ON EXECUTABLE MODELS OF MOLECULAR EVOLUTION

*Marek Kwiatkowski** *and Ian Stark*

School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
Marek.Kwiatkowski@eawag.ch, Ian.Stark@ed.ac.uk

ABSTRACT

Systems biology can provide unique insights into molecular evolution, but most existing work in this domain uses one-off methods and tools. In this paper, we advocate development of generic frameworks for evolutionary systems biology and discuss in some detail key characteristics it should possess to be most applicable and useful. We offer one such framework ourselves, and evaluate it with a case study of a mitogen-activated protein kinase cascade.

1. INTRODUCTION

Recent years have witnessed an increased interest in addressing evolutionary questions using the techniques of systems biology [1, 2]. At the heart of these efforts lies the realisation that the key evolutionary duality between genotype and phenotype is akin to the relationship of a formal model to the outcome of its execution. In other words, a formal model can serve as a proxy for genotype and its execution as a proxy for development, i.e. the determination of phenotype. With reactive computational models used by systems biology one can thus gain insight into the mechanics of development and the causal links between genotype and phenotype, all of which are inaccessible to the established purely mathematical formalisms (cf. [3]).

In this paper we focus on evolutionary systems biology of cellular processes, in particular signalling and regulatory pathways. Typically, the first step of a theoretical research project in this domain consists of setting up a class of formal models of molecular networks. Once the appropriate class of models is defined, one can perform *in-silico* evolution according to a predefined fitness function and mutation schemes, explore the distribution of evolutionary properties of interest over this class (usually by sampling), or both. Recent examples of such work include: simulated evolution of the MAPK cascade [4] and a chemotaxis pathway [5], classification of small networks according to their response to a standardised signal [6], analysis of the causes and mechanisms of canalisation [7] and redundancy [8] in regulatory networks.

All studies reported above employ different classes of models of molecular networks, usually based on the notion of a graph (or, equivalently, a matrix), and purpose-built to study a concrete biological problem. The important exception is the MAPK study of Dematté et al. [4], who used idealised computer programs instead of graph nodes to represent molecules (cf. §2), and presented their work as a generic framework rather than one-off method. We believe that such standardised approaches have the potential to greatly advance evolutionary systems biology, for they provide a common platform on which to study and compare evolution of different systems. In this paper we analyse key requirements that such a platform should satisfy and propose our own prototype.

Let us loosely define a *framework for evolutionary systems biology* (*framework* for short) to mean a modelling formalism for molecular networks together with formal notions of model transformation (i.e. mutation) and model execution (i.e. development). While this definition may appear narrow in view of the great diversity of evolutionary research, it does capture a significant portion of the evolutionary system biology studies of which we are aware, including all works cited above. In our opinion, in order to be maximally applicable and useful, a framework should satisfy the following requirements:

(1) Be *agent-centric* rather than reaction-centric,

(2) Support *dynamic complex formation*,

(3) Execute individual models *deterministically*, but

(4) Admit a *variety of secondary execution modes*.

The agent-centric (e.g. boolean networks [9]) and reaction-centric (e.g. kappa [10]) modelling approaches differ in what they consider to be first-citizen entities: biochemical substrates or biochemical reactions. The requirement that a framework be agent-centric is an expression of the fact that all reactions arise from the properties of individual molecular agents, in particular DNA/RNA sequences and 3D shapes. The step from agents to reactions, therefore, is already a non-trivial part of the genotype-phenotype translation. Moreover, and most importantly, evolutionary transformations of reactions always proceed through alterations of individual substrates, and hence a definition of faithful mutation schemes in a reaction-centric framework is likely to be problematic. While it is certainly not impossible in principle to overcome these difficulties, we wish to stress that they have to be explicitly addressed by any reaction-based approach.

