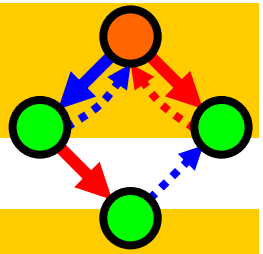
Chaitin's razor: The smallest program that calculates the observations is the best theory.

Artificial Biochemistry



Model Construction and Validation

Luca Cardelli

Microsoft Research

The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

Trento, 2006-05-22..26

www.luca.demon.co.uk/ArtificialBiochemistry.htm

Model Construction

On the Nature of Modeling

- In their own words...
  - Sydney Brenner: When you want to have a predictive science, you have to be able to calculate.
  - Denis Noble: There will probably therefore be no unique model that does everything at all levels. … One of the first questions to ask of a model therefore is what questions does it answer best.
  - Hamid Bolouri & Eric H. Davidson: Abstract models have relatively few parameters and so … it is easier to explore their behavior and build models with them. … In contrast, more detailed models suffer from an explosion in the number of their parameters.
  - Al Hershey: Influential ideas are always simple. Since natural phenomena need not be simple, we master them, if at all, by formulating simple ideas and exploring their limitations.
  - Martin H Fischer: Facts are not science – as the dictionary is not literature.
    - Hiroaki Kitano: Molecular biology has uncovered a multitude of biological facts … but this alone is not sufficient for interpreting biological systems. … A system-level understanding should be the prime goal of biology.

On the Nature of Data

Complexity and Scientific Laws

MY STORY BEGINS in 1686 with Gottfried W. Leibniz's philosophical essay *Discours de métaphysique* (*Discourse on Metaphysics*), in which he discusses how one can distinguish between facts that can be described by some law and those that are lawless, irregular facts. Leibniz's very simple and profound idea appears in section VI of the *Discours*, in which he essentially states that a theory has to be simpler than the data it explains, otherwise it does not explain anything. The concept of a law becomes vacuous if arbitrarily high mathematical complexity is permitted, because then one can always construct a law no matter how random and patternless the data really are. Conversely, if the only law that describes some data is an extremely complicated one, then the data are actually lawless.

G.Chaitin: The Limits of Reason
Scientific American, March 2006

The Pragmatic View

- A model is always wrong
  - Unless it is quantum mechanics, and even then...

- But it is a tool:
  - A tool for calculating predictions
  - A tool for calculating refutations

    - (Sydney Brenner: When you want to have a predictive science, you have to be able to calculate.)

