# Simulating what cannot be simulated

Olaf Wolkenhauer*

Dagstuhl Position Statement

June 25, 2002

**Abstract**

The background to these notes is the expectation that mathematical modelling and simulation is going to play increasingly important role in the understanding of the organization and control of genetic-, metabolic-, and signalling pathways. I am going to argue that for modelling and simulation to help our understanding of cellular dynamics, the current practice of experimental design has to change - away from a 'mining' approach towards a signal- and systems-oriented methodologies. The emergence of systems biology has therefore as much to do with the development of new techniques as it is relying on a new 'way of thinking' about cellular systems. This argument leads us to the fact that there exist a principal limit (or uncertainty principle) to what we can achieve in simulation and with the machine metaphor in particular. Amongst the many modelling paradigms suggested for cellular systems, it is clear that none is accurate and yet general. I am therefore to discuss a conceptual framework that generalizes a number of models (including Bayes nets, state-space models, Boolean networks) and should allow us to discuss the previous issues in a formal framework.

**Keywords:** Systems and control theory, machine metaphor, cellular systems and dynamics.

## 1 A Systems Approach to Cellular Dynamics

Data analysis in the context of bioinformatics, to this day, has had little to do with modelling and simulation as a means to understand cellular processes. Transcriptomics and microarray data analysis in particular, provide an example of the current 'mode of operation' in which the primary objective has been 'gene hunting', i.e., the identification of (un)expected genes in a known context.

In this section, I try to argue for a systems approach to represent cellular systems. I believe that instead of trying to identify genes as causal agents for some change, function or phenotype, we should relate observations to *sequences of events* as it is systems dynamics that give rise to biological form and function. In terms of the analysis of experimental data, this implies a shift of focus from classification towards system identification.

*Remark.* The arguments in this section are not directed against comparative studies in the analysis of experimental data. Basic classification of array data, e.g. comparing cancerous with non-cancerous tissue, to identify drug targets or genes is of course a useful approach to diagnostics. The position taken here is to encourage the analysis of experimental data in support of modelling the dynamics and control of genetic pathways. In comparative studies and classification the results are usually a small piece in the jigsaw which represents

*Department of Biomolecular Sciences and the Department of Electrical Engineering & Electronics. Address: Control Systems Centre, UMIST, Sackville Street, P.O. Box 88, Manchester M60 1QD, UK. E-mail: o.wolkenhauer@umist.ac.uk, Internet: www.umist.ac.uk/csc/people/wolkenhauer.htm

our knowledge about cellular processes. In contrast, a systems biological approach employs system identification, modelling and simulation in an attempt to represent the underlying processes directly.

## 1.1 Intra- and Inter-Cellular Dynamics

The post-genome era of the life sciences witnesses a shift of focus away from molecular characterization and the classification of genes. In this section, I distinguish the organizational, descriptional, and experimental levels at which we investigate cellular systems in the post-genome era. The definition of these levels leads us directly to the key biological questions and research challenges.

The principal approach of modern biomedical research is to study cells, tissue, organs, organisms from the perspective of genes and gene interactions. With regard to these *organizational levels*, the two fundamental questions investigated are:

1. How do *genes* act and interact within the context of the *cell* as to bring about structure and function?
2. How do *cells* act and interact within the context of an *organ*, *tissue* or *organism* to generate organized and functional wholes?

The first question, considering *intra-cellular* dynamics, is related to, for example, transcriptional control. The second question, considering *inter-cellular* dynamics, is related to, for example, the development of an organism or a bacterial colony.

