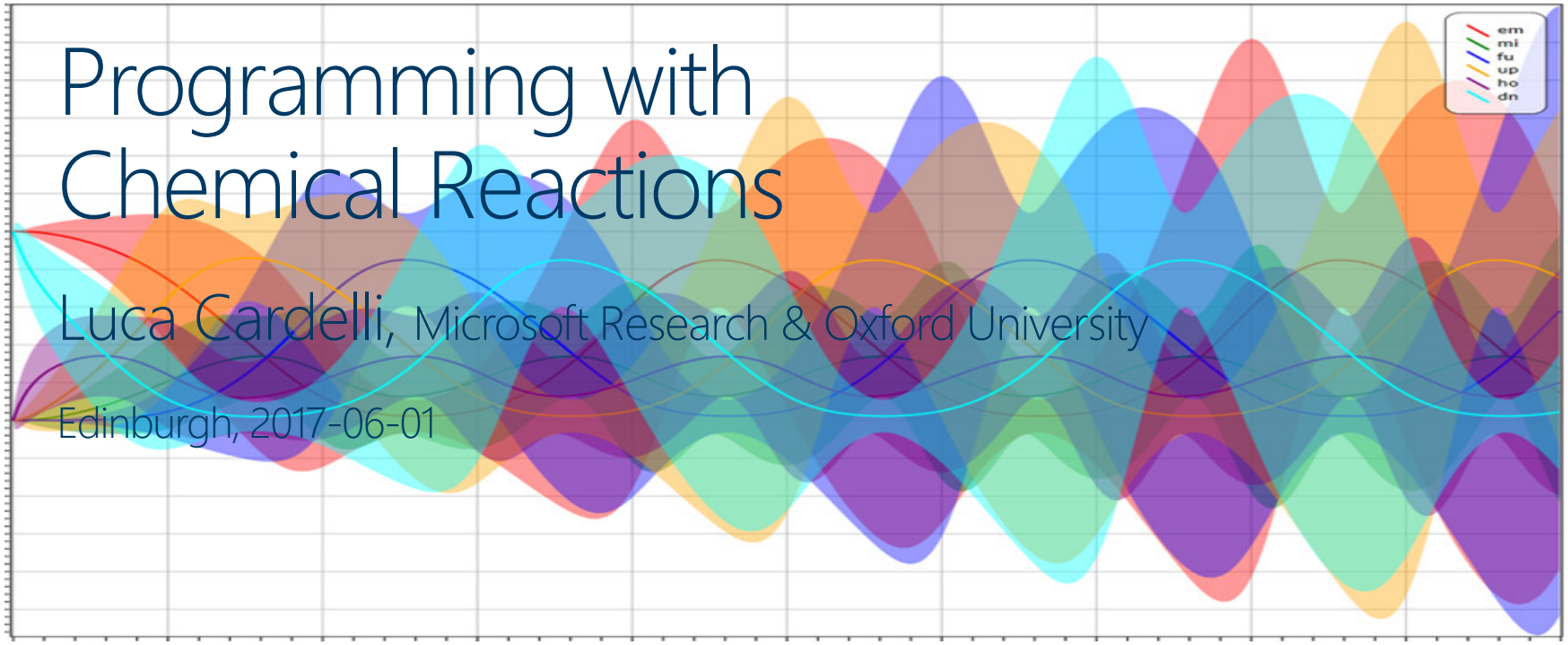
Microsoft Research

UNIVERSITY OF OXFORD

Programming with Chemical Reactions



Luca Cardelli, Microsoft Research & Oxford University

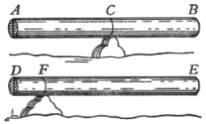
Edinburgh, 2017-06-01

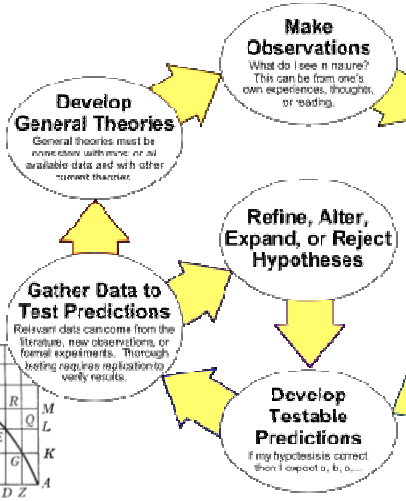
Discovery through Synthesis

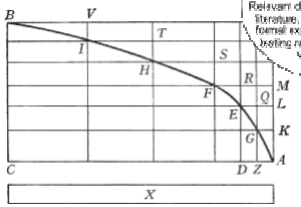
State of the Art Yesterday - Discovery

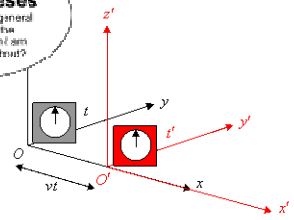
- The Scientific Method ~ 1638

- 1 Guy









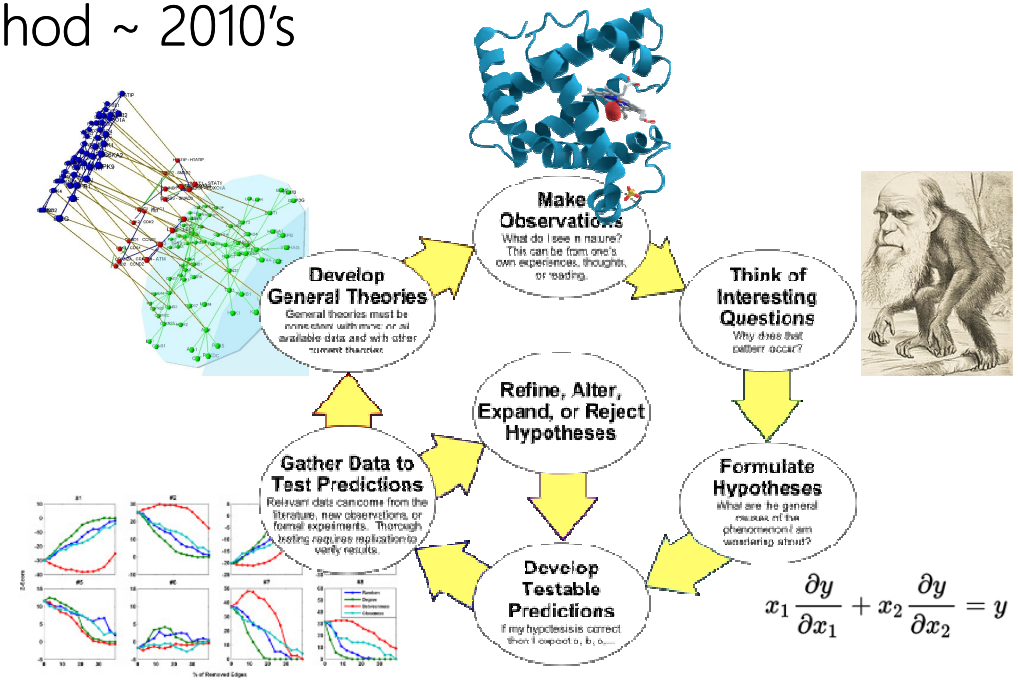
State of the Art Today - Discovery

- The Scientific Method ~ 2010's

- 1 Lab

1 protein = 30 people / 30 years

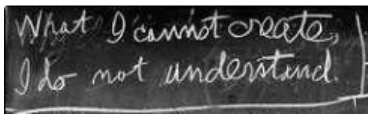
Humans have >250,000 proteins ☹

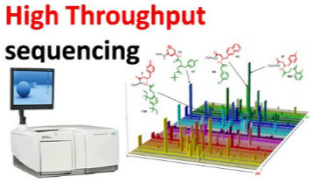


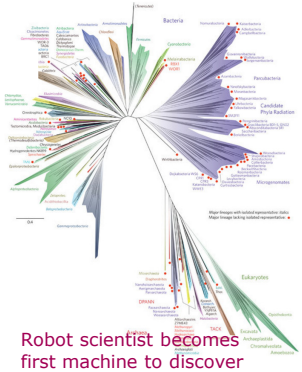
New Approach – Discovery + Synthesis

- The Scientific Method ~ 2020's

- 1 Software







Robot scientist becomes first machine to discover new scientific knowledge

Ross King

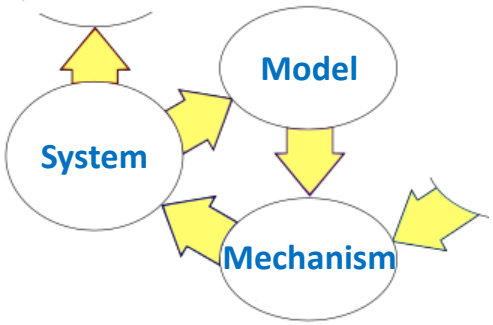
Falsification + Verification

Discovery in complex systems requires increased intervention - synthesis

Read nature, but also write nature.

New Approach – The Inner Loop

- A *model* is refined by testing *mechanisms* within *systems*



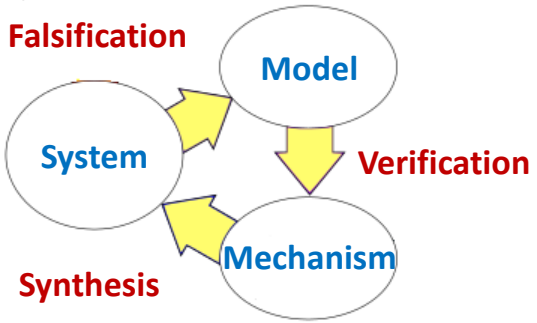
- Today: publication does not accurately reflect execution
  - Model: poorly-maintained matlab script
  - Mechanism : poorly-described manual protocols in the lab
  - System: poorly-characterized and hardly "resettable"
  - ⇒ Crisis in biology: experiments are done once and are hard to reproduce

    http://www.nature.com/news/reproducibility-1.17552

New Approach – The Inner Loop

- Tomorrow, automation

Nodes
- Model: unambiguous (mathematical) description (CompBio)
- Mechanism: standardized (engineered) parts and protocols (SynthBio)
- System: characterized (biological) organism and foundries (SysBio)



Arcs
- Verification: simulation / analysis / model checking / theorem proving
- Synthesis: exponential technological growth – sit back and enjoy
- Falsification: lab automation / statistical inference / model reduction

Lifecycle
- Performance evaluation/optimization: of model+protocol+system combined
