Science of Computer Programming ••• (••••) •••–•••

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

Simulation of multicellular populations with Petri nets and genome scale intracellular networks

Kamil Kędzia [a], Wojtek Ptak [a], Jacek Sroka [a,*], Andrzej M. Kierzek [b]

[a] Institute of Informatics, University of Warsaw, Poland
[b] Certara Quantitative Systems Pharmacology, Blades Enterprise Centre, John Street, Sheffield, S2 4SU, United Kingdom

ARTICLE INFO

Article history:
Received 11 January 2017
Received in revised form 21 September 2017
Accepted 6 October 2017
Available online xxxx

Keywords:
Petri net
Genome scale metabolic network
Multicellular population
Distributed processing
Large graph processing

ABSTRACT

We present a new distributed architecture allowing simulation of living cells in spatial structures. Each cell is represented with a Quasi-Steady State Petri Net that integrates dynamic regulatory network expressed with a Petri net and Genome Scale Metabolic Network (GSMN) where linear programming is used to explore the steady-state metabolic flux distributions in the whole-cell model.

The combination of Petri net and GSMN has already been used in simulations of single cells, but we present an extension to the model and an architecture to simulate populations of millions of interacting cells organised in spatial structures which can be used to model tumour growth or formation of tuberculosis lesions. The crucial element of this solution is the ability of cells to communicate by producing and detecting substances such as cytokines and chemokines. This ability is modeled by allowing cells to share tokens in places called communicators.

To make the simulation of such a huge model possible we use the Spark framework and organise the computation in an agent-based "think like a vertex" fashion as in Pregel-like systems. In the cluster we introduce a special kind of per node caching to speed up computation of the steady-state metabolic flux.

We demonstrate capabilities of the new architecture by simulating communication of liver cells through FGF19 cytokine during the homeostatic response to cholesterol burst. Our approach can be used for mechanistic modelling of the emergence of multicellular system behaviour from interaction between genome and environment.

1. Introduction and objectives

Mechanistic modelling of biological systems is making an increasing impact in research and industry. The ultimate goal is to predict behaviour of the system, given information about its genetic blueprint and environmental conditions to which it is responding. Progress towards this goal is necessary to enable personalised and precision medicine, where diagnostics, disease prevention and therapy will be tailored to the patient's genetic background and lifestyle. In the context of biotechnology and synthetic biology, computer simulation of biological systems is a necessary tool for rational design of genetically engineered cells. With the advent of Next Generation Sequencing it is now possible to obtain the full genome sequence

* Corresponding author.
E-mail addresses: k.kedzia@mimuw.edu.pl (K. Kędzia), w.ptak@mimuw.edu.pl (W. Ptak), j.sroka@mimuw.edu.pl (J. Sroka), andrzej.kierzek@certara.com (A.M. Kierzek).

https://doi.org/10.1016/j.scico.2017.10.002
0167-6423/

Please cite this article in press as: K. Kędzia et al., Simulation of multicellular populations with Petri nets and genome scale intracellular networks, Sci. Comput. Program. (2017), https://doi.org/10.1016/j.scico.2017.10.002

of any species, cell line or individual. Moreover, a large number of other ∼omics approaches enables measurements of proteins, RNAs and metabolites with increasing quantitative accuracy. Availability of these data as well as a legacy of over half a century of molecular biology research further motivate mechanistic modelling of the relationship between genotype, environment and biological systems behaviour.

Intensive research in the field of Systems Biology has already established numerous approaches to represent molecular biology knowledge in the form of molecular interaction network models. Exact Stochastic Simulations provide the most detailed quantitative description, where individual reactive collisions occurring at exact times in single cells are simulated by the Monte Carlo approach of Gillespie [1]. The Ordinary Differential Equation (ODE) formalism applied to study the time evolution of average molecular concentration in cellular population is a workhorse of quantitative modelling [2]. However, it is still not possible to parametrise quantitative dynamic models of whole-cell scale networks and therefore approximate methods are being applied to address the challenge of simulating genotype-phenotype relationship.

Constrained Based Modelling (CBM) has achieved spectacular success modelling metabolism at the full genome scale [3, 4]. It capitalises on the fact that connectivity of a metabolic network is the best-studied sub-system within the full network of molecular interactions of a cell, and can be modelled at quasi-steady state due to time-scale separation between gene regulation (hours) and metabolism (seconds). This enables the application of Linear Programming to explore metabolic flux distributions within Genome Scale Metabolic Networks (GSMN), consistent with Stoichiometric and thermodynamic constraints. The Mixed Integer Linear Programming can be used to incorporate data on enzymatic gene expression [5].