My claim is that it is *system dynamics*, not a genetic program, that gives rise to biological *form* and *function* (see also [2]). For experimental design this means that instead of trying to identify genes as causal agents for some change, function or phenotype, we should relate observations to *sequences of events*. Negative feedback is used in all cells and in metabolic pathways in particular. Control of such processes is achieved through regulatory enzymes that respond to changes in concentrations by increase or decrease in reaction rates. Cellular processes such as cell division, programmed cell death, responses to drugs, nutrients, and hormones are therefore *regulated* by complex *interactions* among large numbers of genes, proteins, and other molecules. To further our understanding of pathways, the fundamental problem is to understand the nature of this regulation. For example, cell signalling or 'signal transduction' is the study of the mechanisms by which biological information is transferred between and within cells. In this field, leading experts have made it clear that dynamic interactions, feedback control, and time delays are the key to an understanding of pathways and while pathway models have been static they need to be 'brought to life' if they are to become a tool or methodology for the biomedical scientist.

From my discussion of the fundamental questions of the life sciences, the following key research challenges arise for modelling and simulation:

1. Dynamic regulation and spatial organization: the need to capture both, spatial as well as temporal aspects simultaneously (spatio-temporal modelling).
2. Intra- and inter-cellular actions and interactions: the need for large-scale and hybrid-systems modelling and simulation.
3. Crossing organizational levels: from cells, to colonies, tissues, organs and organisms, ...
4. Integrating experimental levels: genome, transcriptome, proteome, metabolome and the physiome.
5. Combining data analysis and data management: The need to combine computational tools, developed for specific tasks and different organizational and descriptional levels.

6. Relating formal representations (mathematical models, e.g. Boolean networks and rate-equations). Providing a conceptual framework and theoretical foundations for the previous five points.

As a consequence of the described shift of focus, problems in the post-genome era of the life sciences will not only be experimental or technical but also conceptual. In the 'mining approach' mathematical or statistical models are a means to identify patterns or objects for further analysis and model building. (The latter being in the head of the scientist, not a formal or computational model). To model and simulate dynamic cellular systems as part of hypothesis driven research, the nature of the model and its semantics do matter. The machine metaphor is one popular account of cellular systems and is discussed next.

## 2 The Eternal Life of the Machine Metaphor

The practice of molecular biology and more recently virtual cell projects have once again brought the *machine metaphor* into discussion. Molecular biology has been reductionist to the extreme, focussing on the molecular characterization of cellular components ("the nuts and bolts") and thereby suggesting the cell to be a "chemical factory". Given the extraordinary illustrations that populate modern biology textbooks, it is not surprising that computer scientists naturally translate these concepts into "computations in cells" and thereby revive the machine metaphor.

In the following two sections I am going to argue that there exist a principal limit (or uncertainty principle) to what we can achieve in simulation and with the machine metaphor. I describe why we should look for an alternative to the machine metaphor as any research in this direction may literally become 're-search'.

*Remark.* The arguments in this section should not be seen as a criticism of virtual cell projects. These have of course their practical value in, for example, the testing of algorithms. The position taken here considers some principal limits and therefore more philosophical than practical issues.

### 2.1 Anticipatory and Self-Organizing Cellular Systems

The most concise argument against the machine metaphor, applied to cell systems, was provided by the philosopher Immanuel Kant (see also [2]):

- In *machines* parts exist for each other but not by each other; they work together to accomplish the machine's *purpose*, but their operation has nothing to do with building the machine.
- In *organisms*, each part is at once cause and effect, a means and an end. In organisms the parts not only work together but also generate and maintain the organism and all its parts.

From this follows that while a machine implies a machine maker, an organism is a *self-organising* system. Life is then an *emergent*, rather than an immanent or inherent, property of matter. Although it arises from the material world, it cannot be reduced to it. Or, in other words, a cell is built up of molecules, as a house is with stones. But a soup of molecules is no more a cell than a heap of stones is house.

If a cell is not a computer or doesn't work *like* a machine, one might think that it would not be possible to simulate it in or with a computer. This leads us to the question whether a computer (or Turing machine) can do what a natural system does or whether there are any limitations. Most of Robert Rosen's work was dedicated to demonstrate such limitations, particularly of modelling in the Newtonian realm. Amongst several books (e.g. [5]) in [4] the

concept of realizability in biology and physics is discussed in the context of Church's Thesis. While this discussion considers principal limits at a formal and abstract level, we of course do model and simulate cellular systems. Mathematical modelling and simulation has been very successful in engineering and the physical sciences but we can expect greater challenges for cellular systems. Some of the more practical difficulties and research challenges arising from them are discussed in the following section.

