



Project no. FP6-NEST-2003-1-12789  
ESIGNET  
Evolving Cell Signalling Networks in Silico

Specific Targeted Research Project  
Sixth Framework Programme Priority

Deliverable number 4.1  
Specification of the Properties of Cell Signalling Systems  
Document Describing Formats for Phenotypic Representation

Due date of deliverable: May 2006  
Actual submission date: May 2006

Start date of project: 2005-09-01

Duration: 36 months

Friedrich Schiller University Jena

Revision: final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
<b>PU</b>	Public	X
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission Services)	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Abstract

The overall goal of the ESIGNET project is to study the computational properties of cell signalling networks (CSN) by evolving them using methods from evolutionary computation, and to re-apply this understanding in developing new ways to model and predict real CSNs.

Finding appropriate possibilities to denote and to describe the structure as well the behaviour and resulting properties of CSNs is essential for all subsequent parts of the ESIGNET project. In this report, we summarise different methods and strategies for phenotypic representation and specification of CSNs. Spanning the range from analytical to algebraic and category based approaches, they follow different purposes. Bridging the gap between these description models facilitates a high degree of flexibility in their choice and usage. File format specifications of computer science arisen from these approaches allow to implement interoperable software packages for construction, evolution, analysis, and prediction of CSNs.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Analytical Approaches</b>	<b>3</b>
<b>3</b>	<b>Stochastic Approaches</b>	<b>3</b>
<b>4</b>	<b>Algebraic Approaches</b>	<b>4</b>
<b>5</b>	<b>Category Based Approaches</b>	<b>5</b>
<b>6</b>	<b>Computer Languages</b>	<b>5</b>
<b>7</b>	<b>Bridges between Approaches</b>	<b>8</b>
<b>8</b>	<b>Conclusions</b>	<b>8</b>

# 1 Introduction

The ESIGNET workpackage (WP) 4 includes research activities leading to formal specifications of CSNs and their evaluable properties. In fulfilment of this, WP4 is thought to focus on following steps:

- Develop suitable ways to describe the computational properties of CSNs.
- Define phenotypic representations of a CSN and its dynamical behaviour.
- Implement the phenotypic representation.
- Find a suitable measure for distance between the computational properties of two CSNs.
- Explore fast methods to derive a fitness function given a description of a CSN and the specification of a target CSN.
- Development of user interface software for specification of CSN properties

Objectives within WP4 are specified by five deliverables:

- D4.1.** Document describing the formats for the phenotypic representation of CSNs and their dynamical behaviour.
- D4.2.** Document describing the framework to specify the desired computational properties of the target solution.
- D4.3.** Software that performs simulation, reads specification of CSN properties, and calculates fitness for a given CSN and a target CSN.
- D4.4.** Publication in peer-reviewed journal and/or conference.
- D4.5.** Software to automate the input of the specification of target CSNs.

Modelling phenomena from nature is a basic motivation for development and refinement of mathematics. Beside traditional natural, engineering, and social sciences, modern systems biology requires exhaustive application of mathematical principles. Success in genomics and proteomics can discover interconnections between structure of biomolecules and function of biological systems formed by those biomolecules. CSNs can be seen as a special class of biological systems consisting of interacting proteins and auxiliary substances for the purpose to organise information processing inside living organisms. Identifying what kind of information are encoded by cell signals and finding the way how cell signals are generated, transferred, modified, and utilised should be reflected by mathematical models of CSNs. Furthermore, formal descriptions define a certain level of abstraction. High abstraction levels are suitable for recognition of general conclusions, but they weaken the view into details of system components. Low abstraction levels retrieve details about selected system components, but they produce a huge amount of low structured data. We intend to serve both by a variety of different approaches to model CSNs.

Since this report should correspond to deliverable D4.1, it is structured as follows: The subsequent sections introduce a selection of analytical, stochastic, algebraic, and category based approaches. A separate section is addressed to special purpose computer languages regarding to CSNs. Bridges between different approaches enable some model transformations, partially including increase of abstraction level. They are discussed in the last section.