Storing *Processes*

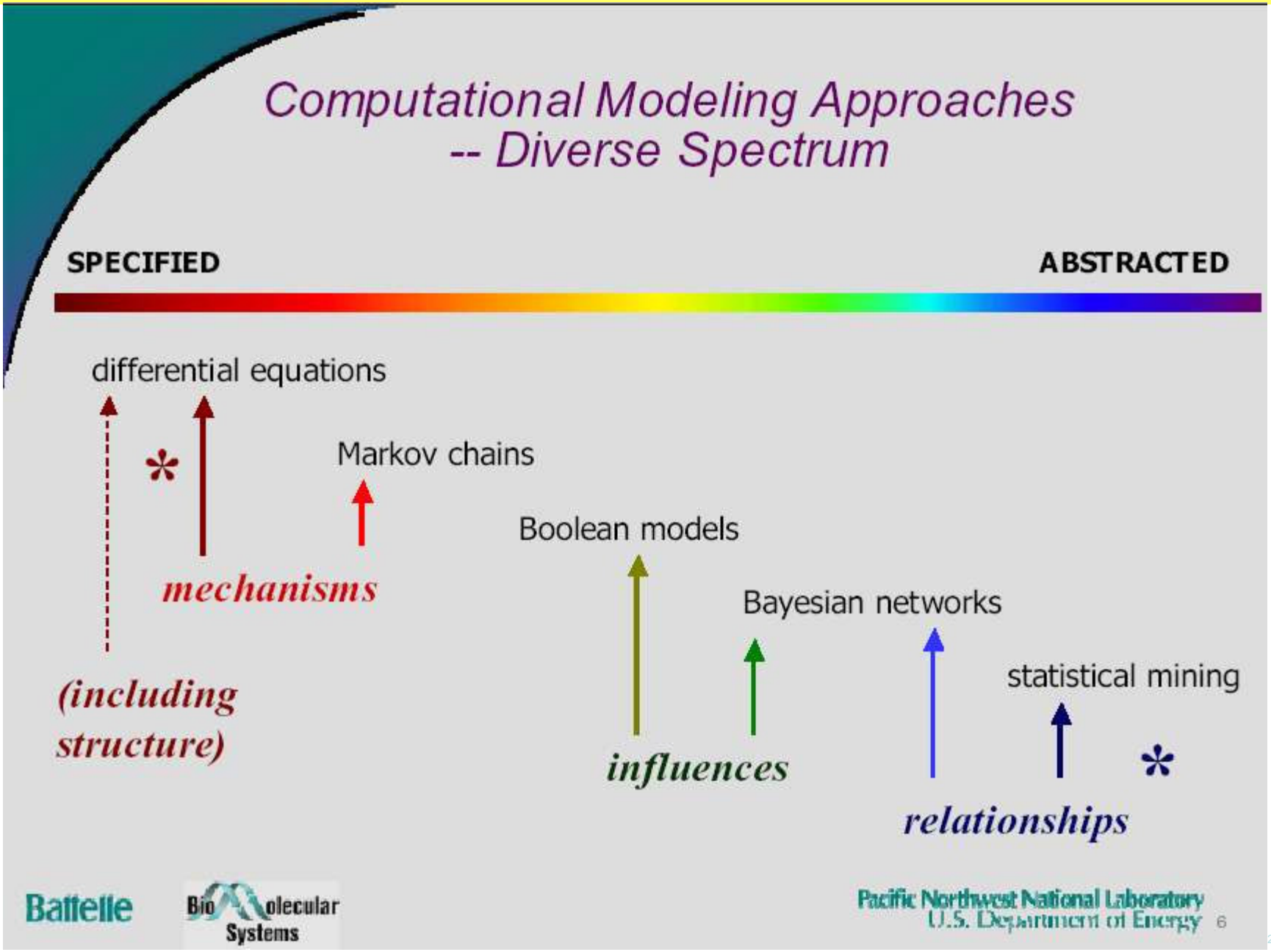
- Today we represent, store, search, and analyze:
  - Gene sequence data
  - Protein structure data
  - Metabolic network data
  - Signaling pathway data
  - …

Cellular Abstractions: Cells as Computation
Regev&Shapiro **NATURE** vol 419, 2002-09-26, 343