- Management: version control, equipment monitoring, data storage

Getting around the inner loop

Nodes

- Models (mathematical): [Oxford]
  - We work on understanding the intrinsic computational capability of matter, as expressed by the "language" of chemical reaction networks
- Mechanisms (technological): [Oxford Physics, MSRC] [previously: Caltech, UW]
  - We engineer nanotechnology constructs that perform computation and control
- Systems (biological): [King's College, MSRC]
  - We search for computational mechanisms in natural systems

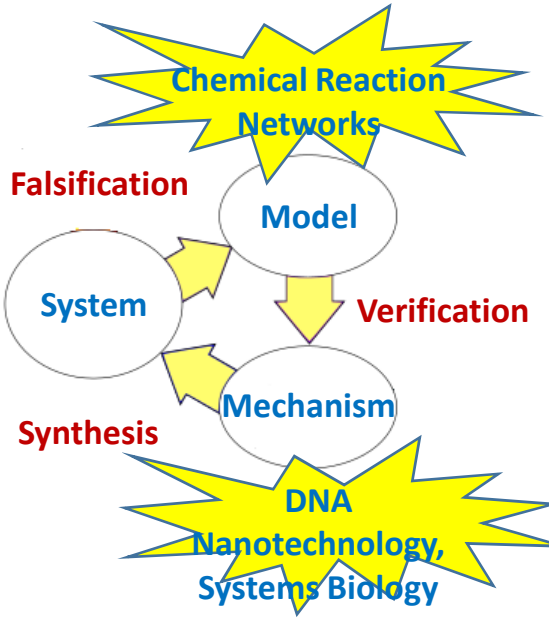
Arcs

- Verification: [Oxford, MSRC]
  - We develop software tools and algorithms for the analysis and simulation of biochemical models.
  - We integrate new algorithms and model classes into our (MSRC) tool suites.
- Synthesis: [Oxford Physics, MSRC] [previously: Caltech, UW] [MSR, Technion?]
  - We develop techniques to "compile" chemical programs into (e.g. DNA) molecules.
- Falsification: [IMT Lucca]
  - We work on advanced algorithms for model reduction of very complex data sets

Lifecycle

- Performance evaluation/optimization: [Oxford, MSRC]
  - We plan to apply hybrid (probabilistic+continuous) modelchecking techniques that we are developing, to verify properties and error bounds of integrated models + lab protocols

Opportunistic



Synthesis through Chemical Reactions

Why are chemical reactions interesting?

$$X + Y \xrightarrow{r} Z + W$$

- A fundamental model of kinetics (i.e. "behavior") in the natural sciences
- A fundamental mathematical structure, rediscovered in many forms
  - Vector Addition Systems, Petri Nets, Bounded Context-Free Languages, Population Protocols, …
- A programming language (coded up in the genome) by which living things manage the processing of matter and information

#1 Discrete (-state) Semantics

- A *state* of the system is a finite multiset of molecules; each molecule belongs to one of a finite set of *species*.
- A fixed finite set of *reactions* over species performs multiset-rewriting over those states.
- Reactions have rates: the state space is a Continuous-Time Markov Chain (a labeled transition system where labels are transition speeds).
- Hence the semantics is discrete and stochastic = atomic theory of matter.

