Systems Biology for the development of microbial cell factories

Eugénio Campos Ferreira

IBB – Institute for Biotechnology and Bioengineering
Centre of Biological Engineering, BioPSE group
Universidade do Minho

Universidade do Minho

Systems Biology approaches for modelling, optimization, and control of microbial cell factories

- Cellular Models for Metabolic Engineering: gene networks
- Inference of Biological Networks
  - From Genome-scale metabolic models
  - From experimental data
  - From literature data mining
- *In Silico* Metabolic Engineering Platforms: Optimization of Microbial strains – OptFlux tool

- Metabolic Engineering can gain major benefits from the systems biology approach
- Systems Biology does not investigate individual cellular components at a time, but the behaviour and relationships of all of the elements in a particular biological system while it is functioning

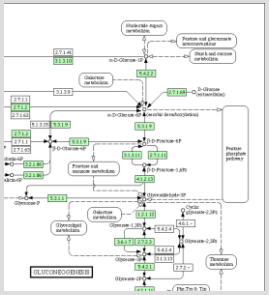
?

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes.

- In order to economically produce desired compounds like antibiotics, therapeutic proteins, food and feed ingredients, fuels, vitamins and other chemicals from microbial cell factories it is generally necessary to retrofit the metabolism
- Metabolic engineering envisages the introduction of directed genetic modifications leading to desirable metabolic phenotypes, as opposed to traditionally used random mutagenesis and screening

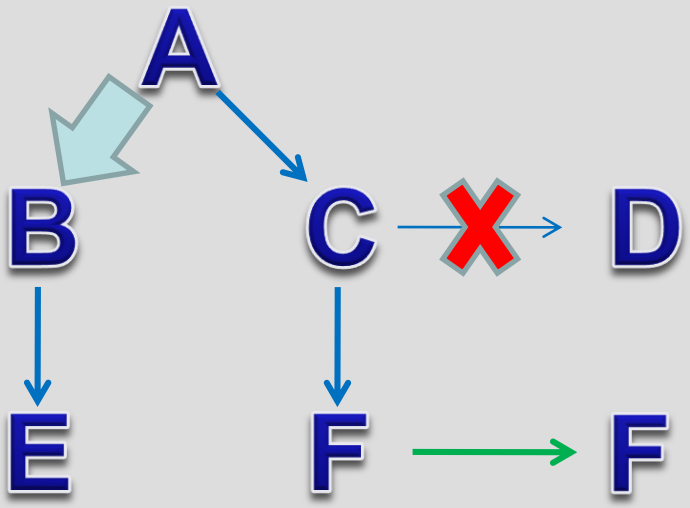
Cell factory

Genome

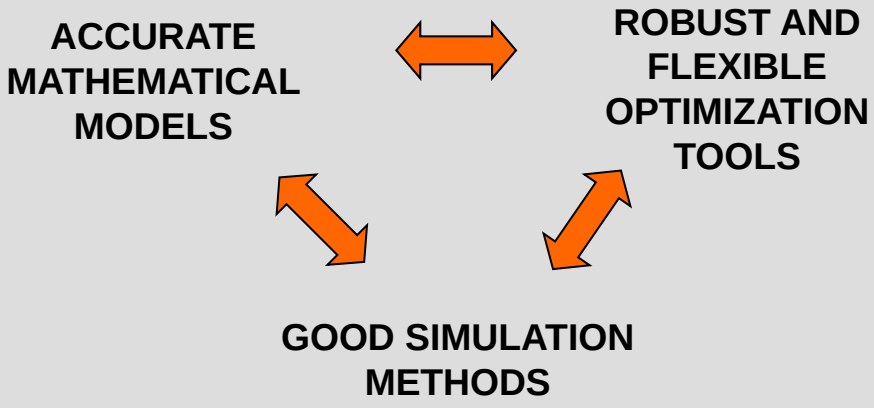


Metabolism / Phenotype

- Gene Deletion
- Gene Addition
- Gene Under/Overexpression
- Manipulation of environmental conditions

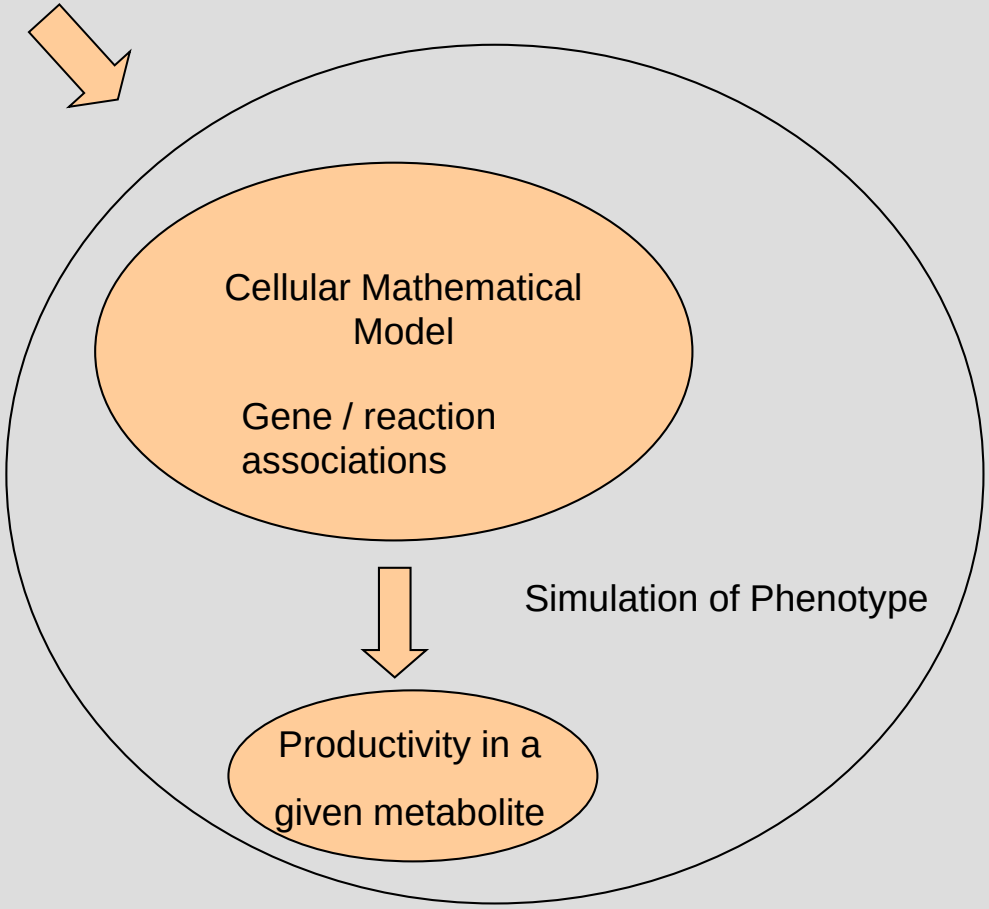


- In metabolic engineering problems, it is often difficult to identify a priori which genetic manipulations will originate a given desired phenotype
- In order to rationally design production strains with enhanced capabilities, it is essential to have:



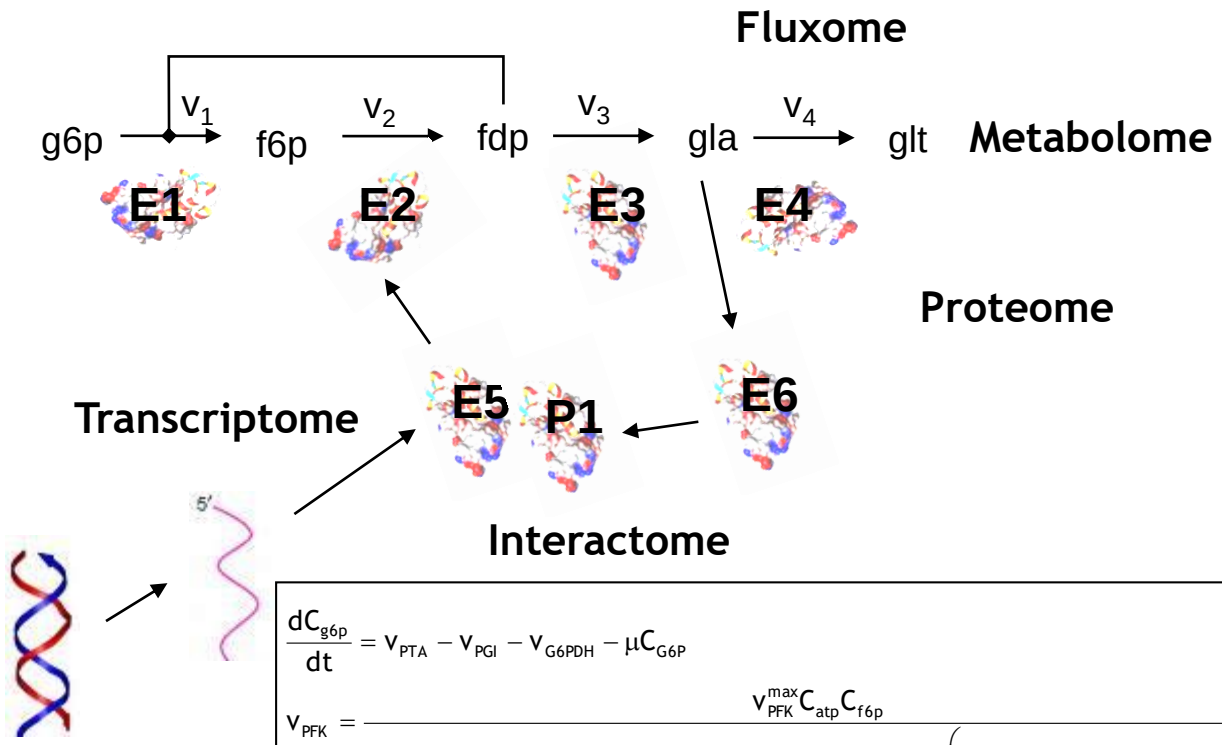
Optimization Algorithm

- Genome modifications and manipulation of environmental conditions
- Objective function – productivity in a given metabolite



Models should comprise different levels of information:

- reactions stoichiometry
- reactions kinetics
- regulatory information

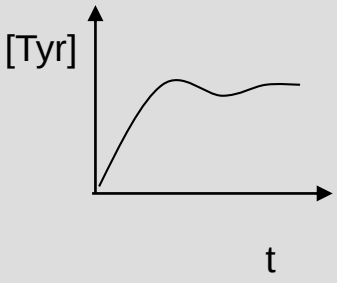


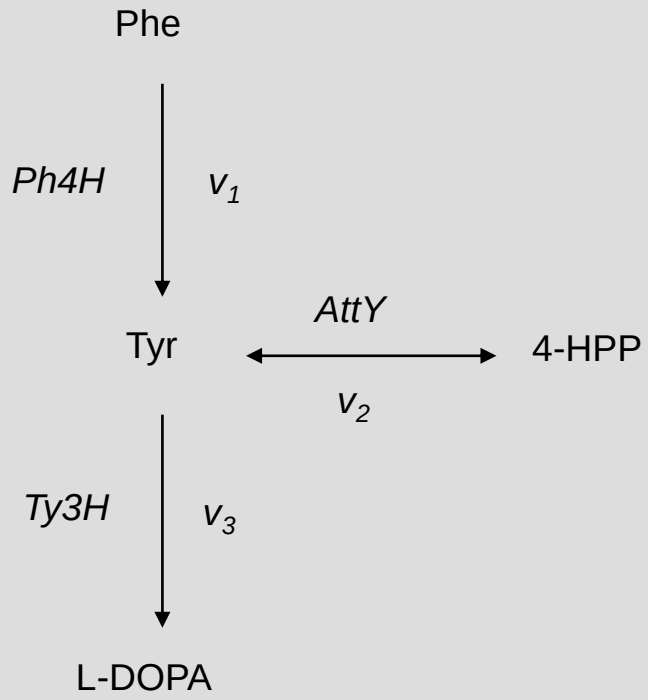
$$\frac{dC_{g6p}}{dt} = v_{PTA} - v_{PGI} - v_{G6PDH} - \mu C_{G6P}$$

$$v_{PFK} = \frac{v_{PFK}^{max} C_{atp} C_{f6p}}{\left(C_{atp} + K_{PFK,atp,s}\left(1 + \frac{C_{adp}}{K_{PFK,adp,s}}\right)\right) + \left(C_{f6p} + K_{PFK,f6p,s}\frac{A}{B}\right)\left(1 + \frac{L_{PFK}}{\left(1 + C_{f6p}\frac{B}{K_{PFK,f6p,s}^{A}}\right)^{nPFK}}\right)}$$