## 2.2 Cellular Weather Forecasting

In this section, I first define the complexity of cellular systems before arguing that modelling and simulation of cellular systems is likely to face the problems that hundreds of years in weather forecasting have not solved.

It is natural to consider the undisputed *complexity* of biological systems[1] as the main difficulty in modelling and simulating cell systems. To be more specific in our subsequent discussion I would therefore first clarify what is understood by the term complexity. I define complexity in the context of cellular systems as

1. A property of an encoding (mathematical model, e.g., its dimensionality, order or number of variables).
2. An attribute of the (sub-)system under consideration, e.g., the number of components, descriptive and organizational levels that ensure its integrity.
3. Our ability to interact with the system, to observe it, i.e., to make measurements and generate experimental data.

On all three accounts, genes, cells, tissue, organs, organisms and populations are individually and as a functional whole a complex system. Warren Weaver defined *disorganized complexity* as a problem in which the number of variables is very large, and any of these variables is best described as a random process. Here we are at the 'molecular level' and the most successful formal methods in representing phenomena at this level derive from statistical considerations. In the context of the cell, at the 'cellular level', matters are complicated by the fact that organization becomes an essential feature of the processes under consideration. Weaver referred to problems in which a large number of factors are interrelated into a whole as *organized complexity*. The number of variables is too large to be dealt with in the Newtonian realm of physics and mathematical modelling, and the systems are too organized to allow a statistical techniques either.

Modelling cellular systems we find that there exist a trade-off between the accuracy of our representation and its generality. To illustrate this, consider two well know problems: a) modelling the interactions of a large number of genes, and b) modelling transcriptional control in bacteria. We can summarize a comparison of the two problems as follows:

*Boolean gene-network simulation*

- ▷ general (independent of organism, deals with thousands of genes).
- ▷ not predictive for a particular biological system.
- ▷ simple model (runs on any PC).

*Biochemical interaction network*

- ▷ specific (parameter for an organism can be identified, feasible for only few genes).
- ▷ relevant only to a particular system, does not generalize well.
- ▷ complex model (requires very fast computer).

[1] As Schopenhauer said: Anyone can squash a bug but all professors in the world couldn't build one.

The principal limit suggested is in fact an uncertainty principle first outlined by Lotfi Zadeh:

> "As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost exclusive characteristics."

Gene expression and regulation takes place within the context of a cell, between cells, organs and organisms. The inevitable, reductionist approach is to 'isolate' a system, conceptually 'close' it from its environment through the definition of inputs and outputs, we inevitably loose information in this approach. (Conceptual closure amounts to the assumption of constancy for the external factors and the fact that external forces are described as a function of something inside the system). Different levels may require different modelling strategies and ultimately we require a common conceptual framework that integrates different models. For example, differential (mass-action or rate-) equations may provide the most realistic modelling paradigm for a single-gene or single-cell representation but cell-to-cell, and large-scale gene interaction networks could, for example, represented by logical or finite-state models, using agent-based simulation.

In dynamic systems theory, one would usually ignore spatial aspects. This approach is limited because both, space and time are essential to explain the physical reality of gene expression. The fact that the concepts of space and time have no material embodiment; they are not to be in the molecules or their DNA sequence; has been an argument against material reductionism. Although this criticism is in principle correct, alternative methods are in short supply. Without spatial entailment there can be no living cell and it seems that a major challenge for areas like for example systems biology is *the* challenge to capture or account for both, the behavior (dynamics) as well as the organization (structure).

In order to verify theoretical concepts and mathematical models we ought to identify the model from experimental data or at least validate mathematical models with data. In the context of post-genome technologies, the problem of complexity appears then in two disguises:

1. System Dimensionality: hundreds or thousands of variables/genes/cells.
2. Experimental Uncertainty: small samples (few time points, few replicates), imprecision, noise.