Dynamic complex formation is the ability of two agents in a model to form a molecular complex that has not been

*Present address: Swiss Federal Institute of Aquatic Science and Technology (Eawag), Überlandstrasse 133, 8600 Dübendorf, Switzerland.

pre-specified by the modeller. The task of specifying not only the agents that are present in the system's initial state, but also all the *potential* complexes of agents is often impractical for models of cellular signalling, where complexation is ubiquitous, and may be downright impossible in the case of evolutionary modelling, where the modeller would in principle have to account for all evolutionarily possible complexes. This problem is tightly linked to the phenomenon known as *combinatorial explosion*, and our requirement could be rephrased in these terms.

The preference for deterministic over non-deterministic or stochastic dynamics stems from one of the primary applications of the postulated framework, namely sampling of the space of models. As a rule, deterministic investigation of molecular dynamics is orders of magnitude less expensive than the corresponding stochastic one; in our case this disparity translates directly to orders of magnitude differences in practicable sample sizes. Ideally, however, a framework should support various ways of executing a given model, including qualitative as well as continuous and stochastic, in order to facilitate the study of evolution at different points of the accuracy-cost trade-off and detailed analyses of isolated models of particular interest.

2. A CANDIDATE FRAMEWORK

In the late 1990s, A. Regev and E. Shapiro realised that molecular dynamics can be successfully modelled by the so-called *process algebras*—a family of formal languages used in computer science to study concurrent computing systems [11, 12]. The key correspondence they identified was between *processes* (short, idealised computer programs) and molecules, and between the independence of concurrent computing threads and the spatial independence of molecules in the cell. Since Regev's work, many existing process algebras have been used in biochemical modelling, and many other have been designed from scratch to tackle specific aspects of biological complexity.

In this section we present a framework for evolutionary systems biology based on the *continuous* π-calculus ($c\pi$), a process algebra which we have designed specifically for this purpose. Unfortunately, for reasons of space we cannot make the present treatment self-contained and we are forced to make certain simplifications; instead, we attempt to provide sufficient intuitions to follow the rest of the paper and refer to the comprehensive presentation elsewhere [13, 14].

2.1. The continuous π-calculus

A basic notion in $c\pi$ is a *species*, a formal entity corresponding to a kind of a biochemical substrate, and defined as follows:

$$A, B \ ::= \ \mathbf{0} \ \mid \ D(\vec{a}) \ \mid \ \Sigma_{i=0}^{n} \pi_i.A_i \ \mid \ A|B \ \mid \ (\boldsymbol{\nu} M)A$$

Hence, the are five kinds of species: the *inactive form* $\mathbf{0}$; the *recursive call* $D(\vec{a})$ (always accompanied by an equation of the form $D(\vec{b}) \stackrel{\Delta}{=} B$ and behaving roughly like B); the *choice* $\Sigma_{i=0}^{n} \pi_i.A_i$, where π_is are mutually exclusive atomic actions, after which the species becomes the appropriate A_i; the *parallel composition* $A|B$, where A and B proceed concurrently; and finally, the *restriction* $(\boldsymbol{\nu} M)A$, which limits the interaction potential of A to those actions that are not listed in M.

From species modelling individual substrates we build *processes*, which model biochemical solutions:

$$P, Q \ ::= \ c \cdot A \ \mid \ P||Q \quad ,$$

where $c \cdot A$ is species A at concentration c, and $P||Q$ is a mixture of P and Q.

In Regev's abstraction the active sites of molecules are represented as *names*, which are components of the π_i actions and thus of species. This interpretation is followed in $c\pi$, where we specify the properties of names, and consequently of active sites, using an undirected weighted graph termed the *affinity network*. Two names linked in the affinity network by an edge denote two interaction sites that can interact according to the mass-action principle, with the rate constant given by the edge label.

Finally, a complete $c\pi$ model is a triple $(\mathcal{D}, P, N)$, where P is a process describing the initial state of the system, N is an affinity network specifying the complementarity and binding affinities of the active sites of the agents in the model, and $\mathcal{D}$ contains the definitions of all species appearing in P. A prototype software tool has been implemented to extract Ordinary Differential Equations (ODEs) from such models.