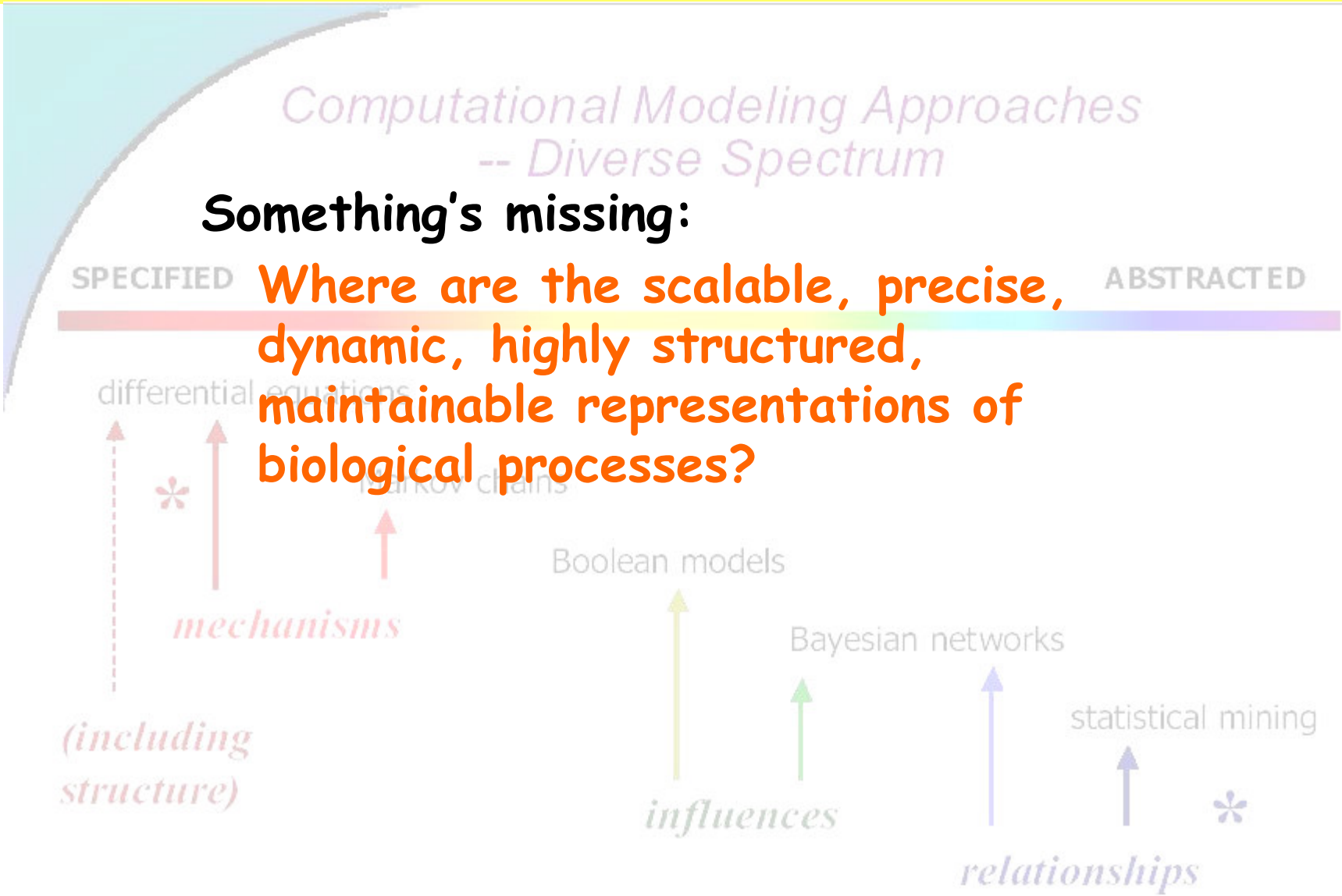
- How can we represent, store, and analyze *biological processes*?
  - Scalable, precise, dynamic, highly structured, maintainable representations for *systems biology*.
  - Not just huge lists of chemical reactions or differential equations.

- In computing…
  - There are well-established scalable representations of dynamic reactive processes.
  - They look more or less like little, mathematically based, programming languages.

A Frequently-Seen Slide



A Frequently-Seen Slide



Something's missing:

Where are the scalable, precise, dynamic, highly structured, maintainable representations of biological processes?

Reactive Systems

- Modeling biological systems
  - Not as continuous systems (often highly nonlinear)
  - But as discrete reactive systems; abstract machines where:
    - States represent situations
    - Event-driven transitions between states represent dynamics
  - The adequacy of describing (discrete) complex systems as reactive systems has been argued convincingly [Harel]

- Many biological systems exhibit features of reactive systems:
  - Discrete transitions between states
  - Deep layering of abstractions ("steps" at multiple levels)
  - Complexity from combinatorial interaction of simple components
  - High degree of concurrency and nondeterminism
  - "Emergent behavior" not obvious from part list

- Still, needs quantitative semantics
  - Stochastic, hybrid, etc. to talk about *rates* (and geometry).