The data we currently have available, do not allow parametric systems identification techniques to build predictive models. A conclusion from this section is that mathematical modelling and simulation of cellular systems may have the same fate as weather forecasting: regardless of the computer power and time available, the predictions remain uncertain. However, even if we will never be able to build accurate predictive models of cellular or genetic systems, systems thinking and the modelling process itself will prove valuable to the biologist, helping him to identify which variables to measure and why. In fact, a common engineering experience is that we learn most from those models that fail! The quest for precision is analogous to the quest for certainty and both – precision and certainty are impossible to attain, at present if not in general.

## 3 Unifying Representations of Cellular Systems

The mathematical models that have been proposed for cellular systems (e.g. transcriptional control, gene interactions, metabolic pathways, etc.) are limited by the uncertainty principle described in Section 2.2. Given the complexity of cellular systems, it seems sensible to contemplate the combination of methodologies and principles, i.e., for example combining continuous- and discrete-time, with continuous/discrete-valued and symbolic dynamics, with rate equations and logical representations, with cost-functional and agent-based models.

In this section, I am to discuss a conceptual framework that generalizes a number of models (including Bayes nets, state-space models, Boolean networks). Such an abstract conceptual framework may allow us to discuss the issues raised in previous sections in a more formal context. The approach I have adopted for my research is in essence an extension of Robert Rosen's work (considered in the context of post-genomic technologies and genomic data) [5, 6, 7, 8].

### 3.1 Relational Biology

One might argue that scientific theories deal with concepts - not reality. Therefore if mathematical models are so formulated as to correspond in some 'useful' way to the real world, modelling and simulation is more an *art* than an objective discovery process (see Figure 1).

In this section I assume causation to be understood as the principle of explanation of *change* in the realm of matter. Causation is therefore a *relationship*, not between components, but between changes of *states* of a *system*. This perspective fits nicely into a systems-theoretic representation of cellular dynamics.

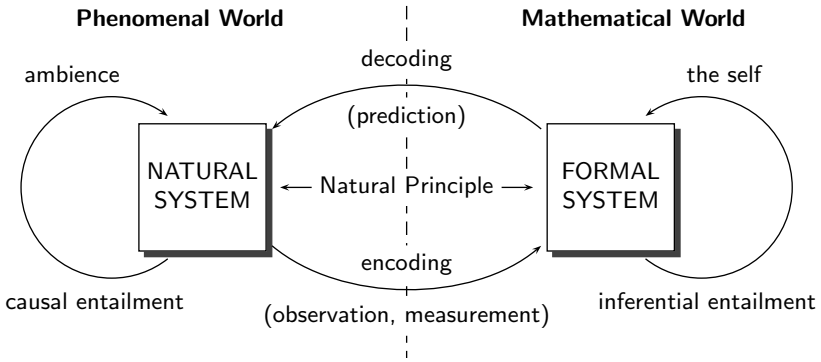



**Figure** 1: Rosen's modelling relation [5] between a natural system $N$ and a formal system $F$. If the modelling relation brings both systems into congruence by suitable modes of encoding and decoding, it describes a *natural principal*. In this case $F$ is a *model* of $N$, that is, $N$ is a realization of $F$.

A *system* is defined by a set or sets of (material or abstract) *objects* and *relations* defined on these sets [3, 7]. We denote these sets with uppercase letters such as $A$ and $B$ with elements $a \in A$ or $A = \{a\}$. The *mapping* $f\colon A \to B$ is then used as a representation of a cellular (sub-)system:

$$\begin{aligned} f : \quad & A \to B, \\ & a \mapsto b = f(a) \ . \end{aligned} \tag{1}$$

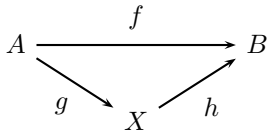
To model a particular cellular process, for example transcription, we have to discuss the selection of subsets of $A$, $B$ and identifying the map $f$. The exponential $B^A = \{f\colon A \to B\}$ denotes the set of all maps from $A$ to $B$ and in our example would therefore represent all transcription processes which are realizable by the cell. The transcription mapping $f$ can be subject to changes, either disturbances such as mutations or deliberate changes to the cell's operating conditions.

