Parameter estimate in biochemical network models

Giuseppe Aprea[1] Grazia Licciardello[2] Vittorio Rosato[3]

[1]Portici Research Center, CRESCO project, Via del Vecchio Macello, 80055 Portici (Naples), Italy. giuseppe.aprea@gmail.com

[2]Science and Technology Park of Sicily, stradale V. Lancia 57, z.i. Blocco Palma I, 95121 Catania, Italy. (Naples), Italy. gralicci@unict.it

[3]Casaccia Research Center, Computinig and Modelling Unit, Via Anguillarese 301, 00123 S.Maria di Galeria, Italy. rosato@casaccia.enea.it

February 11, 2009

ENTE PER LE NUOVE TECNOLOGIE, L'ENERGIA E L'AMBIENTE

CRESCO

Outline

- Introduction:Motivations and Problem Statement
- The genetic algorithm (GA)
- Our GA implementations
- Application to PHA production metabolic network in Pseudomonas Corrugata.

Motivations

- advances in genomic and metabolic profiling have begun to produce **unprecedented amounts of data** that await analysis and interpretation.
- reliable explanations of how processes are regulated require an accurate modeling approach at the systems level; quite often these models of **biochemical networks** rely on several unknown **parameters** which **need to be estimated**.
- the estimate of model parameters from experimental data **remains a bottleneck** for major breakthrough in Systems Biology; it consists in the solution of an **inverse problem** which **requires** the use of an efficient **optimization algorithm**.
- the **genetic algorithm (GA)** is a widely known method yielding reliable values for model parameters with a **large computational demand**.
- **GA is a parallelizzable algorithm**.

Target

- Our aim is to develop a parallel implementation for parameter estimate based on GA to fully deploy the large computational power of a modern grid such as that set up at the Enea Portici site.
- Our implementation relies on:
  - Ecell software from Keio University(Japan) for simulations of biochemical networks
  - LSF - load sharing facility - a job scheduler for (multi)cluster (SGE is supported as well)

Parameter estimate as an optimization problem

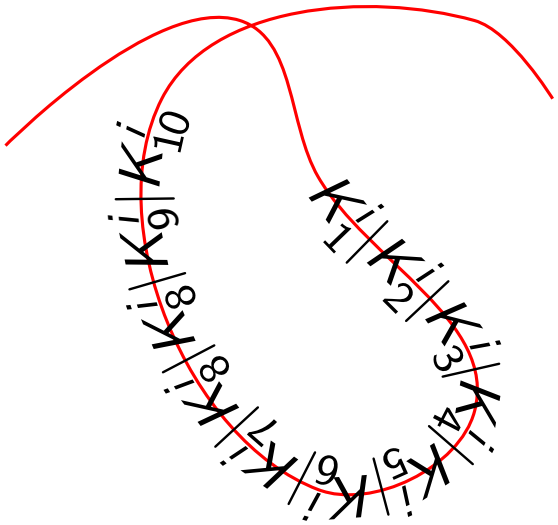
Given

- experimental time course data for some reactants
- a model biochemical network with missing kinetic parameters

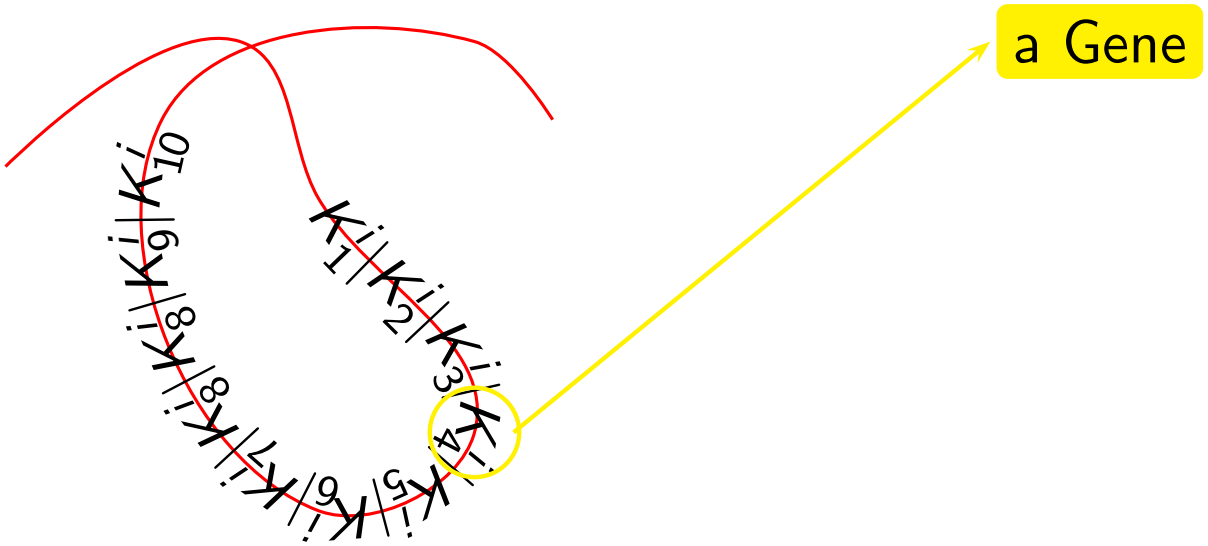
the goal is

evaluating the unknown kinetic constants by minimizing the measure of the distance between experimental and simulated data (cost function).

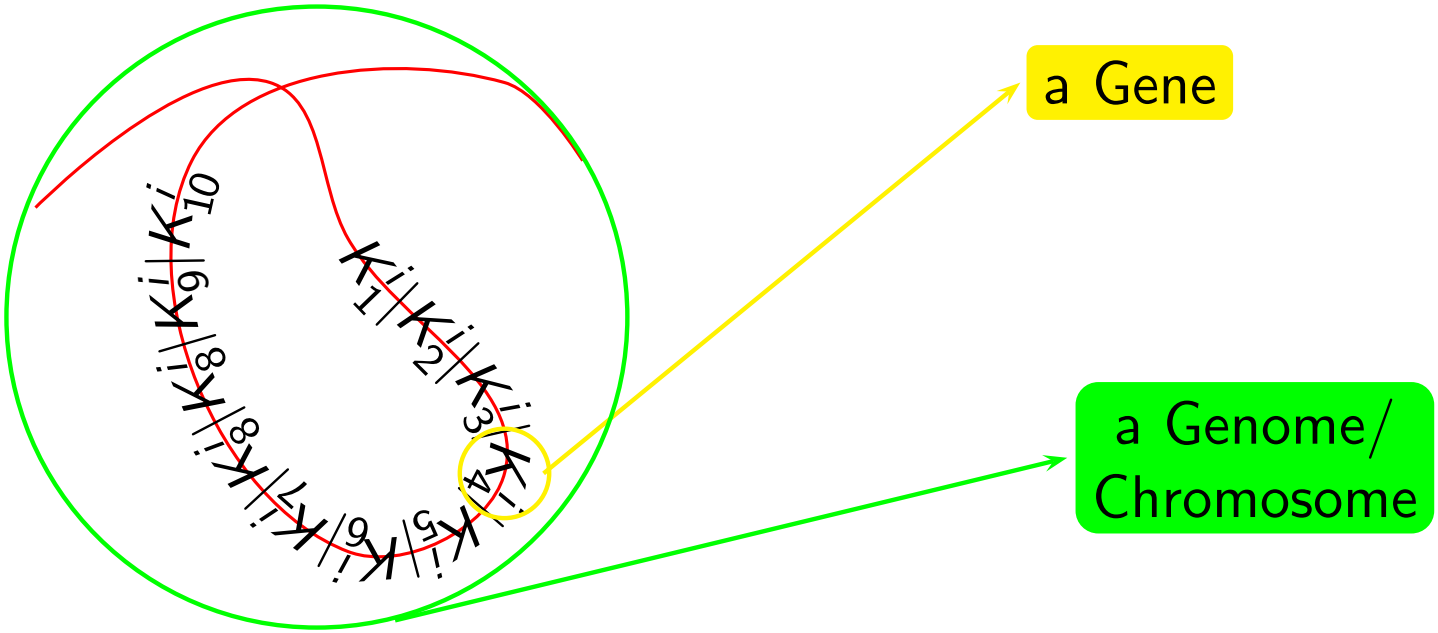
Minimization by analogy with Nature: The genetic algorithm