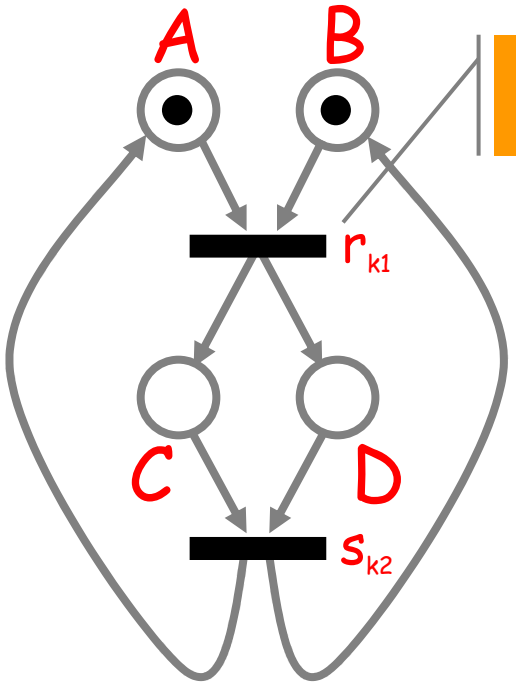
Reaction System vs. Reactive System

A process calculus (chemistry)

$r: A + B \rightarrow_{k1} C + D$

$s: C + D \rightarrow_{k2} A + B$

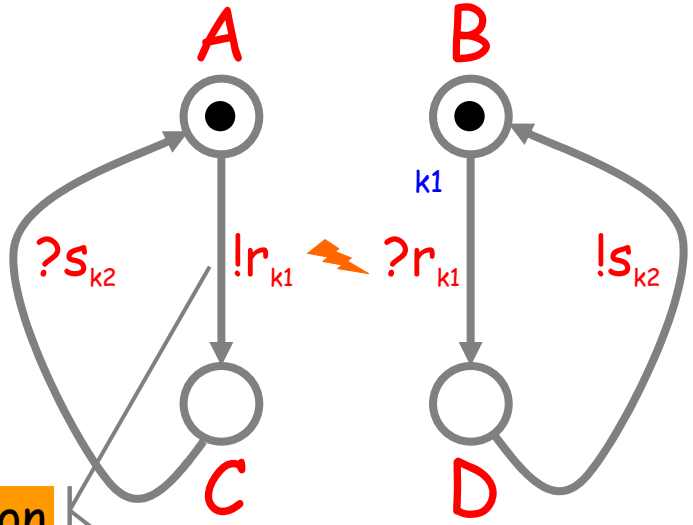
Does A become C or D?



Reaction oriented

1 line per reaction

A different process calculus (π)



Interaction oriented

1 line per component

$A = !r_{k1}; C$

$C = ?s_{k1}; A$

$B = ?r_{k1}; D$

$D = !s_{k2}; B$

A becomes C not D!

The same "model"

Maps to a CTMC	Maps to a CTMC

A Petri-Net-like representation. Precise and dynamic, but not modular, scalable, or maintainable.

A compositional graphical representation (precise, dynamic *and* modular) and the corresponding calculus.

Stochastic Approach

- Relatively recent development on Process Calculi
  - For computer networking simulation and analysis
  - Now for biochemical simulation and analysis

- Continuous Time Markov Chains
  - Finite State Machines, with state transition times exponentially distributed (memoryless)
  - Well studied class of stochastic processes
  - Efficient analysis algorithms for stationary and transient analysis

- High level formalisms mapping to CTMCs
  - Stochastic Petri Nets [Molloy]
  - Markovian Queuing Networks [Muppala & Triverdi]
  - Stochastic Automata Networks [Plateau]
  - Probabilistic I/O Automata [Wu et al.]
  - Stochastic Process Algebras [Herzog et al.] [Hillston]

Importance of Stochastic Effects

- *A* **deterministic** *system:*
  - May get "stuck in a fixpoint".
  - And hence **never oscillate**.

- *A similar* **stochastic** *system:*
  - May be "thrown off the fixpoint" by stochastic noise, entering a long orbit that will later bring it back to the fixpoint.
  - And hence **oscillate**.

Surprisingly enough, we have found that parameter values that give rise to a stable steady state in the deterministic limit continue to produce reliable oscillations in the stochastic case, as shown in Fig. 5. Therefore, the presence of noise not only changes the behavior of the system by adding more disorder but can also lead to marked qualitative differences.