## 2 Analytical Approaches

Grouped under this headline are all modelling approaches that are based on differential equations, one of the most widely used mathematical modelling techniques. In this view, the state of the CSN is expressed in terms of concentrations of its molecular species, which are positive real numbers. Using differential equations also implies that the progression of time is represented as a movement along the positive real axis, so that the concentrations of molecules can be calculated at any given point in time. Therefore, analytical descriptions have to be discretised in order to be simulated on a computer, a task for which software libraries are readily available.

Models employing differential equations can be grouped into three subgroups according to their treatment of space [3]. Pure chemical kinetics systems completely disregard any spatial aspect, while compartmental models couple a set of non-spatial systems in order to achieve a coarse-grained spatial resolution. If fine-grained spatial information is needed, diffusion-reaction systems are the models of choice.

If the system under consideration can be understood as a well-stirred reactor, one can ignore spatial aspects and use ordinary differential equations (ODEs) only. The basic equation describing the change in concentration  $C$  of a molecular species is

$$\frac{dC}{dt} = (\textit{generation}) - (\textit{consumption}),$$

where the term on the right hand side sums up all effects of reactions producing  $C$  minus the effect of all reactions consuming  $C$ . Numerical methods for solving a system of ODEs are well developed, and the formulation of such a system from a biological context is relatively easy. Therefore, ODE models are heavily used in modelling cell signalling networks.

Compartmental models combine a set of ODE models supposed to describe different compartments of the cell, and combine these by adding flux reactions between the compartments. In this framework, space can be coarsely resolved, but the computational advantages of the non-spatial approach are kept. In contrast, partial differential equations (PDEs) address space explicitly. The concentration now depends on both time and space, where movement in space is separated into diffusion and convection:

$$\frac{dC}{dt} = D \frac{\partial^2 C}{\partial x^2} - v \frac{\partial C}{\partial x} + (\textit{generation}) - (\textit{consumption}).$$

This framework allows to model phenomena like wave propagation and pattern formation, but it heavily increases the amount of numerical work required for solving the equations. In the context of CSNs, PDEs are currently used only to model explicitly spatial processes.

## 3 Stochastic Approaches

Due to the small number of molecules involved in some signalling processes, their approximation as continuous processes via differential equations cannot always be valid. In contrast to an analytical approach, stochastic models explicitly account for the uncertainty that is involved in molecular processes, and allow to make predictions not only about the average behaviour of a system, but also about its standard deviation from that behaviour. However, this knowledge comes at a cost: stochastic processes tend to be more expensive to simulate on a computer, and their mathematical analysis is usually not straightforward.

In stochastic modelling, two fields can be distinguished. On the one hand, stochastic simulation algorithms (SSA) have been developed in which reactions have associated kinetic rates depending on the concentration of their substrates, similar to ODE models. Here, the kinetic rate of a reaction is interpreted as the probability of its reaction taking place in a certain time interval. Such approaches, based on the chemical master equation and pioneered by Gillespie [4], are very popular in stochastic biochemical modelling (see [15] for a review).

On the other hand, biochemical systems have been interpreted as Markov chains, in which the state of the chain is represented by the number of molecules present, and reactions are modelled as transitions between these states. As long as there is no feedback in the system, the analysis of Markov chains is well developed and information can be gained on the steady-state probability distribution of the process. Feedback, which is an inherent feature of many CSNs, poses problems for the analysis since a steady-state distribution of the system does not have to exist in this case. An interesting derivation from the usual approach is explored in [14], where a set of interacting Markov chains is used to represent interacting multi-state proteins.