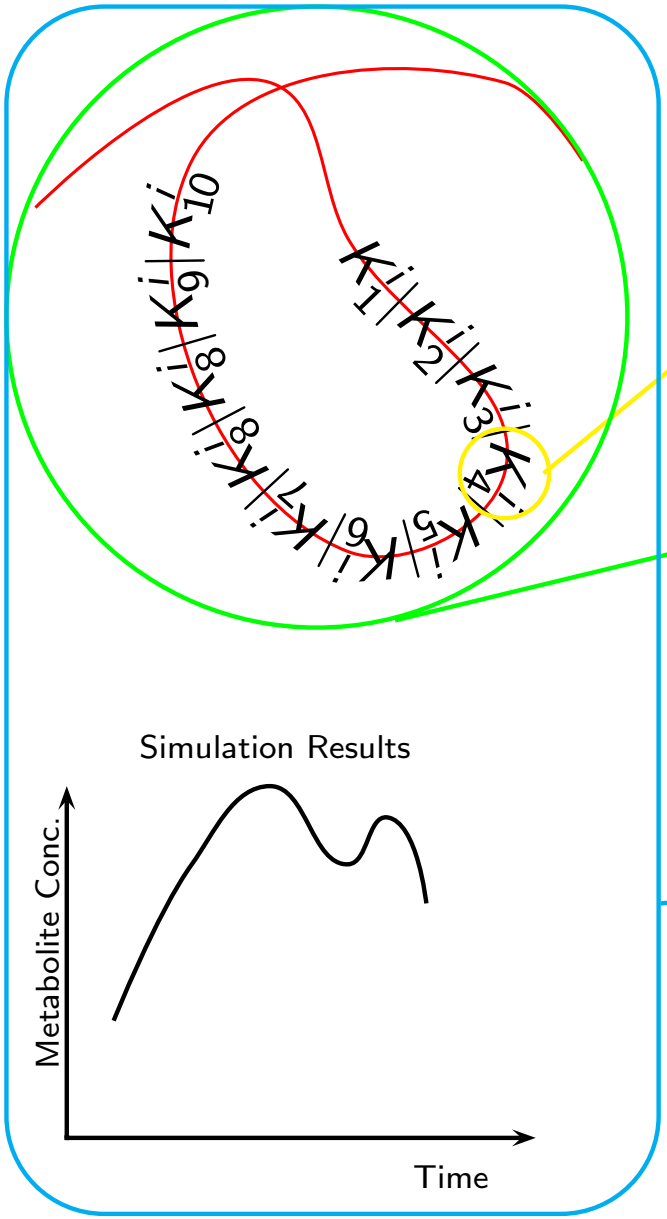
Minimization by analogy with Nature: The genetic algorithm



Minimization by analogy with Nature: The genetic algorithm



Minimization by analogy with Nature: The genetic algorithm



a Gene

a Genome/
Chromosome

an Individual
Several Individuals
form a Population

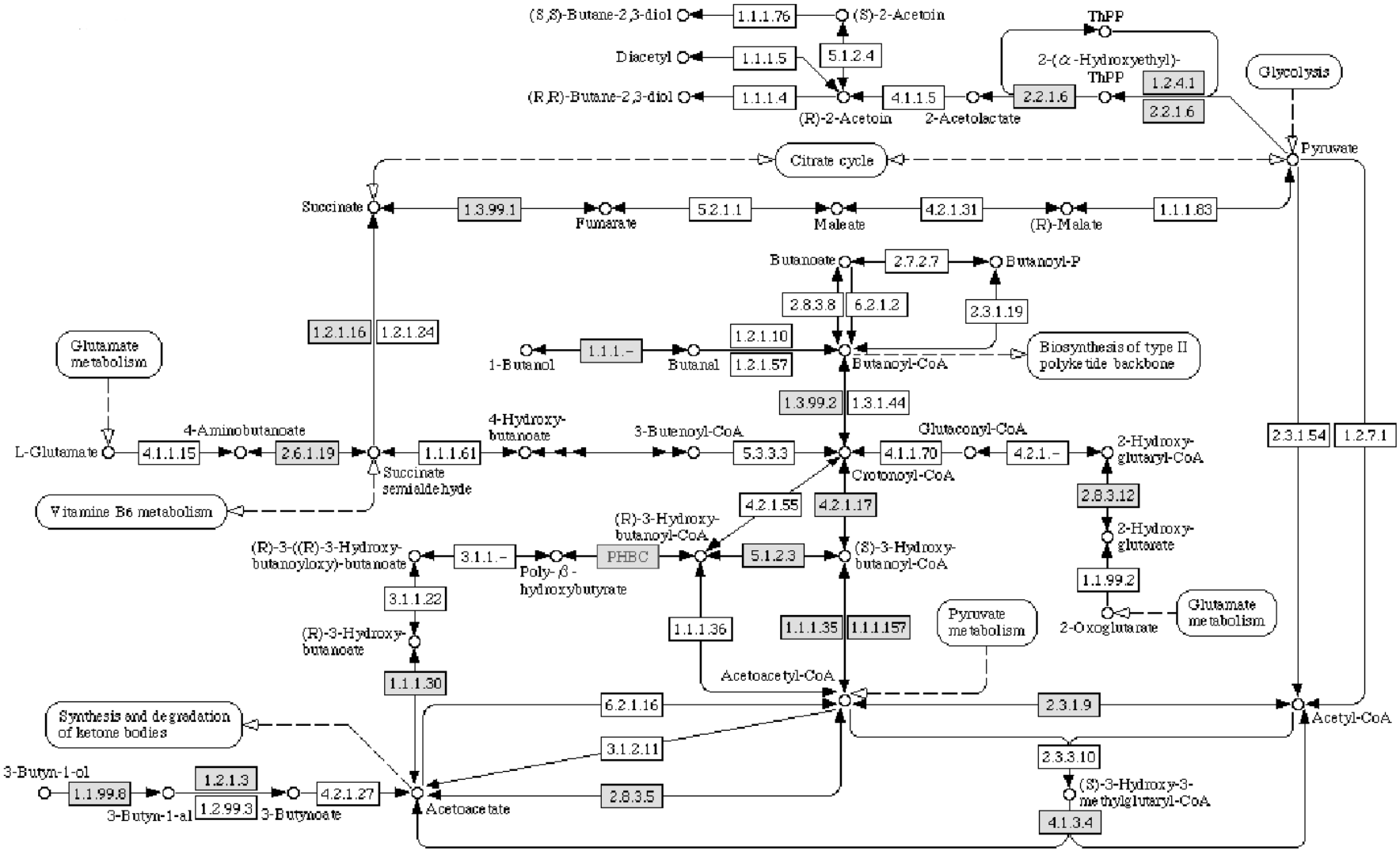
The genetic algorithm: General procedure