2.2. Variation operators

We have designed 11 formal transformations of $c\pi$ models, which we have termed *variation operators*. The operators correspond to mutations commonly occurring in biochemical networks, including gene duplication and loss, evolution of interaction and binding and regulatory changes. Formally, variation operators are inference rules; for example the operator RATE-SITE modelling a mutation of the active site of a molecule is rendered as:

$$\frac{a \in N \qquad f\colon \ N \to \mathbb{R}_{\geq 0}}{(\mathcal{D}, N, P) \longrightarrow (\mathcal{D}, N \odot_f a, P)}\ \text{RATE-SITE}$$

The above rule states that the affinity network of any $c\pi$ model can be altered by changing the connectivity of one name and one name only (a). The new affinities of a are encoded by a real function f, and the new affinity network is formally obtained with the help of a custom operation $\odot$. Observe that the constituent molecular agents ($\mathcal{D}$) and the initial state of the system (P) remain unchanged in this transformation and that all modified affinities involve a. Hence, the RATE-SITE operator precisely models the mutation of a single active site.

For full discussion of the remaining ten variation operators we refer once more to their presentation in [14].

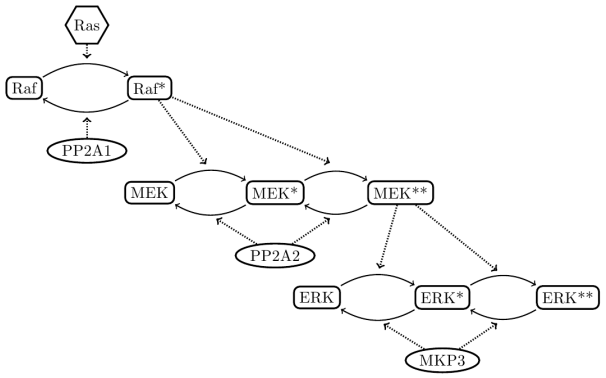

Figure 1. Structure of the MAPK cascade. Boxes are kinases, ovals are phosphotases, and Ras is the input signal; asterisks denote phosphorylation levels.

3. EVOLUTION OF THE MAPK CASCADE: A CASE STUDY

3.1. System

The mitogen-activated protein kinase (MAPK) cascades are important components of many signal transduction pathways. Found in all eukaryotes, they help to control a number of cellular processes, most notably cell growth and cell division. Here, we restrict our analysis to a subfamily of MAPK architectures considered in [15] (Figure 1), and use protein names specific to the human MAPK. The initial signal (Ras) promotes the activation (phosphorylation) of the order 3 protein kinase (Raf). Once activated, Raf (now $\texttt{Raf}^*$) acts as a catalyst for the phosphorylation of the order 2 kinase (MEK). Doubly phosphorylated MEK activates (again, twice) the order 1 kinase (ERK), whose fully activated form ($\texttt{ERK}^{**}$) is considered the output signal of the cascade. Every kinase has a corresponding phosphatase, which performs the opposite action, namely dephosphorylates its target. This multi-tiered architecture (i.e. three different kinases) promotes sensitivity to the input signal and reduces response time; as a result, the pathway is a fast, sensitive, amplifying relay. The MAPK cascade is among the most often modelled and best-understood signalling systems and often serves—as it does here—as a benchmark for new systems biology techniques.

3.2. Setup

We modelled the cascade in $c\pi$ using 12 species definitions corresponding to 12 distinct protein species in Figure 1, an affinity network with 16 nodes modelling 16 active sites on the surfaces of these proteins, and initialised the model with a process containing Ras, the three inactive kinases (Raf, MEK and ERK) and the three phosphotases (PP2A1, PP2A2 and MKP3) in biologically realistic relative amounts. Incidentally, this model can be built incrementally from the null model by applications of variation operators alone.

