Advances in genomics and -omic sciences have begun to produce unprecedented amounts of data that await for 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. 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 which yield reliable values for model parameters with a large computational demand.

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 and Globus are supported as well)

The genetic algorithm: Introduction

Very often modelling of biochemical network is affected by parameters: kinetic constants, initial concentrations, ...

According to GA analogy, unknown parameters like the genes of a DNA string, undergoing six stages of selection and mutation to finally select the fittest individuals, i.e. the best experiment...

The genetic algorithm: Standard Operating Procedure

In the following picture the standard GA procedure is shown. In order to efficiently parallelize GA we determined some units (“ecell session” and “GA operators”) which can be run on a single CPU.

1) Generate a random initial genotype set.
2) Calculate each individual’s fitness \( f(x_i) \), and \( f(x_i)/\lambda \), where \( \lambda \) is a set of experimental observations and \( f(x_i)/\lambda \) is the corresponding simulated quantities; \( \lambda \) represents the unknown parameters.
3) Select individuals from the current population to generate an offspring with a small variance.
4) GA operators
5) To the next step with the new population, or stop if exit criteria (threshold for the cost function or maximum generation limit) are satisfied.

Parallel implementation (single stage)

In this first implementation we have one repeated stage of parallelism. A master process spreds N ecell sessions, each corresponding to the evaluation of the fitness for an individual, to N different CPUs. Then it takes the result and proceeds with the remaining steps (S-D) of GA. This procedure is conceived to be used on a single (better if homogeneous) cluster with many CPUs since all the results from them are coordinated by a single node and latency periods to collect informations from different clusters might be relevant.

Parallel implementation (double stage)

In this second implementation we have two repeated stages of parallelism. A master process spreds different GA sessions, each with a smaller population, to N different CPUs. Each GA session works as explained in the case of the single stage procedure. When all GA sessions have finished, the master node takes the best individuals from each of them and selects a new starting population and the process is repeated again. This procedure is thought to be used on architectures with small distributed clusters with not many CPUs (nodes that we use reduced populations). In this case the central node has less communication with the secondary GA sessions and the procedure is not plagued by latency periods of inter cluster communications. The information exchange performed by the master node is important to avoid a secondary GA session be trapped in a local solution since the small size of its population.

Examples

We used our implementations for GA in the case of the Circadian Cycle metabolism. We used ecell-model-editor, which is an Ecell tool, for drawing the biochemical network model.

This is the evolution of the cost function (N CPUs, generations=100, population=100, \( 10^3 \) total individuals):

Comparison between evolution of the cost function of single and double stages procedures:

Although the double stage procedure yields good results, the single-stage one usually performs better. That is mainly due to the use of a larger population which is a very important parameter in GA.