1. Generate a random initial population
2. Calculate each individual fitness $\sum_{j,t}[\xi_j(t) - f_j(x, k_i, t)]^2$ where $\xi_j(t)$ are a set of experimental observations and $f_j(k_i, t)$ are the corresponding simulated quantities (k_i represents the unknown parameters, t is time)
3. Individuals from the current population are selected to generate an offspring
4. Mutate each Chromosome in the offspring with a small variance
5. Go to step 2 with the new population, or stop if exit criteria are satisfied

Implementations of GA on Enea Grid

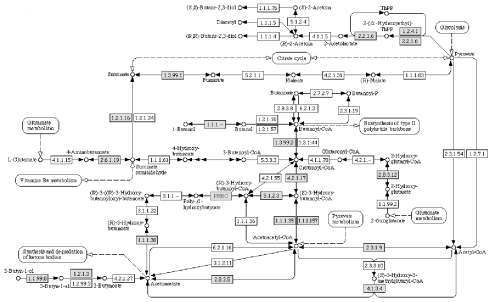
In order to efficiently parallelize GA we need to determine some self-consistent procedure units which can be run on a single CPU.

Algorithm Units: Model Simulation



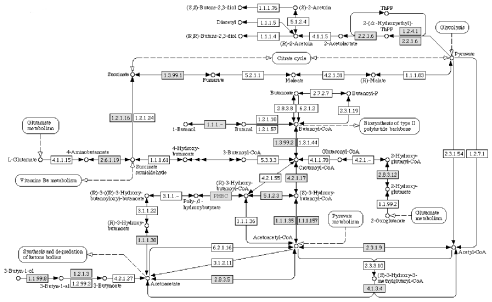
Algorithm Units: Model Simulation

Metabolic Model

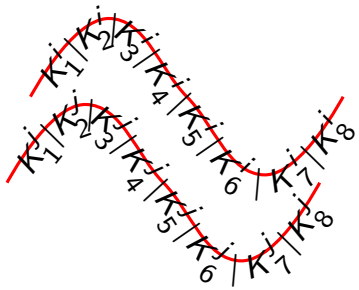


Algorithm Units: Model Simulation

Metabolic Model

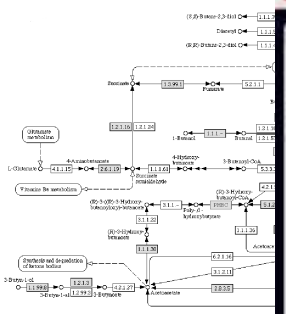


Parameters Guess

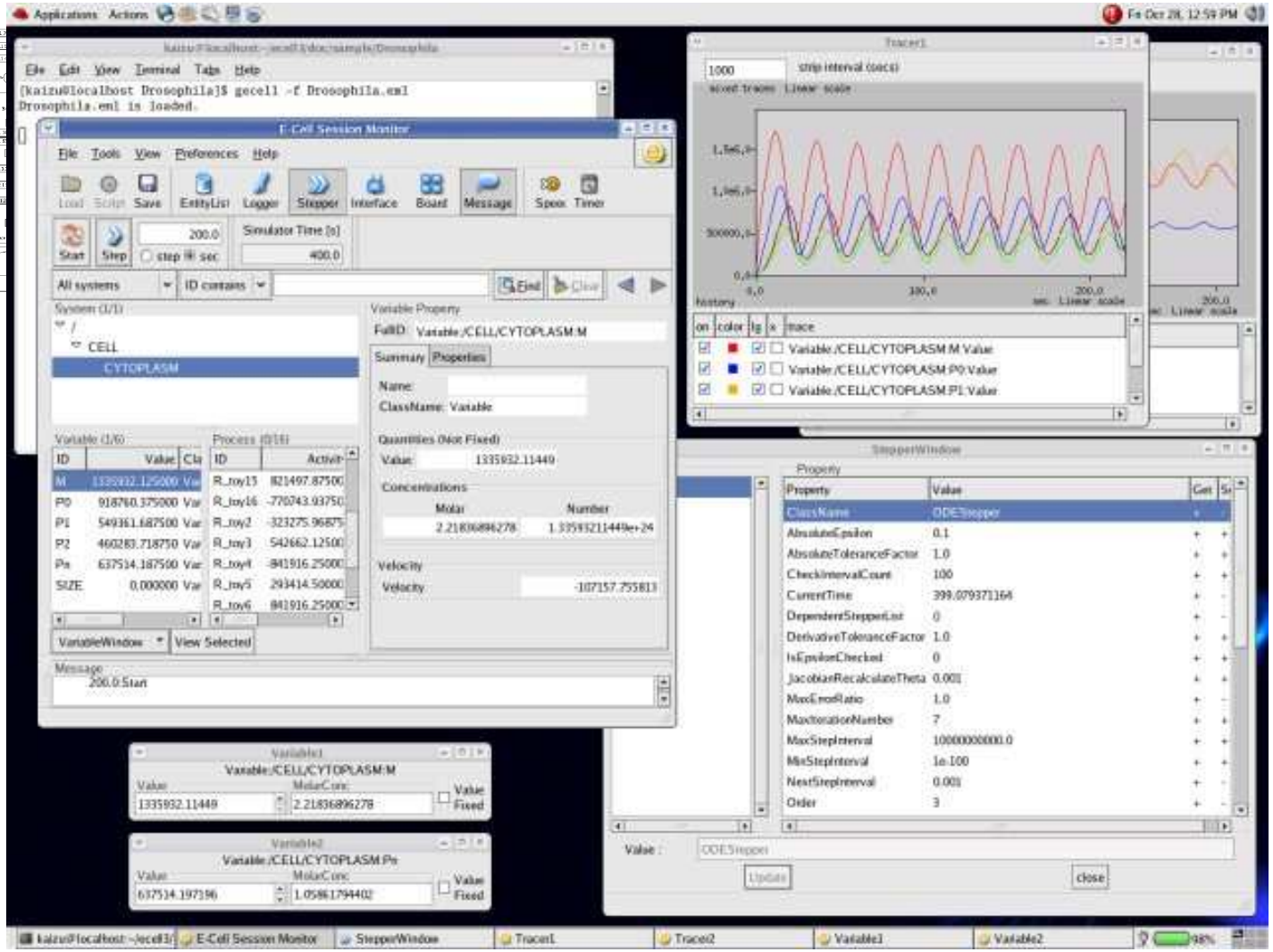


Algorithm Units: Model Simulation

Metabolic Model

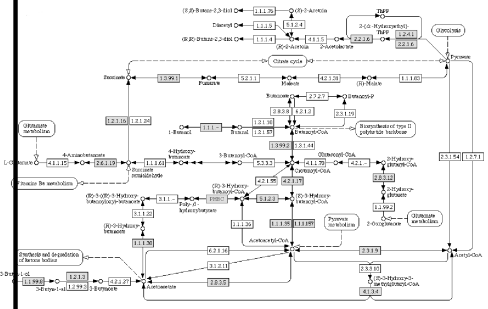


Parameters Guess

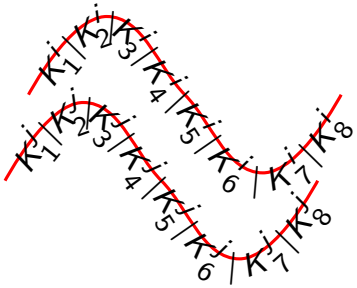


Algorithm Units: Model Simulation

Metabolic Model

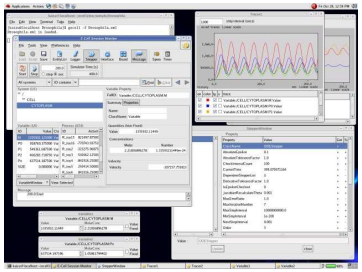


Parameters Guess



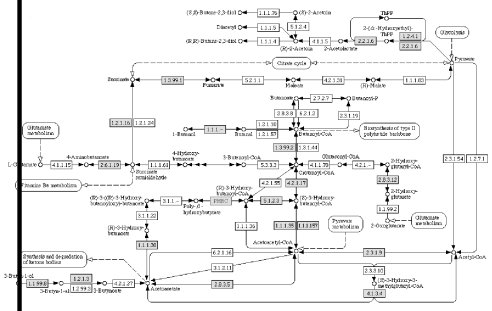
Simulator

E-Cell

