time Estimation

Model example: 9 species, 41 parameters, 22 reactions, 9 ODEs

(Swat et al, 2004)

Single Numerical
Simulation

COMPUTATION DEVICE:

Intel Pentium 2.0 Ghz CPU
with 1GB RAM
Integrator: Stiff
Time units: 1000

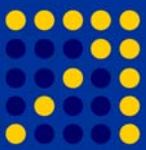
Evolutionary
Computation for
Parameter
Estimation (50000
individuals for 100
generations)



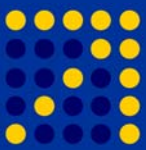
4 seconds



231 days



- The use of a High Performance Computing platform like grid for the computation of a **large number of independent ODE systems solution** is possible;
- The porting of the ODE solver system on the grid has been successfully performed by the creation of an **infrastructure** able to **distribute the computation efficiently**;
- The **parameter estimation engine** works on the top of a set of scripts for:
 - Job submission;
 - Monitoring of the computation;
 - Retrieval and integration of the results.



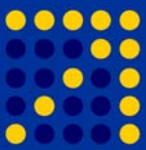
- Development of a **system for the parameter estimation** in order to find the best parameter set by computing many different simulations with the Evolutionary Strategy algorithm using the **grid platform** to overcome the computation complexity coming from:
 - the high number of parameter combination values;
 - the high number of simulations needed to fit data;
- Difference from the other grid-based parameter estimation approaches:
 - type of algorithm used: **Evolutionary Strategy Algorithm**
 - grid platform on which the computation is performed: **grid platform based on gLite.**



- Key parameter for distribution: **number of simulations for each job;**
- The number of equations which have to be simulated in a specific job is related to the computation time needed for each simulation:
 - To better exploit the computation resources we set the number of simulations for each job to **500 ODE systems;**
 - The estimated time for **each job** is about **30 minutes** and considering the queue time the **global computation** needs **almost a week.**



- The parameter estimation is controlled by a **set of scripts** that are responsible for the submission, the monitoring and the retrieval of the results for each job;
- The system works for **generation step**:
 - All the jobs of a generation are sent to the grid;
 - When the results are retrieved the software integrates them in order to create a new generation;
 - The new generation is re-submitted to grid;
- The **ODE solver system** is **deployed** on the **grid node** at job execution time and the results are retrieved from the User Interface where the data are integrated to generate following populations.



- We present a **grid-oriented approach to solve ODE systems** describing cell cycle models, in order to make the numerical simulations of the biological process easier and more accurate;
- We choose to perform **parameter estimation** using a High Performance Computing platform like the grid because the system is designed with the aim to estimate the best model by computing many different simulations of each model;
- The implemented system is useful to **manage the mathematical information** related to cell cycle models and to simulate the whole process **using the grid platform.**



- **This project has been supported by:**

- Italian FIRB-MIUR projects “Italian Laboratory for Bioinformatics Technologies - LITBIO” and “ITALBIONET”
- european “Specific Support Action BioinfoGRID” and “EGEE” projects

- **People from Bioinformatics Group at Institute for Biomedical Technologies, CNR, Milan**

- Luciano Milanesi
- Ivan Merelli
- Ettore Mosca