Regulation of cellular processes is dynamic and cannot be fully described by steady state linear models used in CBM. A number of qualitative simulation approaches have been formulated, such as analysis of steady states in logical hypergraphs [6] and exhaustive enumeration of states in ODE models reduced to a Boolean framework [7]. One of the promising recent approaches to qualitative modelling of cellular dynamic processes is Monte Carlo exploration of the alternative molecular transition sequences constrained by network connectivity expressed in Petri net formalism [8,9].

The Quasi-Steady State Petri Net (QSSPN) [9] algorithm has been developed to integrate Petri net models of gene regulatory and signalling networks in the cell with steady state models of GSMN. The quasi-steady state approximation is used, where for every state of dynamic, Petri net model the steady-state metabolic flux distribution of GSMN is explored with the CBM approach. The QSSPN has been first applied to the integration of a liver-specific GSMN with a qualitative model of a large-scale regulatory network responsible for homeostatic regulation of bile acid synthesis. Recently, a new version of QSSPN algorithm and solver have been published as a part of Multi-Formalism Network Simulator (MUFINS) [10].

The approaches described above are applied to model dynamics of molecular interaction network in single cells. Another important layer of living system organisation consists of interactions between cells spatially arranged in tissues of multicellular organisms. In cases of many diseases, such as tuberculosis or cancer, spatial arrangement of cells in granulomas or tumours is an important factor to be considered in rational drug design. Agent-based approaches have already been applied in the context of both healthy and diseased tissue behaviour. The current frontier is the integration of AB models with molecular networks in multiscale, hybrid simulation scenarios. In particular, GSMN simulation has been recently integrated with AB modelling of tuberculosis granulomas [11]. This has been a motivation for integration of the general QSSPN representation of genome scale molecular networks with AB modelling.

In this paper we extend QSSPN to multicellular systems and present Agent Based Quasi-Steady State Petri Net (AB-QSSPN), a prototype tool which provides a mechanistic link between genotype and behaviour of multicellular system. The multiscale nature of biological systems is one of the major challenges of their mechanistic modelling, especially in medical applications. The physiological state at health and disease emerges from events occurring at molecular, cellular, multicellular (tissue), organ and whole-body levels. Some of these processes, such as tumour growth or formation of tuberculosis lesions require spatial modelling of structures formed by millions of cells. Our distributed architecture enables multi-cellular simulations at this scale.

AB-QSSPN integrates simulation of agent-based models of multicellular systems and qualitative QSSPN models. To the best of our knowledge this is the first demonstration of the feasibility of a simulation of a multicellular system and its gene regulatory network, signalling and whole-cell scale metabolic reactions operating within each individual cell. To make this simulation possible, we capitalised on the MapReduce [12] model and its open source implementation Hadoop [13]. These recent developments in distributed computing enable reliable computation on clusters of commodity computers. As we are interested in iterative simulations we use Apache Spark [14], a new programming system which is the de-facto successor to MapReduce. It neatly follows a shift in the hardware used in data processing centres where the processing nodes are equipped with an increasing amount of RAM memory. Spark is based on the Resilient Distributed Datasets abstraction where the data by default, if possible, is kept in memory. In this way Spark avoids unnecessary spills to disk which are by orders of magnitude slower than in-memory operations. This allows Spark to outperform Hadoop in iterative computation similarly as HaLoop, Twister and Pregel/Giraph do. On the other hand, Spark is similar to MapReduce in the way data is grouped and shuffled between nodes. Its programming model extends MapReduce with additional operations like joins and sorts and allows explicit control of data distribution with repartitions. Spark dataflows can also be composed of more steps than just map and reduce. The grouping of those steps between the group by/shuffle phases is optimised by the framework. These features can be used in future versions of our system to extend it with post-processing of the results.

This paper is an extended version of [15] and presents a new caching architecture that improves efficiency. Here we also propose further inter-cellular communication mechanisms and discuss their experimental evaluation. We demonstrate the capabilities of our distributed algorithms by simulating the communication of liver cells through FGF19 cytokine during

Please cite this article in press as: K. Kędzia et al., Simulation of multicellular populations with Petri nets and genome scale intracellular networks, Sci. Comput. Program. (2017), https://doi.org/10.1016/j.scico.2017.10.002