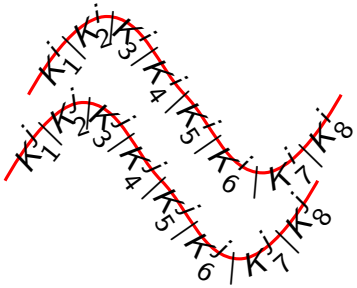


Algorithm Units: Model Simulation

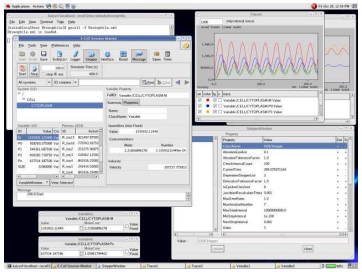
Metabolic Model

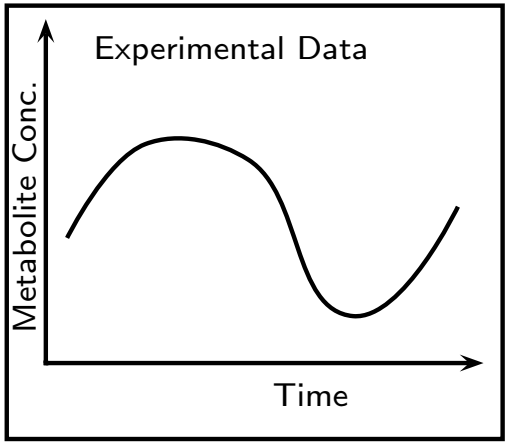


Parameters Guess

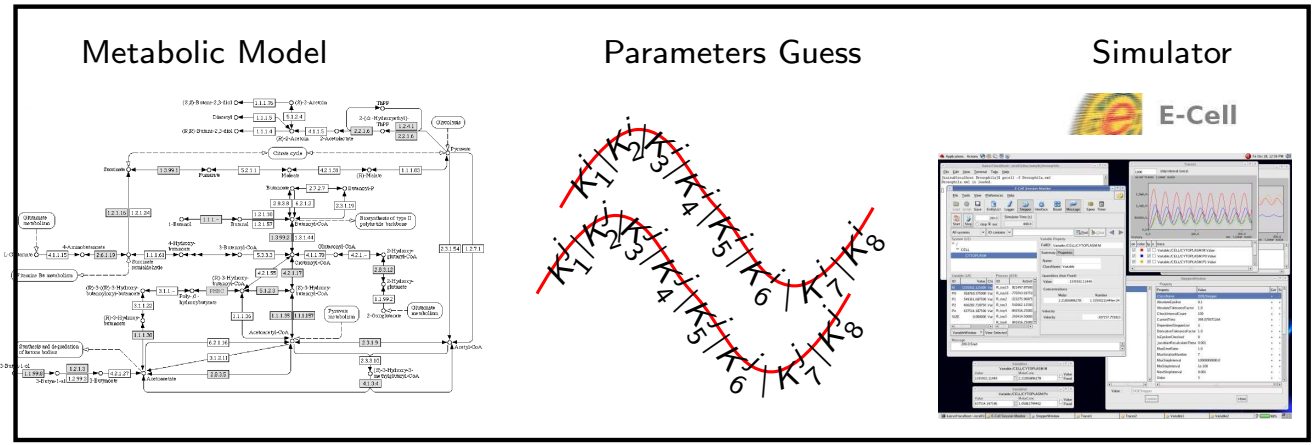


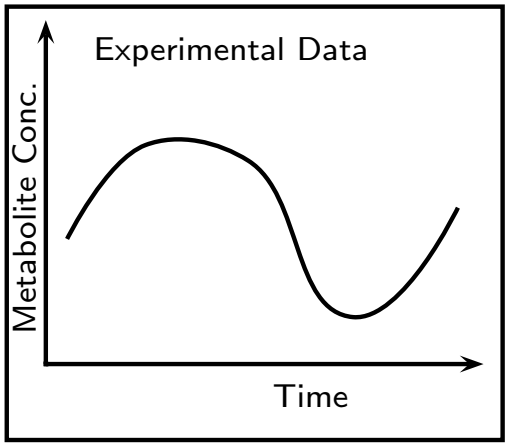
Simulator

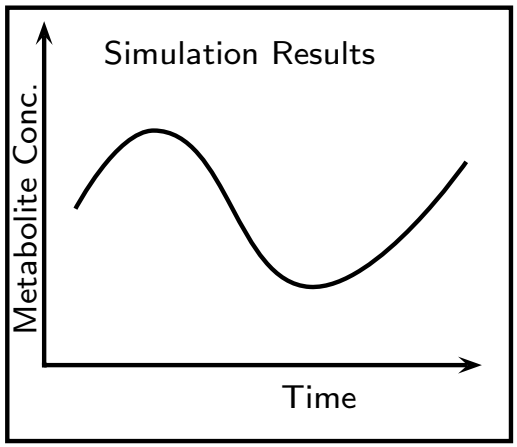




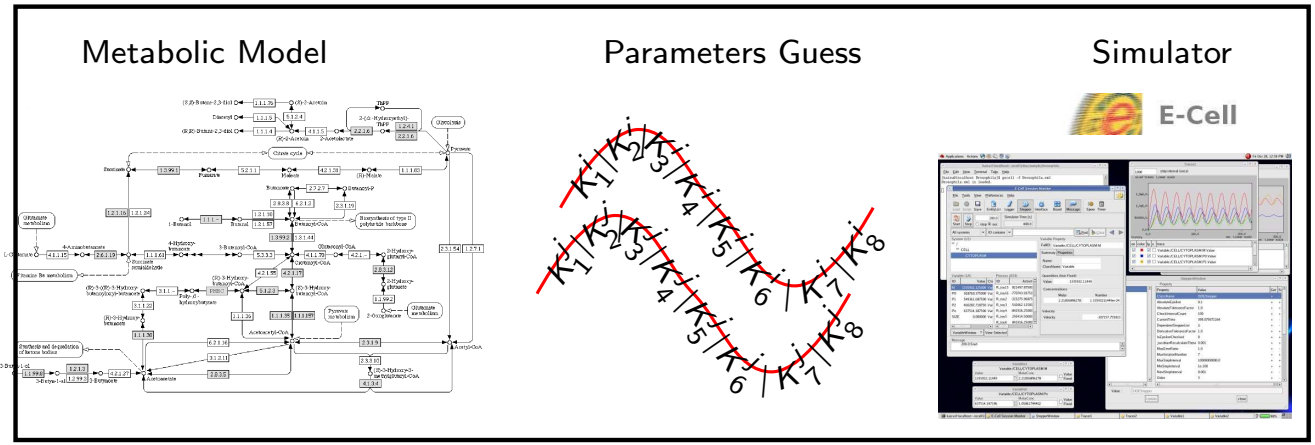
Algorithm Units: Model Simulation

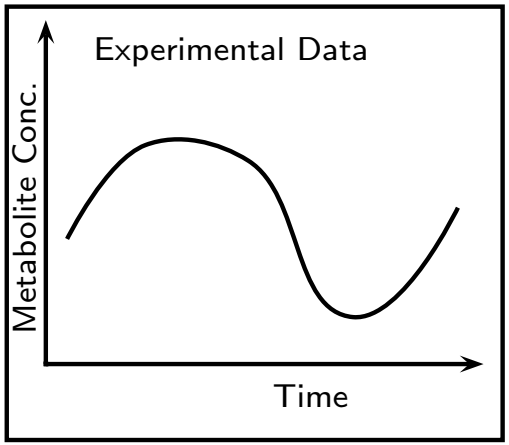


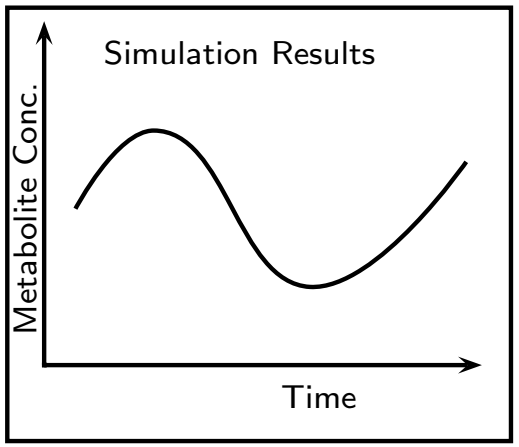


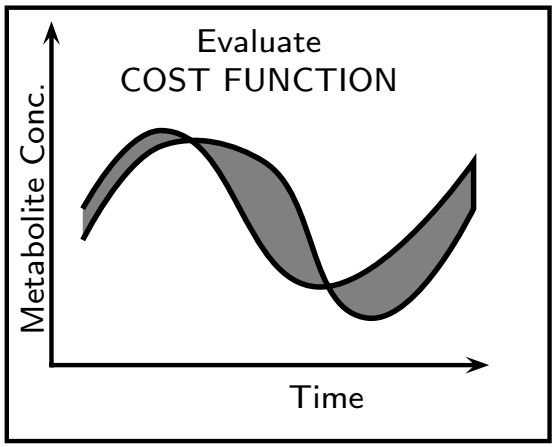


Algorithm Units: Model Simulation

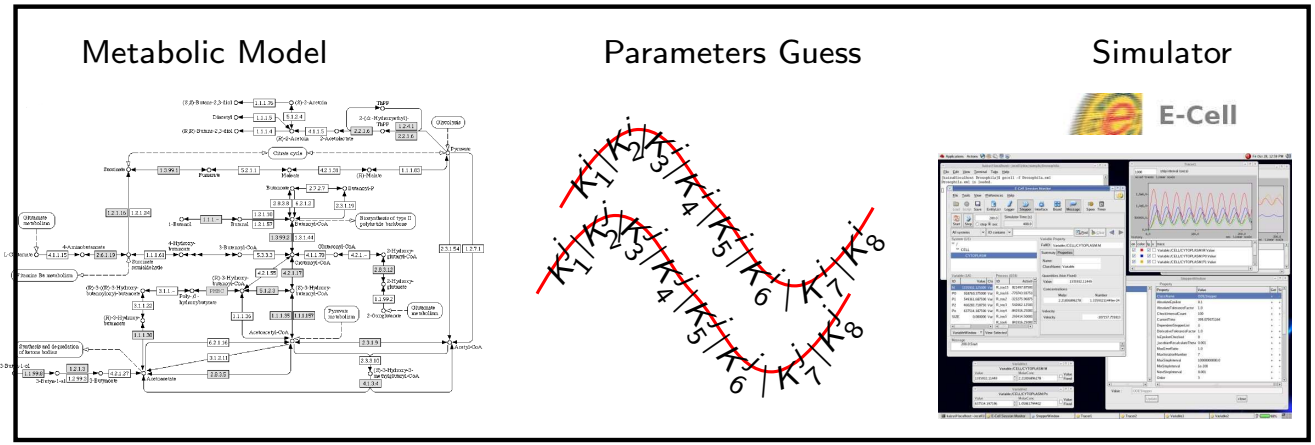


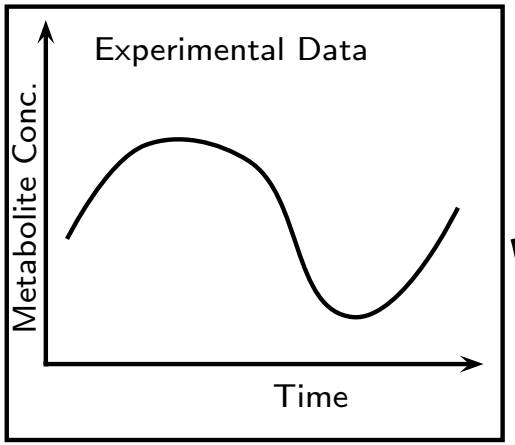


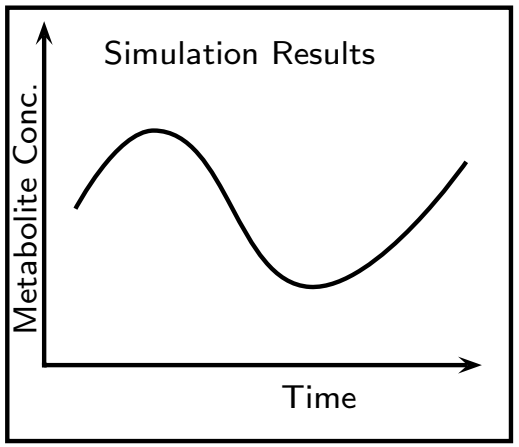