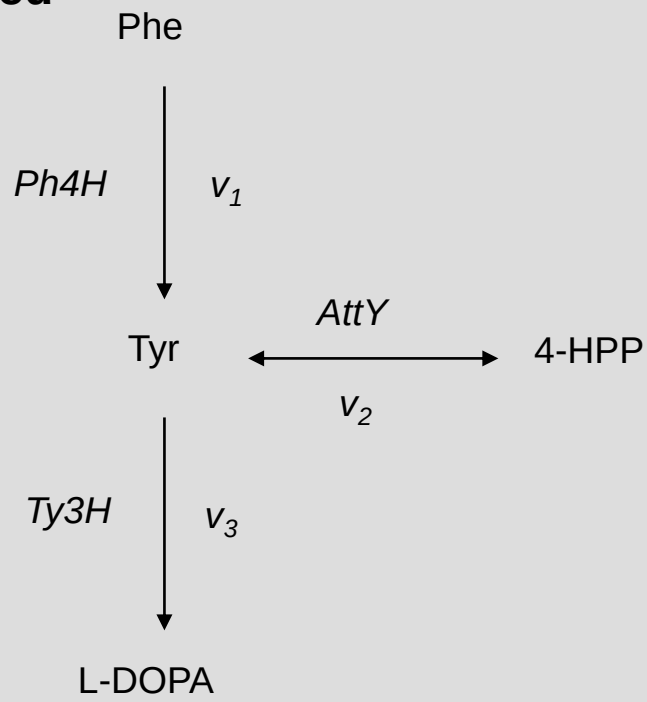
- There are several ways to represents the **chemical conversions** associated to metabolic reactions
- Kinetic or mechanistic models use **deterministic** differential equations relating the amount of reactants with the quantity of products, according to a given reaction rate and other parameters
- Given an initial state, the trajectory of metabolite can be obtained by numerical simulation

$$\frac{d[Tyr]}{dt} = \ldots$$





- A (pseudo)**steady state condition is usually assumed** inside the cell



$$\frac{d[Tyr]}{dt} = v_1 - v_2 - v_3$$

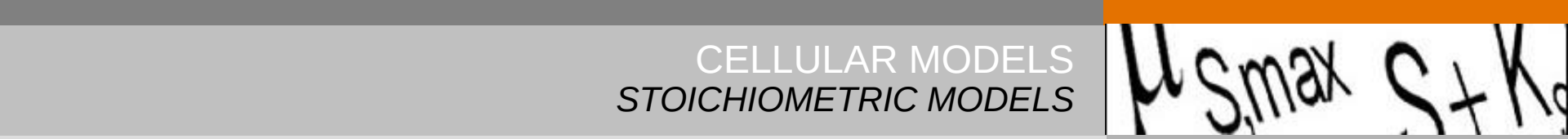
$$v_1 - v_2 - v_3 \approx 0$$

This procedure is repeated for *all* considered metabolites and will originate the so-called stoichiometric model

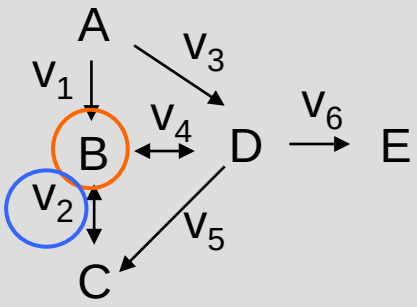
The result is a Linear Equations system described by stoichiometric matrix *S.*

$$S\,v = 0$$

$$\beta_j \le v_j \le \alpha_j$$



For an identified reaction set:

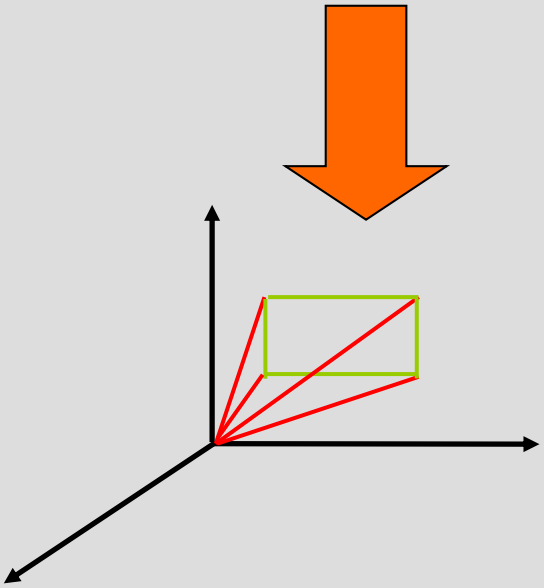


$$S = \begin{array}{c} \begin{matrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & \dots & \textit{flux } n \end{matrix} \\ \begin{bmatrix} -1 & 0 & -1 & 0 & 0 & 0 & \dots & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & \dots & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 & \dots & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{l} A \\ B \\ C \\ D \\ E \\ \dots \\ \textit{metabolite } m \end{array}$$

$$0 \le v_1 \le +\infty$$
$$-\infty \le v_2 \le +\infty$$
$$0 \le v_3 \le +\infty$$
$$-\infty \le v_4 \le +\infty$$
$$0 \le v_5 \le +\infty$$
$$0 \le v_6 \le +\infty$$
$$\dots$$
$$\beta \le v_n \le \alpha$$

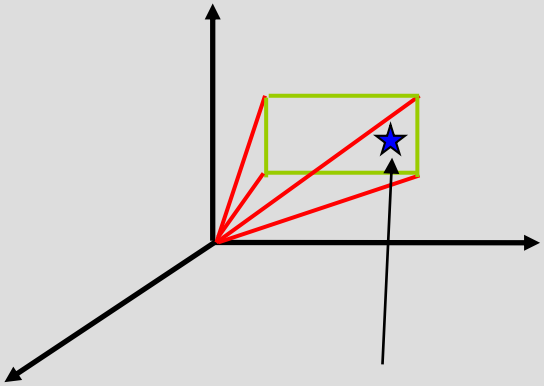
- Stoichiometric models typically have more fluxes than balanced metabolites.
- The equation system, S • v = 0, then has more variables than equations. This is a so-called under-determined equation system with infinitely many solutions:

$$S\,v = 0$$
$$\beta_j \le v_j \le \alpha_j$$

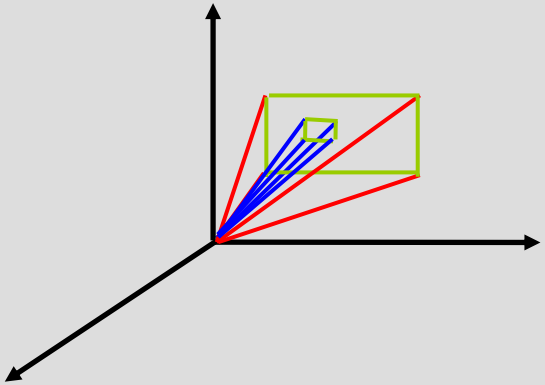


Hyper-cone containing ALL solutions as defined by the stoichiometry of the organism

- How can we reduce the cone of solutions?
  - By optimizing a given criterion – FBA, MoMA, ROOM…
  - By the introduction of <u>regulatory information (ex: Gene Networks)</u>



Particular Solution



Reduction of the solutions space

FBA: Flux Balance Analysis
ROOM: Regulatory On/Off *Minimization*
MoMA: Minimization of Metabolic Adjustment

- The idea is to find *one* solution to the under-determined system

  S • v = 0 by **optimization of a given criterion**.

Maximize:

$$z = c^T v = v_{prod}$$

c = row vector containing weights specifying what combination of fluxes to optimize

Subject to:

$$S\,v = 0$$

Constraints from stoichiometry

$$\beta_j \le v_j \le \alpha_j$$

LINEAR PROGRAMMING PROBLEM!

- A vector containing the values of each individual metabolic flux is obtained

BUT WHAT SHOULD WE OPTIMIZE?

- Studies in several organisms demonstrated that their metabolic network has evolved for optimization of the specific growth rate under several carbon source limiting conditions

- Thus, for simulating cellular behaviour, the most common objective function is the maximization of biomass production (BPCY: Biomass-Product Coupled Yield)

Ibarra et al (2002), Nature

- For mutants and organisms grown on unusual carbon sources the hypothesis of optimal growth is not always real

- Such strains may undergo minimal redistribution of fluxes with respect to the wild-type strains (MoMA)

- The problem is the search of a flux set (x) that has a minimal distance from the wild-type flux vector (w) obtained with FBA.

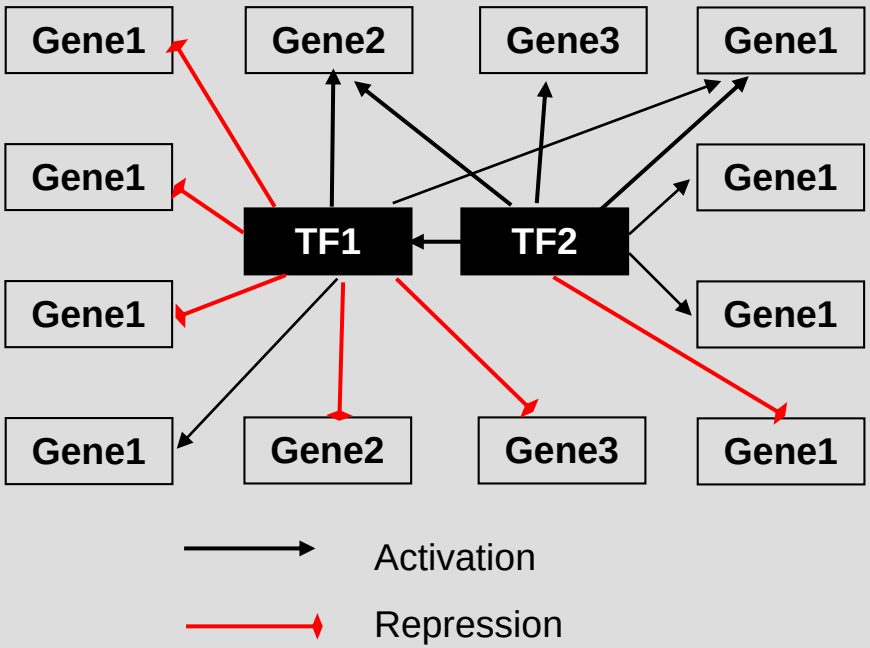
- The distance between w and x is given by the Euclidean distance:

$$D(w,x) = \sqrt{\sum_{i=1}^{N} (w_i - x_i)^2}$$

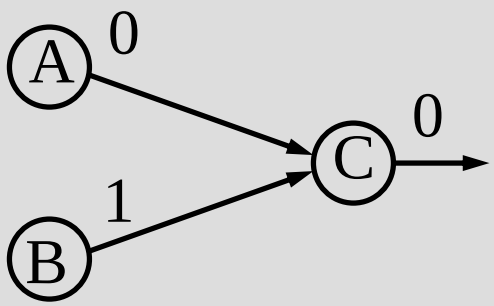
- The minimization of that distance can be formulated as a QP problem

Segre *et al.* (2002), PNAS

- Gene Regulatory Networks represent regulatory elements and their interactions
- A regulatory network will direct the activation or repression of a set of genes in response to a specific environmental stimulus, like O_2 or pH



- The simulation of a genetic network can be performed in several ways
- The simplest one is to consider Boolean Networks, where ON/OFF gene states are assumed.

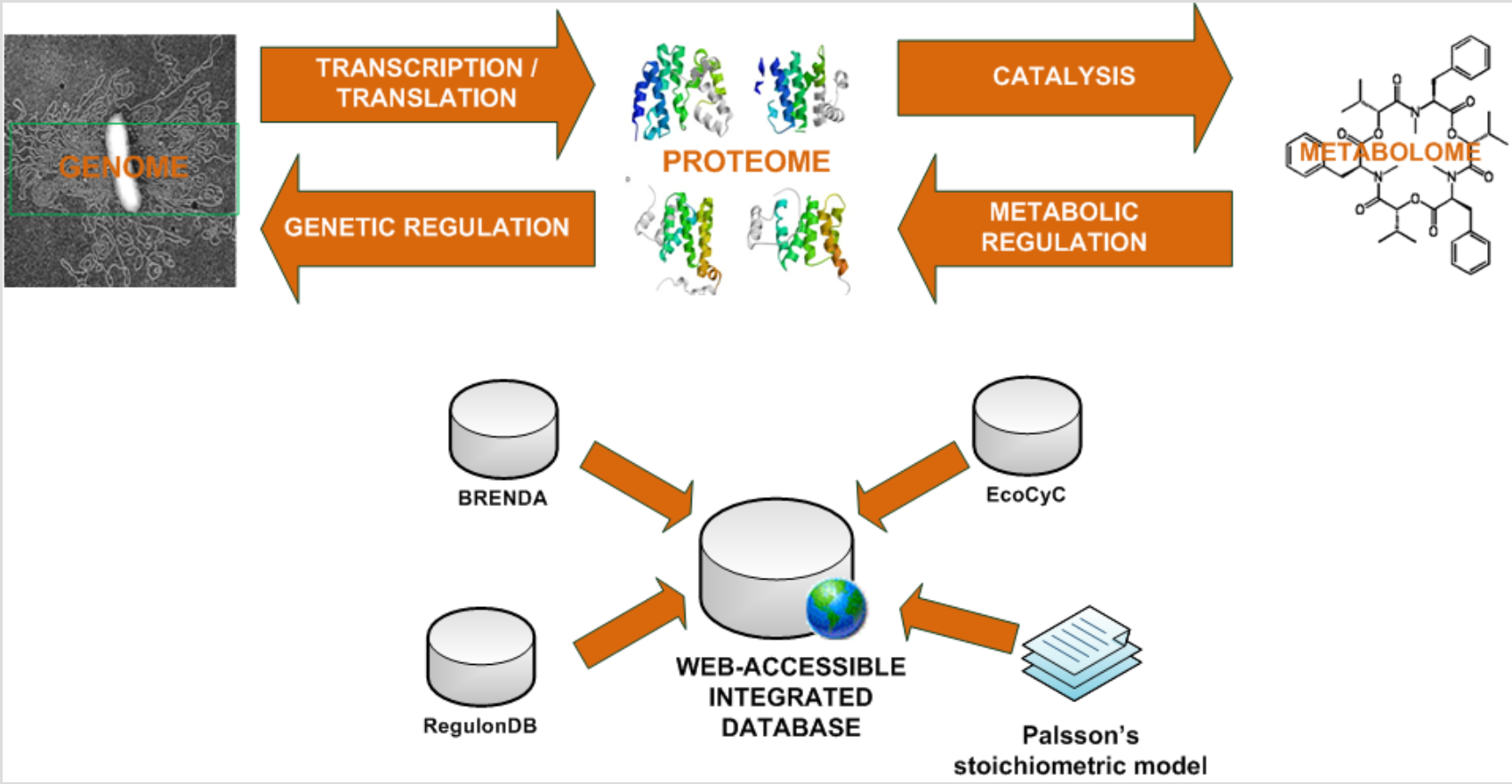


$$C = A \text{ AND } B$$

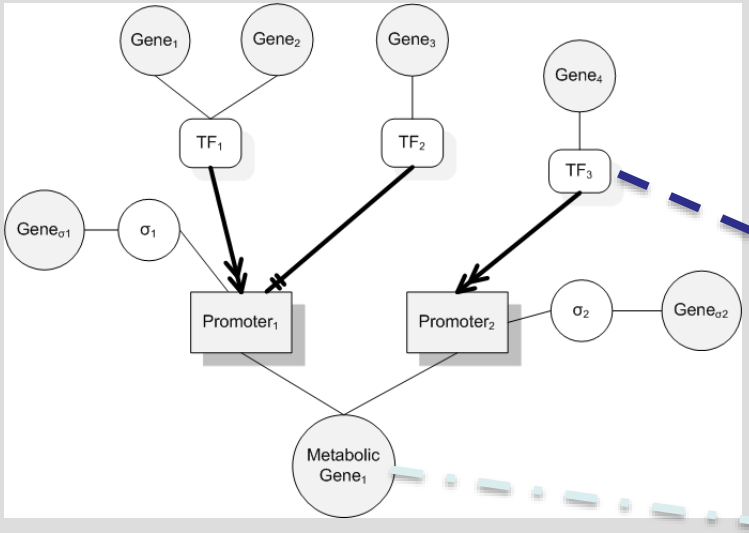
- This approach:
  - Allows analysis at the network-level
  - Provides useful insights in network dynamics