## 4 Algebraic Approaches

Algebraic approaches mainly come from theoretical computer science. They have in common the assumption of a finite or recursive enumerable number of objects. Each object is considered as the smallest unit that can be handled by the system model. Both biomolecules and processes can form objects depending on the type of the algebraic approach. Interactions between objects or additional parameters about objects are specified by sets (relations) of acceptable system configurations. The whole description is based on discrete transitions. This allows structural and comparative analysis of system composition and behaviour independent of numeric simulation results. In terms of their computational properties, term rewriting systems, state based systems, and process calculi can be distinguished.

### Term Rewriting Systems

Controlled term rewriting is a basic principle of information processing. Biomolecules, their polymeric subunits or groups of similar biomolecules are interpreted as objects encoded by character strings (terms). Sets of term rewriting rules describe possible interactions among objects and system components like pathways or membrane structures. Each application of a rule performs a discrete step of a process. The terms as a whole contain all information about the system status. Term rewriting processes can run in a massively parallel manner considering nondeterministic recombinations. Classes of *grammar systems* and *P systems* ([12]) exemplify representations of CSNs.

### State Based Systems

While term rewriting systems hold the whole system status only using terms, state based systems model a finite external storage unit that can toggle between predefined values. System steps are performed by state transitions following transition functions or tables. Beside the current state, they often evaluate additional stepwise input data coming from outside and/or from an underlying persistent memory. The system cannot leave its state space. Sequences of state

transitions encode processes, their effects, and computations. States are interpretable in the sense of semantics, e.g. forming final states or error states. One of the oldest state based systems is the universal *Turing Machine*. Classes of *abstract machines* and *X machines* [2] exemplify representations of CSNs.

## Process Calculi

Process calculi facilitate formal descriptions of concurrent systems. They are constructed in a way to model interaction, communication, and synchronisation between groups of independent processes (also called agents). Objects from which process calculi are composed of are derived directly from the processes. Assigned properties complete a process denotation. Dependency structures between processes form a network topology with dynamical behaviour. Following certain transformation rules, the network of processes can be analysed. An example for a general purpose process calculus is *CSP* (Communicating Sequential Processes, [6]). Instances of *Petri nets* ([13]),  $\pi$  *calculus* ([10]), and the *ambient calculus* ([1]) exemplify representations of CSNs.

## 5 Category Based Approaches

The theory of categories provides a general theory of mathematical structures. CSNs, their components, behaviour, and properties can be seen as examples for abstractions of selected mathematical structures. In contrast to traditional algebraic approaches, relations between system objects are defined by comparison operators inside or between classes of similar system elements (categories). Morphisms, homologies, and functors act as comparison operators. Processes are encoded by sequences of natural transformations. Facets of similarities between biomolecules, their properties, and processing mechanisms can be modeled by category theory. This leads to systems of *description logics*, *formal semantics* as well as *domain theory* and methods for *model checking*, *formal specification and verification* on various levels of abstraction ([9]).

## 6 Computer Languages

If the ever-growing number of cell signalling models is to be handled and put to good use, the modelling community will have to reach standards for describing, storing and exchanging them. Such a format has to facilitate analysis, visualisation, and simulation, and it has to provide easy ways of refinement and incorporation of new knowledge. So far, two approaches have emerged, resulting in the model-description languages SBML (Systems Biology Markup Language) [7] and CellML [8], both based on the XML markup language.

In SBML, a biochemical network is described in terms of the molecules taking part in it - termed species - and the reactions taking place between them. The present amount of each species can be expressed either in terms of its concentration or of the number of molecules present. Each reaction has an associated kinetic law, which defines the rate of the reaction depending on the present amount of its substrates. Additionally, the model can be subdivided into a set of compartments to include a spatial component. In CellML, a more general approach is taken, in which a model consists of components and connections between components. Each component can contain variables and a reaction between them, and connections are used to transfer the value of variables from one component to another.