Mechanisms of noise-resistance in genetic oscillators

Jose´ M. G. Vilar, Hao Yuan Kueh, Naama Barkai, Stanislas Leibler

PNAS **April 30, 2002** vol. 99 no. 9 p.**5991**

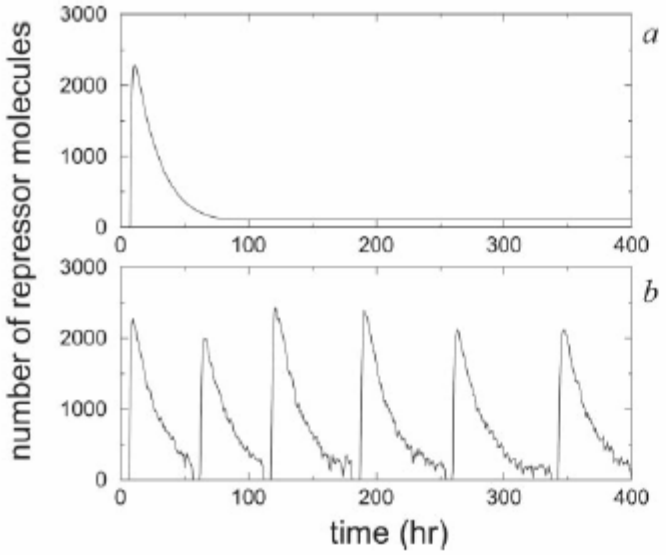


Fig. 5. Time evolution of R for the deterministic Eq. [1] (*a*) and stochastic (*b*) versions of the model. The values of the parameters are as in the caption of Fig. 1, except that now we set $\delta_R = 0.05\ h^{-1}$. For these parameter values, $\tau < 0$, so that the fixed point is stable.

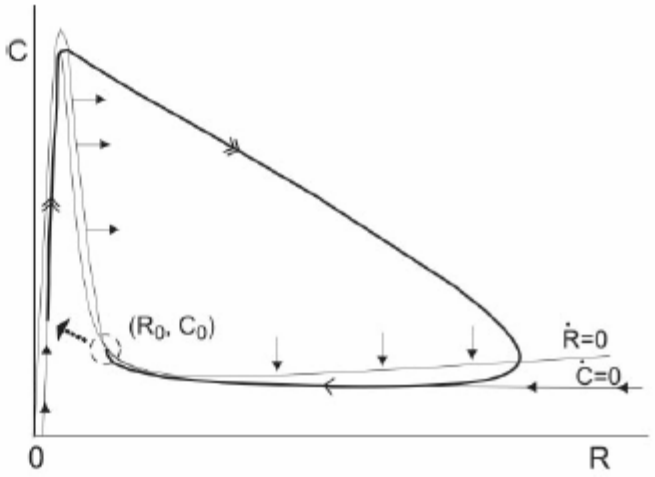


Fig. 6. Phase portrait as in Fig. 4 but for a situation in which the system falls into the stable fixed point (R_0,C_0). The dotted arrow to the left of the fixed point illustrates a perturbation that would initiate a single sweep of the (former) oscillatory trajectory.

Model Validation

Methods

- Model Construction (writing things down precisely)
  - Formalizing the notations used in systems biology.
  - Formulating description languages.
  - Studying their kinetics (semantics).

- Model Validation (using models for postdiction and prediction)
  - Simulation from compositional descriptions
    - Stochastic: quantitative concurrent semantics.
    - Hybrid: discrete transitions between continuously evolving states.
  - "Program" Analysis
    - Control flow analysis
    - Causality analysis
  - Modelchecking
    - Standard, Quantitative, Probabilistic

Model Validation: Simulation

- Basic stochastic algorithm: Gillespie
  - Exact (i.e. based on physics) stochastic simulation of chemical kinetics.
  - Can compute concentrations and reaction times for biochemical networks.