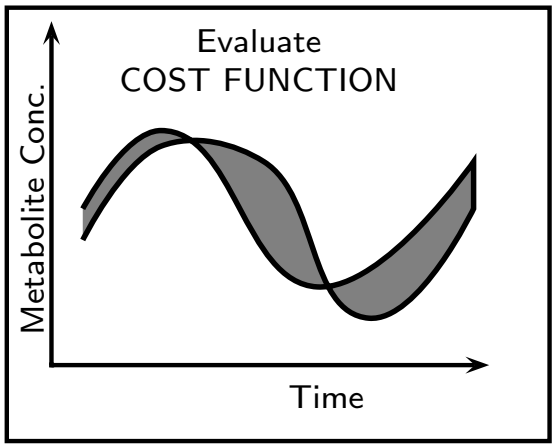


Algorithm Units: Model Simulation



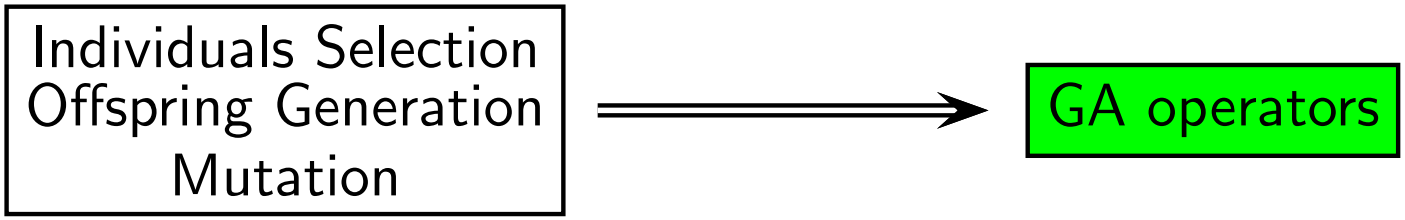






All these operators ⇒ ecell session

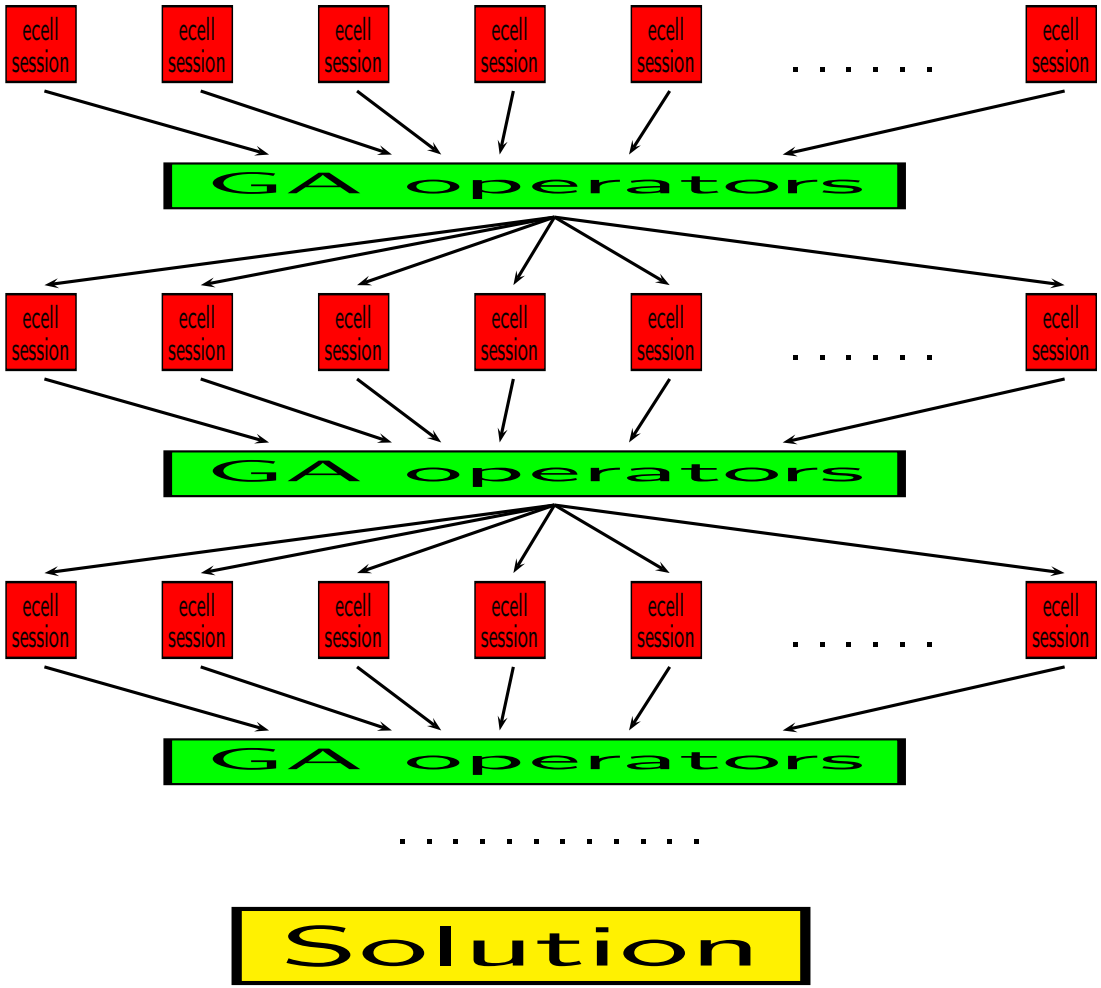
Algorithm Units: GA operators



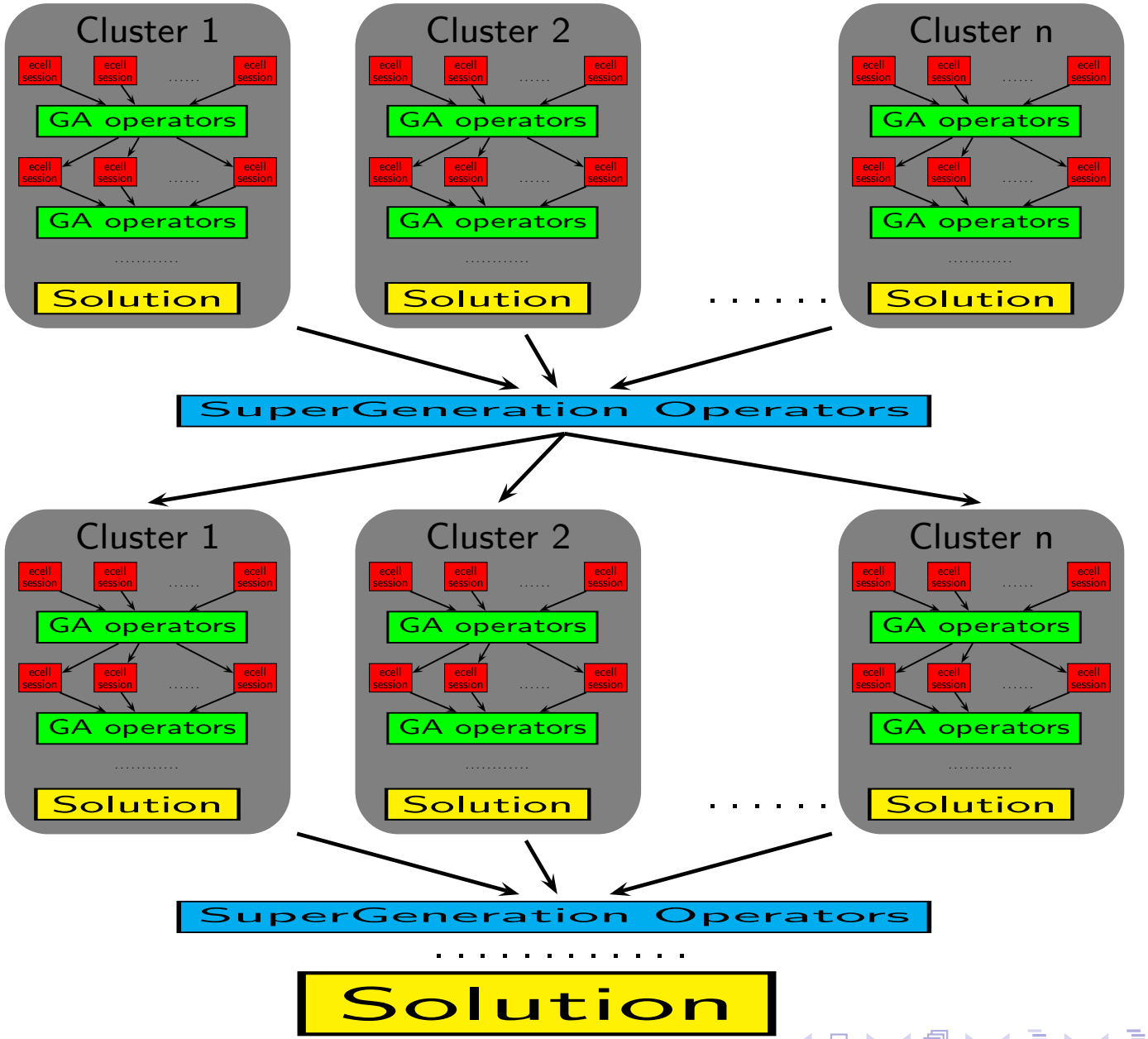
Advices

- first GA implementation (single stage of parallelism) is thought to run on a single cluster architecture (SCA), characterized by many computational nodes tightly interconnected through a low-latency, high-bandwidth network;
- second GA implementation (double stage of parallelism) is thought to run on a GRID (or multicluster) architecture), characterized by a number of SCAs linked through a Wide Area Network.