The RATE-SITE variation operator was then used to generate $16 \times 2^{16} = 1048576$ variants of this model, corresponding to effects of all 65536 qualitatively different mutations for each of the 16 sites, and thus constituting a sample of models of the evolutionary neighbourhood of the initial system. The resulting models were translated to sets of ODEs and solved on a parallel cluster.

The time-series $e = (e_i)_{i=0}^{720}$ of the output signal of the cascade (concentration level of $\texttt{ERK}^{**}$) was the basis for two types of analysis: fitness distributions and signal classification. Under our assumptions regarding the input signal (Ras), a functional cascade should produce a peak of the $\texttt{ERK}^{**}$ concentration at around 200 time units. Inspired by the work of Dematté et al. [4], we defined the fitness of a cascade as

$$\mathsf{fitness(e)} \stackrel{\text{df}}{=} \sum_{i=0}^{135} e_i - \sum_{i=302}^{720} e_i \quad ,$$

where the cut-off time points 135 and 302 correspond to the exponentially decaying input signal reaching 1/16 and 1/256 of the initial strength. Thus, the formula above rewards (the first sum) a quick and strong response to the initially strong signal and simultaneously punishes (the other sum) a late or incomplete reaction to the input signal falling to negligible levels.

In addition to the assessment of fitness, the general shape of the output signal was classified into four categories: peak (signal starts low, reaches a high level, then falls low again), switch (starts low, reaches a high level and remains there), oscillatory (two or more consecutive peaks) and noise (all other signals); here we are inspired by a study of Soyer et al. [6]. The classification itself was performed using LTL model checking [16], a common analysis technique for process algebras, here adapted to the degenerate case of fully linear transition systems (i.e. time series).

3.3. Results

Over 45% of the analysed cascade variants exhibit the switch phenotype, while a further 7% maintain the peak characteristic, suggesting a degree of robustness of the ostensibly fragile and refined MAPK architecture (cf. Figure 1). Not a single one of the variants showed oscillatory dynamics.

The analysis of fitness revealed that a great majority (almost 96%) of mutations were deleterious, with a significant fraction of mildly deleterious ones, which is in broad agreement with a body of experimental and theoretical data on distributions of mutational effects. Among the 16 fitness distributions corresponding to collections of mutants of each of the 16 active sites in the base model, many exhibited pronounced low-fitness peaks (Figure 2). We readily established that the cause of the peaks was a loss of one or more functional phosphotases, either by a direct loss of function, or by engaging in spurious interactions with other molecules. On the other hand, higher than base fitness was usually due to the acquisition of a feed-forward architecture (e.g. $\texttt{Raf}^*$ catalysing the phosphorylation of ERK).

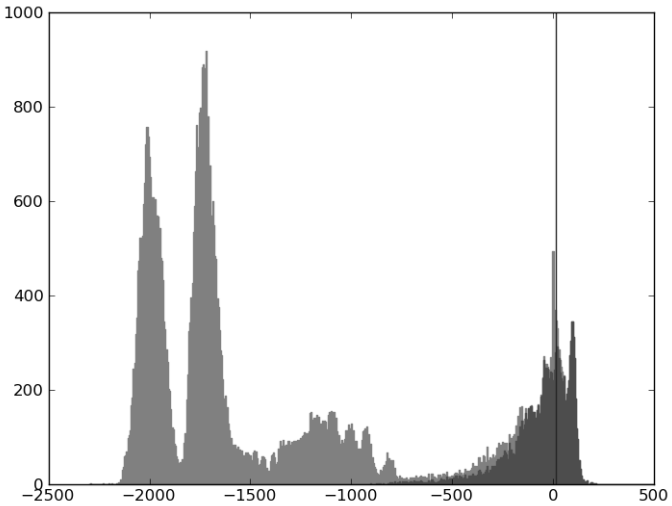

Figure 2. Distribution of fitness of mutants of the active site in MEK. The darker area contains the peak variants; the vertical line shows the position of the base model.