Programming Examples

spec	*program*
$Y = 2X$	X -> Y + Y
$Y = \lfloor X/2 \rfloor$	X + X -> Y
$Y = X1 + X2$	X1 -> Y X2 -> Y
$Y = \min(X1, X2)$	X1 + X2 -> Y

Advanced Programming Examples

spec

Y = max(X1, X2)

program

X1 -> L1 + Y
X2 -> L2 + Y
L1 + L2 -> K
Y + K -> 0

max(X1,X2)=
(X1+X2)-min(X1,X2)

(but is not computed "sequentially")

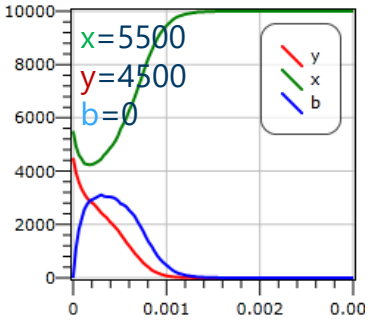
Approximate Majority

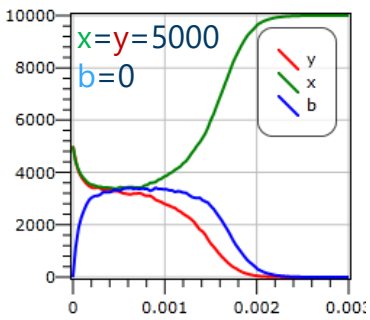
(X,Y) :=
if X≥Y then (X+Y, 0)
if Y≥X then (0, X+Y)

X + Y -> Y + B
Y + X -> X + B
B + X -> X + X
B + Y -> Y + Y

A Consensus Algorithm

- Approximate Majority (AM) Algorithm
  - Uses a third "undecided" population b
  - Disagreements cause agents to become undecided
  - Undecided agents agree with any non-undecided agent

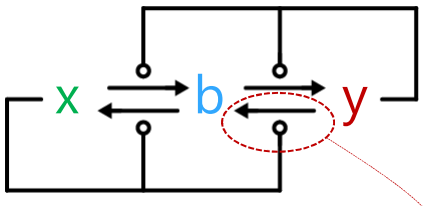




Dana Angluin · James Aspnes · David Eisenstat

A Simple Population Protocol for Fast Robust Approximate Majority

catalysis



chemical reaction network

$x + y \to^r y + b$

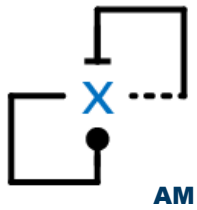
$y + x \to^r x + b$

$b + x \to^r x + x$

$b + y \to^r y + y$

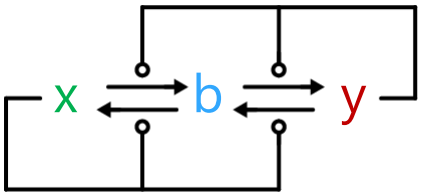
activation

inhibition



A Biological Implementation

Approximate Majority (AM)



1) Bistable
   Even when initially x=y (stochastically)

2) Fast (asymptotically optimal)
   O(*log n*) convergence time

3) Robust to perturbation
   above a threshold, initial majority wins *whp*

Dana Angluin · James Aspnes · David Eisenstat

A Simple Population Protocol for Fast Robust Approximate Majority

2007

Epigenetic Switch

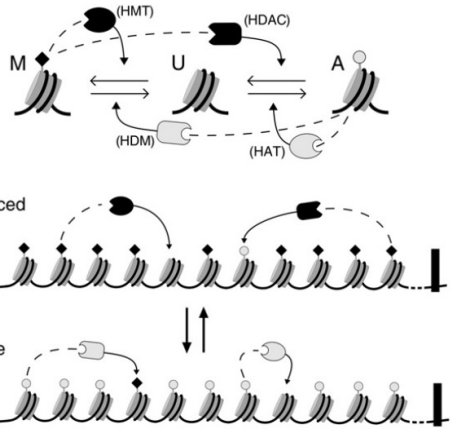
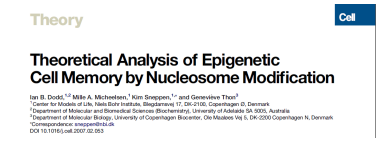


Figure 1. Basic Ingredients of the Model


Theory

Cell

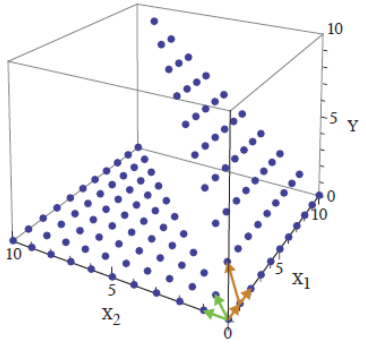
Theoretical Analysis of Epigenetic Cell Memory by Nucleosome Modification

2007

What can we compute this way?

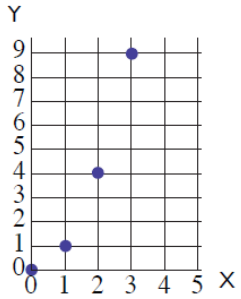
- The semilinear functions
  - Those whose graph is a finite union of linearly-bounded regions

$f(x_1, x_2)=x_2$ if $x_1>x_2$ and 0 otherwise



$\{n_1 \cdot (1,1,0) + n_2 \cdot (0,1,0) \mid n_1, n_2 \in \mathbb{N}\} \cup$
$\{(1,0,0) + n_1 \cdot (1,1,1) + n_2 \cdot (1,0,0) \mid n_1, n_2 \in \mathbb{N}\}$

$f(x) = X^2$



not semilinear

Chen, Doty, Soloveichik, "Deterministic Function Computation with Chemical Reaction Networks" (2013)

But also Register Machines (almost…)

i: INC R_1; JMP j	$PC_i \rightarrow R_1 + PC_j$
i: DEC R_1; JMP j	$PC_i + R_1 \rightarrow PC_j$
i: IF R_2>0 {INC R_1; JMP j}	$PC_i + R_2 \rightarrow R_2 + R_1 + PC_j$
i: IF R_2=0 …	??? Whatever trick we use will have some error

- Turing-complete up to an arbitrarily small error
  - The error bound is set in advance uniformly for any computation of arbitrary length (because we cannot know how long the computation will last), and the machine will progressively "slow down" to always stay below that bound.