Parallel implementation (single stage)



Parallel implementation (double stage)



Application: Pseudomonas PHA production

Pseudomonas Corrugata is a Gram-negative bacteria, ubiquitous; it has been recognized as the causal agent of tomato pith necrosis. Several other interesting properties have been studied:

- some strains produce toxic peptides (LPD) which are thought to protect roots from fungi and phytophagous nematodes;
- some strains P. corrugata obtained from contaminated sites (benzene, toluene, ethylbenzene, m-xylene, p-xylene, o-xylene, naphthalene, phenol e p-cresol, fuel oils components and 4-chloroaniline) have shown degradation activity and thus suggest the use of P. corrugata in bioremediation.

Application: Pseudomonas PHA production

- P. corrugata can produce mcl-PHAs under some stress conditions. PHAs have the same characteristics as synthetic polyesters. In addition, they are biodegradable and thus have great potential for industrial and medical applications. P. corrugata can produce PHAs not only from pure sources(expensive) but also from renewable, low-cost sources such as biodiesel , glycerol, used cooking oils and soy molassa.

Model equations for PHA production (Khanna,Srivastava)

Legend:
$R(t) -> $ Biomass (includes Product), $P(t) -> $ Product(PHA),
$S_1(t) -> $ Nutrient (oleic acid), $S_2(t) -> $ Nitrogen Source.

$$\frac{1}{R}\frac{dR}{dt} = \mu \left[\frac{(S_1)^{n_1}}{(S_1)^{n_1} + (K_{S_1})^{n_1}}\right]\left[\frac{(S_2)^{n_2}}{(S_2)^{n_2} + (K_{S_2})^{n_2}}\right]\left[1 - \left(\frac{S_1}{S_{m_1}}\right)^{a_1}\right] \cdot \left[1 - \left(\frac{S_2}{S_{m_2}}\right)^{a_2}\right]$$

$$\frac{1}{R}\frac{dP}{dt} = K_1\mu + K_2$$

$$\frac{1}{R}\frac{dS_1}{dt} = -(\alpha\mu + \gamma)$$

$$\frac{1}{R}\frac{dS_2}{dt} = -(Y_{R/S_2}\mu + M_{S_2})$$

Model equations: details

where:

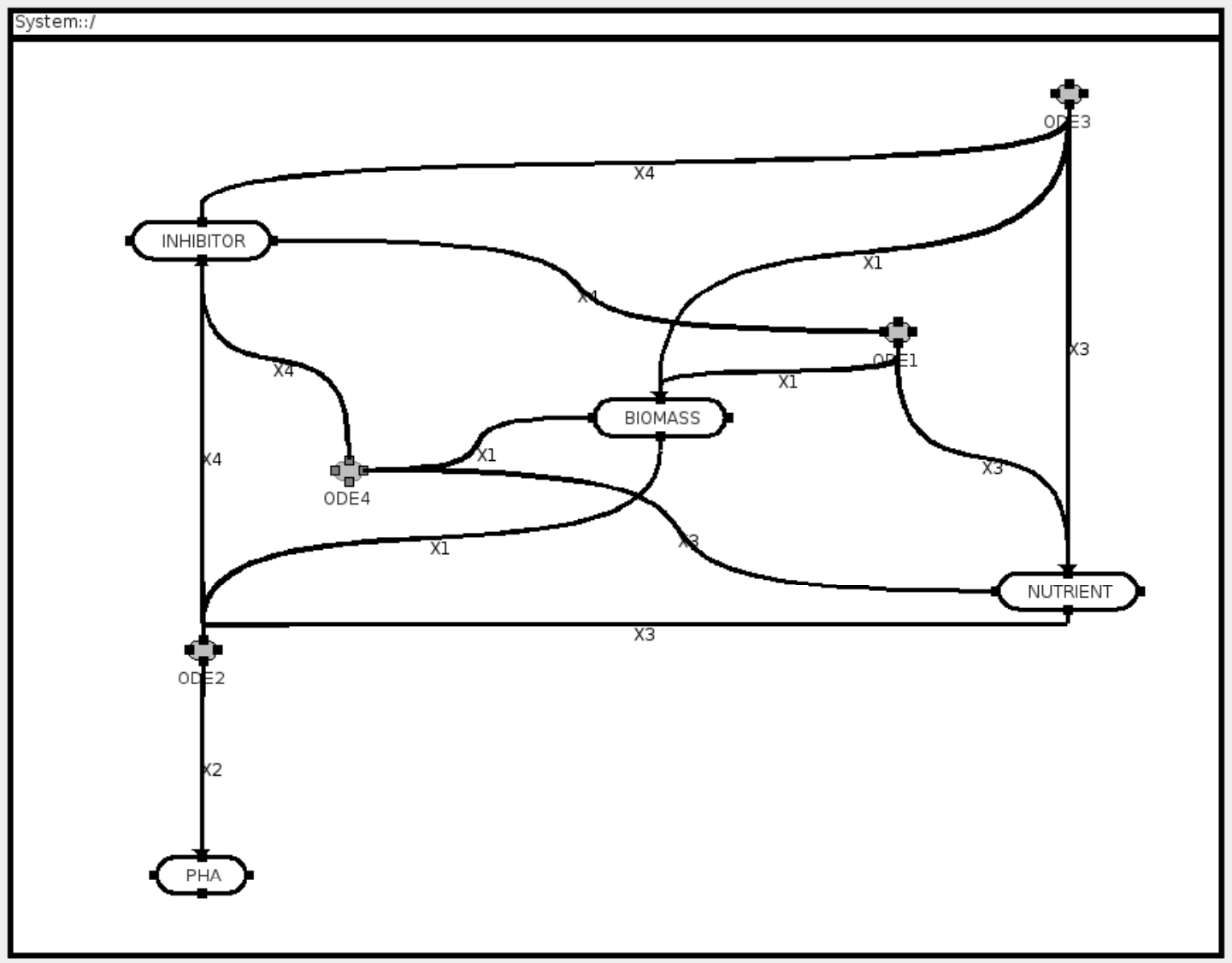
$$\mu\left[\frac{(S_i)^{n_i}}{(S_i)^{n_i}+(K_{S_i})^{n_i}}\right]->$$

A contribute to the specific growth rate of the micro-organism is expressed as a function of limiting nutrient (S_i) concentration by a sigmoidal relationship where K_S is the saturation constant.