4. DISCUSSION

We have offered a framework for evolutionary systems biology based on a custom-built process algebra, possessing all four properties discussed in §1. All of them conferred significant advantages during the analysis of the MAPK cascade. Thanks to the ability to dynamically form complexes, it was not necessary to specify any of the transitional aggregations of proteins (of which there are 10 in the base model, and up to 25 in the variants). The use of ODE dynamics ensured that the batch of over 1 million models, each in tens of variables, could be efficiently processed. We have used two different analysis techniques, yielding different yet complementary insights, without the need to redefine the underlying models. Finally, the agent-centred perspective has enabled us to vary individual molecules while remaining agnostic on the impact this variation may have on molecular interactions.

Our study revealed that our process algebra-based framework, while very promising overall, has also several drawbacks. Most importantly, the π-calculus and its derivatives tend the be too expressive for biochemical modelling, in the sense that it is possible to give well-formed models that do not meaningfully represent any biochemical network. It is a particularly acute issue in the case when model transformations have to be defined, for it is of utmost importance that they generate no nonsensical models as variants of a legitimate one. These considerations underlie the very conservative character and the considerable degree of syntactical complexity of our variation operators [14].

We conclude that process algebras may indeed provide the basis for a mature framework for evolutionary systems biology. The biggest immediate challenge in this line of research is surely implementing dynamic complex formation in a setting which does not admit biologically irrelevant models.

References

[1] L. Loewe, "A framework for evolutionary systems biology," *BMC Systems Biology*, vol. 3, 2009.

[2] A. Wagner, "Neutralism and selectionism: a network-based reconciliation," *Nat. Rev. Gen.*, vol. 9, pp. 965–974, 2008.

[3] J. Fisher and T. Henzinger, "Executable cell biology," *Nat. Biotechnol.*, vol. 25, pp. 1239–49, 2007.

[4] L. Dematté et al., "A formal and integrated framework to simulate evolution of biological pathways," 2007, vol. 4695 of *Lect. N. Bioinformat.*, pp. 106–120.

[5] O. Soyer et al., "Simulating the evolution of signal transduction pathways," *J. Theor. Biol.*, vol. 241, pp. 223–232, 2006.

[6] O. Soyer et al., "Signal transduction networks: Topology, response, and biochemical reactions," *J. Theor. Biol.*, vol. 238, pp. 416–425, 2006.

[7] Mark L. Siegal and Aviv Bergman, "Waddington's canalization revisited: Developmental stability and evolution," *P. Natl. Acad. Sci. USA*, vol. 99, no. 16, pp. 10528–10532, 2002.

[8] E. Borenstein and D. Krakauer, "An end to endless forms: Epistasis, phenotype distribution bias and non-uniform evolution," *PLoS Comput. Biol.*, vol. 4, 2008.

[9] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, pp. 437–467, 1969.

[10] V. Danos and C. Laneve, "Formal molecular biology," *Theor. Comput. Sci.*, vol. 325, pp. 69–110, 2004.

[11] A. Regev, *Computational Systems Biology: A Calculus for Biochemical Knowledge*, Ph.D. thesis, Tel Aviv University, 2002.

[12] A. Regev and E. Shapiro, "Cellular abstractions: Cells as computations," *Nature*, vol. 419, pp. 343, 2002.

[13] M. Kwiatkowski and I. Stark, "The continuous π-calculus: A process algebra for biochemical modelling," 2008, vol. 5307 of *Lect. N. Bioinformat.*, pp. 103–122.

[14] M. Kwiatkowski, *A formal computational framework for the study of molecular evolution*, Ph.D. thesis, The University of Edinburgh, 2010.

[15] C. Y. Huang and J. E. Ferrell, "Ultrasensitivity in the mitogen-activated protein kinase cascade," *P. Natl. Acad. Sci. USA*, vol. 93, pp. 10078–10083, 1996.

[16] C. Baier and J.-P. Katoen, *Principles of Model Checking*, MIT Press, 2008.