Parameter estimate in biochemical network models

Giuseppe Aprea\textsuperscript{1}  Grazia Licciardello\textsuperscript{2}  Vittorio Rosato\textsuperscript{3}

\textsuperscript{1}Portici Research Center, CRESCO project, Via del Vecchio Macello, 80055 Portici (Naples), Italy. giuseppe.aprea@gmail.com

\textsuperscript{2}Science and Technology Park of Sicily, stradale V. Lancia 57, z.i. Blocco Palma I, 95121 Catania, Italy. gralicci@unict.it

\textsuperscript{3}Casaccia Research Center, Computing and Modelling Unit, Via Anguillarese 301, 00123 S.Maria di Galeria, Italy. rosato@casaccia.enea.it

February 11, 2009
Outline

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

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

the goal is to evaluate 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

Parameter estimate
Minimization by analogy with Nature: The genetic algorithm

Parameter estimate

G. Aprea et al.
Minimization by analogy with Nature: The genetic algorithm

a Gene

a Genome/Chromosome
Minimization by analogy with Nature: The genetic algorithm

- a Gene
- a Genome/Chromosome
- an Individual
- Several Individuals form a Population

Simulation Results

Metabolite Conc. vs Time

Parameter estimate
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
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

G. Aprea et al.  Parameter estimate
Algorithm Units: Model Simulation

Metabolic Model

Parameters Guess

Simulator

G. Aprea et al.   Parameter estimate
Algorithm Units: Model Simulation

Metabolic Model

Parameters Guess

Simulator

Experimental Data

Metabolite Conc.

Time

G. Aprea et al. Parameter estimate
Algorithm Units: Model Simulation

Metabolic Model

Parameters Guess

Simulator

Experimental Data

Simulation Results

G. Aprea et al. Parameter estimate
Algorithm Units: Model Simulation

Metabolic Model

Parameters Guess

Simulator

Experimental Data

Metabolite Conc. vs Time

Metabolite Conc. vs Time

Simulation Results

Evaluate COST FUNCTION

G. Aprea et al. Parameter estimate
Algorithm Units: Model Simulation

Metabolic Model

Parameters Guess

Simulator

Experimental Data

Metabolite Conc.

Time

Simulation Results

Evaluate
COST FUNCTION

Metabolite Conc.

Time

All these operators $\Rightarrow$ ecell session

G. Aprea et al.  Parameter estimate
Algorithm Units: GA operators

Individuals Selection
Offspring Generation
Mutation

GA operators

Parameter estimate
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)

GA operators

GA operators

GA operators

Solution
Parallel implementation (double stage)

Cluster 1

GA operators

Solution

Cluster 2

GA operators

Solution

Cluster n

GA operators

Solution

SuperGeneration Operators

Solution

SuperGeneration Operators
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.
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:
\( R(t) \rightarrow \) Biomass (includes Product), \( P(t) \rightarrow \) Product(PHA), 
\( S_1(t) \rightarrow \) Nutrient (oleic acid), \( S_2(t) \rightarrow \) Nitrogen Source.

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

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

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

\[
\frac{1}{R} \frac{dS_2}{dt} = - (Y_{R/S_2} \mu + M_{S_2}) 
\]
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.
Model fitting (in progress...)

- Biomass [Exp] vs Time [h]
- Biomass [In Silico] vs Time [h]
- PHA [Exp] vs Time [h]
- PHA [In Silico] vs Time [h]
- Oleic Acid [Exp] vs Time [h]
- Oleic Acid [In Silico] vs Time [h]
- Nitrogen [Exp] vs Time [h]
- Nitrogen [In Silico] vs Time [h]
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...).


Acknowledgments

Frascati Enea-INFO group

- Giovanni Bracco
- Salvatore Podda
- Carlo Sció