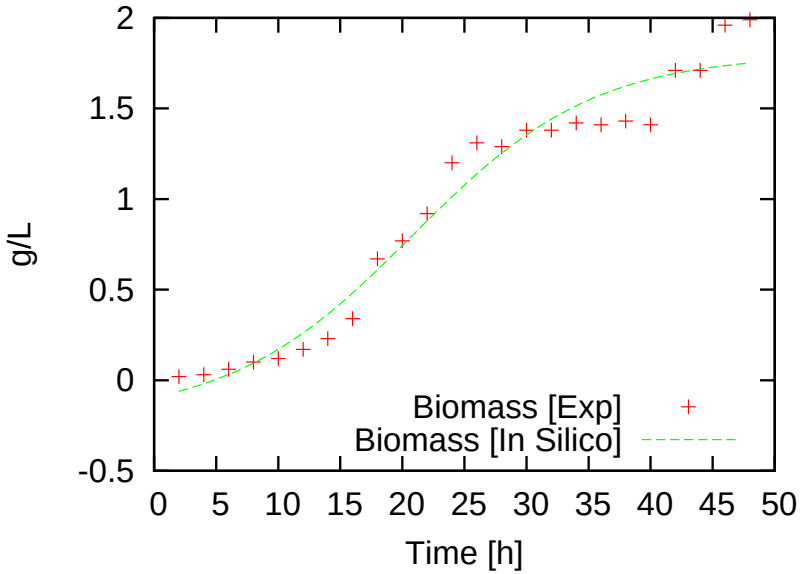
$$\left[1-\left(\frac{S_i}{S_{m_i}}\right)^{a_i}\right]->$$

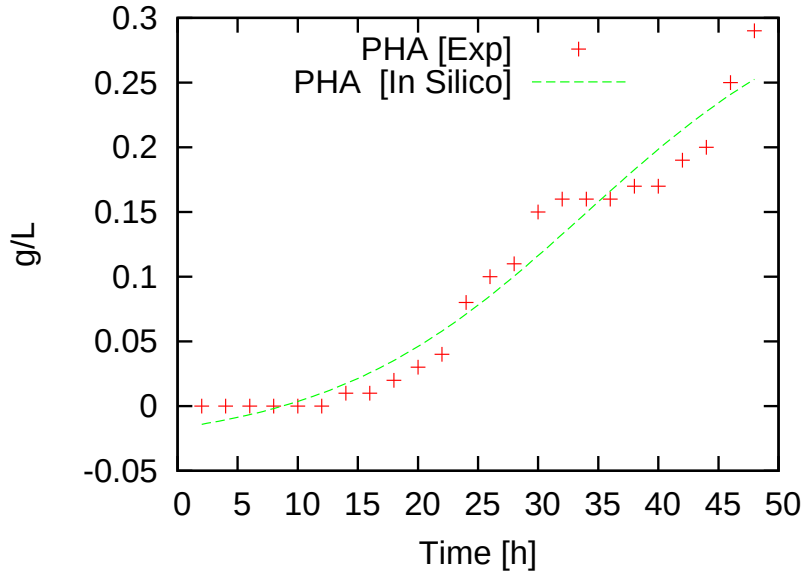
Contribute from substrate inhibition kinetics.

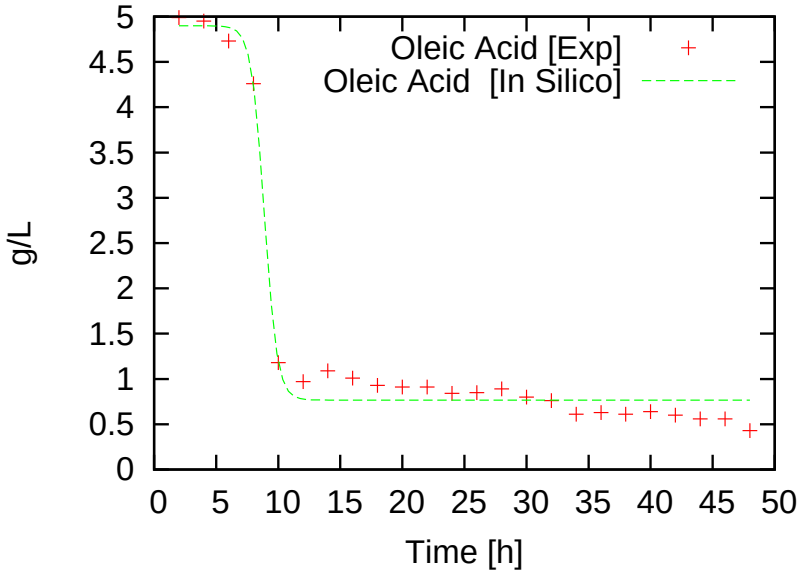
Network building with Ecell

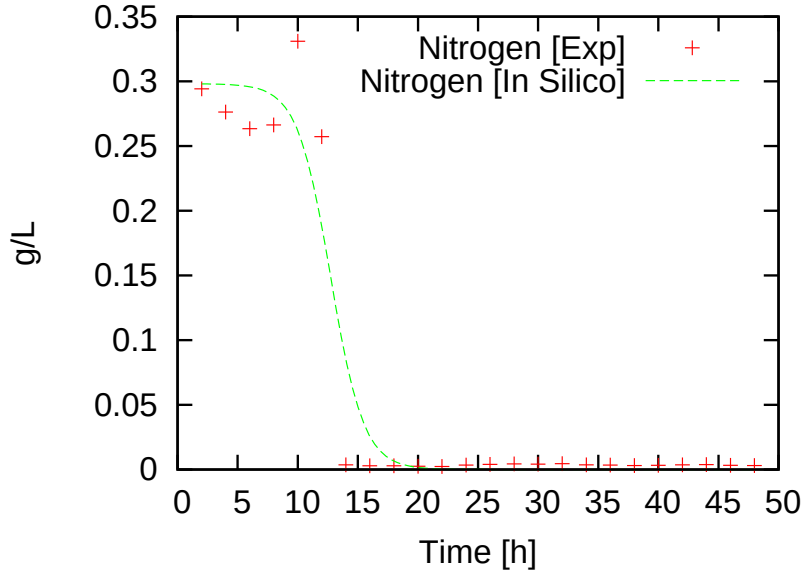


Model fitting (in progress...)









Conclusions and Perspectives

- a single and a double stage parallelization implementation of GA have been implemented on Enea Grid; they offer a sizeable time saving and advantages for the solution-space exploration
- the code prepared to send jobs on a grid using LSF scheduler is going to be included in the next Ecell releases.
- P.Corrugata DNA sequencing and expression profiling is planned in order to build a much more exhaustive model of the metabolism(network inference, FBA. . .).

References I

- S. Kikuchi, D. Tominaga, M. Arita,K. Takahashi and M. Tomita, Dynamic modeling of genetic networks using genetic algorithm and S-system. BIOINFORMATICS, Vol. 19 no. 5 (2003), 643-650
- S. Tsutsui, M. Yamamura and Higuchi,T. Multi-parent recombination with simplex crossover in real coded genetic algorithms. Proceedings of the Genetic and Evolutionary Computation Conference, (1999), 657-664.
- S. Khanna and A. K. Srivastava, A Simple Structured Mathematical Model for Biopolymer (PHB) Production. Biotechnology Progress, Vol. 21 no. 3 (2005), 830-838

Acknowledgments

Frascati Enea-INFO group

- Giovanni Bracco
- Salvatore Podda
- Carlo Sció