- David Soloveichik, Matt Cook, Erik Winfree, Shuki Bruck, "Computation with Finite Stochastic Chemical Reaction Networks".
  [Natural Computing, (online Feb 2008), or Technical Report: CaltechPARADISE:2007.ETR085: .pdf]

Programming Discrete Distributions

$$\tilde{\pi}(y) = \begin{cases} \frac{1}{6}, & \text{if } y = 2 \\ \frac{1}{3}, & \text{if } y = 5 \\ \frac{1}{2}, & \text{if } y = 10 \\ 0, & \text{, otherwise} \end{cases}$$

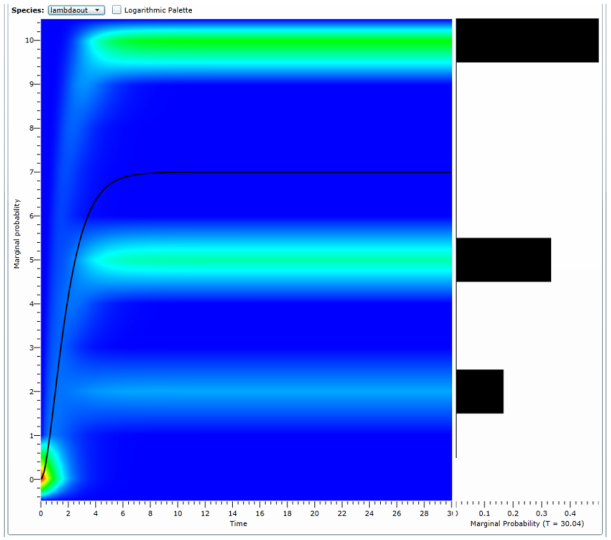
Consider the following CRN:

$$\tau_1 \colon \lambda_Z \xrightarrow{\frac{1}{6}} \lambda_{1,1}$$
$$\tau_2 \colon \lambda_Z \xrightarrow{\frac{1}{3}} \lambda_{2,2}$$
$$\tau_3 \colon \lambda_Z \xrightarrow{\frac{1}{2}} \lambda_{3,3}$$
$$\tau_{1,1} \colon \lambda_1 + \lambda_{1,1} \rightarrow \lambda_{1,1} + \lambda_{out}$$
$$\tau_{2,2} \colon \lambda_2 + \lambda_{2,2} \rightarrow \lambda_{2,2} + \lambda_{out}$$
$$\tau_{3,3} \colon \lambda_3 + \lambda_{3,3} \rightarrow \lambda_{3,3} + \lambda_{out}$$

With initial configuration x_0:

$$x_0(\lambda_z) = 1, x_0(\lambda_1) = 2, x_0(\lambda_2) = 5, x_0(\lambda_3) = 10,$$
$$x_0(\lambda_{1,1}) = x_0(\lambda_{2,2}) = x_0(\lambda_{3,3}) = 0,$$

CME solution for λ_{out} :



Programming Discrete Distributions with Chemical Reaction Networks.
Luca Laurenti, Luca Cardelli, Marta Kwiatkowska.
Natural Computing Journal

Calculus for Distributions

$$P := (P + P) \,|\, min(P, P) \,|\, k \cdot P \,|\, (P)_D : P \,|\, one \,|\, zero$$

$$D := p \,|\, p \cdot c_i + D$$

where $k \in \mathbb{Q}_{\geq 0}$, $p \in \mathbb{Q}_{[0,1]}$ *are rational and* $V = \{c_1, ..., c_n\}$ *is a set of variables with values in* $\mathbb{N}$.

- P is a pmf, obtained as composition of $zero$ and one

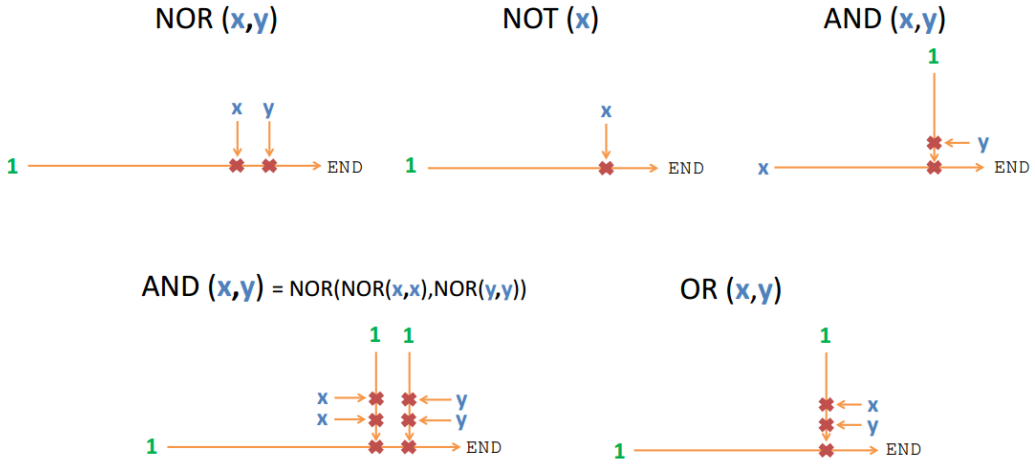
$$\pi_{one}(y) = \begin{cases} 1, & if\ y = 1 \\ 0, & otherwise \end{cases} \qquad \pi_{zero}(y) = \begin{cases} 1, & if\ y = 0 \\ 0, & otherwise \end{cases}$$

- $(P_1)_p$: P_2 is the **convex combination** of P_1 and P_2. That is, P is equal to P_1 with probability p and to P_2 with probability $1 - p$

- $V = \{c_1, \dots, c_n\}$ are called **environmental variables**. They model external inputs that can influence the probability of the formulas

Computing with DNA walkers

- Walkers walk along tracks
  - Taking discrete stochastic steps
  - Blocking other walkers



  - It is envisioned that DNA walkers would carry along other chemicals to specific locations, where they would cause them to interact in a precise sequence, therefore implementing a precisely programmed assembly line of chemical reactions. Logic on the tracks would make this assembly process conditional on e.g. environmental inputs.

Computing with DNA walkers

- We model these walkers with stochastic Petri nets
  - I.e. the same mathematical model (CTMC) as chemical reactions

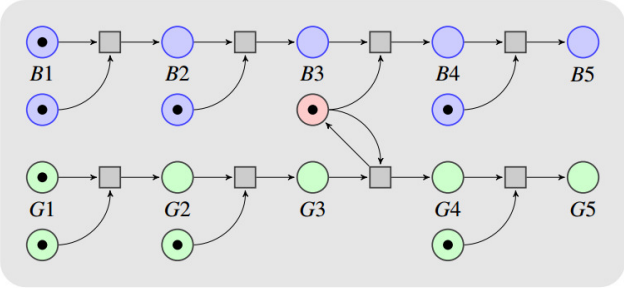


Figure 5: Two tracks, green and blue, with a blocking junction on the third anchorage of each track (G3 and B3). If the blue walker arrives at the junction first, it can block the green track by using up the token of the shared node (shown in red). Blocking is not symmetric: the blue walker can block the track for the green walker, but not vice versa.

The Formal Language and Design Principles of Autonomous DNA Walker Circuits.
Michael A. Boemo, Alexandra E. Lucas, Andrew J. Turberfield, Luca Cardelli.
ACS Synthetic Biology.