Although CellML is following a slightly more general approach, it is not as widely used as

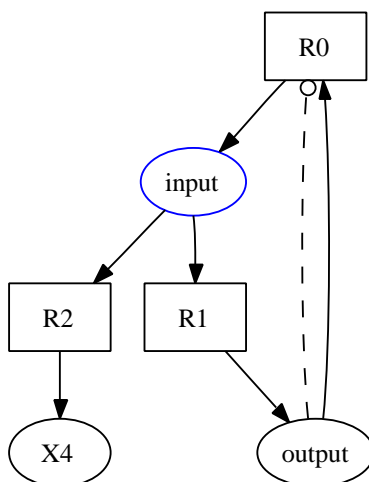


Figure 1: The digraph representation of a CSN model computing the squareroot of a real number.

SBML, for which a large collection of software tools is available (see [www.sbml.org](http://www.sbml.org) for a list of these tools). Additionally, the first model repositories have started to use SBML as a representation language, e.g. see the BIOMODELS database at [www.ebi.ac.uk/biomodels](http://www.ebi.ac.uk/biomodels). Therefore, SBML can be seen as the first emerging specification standard for biological models at the cellular level. Due to its widespread acceptance, we aim to use SBML as input/output format for the software developed within the ESIGNET project.

## SBML: An Example

To illustrate the way in which biochemical models are represented in SBML, we present an evolved network that computes the square root of a real positive number (see figure 1). It consists of three chemical species, of which the third,  $X_4$ , is a redundant artifact of the evolutionary algorithm that created the network. The concentration of species *input* serves as input into the network, and the steady state concentration of *output* represents the result of the computation. The node *input* is set to be constant. Written as differential equations, the model is given in the following form:

$$\begin{aligned} \frac{dinput}{dt} &= 0 \\ \frac{dX_4}{dt} &= k_2 \cdot input \\ \frac{doutput}{dt} &= k_1 \cdot input - k_0 \cdot output^2 \end{aligned}$$

While that first equation states that *input* is constant, the second one constitutes a redundant production of  $X_4$  and can be ignored. The third equation bears the desired result: it can easily be checked that for  $k_0 = k_1$ , the value of *output* actually approaches  $\sqrt{input}$  in the steady state.

In SBML, the model is described by a list of its components (compartments, species) and interactions between them (reactions). Specifically, the squareroot model described above is given in figure 2.

Of the large array of software tools available for SBML, most can only deal with a subset of the language. Currently, this is also true for our software specification, which cannot deal

```

<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level2" level="2" version="1">
  <model>
    <listOfCompartments>
      <compartment id="uVol" size="1"/>
    </listOfCompartments>
    <listOfSpecies>
      <species id="input" compartment="uVol" initialConcentration="16"
        boundaryCondition="true"/>
      <species id="output" compartment="uVol" initialConcentration="0"/>
      <species id="X4" compartment="uVol" initialConcentration="14"/>
    </listOfSpecies>
    <listOfReactions>
      :
      <reaction id="R1" reversible="false">
        <listOfReactants>
          <speciesReference species="input"/>
        </listOfReactants>
        <listOfProducts>
          <speciesReference species="output"/>
        </listOfProducts>
        <kineticLaw>
          <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
              <times/>
              <ci> k </ci>
              <ci> input </ci>
            </apply>
          </math>
          <listOfParameters>
            <parameter id="k" name="0_100000000_" value="7.0370457"/>
          </listOfParameters>
        </kineticLaw>
      </reaction>
      :
    </listOfReactions>
  </model>
</sbml>

```

Figure 2: SBML source code for the squareroot model. Reactions  $R_0$  and  $R_2$  are skipped for brevity.

with events and ignores function definitions. However, all the main features are included (see the upcoming ESIGNET report on software specifications).

In order to gain access to established and new analysis techniques from computer science, we also consider algebraic representations of CSNs, based on the notion of P systems mentioned above. First results on a system capable of describing various features of cell signalling have been reached and are reported on in [5].