- Stochastic Process Calculi
  - BioSPi [Shapiro, Regev, Priami, et. al.]
    - Stochastic process calculus based on Gillespie.
  - BioAmbients [Regev, Panina, Silverma, Cardelli, Shapiro]
    - Extension of BioSpi for membranes.
  - Case study: Lymphocytes in Inflamed Blood Vessels [Lecaa, Priami, Quaglia]
    - Original analysis of lymphocyte rolling in blood vessels of different diameters.
  - Case study: Lambda Switch [Celine Kuttler, IRI Lille]
    - Model of phage lambda genome (well-studied system).
  - Case study: VICE [U. Pisa]
    - Minimal prokaryote genome (180 genes) and metabolism of *whole* VIrtual CEll, in stochastic π-calculus, simulated under stable conditions for 40K transitions.

- Hybrid approaches
  - Charon language [UPenn]
    - Hybrid systems: continuous differential equations + discrete/stochastic mode switching.
  - Etc.

Model Validation: "Program" Analysis

- Causality Analysis
  - *Biochemical pathways*, ("concurrent traces" such as the one here), are found in biology publications, summarizing known facts.
  - This one, however, was automatically generated from a program written in BioSpi by comparing traces of all possible interactions. [Curti, Priami, Degano, Baldari]
  - One can play with the program to investigate various hypotheses about the pathways.

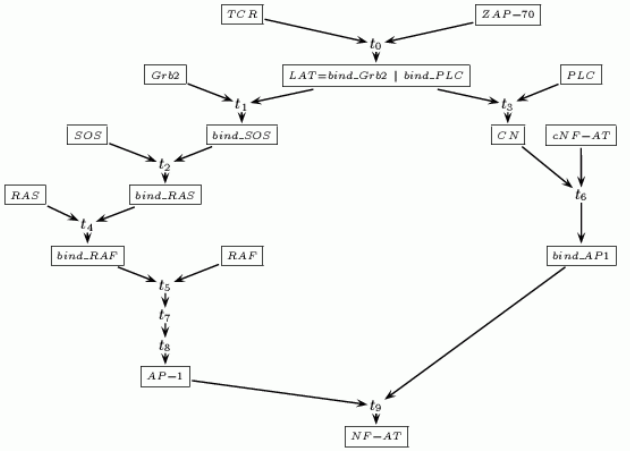


Fig. 2. A computation of *Sys*. For readability, the processes, enclosed in boxes, have no address. Causality (both on transitions and processes) is represented by the (Hasse diagram resulting from the) arrows; their absence makes it explicit concurrent activities.

- Control Flow Analysis
  - Flow analysis techniques applied to process calculi.
  - Overapproximation of behavior used to answer questions about what "cannot happen".
  - Analysis of positive feedback transcription regulation in BioAmbients [Flemming Nielson].

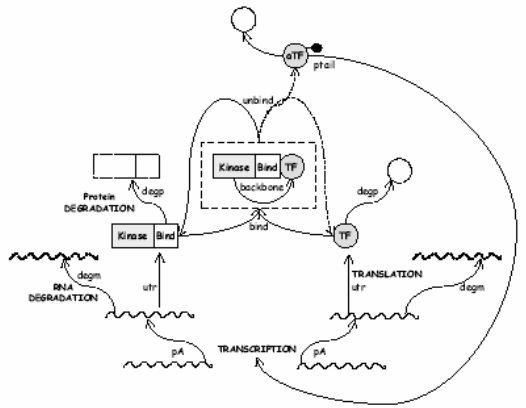


Fig. 1. Graphical presentation of Transcriptional Regulation by Positive Feedback [25].

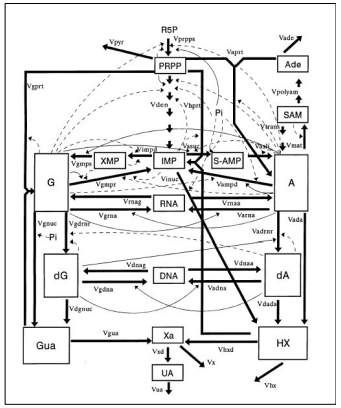
- Probabilistic Abstract Interpretation
  - [DiPierro Wicklicky].