#2 Continuous (-state) Semantics

- A state of the system is a (real-valued) concentration for each species.
- A fixed finite set of reactions act (continuously) on such states.
- The Law of Mass Action describes how the system evolves in continuous time.
  - Each reaction acts with a "speed" that is proportional to the product of the concentrations on its left-hand-side, multiplied by its rate.
  - Each species concentration increases or decreases according to the sum of the effects of all the reactions.
- Computing Kinetics (outcomes over time)
- Computing Equilibria (steady-state outcomes)

Sniffers, buzzers, toggles and blinkers

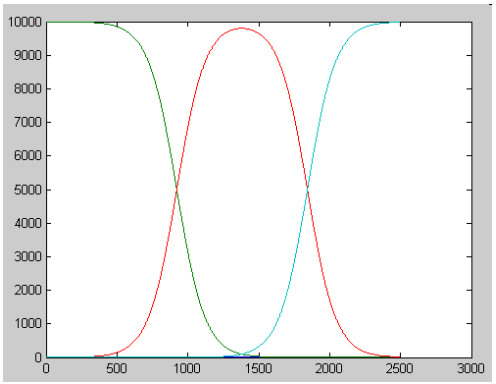
- Sigmoidal (***buzzer***)
- Perfectly adapted (***sniffer***)
- Positive feedback
  - – Mutual activation (***one way switch***)
  - – Mutual inhibition (***toggle switch***)
- Negative feedback
  - – homeostasis
  - – oscillations (***Blinker***)

Tyson JJ - *Sniffers, buzzers, toggles and blinkers.* Curr Opin Cell Biol. 2003 Apr;15(2):221-31.

http://www.inf.ed.ac.uk/teaching/courses/csb/CSB_lecture_dynamic_signalling_and_gene_expression.pdf

Making Waves

How to produce a *symmetric* wave?



A+B -> B+B
B+C -> C+C

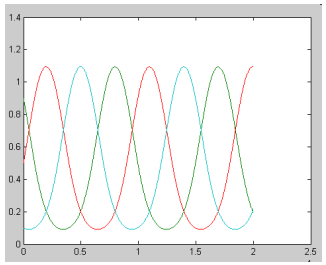
dA/dt = -AB
dB/dt = AB-BC
dC/dt = BC

Synthesizing programs such as this from specifications

Syntax-Guided Optimal Synthesis for Chemical Reaction Networks. Luca Cardelli, Milan Ceska, Martin Fränzle, Marta Kwiatkowska, Luca Laurenti, Nicola Paoletti, Max Whitby. Computer Aided Verification, CAV'17.

Making Clocks

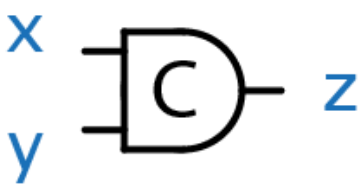
- Large literature going back to Lotka in the 1920's
- *Minimal* oscillators still a topic of interest
  - How many species? How many reactions? How symmetrical?
  - How sensitive to parameters?
  - Free running or self-regulating (limit-cycle)?
- Ex: one built with DNA strand displacement



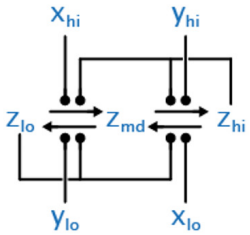
A+B -> B+B
B+C -> C+C
C+A -> A+A

Niranjan Srinivas, James Parkin, Georg Seelig, Erik Winfree, David Soloveichik, "Enzyme-free nucleic acid dynamical systems".
[Preprint: bioRxiv: .pdf paper and .pdf supplementary information]

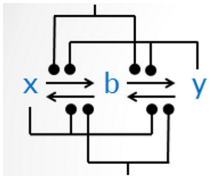
Making Handshakes



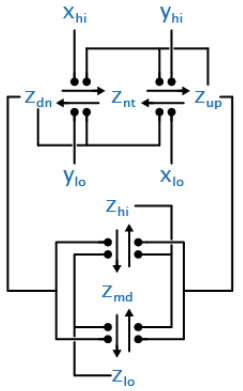
- Muller C-Element
  - When x = y then z = x = y, otherwise z remembers its last state.



Core C-Element



cf. AM with external set/reset inputs



Full C-Element with output rectified by another AM

Chemical Reaction Network Designs for Asynchronous Logic Circuits.
Luca Cardelli, Marta Kwiatkowska, Max Whitby.
Natural Computing Journal.

Steady-State Multiply (and Divide)

[X] := [A]*[B] (at steady state)

$$A + B \xrightarrow{k_1} A + B + X,$$

$$X \xrightarrow{k_2}.$$

H. J. Buisman et al. Computing Algebraic Functions with Biochemical Reaction Networks

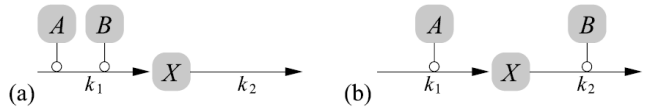


Figure 2. Catalytic reaction networks for (a) multiplication and (b) division.

$$\dot{x} = k_1 ab - k_2 x,$$

whose solution is

$$x = \frac{k_1 a_0 b_0 - (k_1 a_0 b_0 - k_2 x_0) e^{-k_2 t}}{k_2},$$

with stable steady state

$$\hat{x} = \lim_{t \to \infty} x = \frac{k_1}{k_2} a_0 b_0.$$

Computing Algebraic Functions

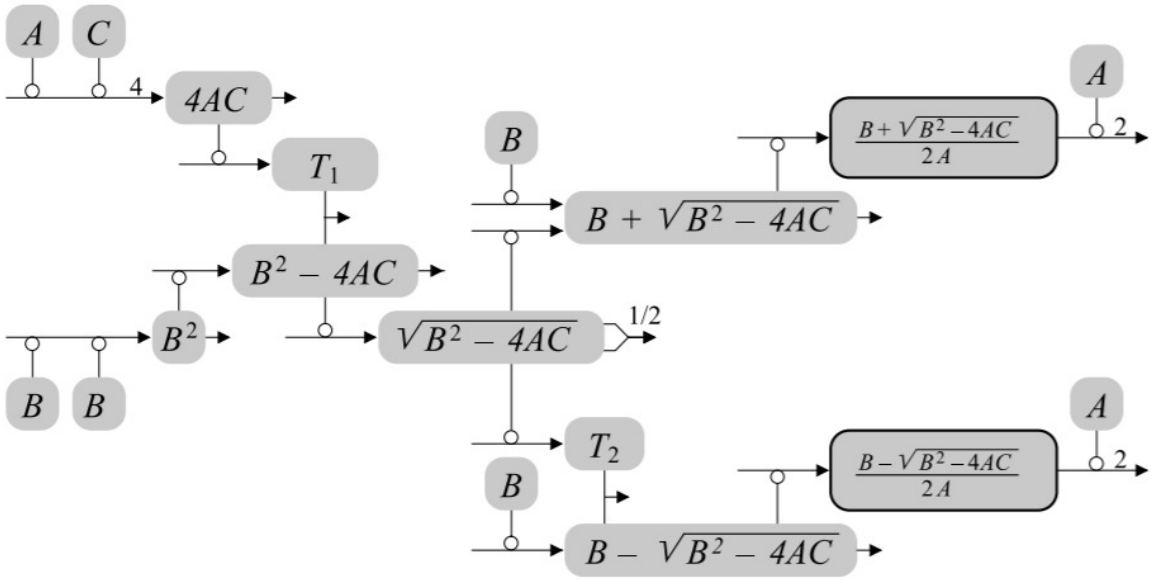


Figure 8. The quadratic formula for finding (the positive real parts of) the roots of $ax^2 - bx + c = 0$. Each of the species in the network has been given a name that represents its steady state concentration. The output species of the computation are highlighted with a black border.