CELLULAR MODELS
GENE NETWORKS

WORK IN PROGRESS:
Reconstruction of *E. coli* regulatory network and integration with stoichiometric model



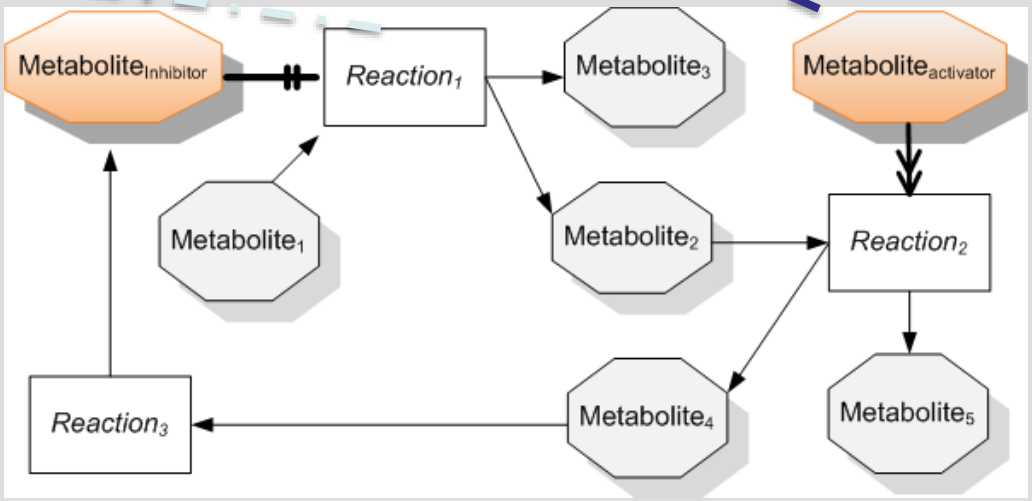
WORK IN PROGRESS: Reconstruction of *E. coli* regulatory network and integration with stoichiometric model



Metabolic Transcriptional Regulation

Inhibition / Activation

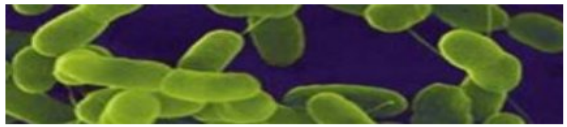
Genetic Regulation



CELLULAR MODELS
GENE NETWORKS

WORK IN PROGRESS: Information System of Biochemical and Regulatory Data on *Escherichia coli*

Research Activities on Bioinformatics



Home | Projects | Collaborations | Publications | Resources | Members | Contacts

Welcome analia! — Logout

Regulatory Models

This is our datatabase of regulatory models for Escherichia coli. Its primary goal is to assist researchers in their daily activities, providing the means to manage the catalogue as well as advanced analysis features. Database management covers the insertion, update and removal of information at each entity level, i.e., genes, transcription factors, pathways, reactions, regulatory events and so on. While providing some means of consultation and search these options are not meant for analysis purposes. Research needs in terms of data crossover and analysis led to the construction of more sophisticated and specific reports.

Database Management

- External Factors
- Genes
- Metabolites
- Pathways
- Reactions
- Regulatory Events
- Ri functions
- Sigma factors
- Transcription Factors

Reports

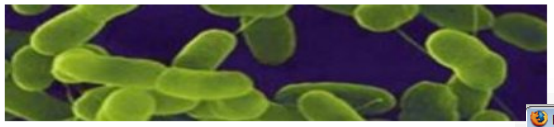
- Encoding Genes
- Metabolites
- Reactions
- Transcription Factors

© All rights reserved.

CELLULAR MODELS
GENE NETWORKS

WORK IN PROGRESS: Information System of Biochemical and Regulatory Data on *Escherichia coli*

Research Activities on Bioinformatics



Home | Projects | Collaborations | Publications | Resources | Members | Contacts

Welcome analia! — Logout

Regulatory Models

Database Management

- External Factors
- Genes
- Metabolites
- Pathways
- Reactions
- Regulatory Events
- Ri functions
- Sigma factors
- Transcription Factors

Reports

- Encoding Genes
- Metabolites
- Reactions
- Transcription Factors

Reactions

New Record | Regulated Gene | Equation | Export TXT | Print | Refresh |

Search

2008-5-15 (14:08)

Pages

Records: 0 - 20
Total of Records: 931

Pathway	Reaction	ECnumber	Equation
Alanine and aspartate metabolism	ALAR	EC-5.1.1.1	[c]ala-L <==> ala-D
Alanine and aspartate metabolism	ALARi	EC-5.1.1.1	[c]ala-L --> ala-D
Alanine and aspartate metabolism	ALATA_L	EC-2.6.1.2	[c]akg + ala-L <==> glu-L + pyr
Alanine and aspartate metabolism	ASNN	EC-3.5.1.1	[c]asn-L + h2o --> asp-L + nh4
Alanine and aspartate metabolism	ASNS1	EC-6.3.5.4	[c]asp-L + atp + gln-L + h2o --> amp + asn-L + glu-L + h + ppi
Alanine and aspartate metabolism	ASNS2	EC-6.3.1.1	[c]asp-L + atp + nh4 --> amp + asn-L + h + ppi
Alanine and aspartate metabolism	ASPT	EC-4.3.1.1	[c]asp-L --> fum + nh4
Alanine and aspartate metabolism	ASPTA	EC-2.6.1.1	[c]akg + asp-L <==> glu-L + oaa
Alanine and aspartate metabolism	DAAD	EC-1.4.99.1	[c]ala-D + fad + h2o --> fadh2 + nh4 + pyr

http://sysbio.di.uminho.pt - Record Visualisation - Mozilla Firefox

Visualising record at table reactions

Update | Remove | Print | Close

Pathway: Alanine and aspartate metabolism
Reaction: ALAR
ECnumber: EC-5.1.1.1 BRENDA
Equation: [c]ala-L <==> ala-D

Reagents: L-Alanine

Products: D-Alanine

Reversible: 1
Boolean rule: CRP AND NOT Lrp
Comment: Activity of alanine racemase in E. coli is due to two distinct gene products. One alanine racemase (Alr) is constitutive; it is encode by alr. The other DadX is induced by D- or L-alanine and repressed by glucose; it is , and is encoded by the dadX. Alr is less abundant than DadX. Boolean rule de acordo com as posições de ligação do CRP e Lrp (ecocyc)

Encoding Genes:

- alr - b4053 EcoCyc
- dadX - b1190 EcoCyc

Regulatory Genes:

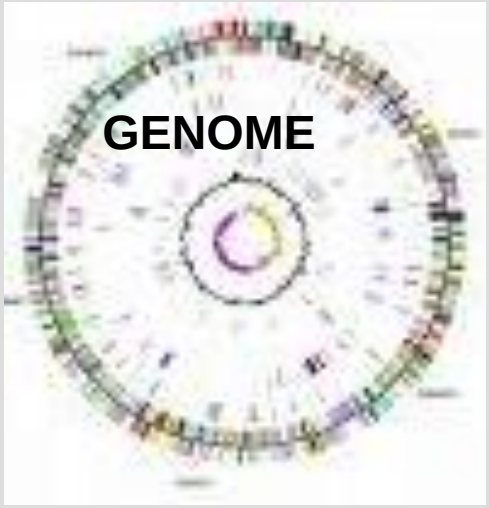
- crp - b3357 EcoCyc
- lrp - b0889 EcoCyc

2008-5-15 (14:09)

Done

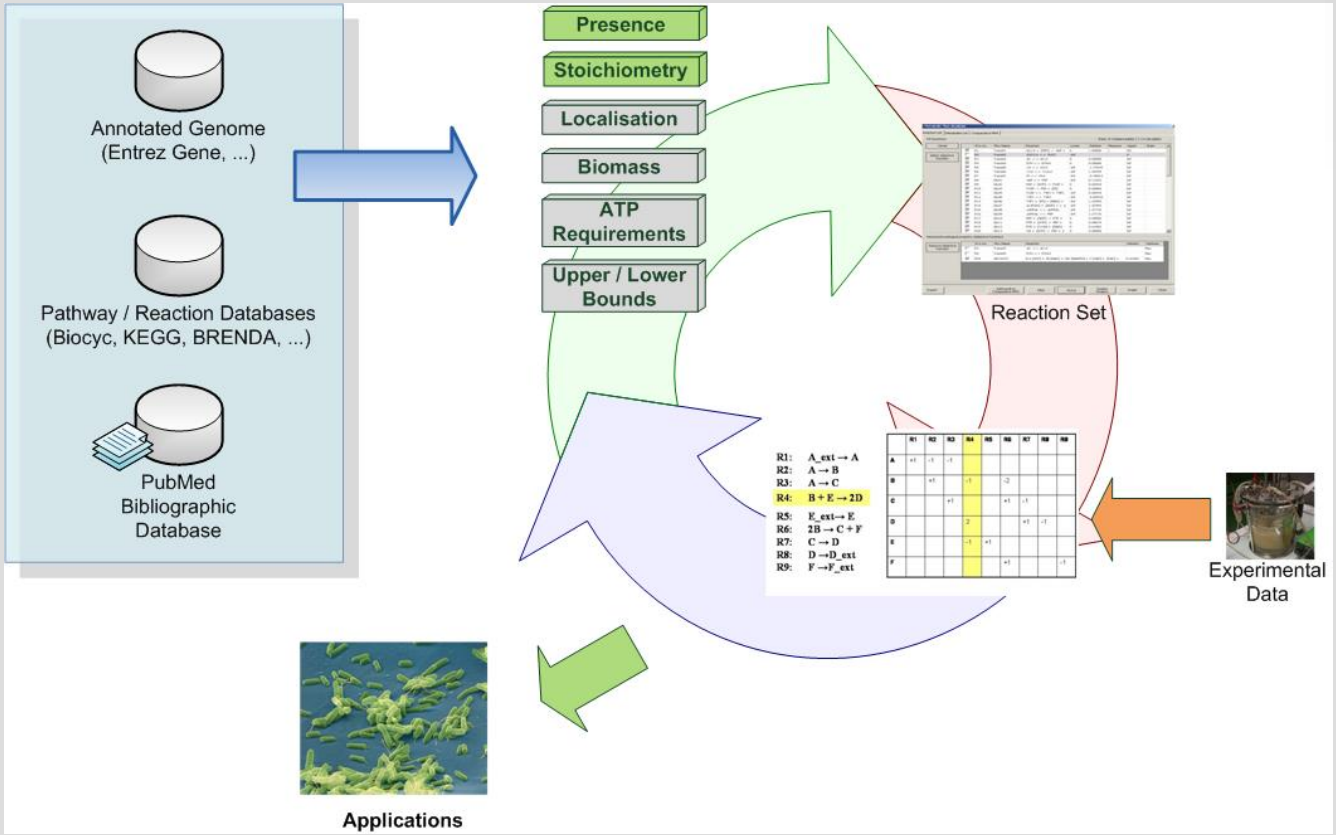
INFERENCE OF BIOLOGICAL NETWORKS

- HOW CAN WE BUILD THE MODELS IN AN AUTOMATED WAY?
- Ideally, it should be possible to extract all the knowledge necessary to construct biological models from the information obtained during genome sequencing
- However, the knowledge extracted is still very limited…



- Presently, the methodology of obtaining stoichiometric models from genome annotation is quite developed

INFERENCE OF BIOLOGICAL NETWORKS
GENOME-SCALE METABOLIC MODELS



WORK IN PROGRESS:
Reconstruction of Metabolic Networks of:
-H. pylori
-K. lactis
-Streptococcus faecalis