Model Validation: Modelchecking

- Temporal
  - Software verification of biomolecular systems (NA pump) [Ciobanu]
  - Analysis of mammalian cell cycle (after Kohn) in CTL. [Chabrier-Rivier Chiaverini Danos Fages Schachter]
    - E.g. is state S_1 a necessary checkpoint for reaching state S_2?

- Quantitative: Simpathica/xssys [Antioniotti Park Policriti Ugel Mishra]
  - Quantitative temporal logic queries of human Purine metabolism model.

```
Eventually(Always (PRPP = 1.7 * PRPP1)
                  implies
                  steady_state()
                  and Eventually(Always(IMP < 2 * IMP1))
                  and Eventually(Always(hx_pool < 10*hx_pool1)))
```



- Stochastic: Spring [Parker Normal Kwiatkowska]
  - Designed for stochastic (computer) network analysis
    - Discrete and Continuous Markov Processes.
    - Process input language.
    - Modelchecking of probabilistic queries.

Model Validation: Perturbation

- Perturbation
  - Changing the inputs
    - Environment perturbation
  - Printing the outputs
    - Fluorescent tags etc.
  - Turning off subsystems
    - Gene knockout
    - RNA interference
  - Replacing subsystems
    - Activator bypass

- General Inspection ("Debugging") Techniques
  - "Code walking": what's in the program
    - Genome sequencing
  - "Stack dumping": what's running now
    - Transcriptional States - mRNA micro-array assays
  - "Core dumping": what's being produced
    - Translational States - Proteomics

Model Maintenance

Model Maintenance

- Large models are just like
  - Large programs
  - Large theorems
- That is
  - They need to be maintained
  - They have to be written in a language that facilitates maintenance
- Which means
  - Models must be easy to read, not to write
  - Models must be compact

Chemical Models Explode

- Biochemistry (unlike much of chemistry) is combinatorial
  - Biochemical systems have many regular repeated components
  - Components interact and combine in complex combinatorial ways
  - Components have local state
  - A biochemical system is vastly more compact that its potential state space

- Chemical (and, consequently, ODE) descriptions blow up
  - Each "state" of a molecule or complex becomes a "chemical species"
  - This may lead to exponential explosion in the model description (stoichiometric matrix)
  - Because the state space gets explicitly represented in the model

- There is a better way:
  - Describe biochemical systems compositionally
  - Each molecule with its own state and interactions
  - ... as Nature intended...

The Plan

What Reactive Systems Do For Us

We can write things down precisely

- We can modularly describe high structural and combinatorial complexity ("do programming").

We can calculate and analyze

- Directly support simulation.
- Support analysis (e.g. control flow, causality, nondeterminism).
- Support state exploration (modelchecking).

We can visualize

- Automata-like presentations.
- Petri-Net-like presentations.
- State Charts, Live Sequence Charts [Harel]
  - Hierarchical automata.
  - Scenario composition.

We can reason

- Suitable equivalences on processes induce algebraic laws.
- We can relate different systems (e.g. equivalent behaviors).
- We can relate different abstraction levels.
- We can use equivalences for state minimization (symmetries).

Disclaimers

- Some of these technologies are basically ready (medium-scale stochastic simulation and analysis, medium-scale nondeterministic and stochastic modelchecking).
- Others need to scale up significantly to be really useful. This is our challenge.

Many approaches, same basic philosophy, tools being built:

$\Rightarrow$ *Proc. Computational Methods in Systems Biology* [2003-2005]

Summary

- Model Construction
  - Various classical approaches, from Bayesian Networks (phenomenological) to Molecular Dynamics (mechanistic).
  - New approaches based on Reactive Systems (mechanistic).
- Model Validation
  - Various techniques from computing are novel to biology.
- Model Scaling and Maintenance
  - It is a major issue, and it will get worse.
  - A classical "software engineering" problem. Now "model engineering" ?
- Stochastic Approach
  - Between discrete and continuous.
  - Between deterministic and nondeterministic.
  - Exposes new phenomena not evident in any of the above.

Q?