## 7 Bridges between Approaches

The aforementioned approaches for representation of CSNs unify different aspects of the view to biological systems. Since each approach is of particular interest to answer specific questions, a variety of system descriptions and denotations is used within the project. Most of the presentations can be transformed into each other by performing well defined transformation algorithms. For instance, SBML can be simulated via ODEs or SSAs, and there is a conversion of P systems into SBML (see [11]). Since each approach is assigned to an own paradigm and a specific balance between details of system components, network topology, object properties, transition methodology, process-related behaviour, and underlying assumptions, changing representation can lead to loss of information or requires additional information. This reflects different abstraction levels between the paradigms and intentions of the models.

## 8 Conclusions

Each of the sketched approaches focuses on specific facets of interest in modelling and analysis of CSNs. Ideas and resulting formalisms of description (formal systems and languages) connected with the approaches are well established in the scientific community. Although some of them are known for several decades, new applications and areas of usage arise constantly. This emphasises the importance of fundamental research for upcoming fields of scientific interest.

The aforementioned representations have shown their practicability in a plethora of application scenarios. By devising specialised instances, they can easily be adapted and handled as flexible tools. Depending on the current area of study, the range of representations can be used in different phases and workpackages. We abstain from giving a detailed introduction into denotations and nomenclatures of the representations here, but rather refer to primary literature mentioned as well as to publications in the context of the ESIGNET project and its members.

## References

- [1] L. Cardelli, A.D. Gordon. *Mobile Ambients*. LNCS vol. 1378, Springer-Verlag London, 1998
- [2] S. Eilenberg. *Automata, Languages, and Machines*. Academic Press New York, 1976
- [3] N.J. Eungdamrong, R. Iyengar. *Modeling Cell Signaling Networks*. *Biology of the Cell* 96, pp. 355-362, 2004
- [4] D.T. Gillespie. *Exact stochastic simulation of coupled chemical reactions*. *Journal of Physical Chemistry* 22, pp. 403-434, 1977
- [5] T. Hinze, T. Lenser, P. Dittrich. *A Protein Substructure Based P System for Description and Analysis of Cell Signalling Networks*. Submitted to the 7th Workshop on Membrane Computing, Leiden, 2006
- [6] C.A.R. Hoare. *Communicating Sequential Processes*. Prentice Hall International, 2004 (first edition 1985)

- [7] M. Hucka, A. Finney, B.J. Bornstein, S.M. Keating, B.E. Shapiro, J. Matthews, B.L. Kovitz, M.J. Schilstra, A. Funahashi, J.C. Doyle, H. Kitano. *Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project*. *Systems Biology* 1(1), pp. 41-53, 2004
- [8] C.M. Lloyd, M.D.B. Halstead, P.F. Nielsen. *CellML: its future, present and past*. *Progress in Biophysics and Molecular Biology* 85(2-3), pp. 433-450, 2004
- [9] S. Mac Lane. *Categories for the Working Mathematician*. Springer-Verlag Berlin, New York, 1998
- [10] R. Milner. *Communicating and Mobile Systems: the Pi-Calculus*. Cambridge University Press, 1999
- [11] I. Nepomuceno, J.A. Nepomuceno, F.J. Romero-Campero. *A Tool for Using the SBML Format to Represent P Systems which Model Biological Reaction Networks*. *Proceedings of the Third Brainstorming Week on Membrane Computing*, pp. 219-228, 2005
- [12] G. Păun. *Membrane Computing: An Introduction*. Natural Computing Series. Springer-Verlag Berlin, Heidelberg, 2002
- [13] J.L. Peterson. *Petri Net Theory and the Modelling of Systems*. Prentice Hall, 1961
- [14] M.R. Said, A.V. Oppenheim, D.A. Lauffenburger. *Modeling cellular signal processing using interacting markov chains*. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003
- [15] T.E. Turner, S. Schnell, K. Burrage. *Stochastic approaches for modelling in vivo reactions*. *Computational Biology and Chemistry* 28, pp. 162-178, 2004