Rocha et al (2007), Gene Ess Gen Scale

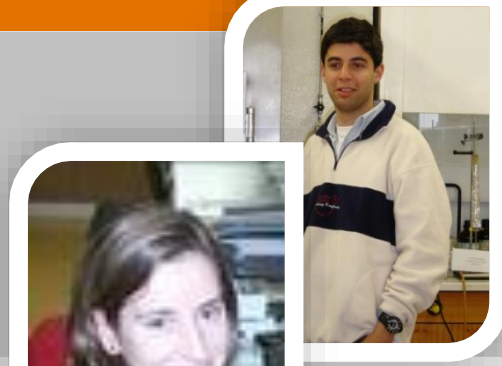
INFERENCE OF BIOLOGICAL NETWORKS
INFERENCE FROM EXPERIMENTAL DATA

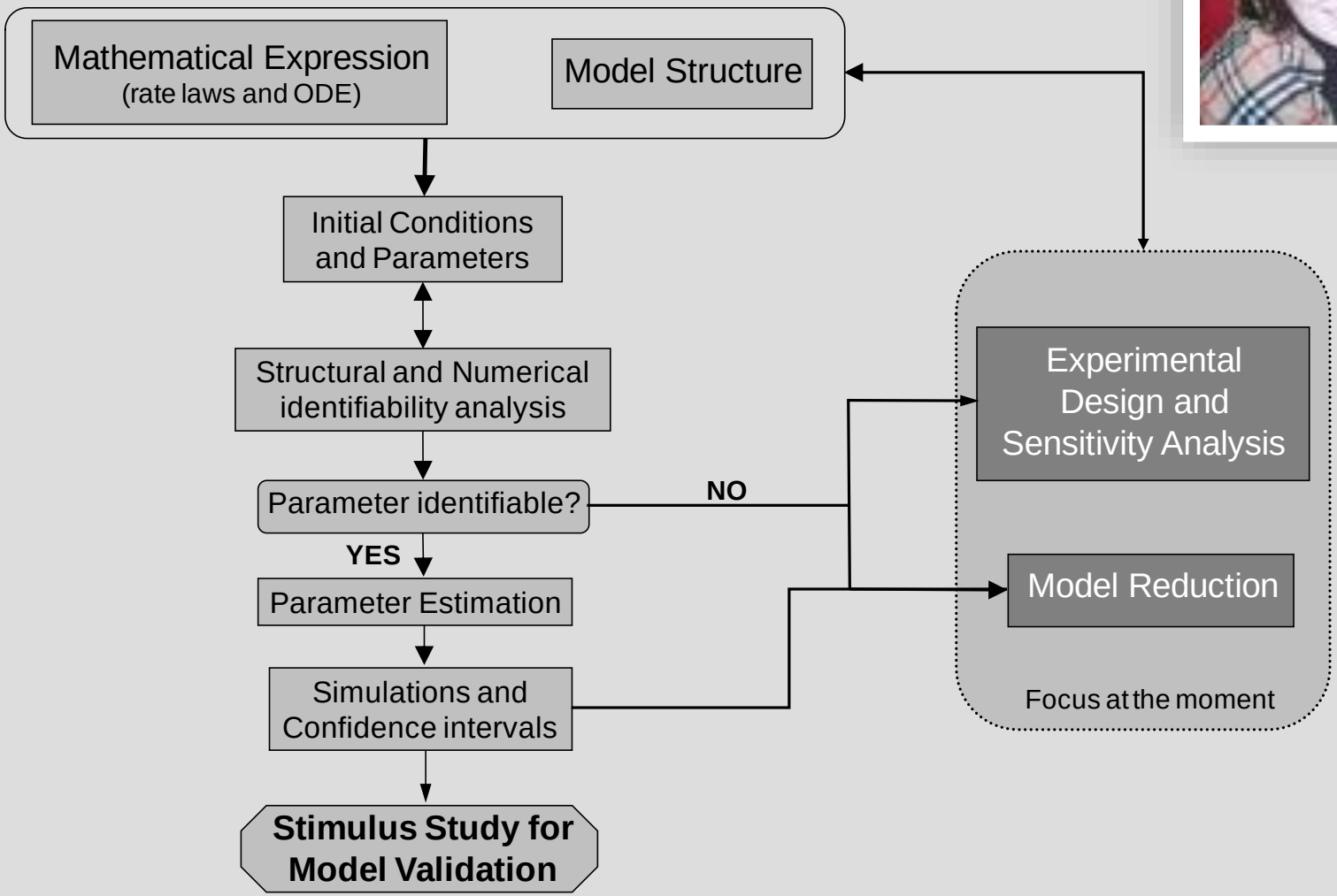
- Inference of Biological Networks can also be performed, from **experimental data**.
  - Flux and metabolomic data allow, in principle, to estimate model parameters for kinetic deterministic metabolic models
  - However, the number of experiments and measurements to be performed is very high!
  - Also, the structure of the kinetic equations has to be imposed a priori
  - In this field, optimal experiment design play an important role

- An alternative is to use **Text Mining** tools to automatically search in the literature for biological relations

INFERENCE OF BIOLOGICAL NETWORKS

WORK IN PROGRESS: Experimental Design for inferring model structure and parameters from flux data

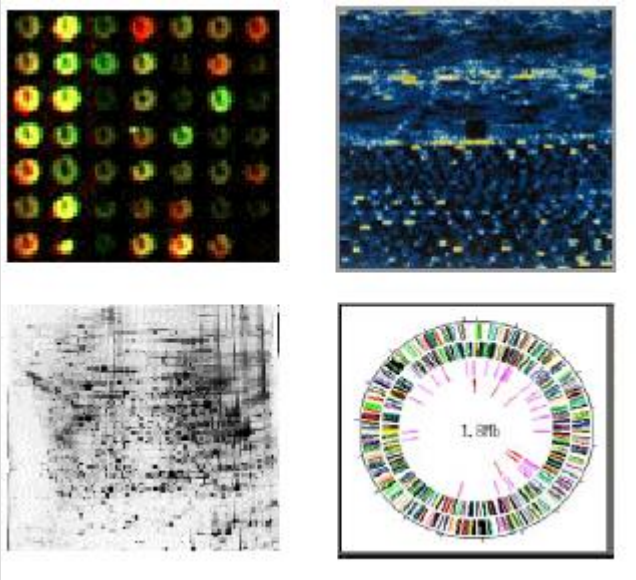




Model Reduction based on dynamic sensitivity analysis

$$\frac{dC_i}{dt} = \sum_j \nu_{ij} r_j - \mu C_i$$

From *in vivo* to *in silico* and back



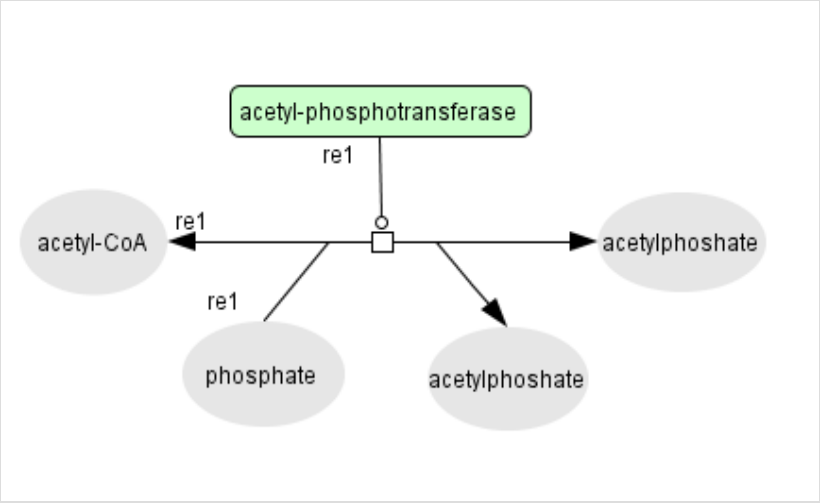
High-Throughput Data

Stoichiometric and/or Dynamic Modelling

Metabolism Complexity

A complete kinetic description
Complexity of dynamic modelling

$$\frac{dC_i}{dt} = \sum_j \nu_{ij} r_j - \mu C_i$$



- **re1$_{PTA}$**= f(metabolites, enzyme, parameters, regulators,…
- **M**odel fluxes and concentrations over time

Drawbacks

- Lots of parameters
- Measured *in vitro* (valid *in vivo*?)
- Nearly impossible to get all parameters at genome scale model

Obstacle for their effective use in optimization and control processes

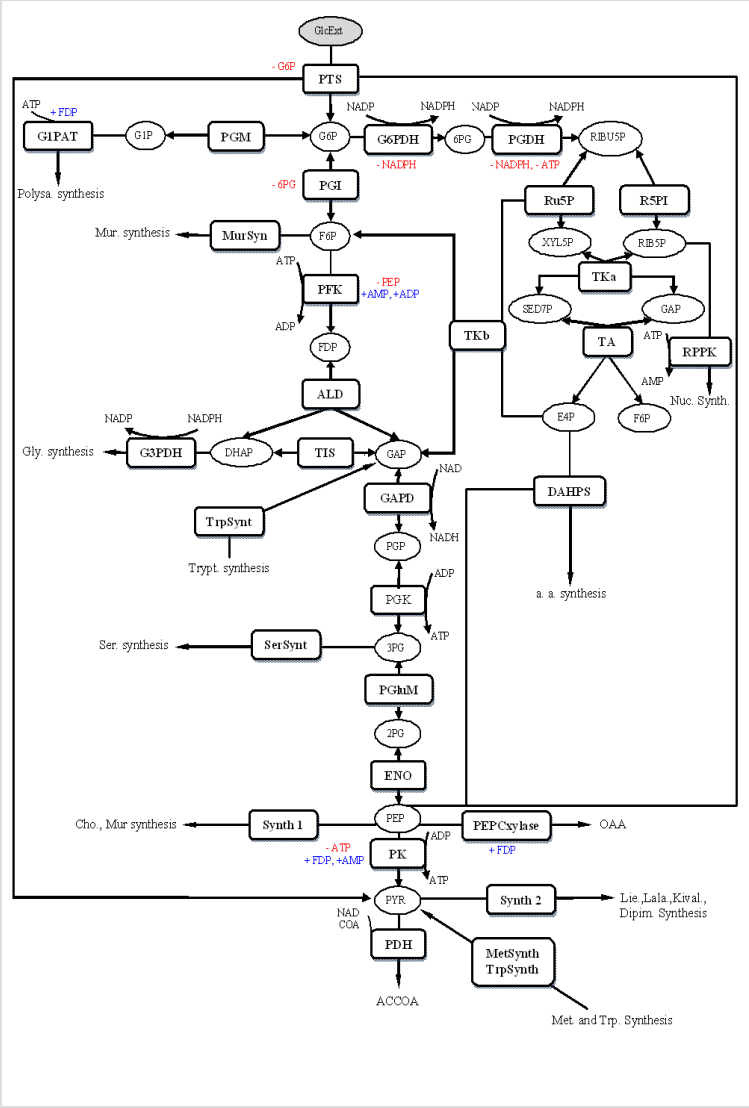
$$re1_{PTA} = \frac{r_{PTA}^{\max}\left(\frac{1}{K_{i,acetyl-CoA}K_{pta,p}}\right)\left(C_{acetyl-CoA}C_p - \frac{C_{acetyl-P}C_{coA}}{K_{pta,eq}}\right)}{1+\frac{C_{acetyl-CoA}}{K_{i,acetyl-CoA}}+\frac{C_p}{K_{i,p}}+\frac{C_{acp}}{K_{i,acp}}+\frac{C_{CoA}}{K_{i,CoA}}+\left(\frac{C_{acetyl-CoA}C_p}{K_{i,acetyl-CoA}K_{pta,p}}\right)+\left(\frac{C_{acp}C_{CoA}}{K_{pta,acetyl-P}K_{i,CoA}}\right)}$$

Motivation for model reduction

$$\frac{dC_i}{dt} = \sum_j \nu_{ij} r_j - \mu C_i$$

Complex *E. coli* dynamic model describing the carbon central metabolism with 116 parameters was used to:

- Identify key parameters that have more impacts on the global systems – **Sensitivity analysis**
- Study a **model reduction** strategy based on univariate analysis of the Euclidean-norm to consider the effect to all metabolites.