Solving Algebraic Equations

Golden Ratio (-conjugate)

Z + Y -> Y + W
W + X -> X + Z
Z + W -> W + W

Init x=y=w=1.0
Init z = 0.0
all rates 1.0

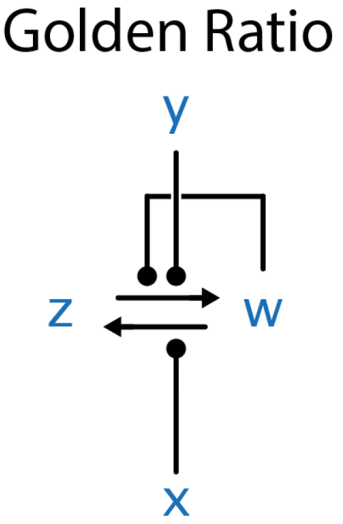
Then (we can easily show analytically by the mass action ODEs that) at steady state:

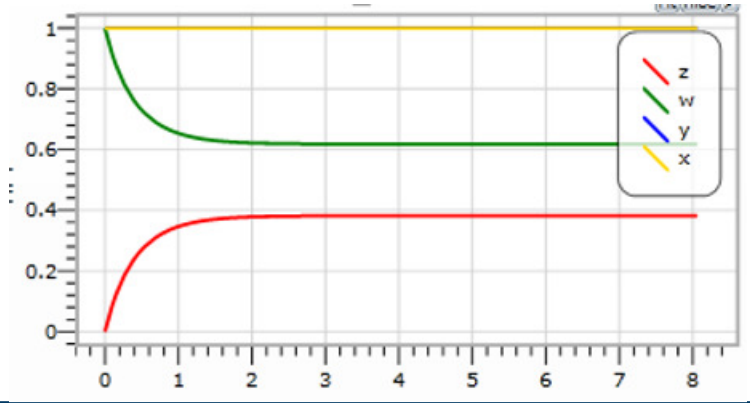
$$1/w = w - 1$$

hence $w = \varphi = 0.61803\ldots$

$$1/\varphi = \varphi - 1$$

All algebraic equations can be solved [Ref]





Finding CRN steady states

- "CRNT" Chemical Reaction Network Theory
  - Martin Feinberg
  - "Static analysis" techniques (on the structure of reactions) based on linear algebra for determining weather a CRN has one or many "positive steady states".
  - Tutorial: https://www.math.wisc.edu/~anderson/RecentTalks/2014/BIRS_TutorialPublic.pdf

Invariance from Initial Conditions

X + Y -> Y + Y

Y -> X

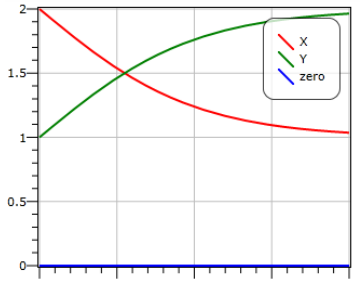
Will produce some X-Y equilibrium, which usually depends on initial values.

But here, for any initial values of X and Y (above 1) the value of X gets fixed to 1
(in general to the ratio of the second reaction rate over the first)

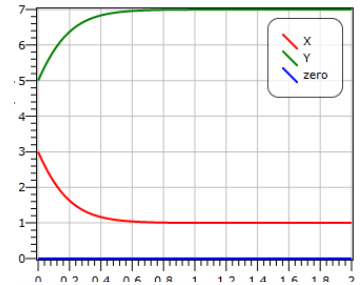
There is a static analysis that will tell you that:

Structural Sources of Robustness in Biochemical Reaction Networks

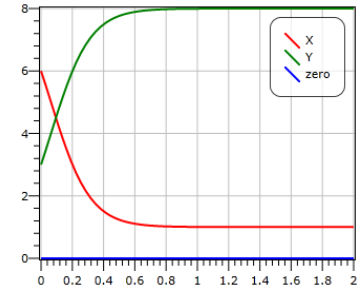
Guy Shinar[1] and Martin Feinberg[2,*]