One way to be more specific about our previous formulation (1), is to view the element $a \in A$ as a sequence of events, expressed through an observed time-series. As $a$ entails $b$, we define the 'input' sequence $a$ and 'output' sequence $b$ as elements of finite-dimensional vector

spaces $A$ and $B$ respectively:

$$A = \{a\colon a = [u_0, u_1, \ldots]\}\,, \qquad u_t \in U,\ U = \mathbb{R}^m,$$
$$B = \{b\colon b = [y_1, y_2, y_3, \ldots]\}\,, \quad y_t \in Y,\ Y = \mathbb{R}^q\ ,$$

The space $A$ describes all possible (finite) input sequences to the system. The system has $m$ independent inputs and $q$ output variables. Mathematical causation is acknowledged by the fact that the first output appears one discrete time step after the first input. If $f$ is further assumed to be linear and constant, we can express the relationship between dependent and independent variables in a form that is familiar from dynamic systems theory [1, 8]. The measurements of input-output data $(a, b)$ describe the system in an *external* sense. The output of a system, in general, depends on both the present input $u_t$ and the past history of the system. To allow us to present *inner relations* we say, therefore, that the present output depends on the *state* of the system, and define the (present) state of the system as that part of the present and past history which is relevant to the determination of present and future outputs. A state is defined subsequently by a set of internal or *state variables* which must not necessarily be directly observable (measurable). The problem of explaining the internal dependencies, which generate the observed behavior, using a mathematical model called the *realization*. This concept is a straightforward extension of the input-output map $f\colon A \to B$ by adding a set of states and two new maps, $g$ and $h$ connecting this state-space with the input and output space:




(2)

As before, $a = [u_0, u_1, \ldots]$, $a \in A$, and $u_t \in U$ are measurements of the input variables at instances of time $t \in \mathbb{Z}_+$. Similar $b = [y_1, y_2, \ldots]$, $b \in B$ and $y_k \in Y$. The measured data we have available for the identification of system parameters is $\{(u_t, y_{t+1})\}$. In the diagram (2), $X$ denotes the *state-space* and the map $g$ is assumed to be surjective (onto $X$), i.e., more than one input sequence $a \in A$ can map into the same state $x \in X$. The output map $h$ is one-to-one (injective); i.e., any $x \in X$ maps to exactly one output sequence $b \in B$.

For the representation (2) to give rise to the observations $\{(u_t, y_{t+1})\}$, we must be able to construct the state-space $X$ and the maps

$$\begin{aligned} &\phi\colon X \times U \ \to\ X \qquad \text{such that} \qquad x_{t+1} = \phi(x_t, u_t) \\ &h\colon X \ \to\ Y \qquad\qquad\qquad\quad\ \ y_t = h(x_t)\ . \end{aligned} \tag{3}$$

In other words, given the present state $x_t \in X$ and input $u_t \in U$ the (nonlinear) map $\phi$ determines the next state and for every state $x$, the output map $h$ determines an output $y_t$. It is usually assumed that $X = \mathbb{R}^n$ and thereby any state can be represented as a point[2] in $X$. Note that the concept of state is a general notion, defining a set of $n$ state-variables such that the knowledge of these variables at some initial point in time $t = t_0$ together with the knowledge of the input for $t \geq t_0$ completely determines the behavior of the system for any time $t \geq t_0$. State variables need not be physically measurable or observable quantities.