Strategy

- Nonlinear ODE model

$$\frac{dX_i}{dt} = \sum_j \nu_{ij} r_j - \mu X_i \qquad X_i(t) = f(t, p_j, X_0)$$

- Computing sensitivity analysis

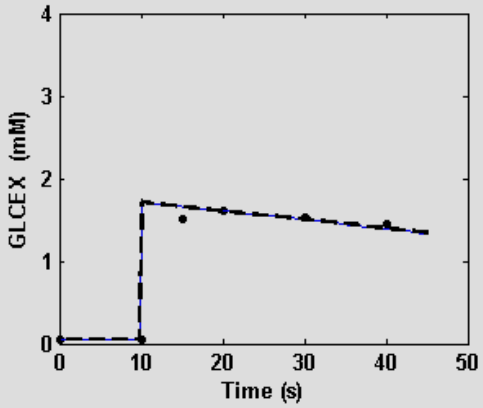
$$S_{i,j}(t) = \frac{X_i(p_j + \Delta p_j) - X_i(p_j - \Delta p_j)}{2\Delta p_j} \times \frac{p_j}{X_i(p)} \approx \frac{\partial \ln X_i(t,p)}{\partial \ln p_j}$$

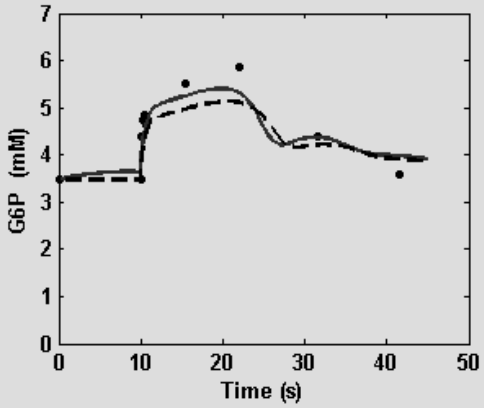
- Dynamic sensitivity analysis based on Euclidean-norm

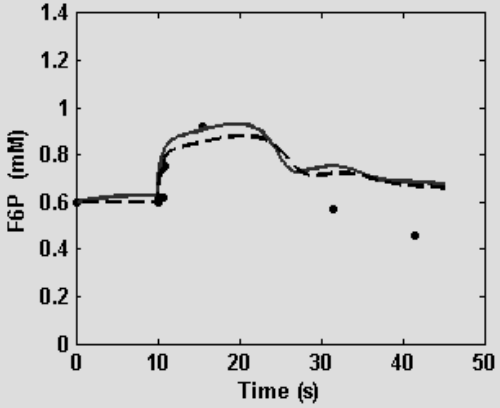
$$OS_j = \frac{1}{n} \sqrt{\sum_{k=1}^{n} \sum_{i=1}^{p} \left| S_{i,j}(t) \right|^2}$$

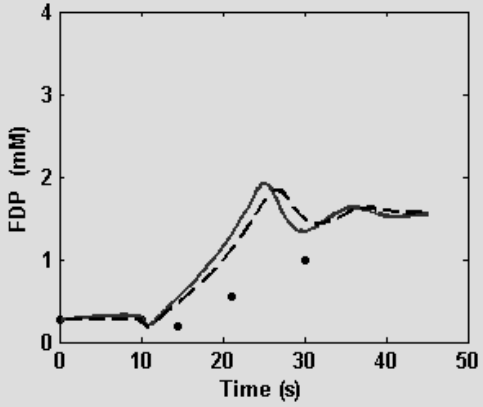
Comparison Original and reduced Model - Metabolite

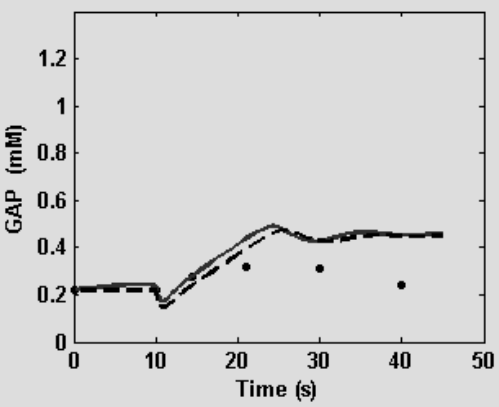
$$\frac{dC_i}{dt} = \sum_j \nu_{ij} r_j - \mu C_i$$

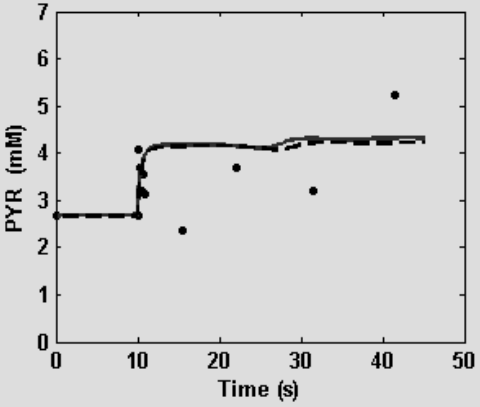










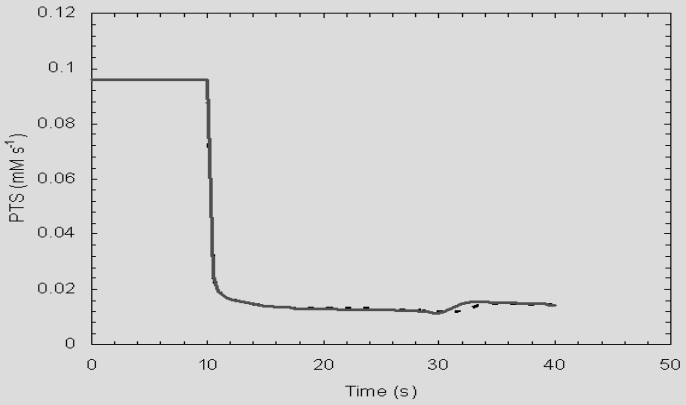


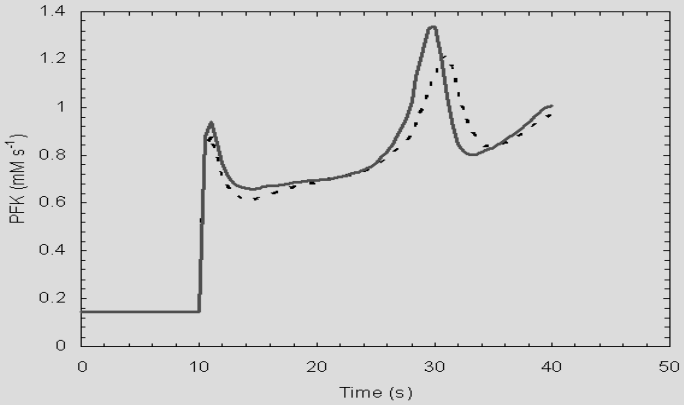
dotted line = original model
Solid line = reduced model

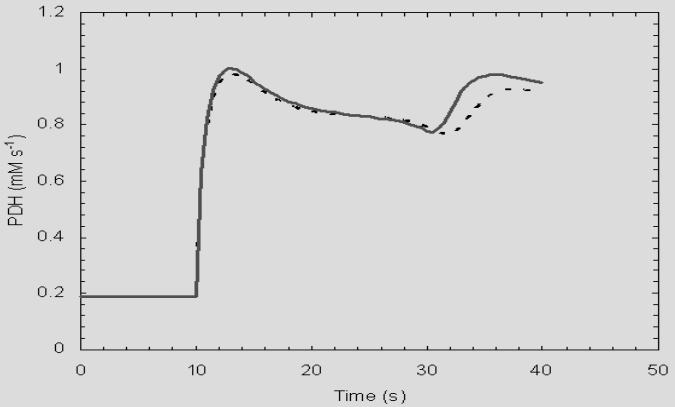
41 (35.3%) parameters were rejected

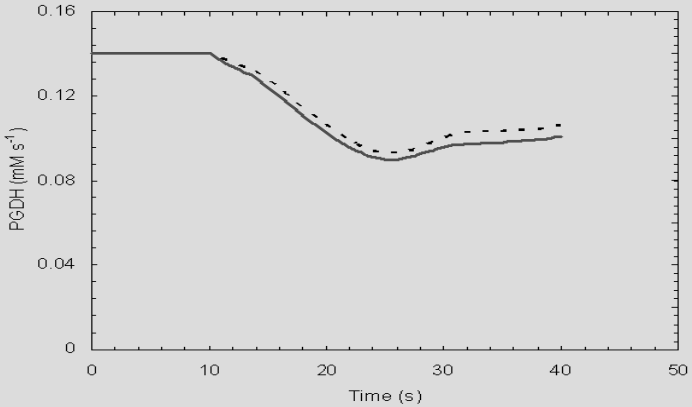
Comparison Original and reduced Model - Fluxes

$$\frac{dC_i}{dt} = \sum_j \nu_{ij} r_j - \mu C_i$$





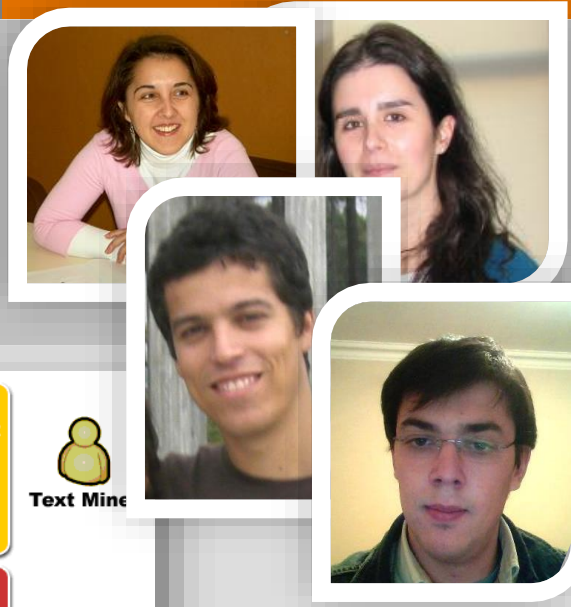


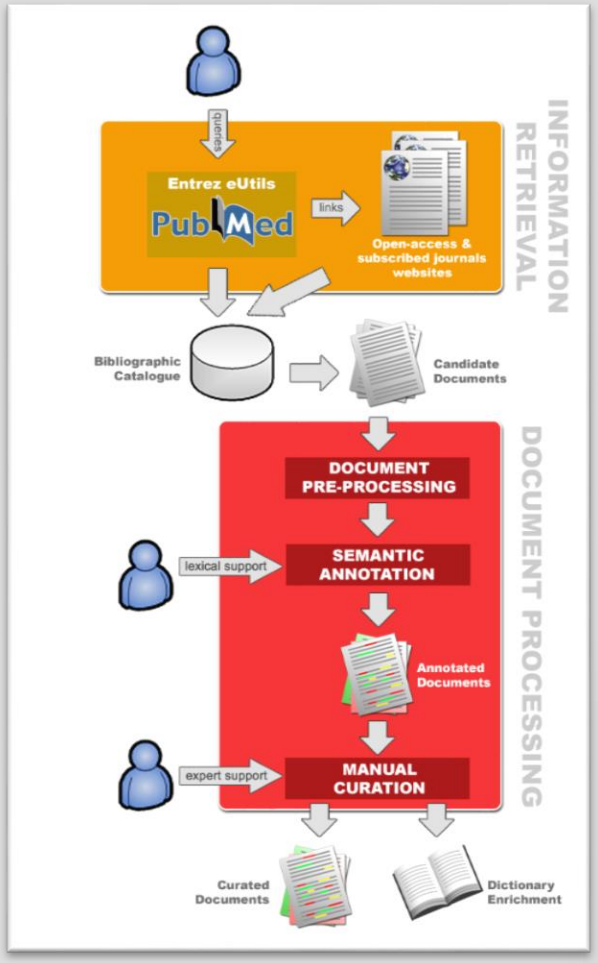


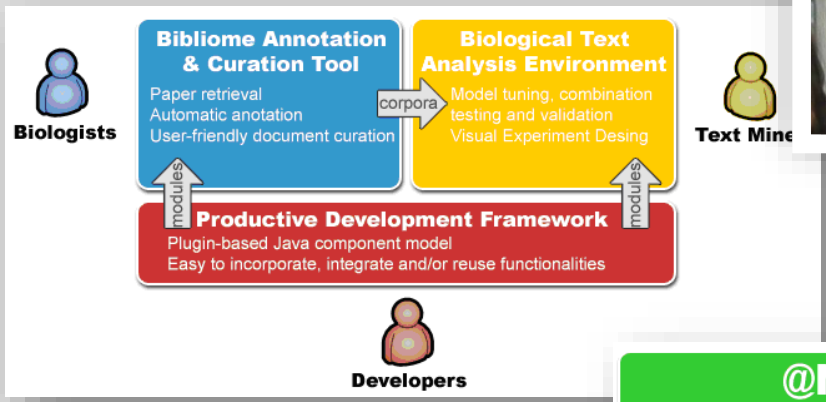
dotted line = original model
Solid line = reduced model

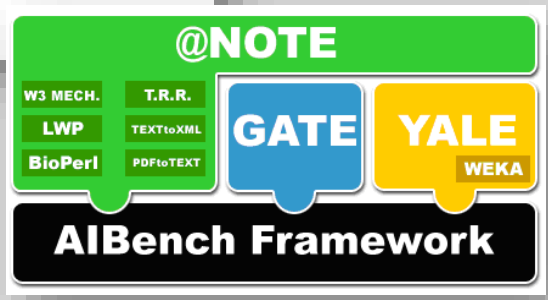
INFERENCE OF BIOLOGICAL NETWORKS

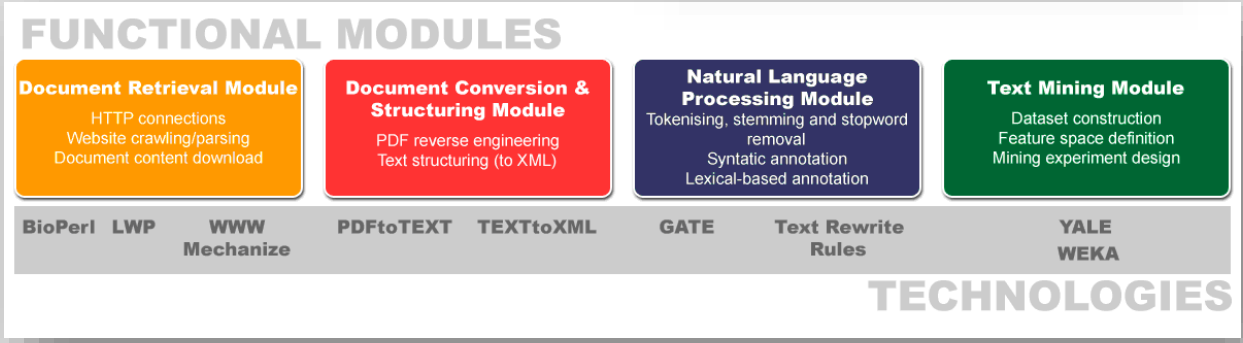
WORK IN PROGRESS: Development of tools for automatically inferring regulatory networks from literature data