$X_0=2$, $Y_0=1$
$X_\infty=1$

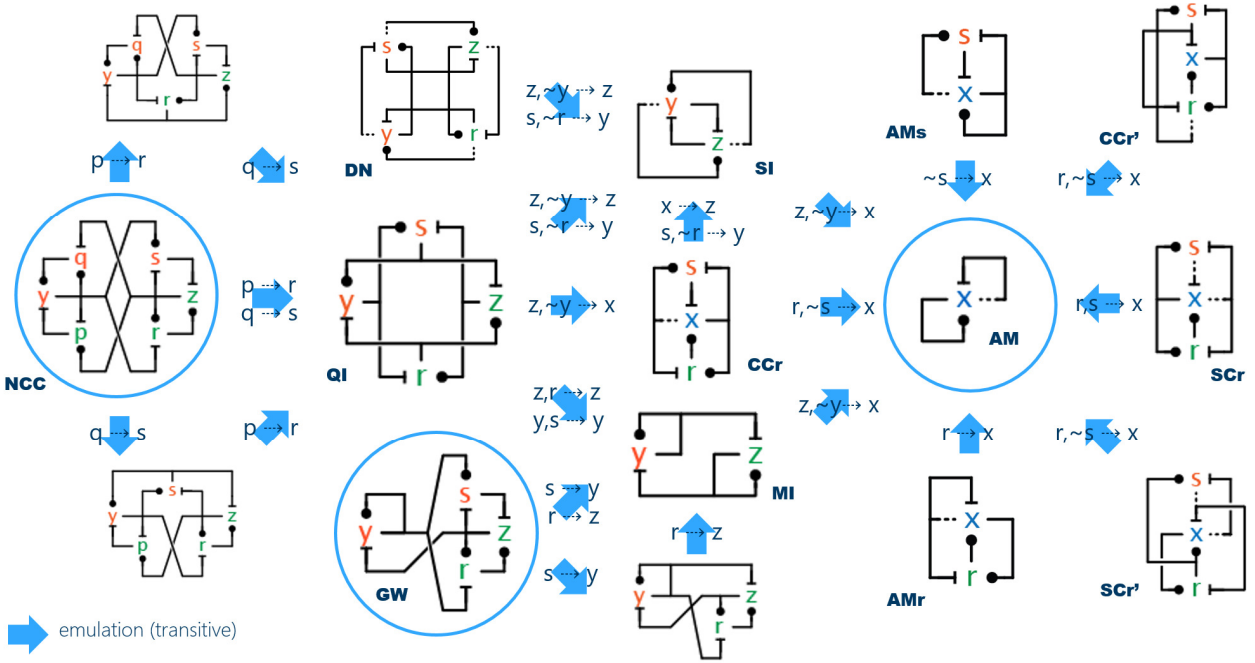


$X_0=3$, $Y_0=5$
$X_\infty=1$



$X_0=6$, $Y_0=3$
$X_\infty=1$

Finding CRN morphisms and bisimulations



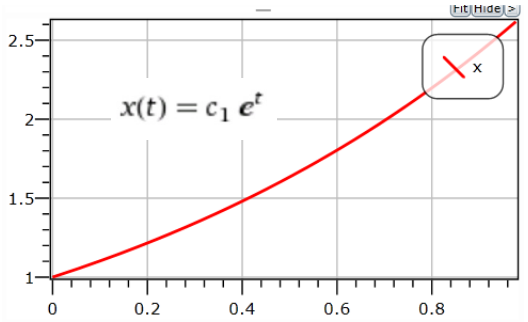
Morphisms of Reaction Networks that Couple Structure to Function (BMC Systems Biology'14)

Comparing Chemical Reaction Networks: A Categorical and Algorithmic Perspective (LICS'16)

Some *Bad* and *Very Bad* Programs

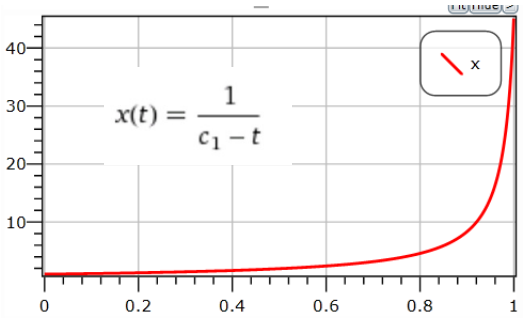
X -> X + X

Violates "only" conservation of mass. (No biggie.)



X + X -> X + X + X

Violates "finite density". (This is bad.)



#3 Wait, there are *two* semantics?

- In a given volume are there
  - (A) A finite number of molecules? or
  - (B) A continuous concentration of <something>?

- Does it make a difference?
  - Related by Avogadro's number: #molecules = concentration * Avogadro
  - But finite density issues: concentration is not unbounded in the discrete model: the program 2X -> 3X will stop when there is no more "space" for molecules

Are these programs equivalent? (YES!)

AM with 4 reactions

X + Y -> Y + B
Y + X -> X + B
B + X -> X + X
B + Y -> Y + Y

AM with 3 reactions

X + Y -> B + B
B + X -> X + X
B + Y -> Y + Y

Same *identical* ODEs => EQUIVALENT

dX/dt = -XY + BX
dY/dt = -YX + BY
dB/dt = 2XY - BX - BY

Are these programs equivalent? (NO!)

- With 3 reactions:
  - {X, Y} -> {B, B} in one step, then stop

```
X + Y -> B + B
B + X -> X + X
B + Y -> Y + Y
```

- With 4 reactions:
  - {X, Y} -> ({X, B} or {Y, B}) -> ({X, X} or {Y, Y}), then stop
  - (no {B, B} final state)

```
X + Y -> Y + B
Y + X -> X + B
B + X -> X + X
B + Y -> Y + Y
```

- Different final states => NOT EQUIVALENT
  - The 3-reaction version fails the requirement that in the end one of the outputs should be the sum of the inputs.

Who is right?

- #1: Believe the discrete nature of atoms (and cells): there are no continuous concentrations
- #2: Believe the analytical power of calculus: a useful approximation in appropriate conditions
- Biology has (quite recently) discovered that #1 must be taken seriously, because of advances in laboratory equipment that allow examining single molecules and single cells.

Chemical Reactions

One programming language
with two (or three) target architectures

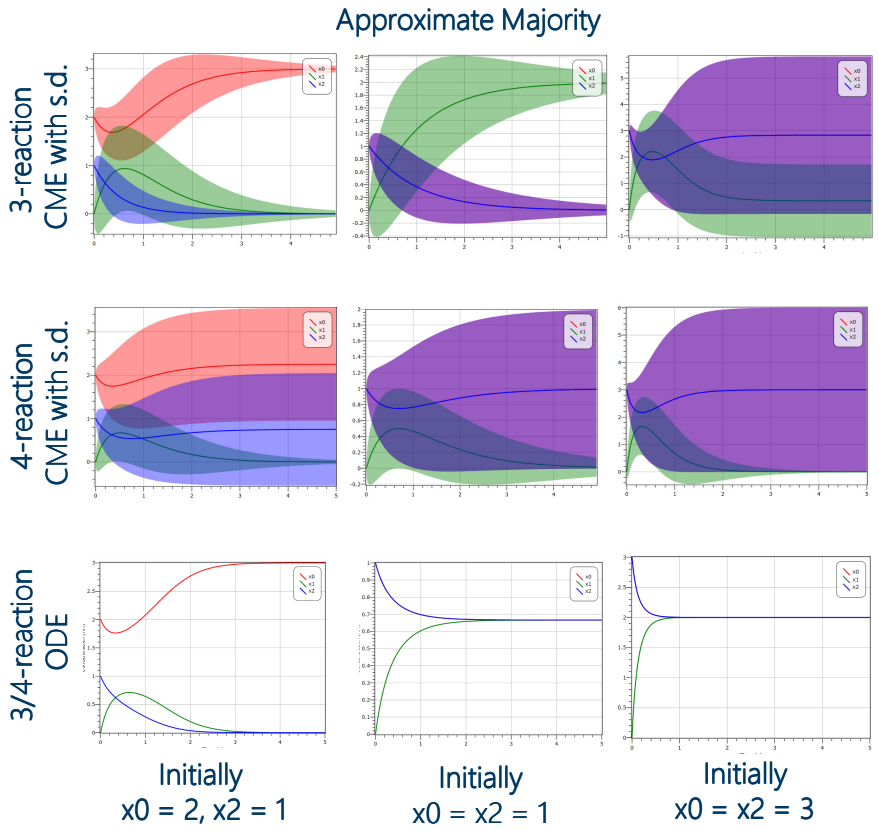
Stochastic Systems

- CME – chemical master equation
- The clock ticks but randomly!

Dynamical Systems

- ODE – ordinary differential equations
- The clock doesn't tick, it swooshes!

(The original AM population protocol
was in discrete time, with a proper clock.)