State-space equations (3) form the basis for two well established conceptual frameworks: automata theory and control theory. An automaton is a discrete-time system with finite input and output sets $U$ and $Y$, respectively. In this context, $\phi$ is referred to as the next-state function. If at any time $t$ the system is in state $x_t$ and receives input $u_t$, then at time $t + 1$ the system will be in state $\phi(x_t, u_t)$. We say the automata is finite if $X$ is a finite set[3]. Automata theory has been used to model numerous systems including 'gene networks'. A

[2] A dynamical system is *finite dimensional* if $X$ is a finite dimensional linear space; it is *finite state* if $X$ is a finite set. If $X$, $U$, and $Y$ are finite sets and the system is discrete time, it is known as a (finite) automaton.

[3] The state of a linear dynamic system, continuous-time or discrete-time evolves in $\mathbb{R}^n$, whereas the state of an automaton resides in a finite set of 'symbols'.

state-space model or rate equations are obtained from (3) by considering (or assuming) a linear time-invariant system:

$$x_{t+1} = Sx_t + Gu_t \tag{4}$$
$$y_t = Hx_t \ , \tag{5}$$

## 4 Conclusions

No conclusions yet - this will hopefully follow from the meeting! My main objective for the Dagstuhl meeting is to learn more about different modelling paradigms and simulation strategies in order to further develop the mathematics of this/my research in systems biology.

## Appendix

For my discussion it will be useful to specify the meaning of some terms that I commonly use or refer to.

**System:** A *system* is defined as a set of objects and relations defined on them.

**Behavior:** A particular time-invariant relation specified for a system is called the *behavior* of the corresponding system.

**Organization:** If a system exhibits a particular behavior, it must posses certain properties producing the behavior. These properties will be called the *organization* of the system. Life resides in organization, not in material objects.

**Dynamics:** If the behavior of the system can change, the behavior is also referred to as the *dynamics*.

**Structure:** If the organization of a system is fixed, the organization is also referred to as its *structure*. For material systems (e.g. a cell) the structure is the specific embodiment of components into physical entities (e.g. molecules).

**Components:** To study natural systems using formal systems, we decompose a system into abstract objects, referred to as *components*. The organization and behavior of natural systems can be studied through interaction inducing a *change*. Discrepancies between behaviors determines its *function* while discrepancies between system structures determine its components.

**States:** In modelling natural systems,we define a set of state-variables such that the knowledge of these variables at some initial point in time together with knowledge of the input to some component determines the behavior of the component/system. The values of the state-variables at any particular point in time define the *state* of the system.

**Events:** Causality manifests itself through changes of states, called *state transitions*. The change of a particular state will be called an *event*. Causation is therefore not a relationship between things but a relationship between changes of states.

**Processes:** Sequences of events (changes of state over time) define a *process*.

**Genomics:** *Genomics* is the field of biological research, taking us from the DNA sequence of a gene to the *activity* of the product (usually a protein) for which it codes.

**Gene Expression:** *Gene expression* is the *process* by which information, coded in the DNA, is converted into proteins (hormones, enzymes, antibodies,...).

**Systems Theory:** *Systems theory* is a family of methodologies to formally represent *organization* and *behavior*.

**Systems Biology:** *Systems biology* aims at a system-level understanding of the *organization* and *control* of genetic-, metabolic-, and signalling pathways.

## References


[1] John L. Casti : *Reality Rules*. John Wiley, Chichester, UK 1992.

[2] Franklin M. Harold : *The Way of the Cell: Molecules, Organisms and the Order of Live*. Oxford University Press, 2001.

[3] George Klir : *Facets of Systems Science*. Plenum Press, 1991.

[4] Robert Rosen : Church's Thesis and its Relation to the Concept of Realizability in Biology and Physics. *Bulletin of Mathematical Biophysics*, Vol. 24, 375–393, 1962.

[5] Robert Rosen : *Life Itself*. Columbia University Press, 1991.

[6] Olaf Wolkenhauer : Systems Biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, Vol. 2, No. 3, 258–270, 2001.

[7] Olaf Wolkenhauer : *Data Engineering*. John Wiley, New York, 2001.

[8] Olaf Wolkenhauer : Mathematical modelling in the post-genome era: understanding genome expression and regulation - a system theoretic approach. *BioSystems*, Vol. 65, 1–18, 2002.