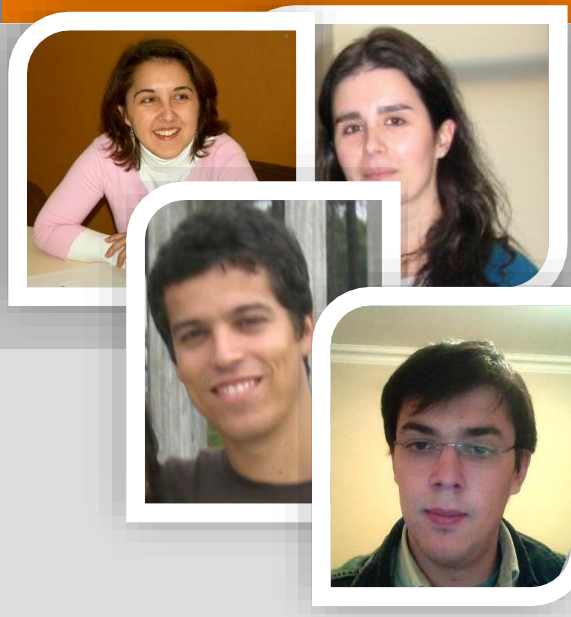


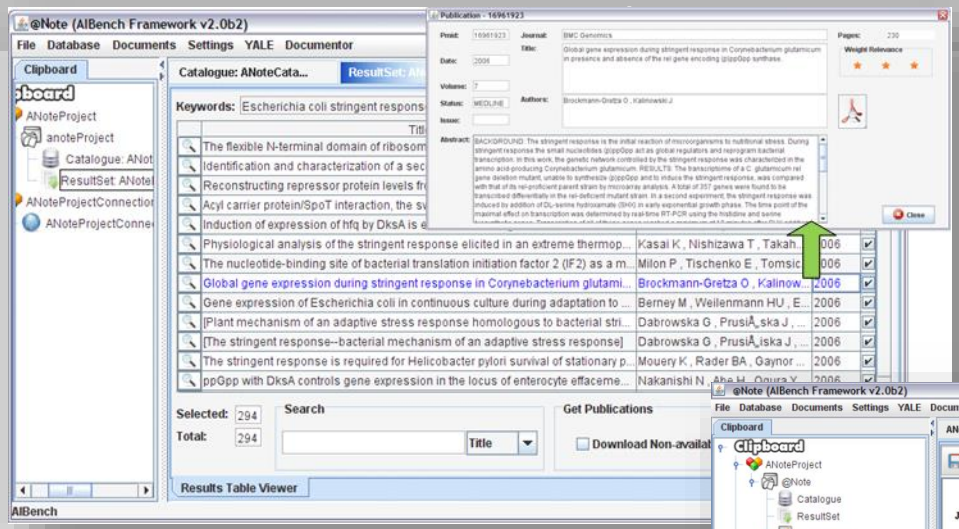


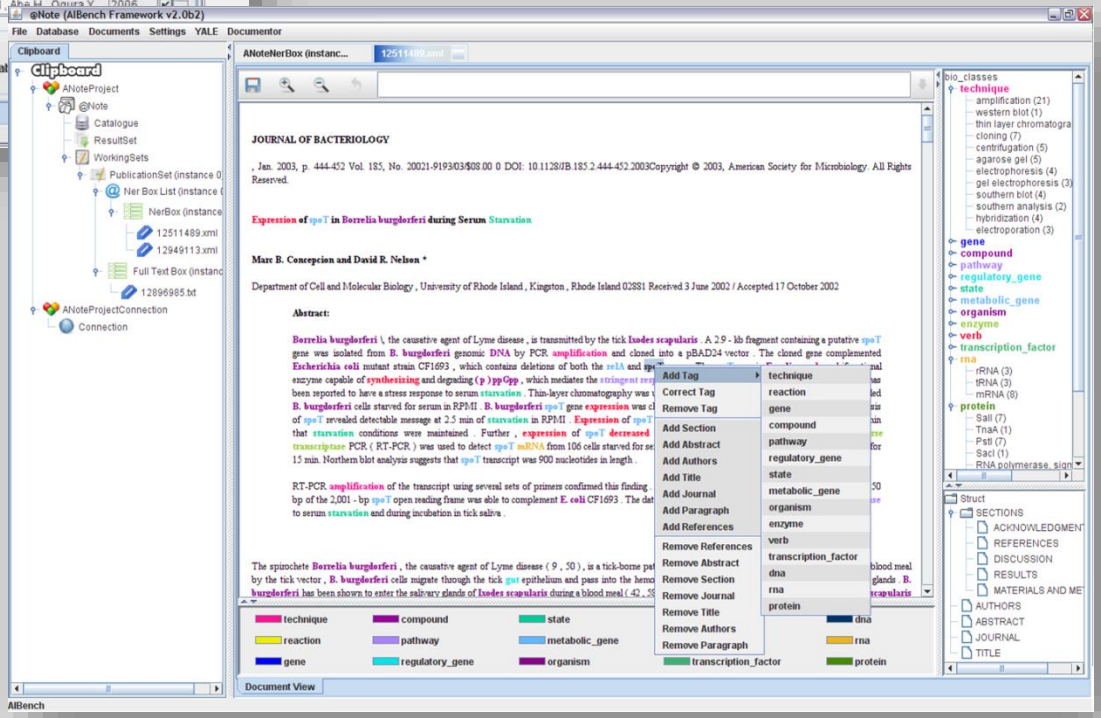


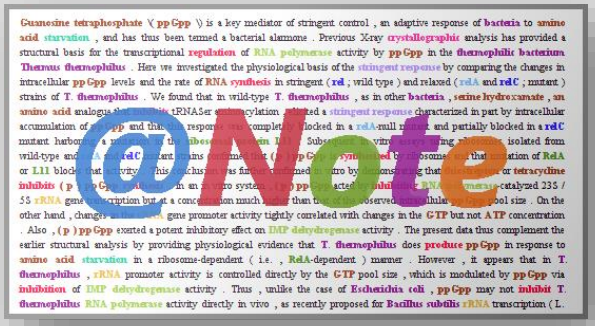
INFERENCE OF BIOLOGICAL NETWORKS

WORK IN PROGRESS: Development of tools for automatically inferring regulatory networks from literature

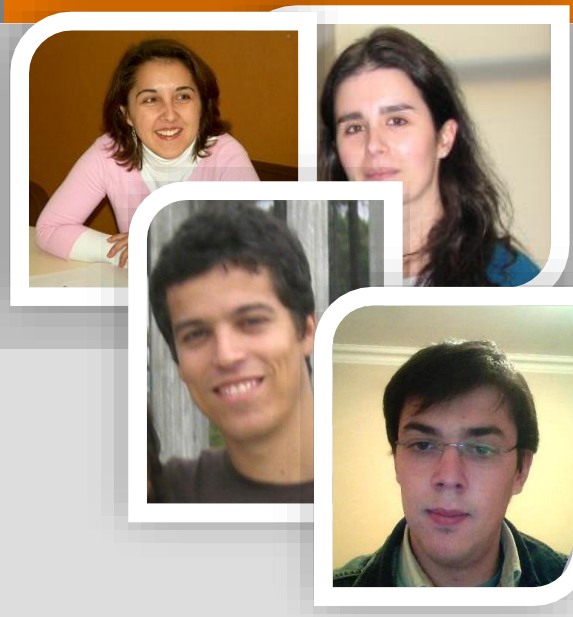




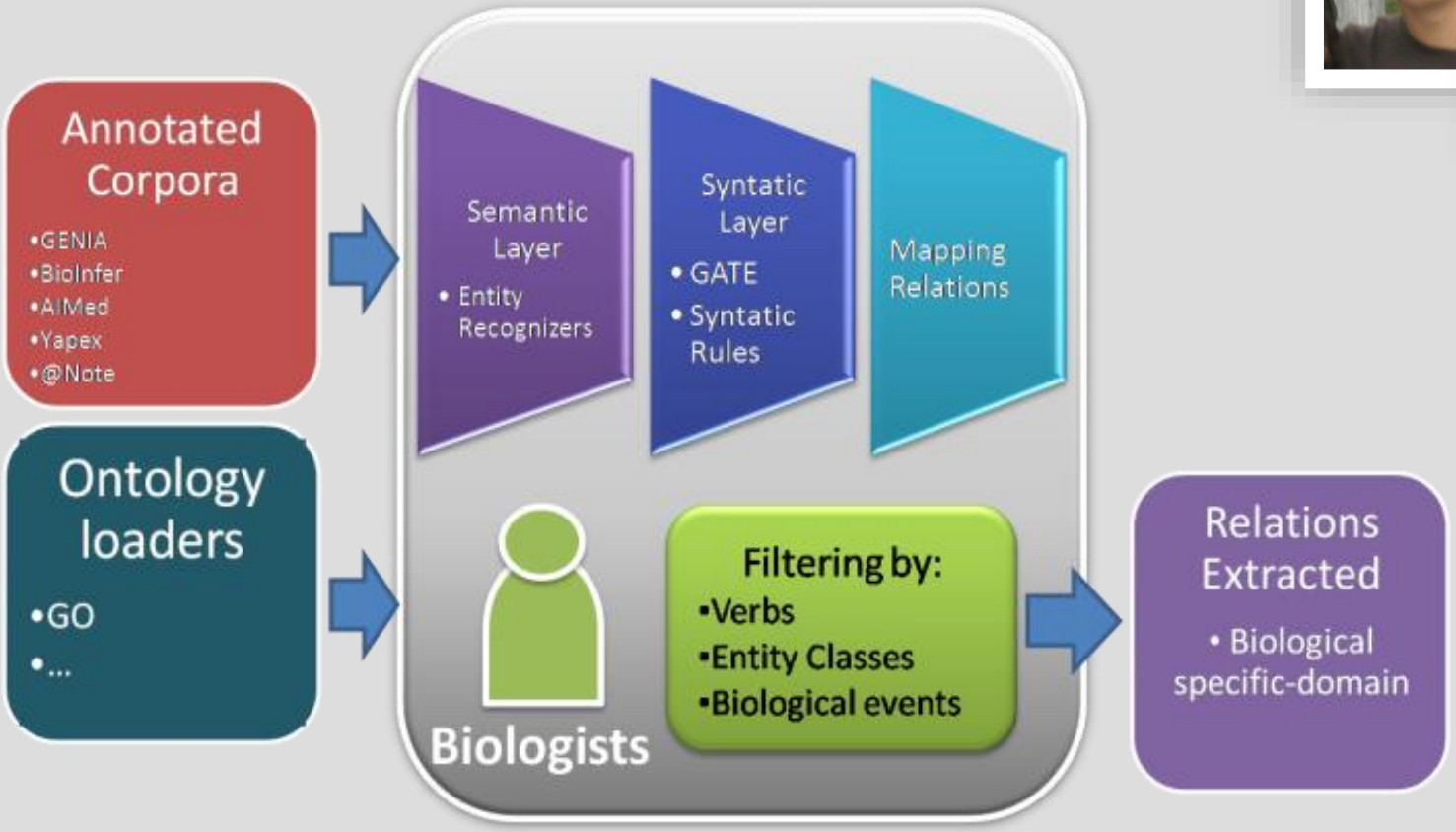




INFERENCE OF BIOLOGICAL NETWORKS

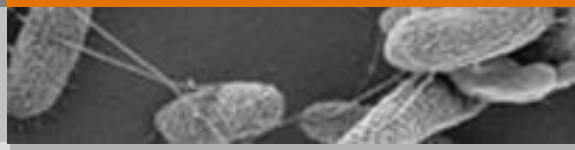


WORK IN PROGRESS: Development of tools for automatically inferring regulatory networks from literature data



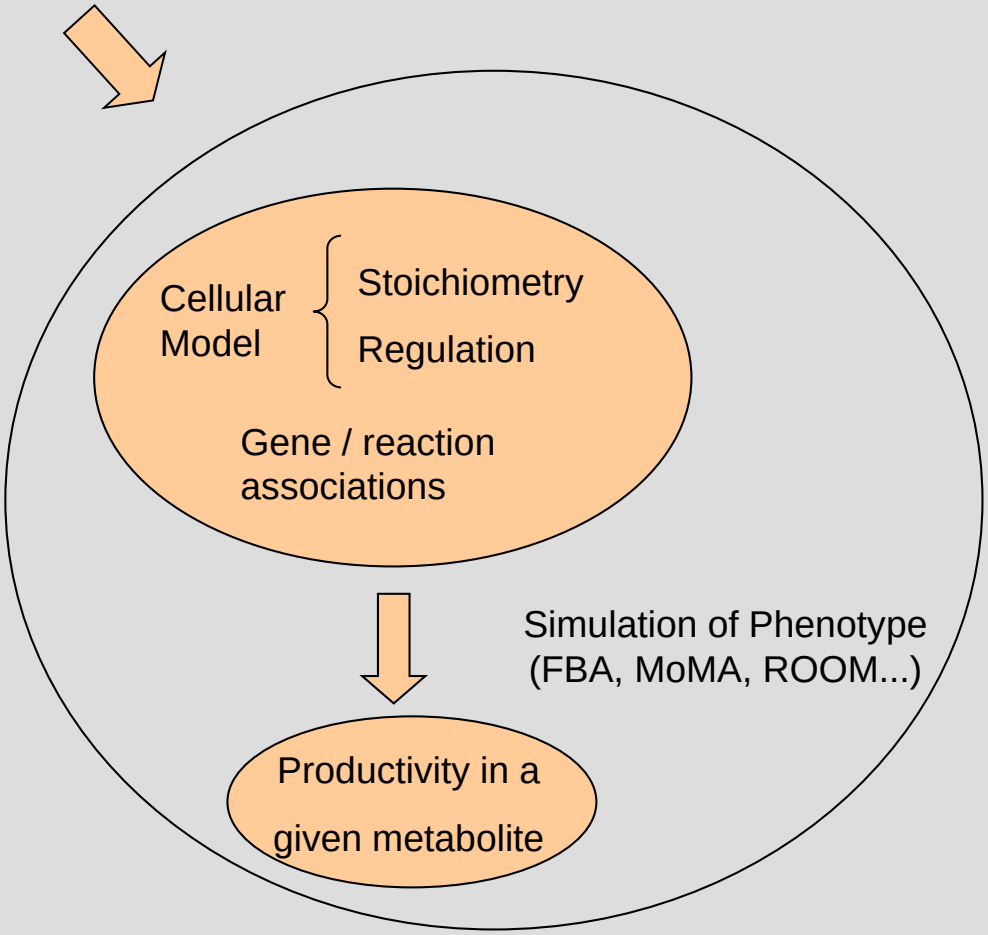
Lourenço et al., J. Biomed Inform. 2009, 42(4):710-720.

OPTIMIZATION TOOLS
OPTIMIZATION PROBLEM



Optimization Algorithm

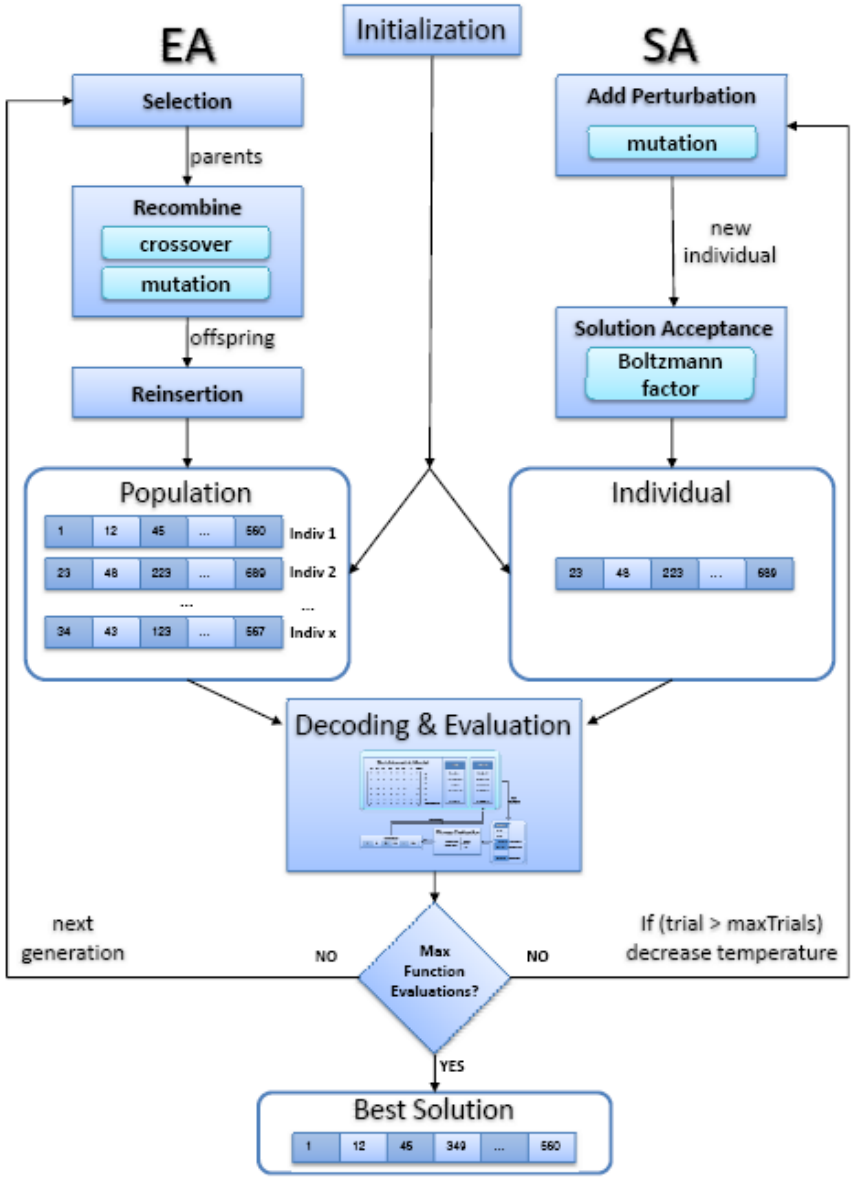
- Manipulation of gene deletions / additions and environmental conditions
- Objective function – productivity in a given metabolite



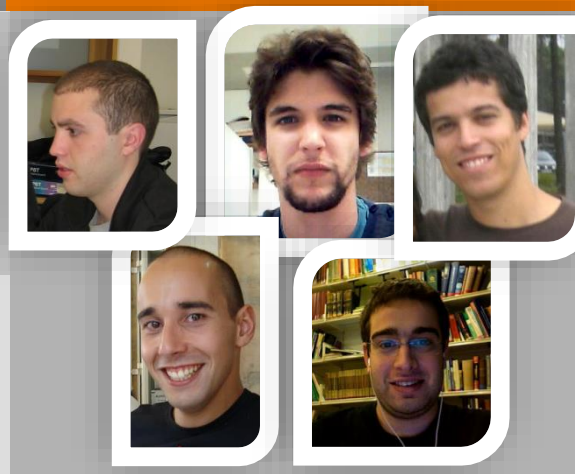
FBA: Flux Balance Analysis
ROOM: Regulatory On/Off *Minimization*
MoMA: Minimization of Metabolic Adjustment

The only reported algorithms are:

- OptKnock (Burgard *et al.*, Biotech Bioeng 2003)
  - Based on MILP
  - Only applicable to relatively small stoichiometric models
- OptGene (Patil *et al.*, BMC Bioinf 2005) & **OptFlux** (Rocha *et al*., BMC Bioinf 2008)
  - Evolutionary Algorithms
  - Applicable to different types of (large-scale) models
- Additional algorithms being applied:
  - Local Search
  - Simulated Annealing



WORK IN PROGRESS: OptFlux – a software for the Optimization of microbial strains



- **OptFlux** is an open-source, user-friendly and modular software aimed at being the reference computational tool for metabolic engineering applications. It allows the use of stoichiometric metabolic models for simulation and optimization purposes.

www.optflux.org

Functional Modules

Model Handling
- Model loading
- Model visualization
- Model Export

Simulation
- Wild Type
- Mutant Strain Design
- Env. Conditions
- Flux Var. Analysis

Optimization
- Global Strain Optimization
- Local Solution Optimization

Preprocessing
- Zero-valued variables identification
- Equivalent variables identification
- Critical Reactions identification

SBML Biovisualizer
Flat Files
CellDesigner SBML

Flux Balance Analysis
Minimization Of Metabolic Adjustment
Regulatory On/Off Minimization

Evolutionary Algorithms
Simulated Annealing
Best Neighbour
Hillclimbing

Methods & Technologies

Strain Optimization

- ✓ Evolutionary algorithms
- ✓ Simulated Annealing
- ✓ Local optimization

Objective function: Maximize the production of a given metabolite

Phenotype Simulation

- ✓ FBA, MoMA, ROOM
- ✓ Boolean net simulation
- ✓ …

Determine the set of fluxes (in steady state)

Environmental Conditions

Genetic conditions (e.g. knockouts)

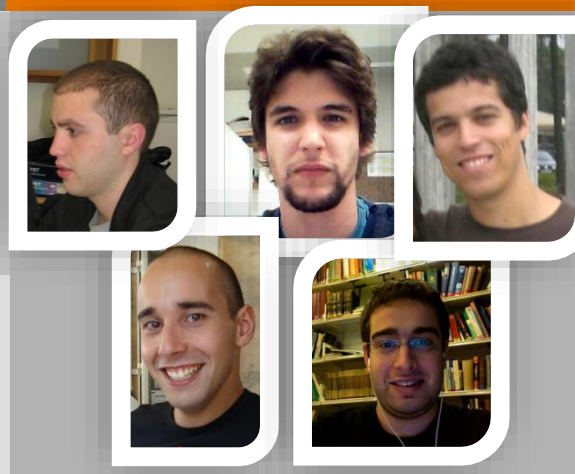
Override Model

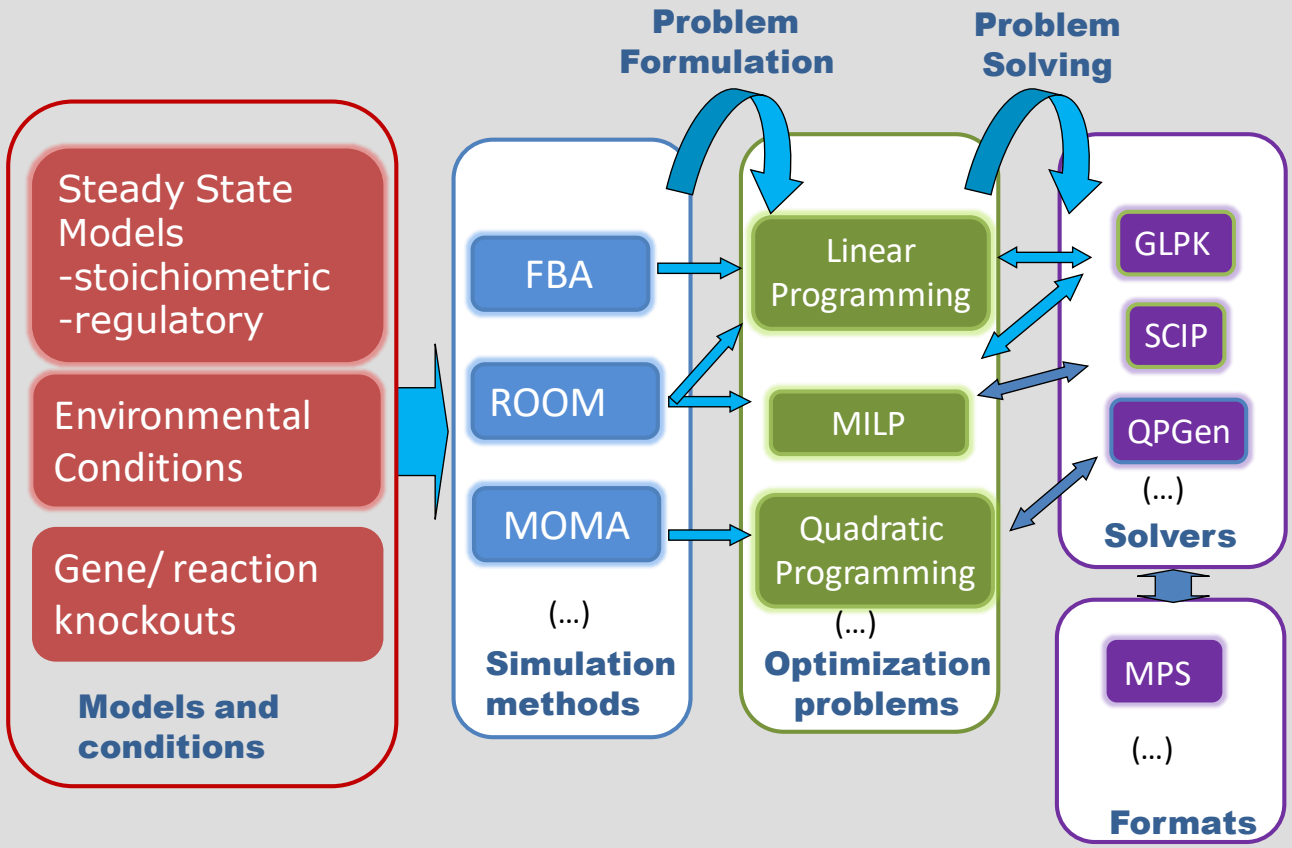
Model

- ✓Stoichiometric
- ✓Regulatory
- ✓…

Original Model

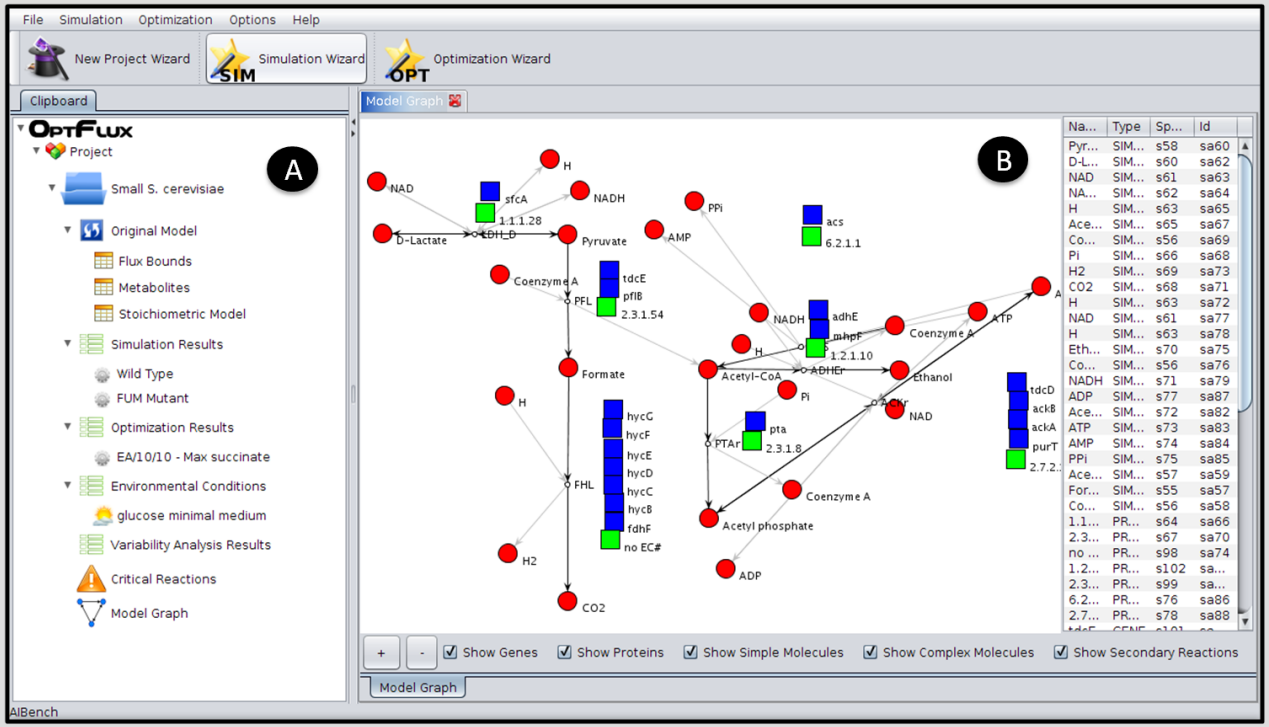
WORK IN PROGRESS: OptFlux – a software for the Optimization of microbial strains

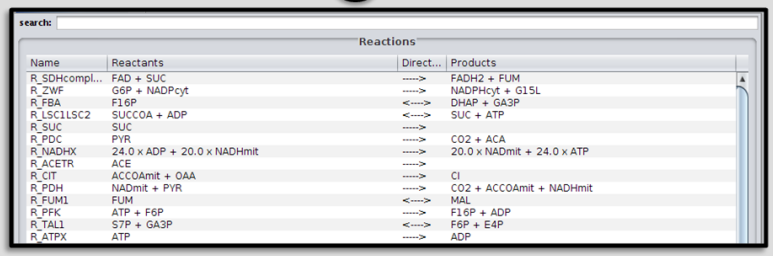


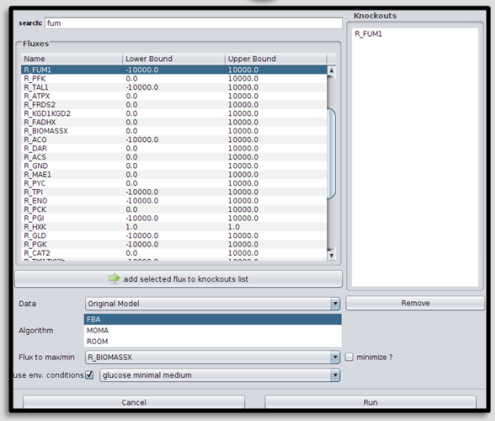


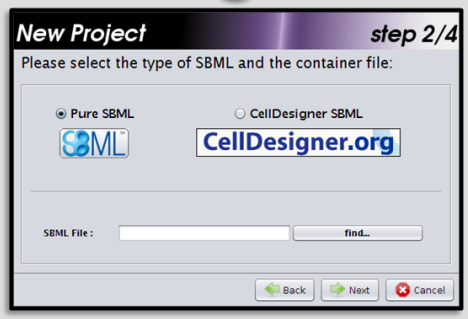
www.optflux.org

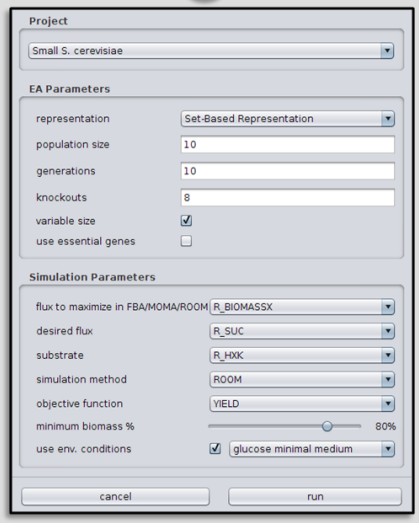
OPTIMIZATION TOOLS
OPTFLUX

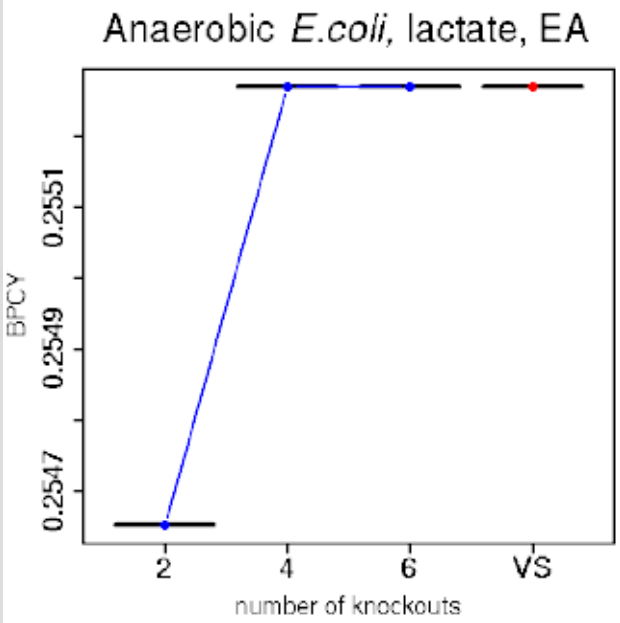


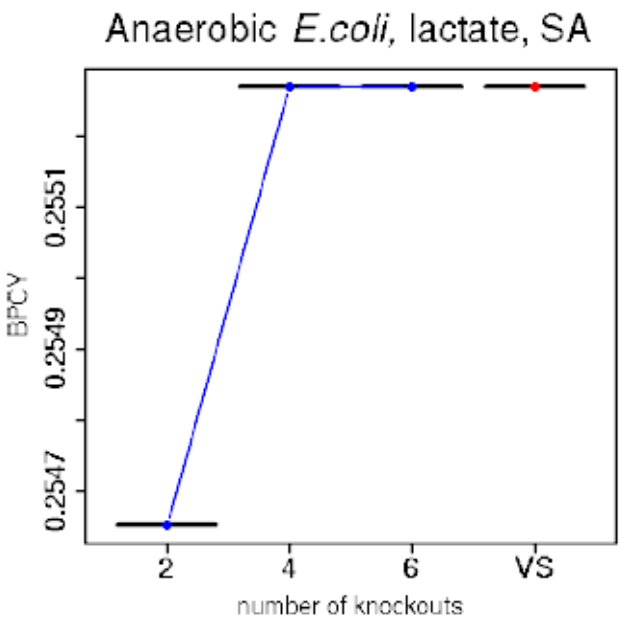


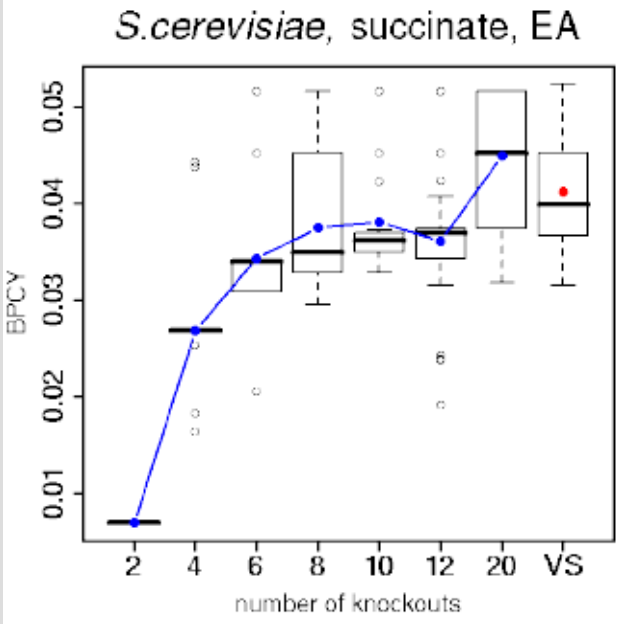


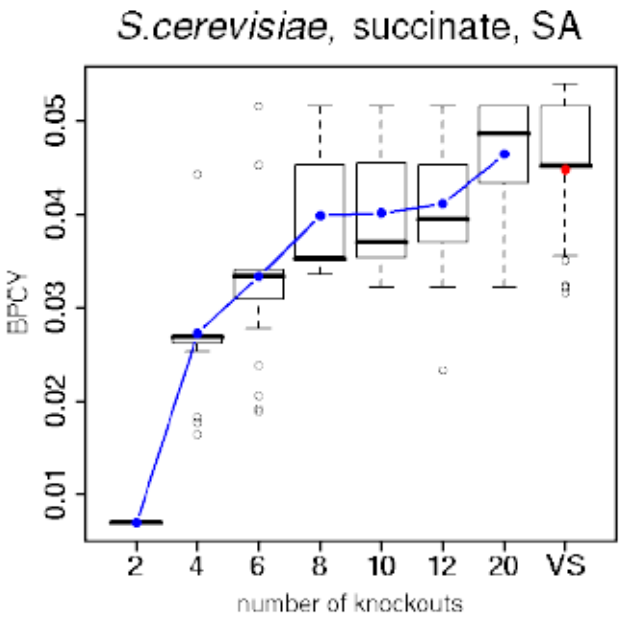












BPCY: Biomass-Product Coupled Yield
VS: Variable size

Acknowledgments - BioPSE group

Faculty:
Eugénio Ferreira
Isabel Rocha
Miguel Rocha
Rui Mendes
Ana Veloso
Anália Lourenço

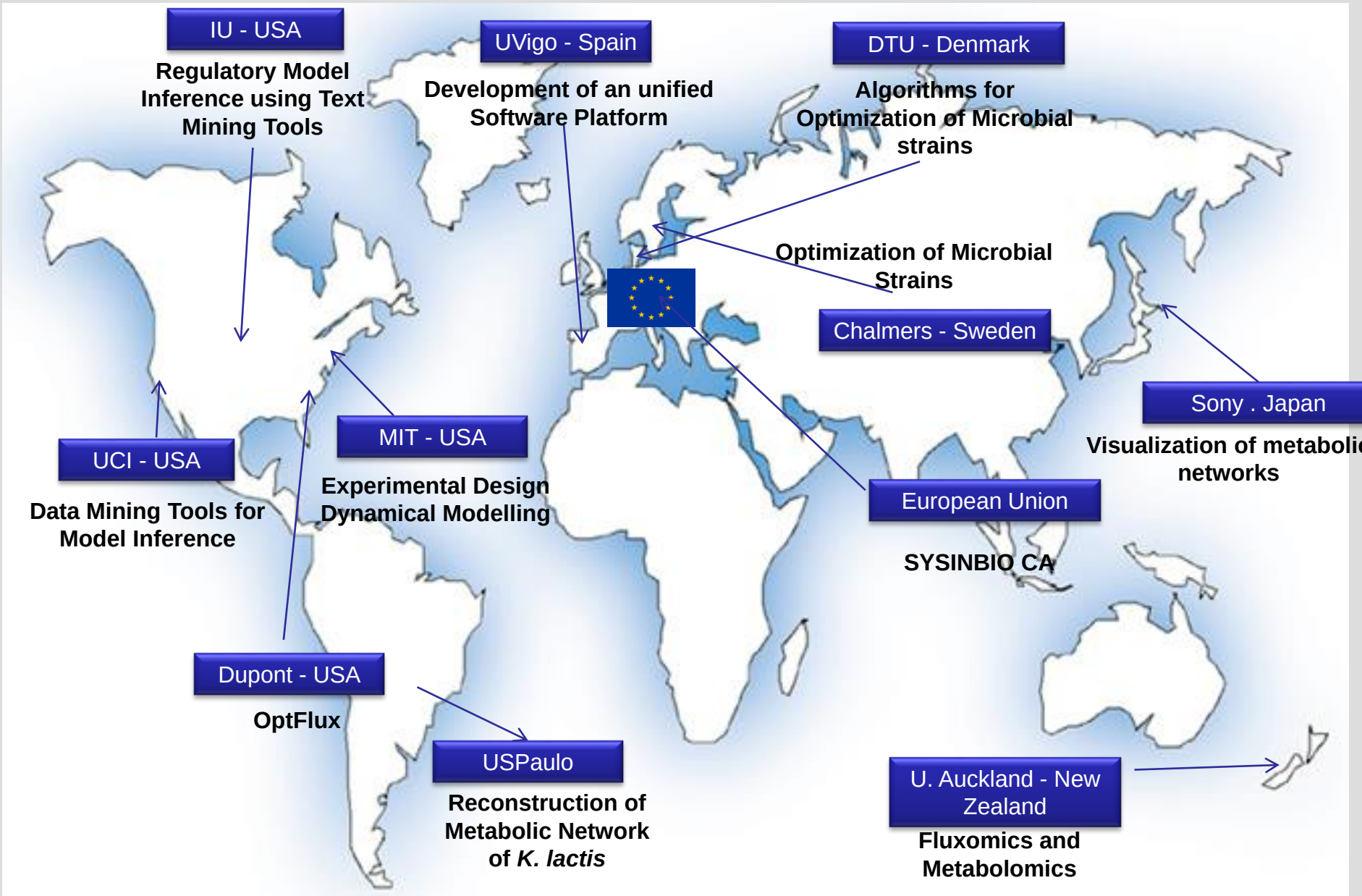
Post-Docs:
Zita Soons

PhD students:
Sónia Carneiro
Rafael Costa
José Pedro Pinto
Daniel Machado
Óscar Dias
Carla Portela
Daniela Correia

Other researchers:
Paulo Maia
Pedro Evangelista
Rafael Carreira
Simão Soares
José P. Faria
Paulo Vilaça
Rui Pereira

IBB – Institute for Biotechnology and Bioengineering
Centre of Biological Engineering
Universidade do Minho

http://biopseg.deb.uminho.pt


IU - USA
Regulatory Model Inference using Text Mining Tools
UVigo - Spain
Development of an unified Software Platform
DTU - Denmark
Algorithms for Optimization of Microbial strains
Optimization of Microbial Strains
Chalmers - Sweden
Sony . Japan
Visualization of metabolic networks
UCI - USA
Data Mining Tools for Model Inference
MIT - USA
Experimental Design Dynamical Modelling
European Union
SYSINBIO CA
Dupont - USA
OptFlux
USPaulo
Reconstruction of Metabolic Network of *K. lactis*
U. Auckland - New Zealand
Fluxomics and Metabolomics