Current Collaborators

Oxford University - Computer Science

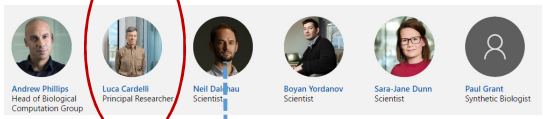
Marta Kwiatkowska
Professor

Luca Laurenti
Royal Society PhD Student

Max Whitby
PhD Student

What/how can we compute with chemical networks and DNA?
How can we verify the properties of engineered molecular systems?
CMSB'15 CMSB'16 QEST'16 DNA'16 DNA'16 HSCC'17 BioSystems'17

MSR Cambridge - Biological Computation



Algorithms & Tools

King's College London - Biology

Attila Csikász-Nagy
Senior Lecturer

Rosa Hernansaiz Ballesteros
MSR PhD Scholarship '15

How do biological switches and oscillators work?
What are the algorithms and how did they evolve?
ScientificReports'16, PLoS'17

Oxford University - Physics

Andrew Turberfield
Professor

Michael Boemo
Postdoc

How can we perform logic with DNA walkers?
ACS SynthBio'16 PhD'16

Lucca Institute for Advanced Studies

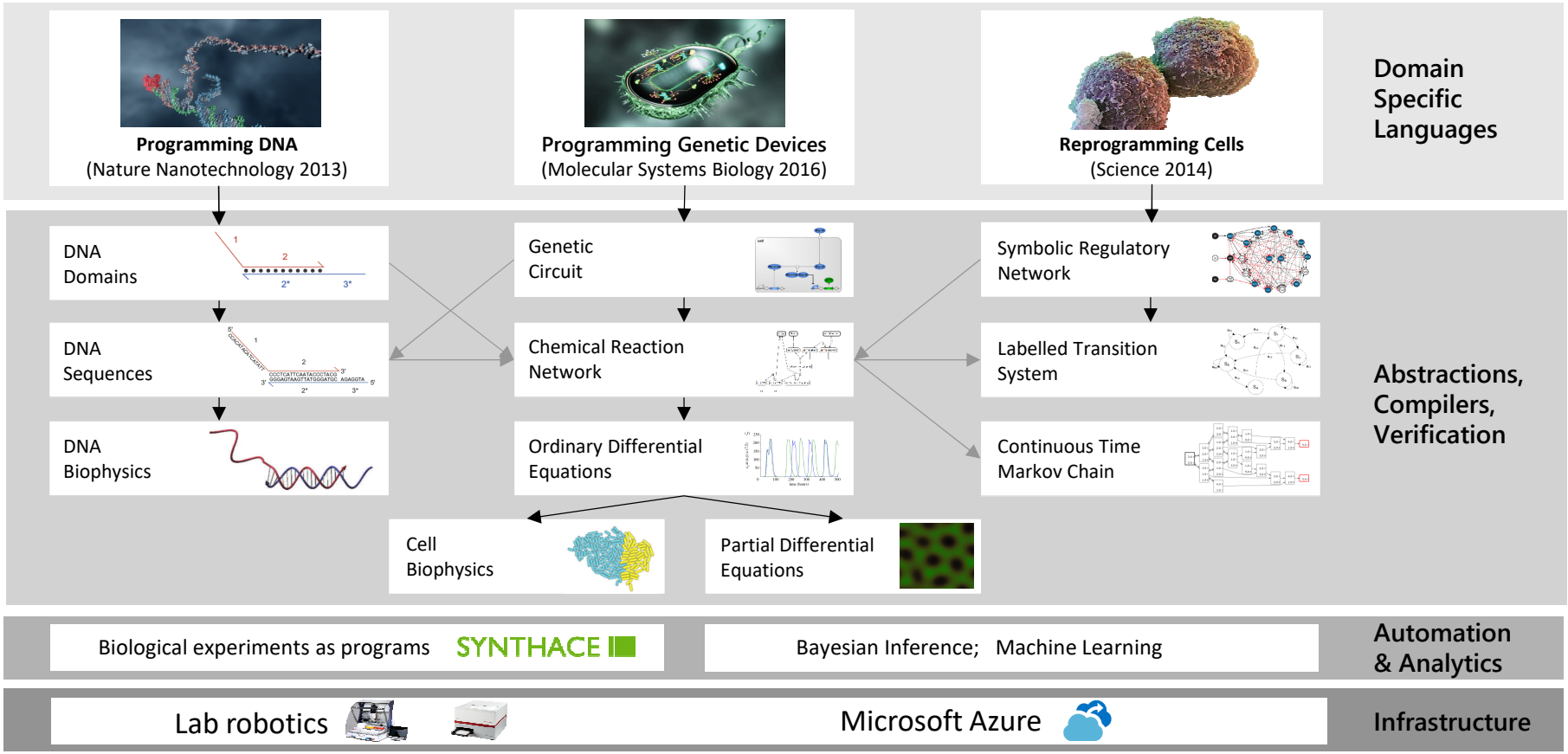
Mirco Tribastone
Associate Professor

Andrea Vandin
Assistant Professor

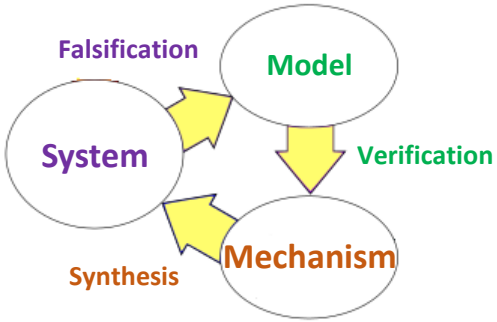
Max Tschaikowski
Assistant Professor

How can we automatically simplify large molecular networks, natural or synthetic, exactly or approximately?
CONCUR'15 POPL'16 LICS'16 TACAS'16 TACAS'17

A platform for programming biology



References



Key papers from some years back

- **The Cell Cycle Switch Computes Approximate Majority** (Scientific Reports'12)
- **Programmable chemical controllers made from DNA** (Nature Nanotech'13)
- **Morphisms of Reaction Networks that Couple Structure to Function** (BMC Systems Biology'14)

- Biological Algorithms
- Nanotechnology
- Model reduction

Color Coded

Recent papers

- **Efficient Switches in Biology and Computer Science** (PLOS Computational Biology'17)
- **ERODE: A Tool for the Evaluation and Reduction of Ordinary Differential Equations** (TACAS'17)
- **Noise Reduction in Complex Biological Switches** (Scientific Reports'16)
- **Chemical Reaction Network Designs for Asynchronous Logic Circuits** (DNA22 '16)
- **The Formal Language and Design Principles of Autonomous DNA Walker Circuits** (ACS Synthetic Biology'16)
- **A Stochastic Hybrid Approximation for Chemical Kinetics Based on the Linear Noise Approximation** (CMSB'15, BioSystems'16)
- **Comparing Chemical Reaction Networks: A Categorical and Algorithmic Perspective** (LICS'16)
- **Programming Discrete Distributions with Chemical Reaction Networks** (DNA22 '16)
- **Approximation of Probabilistic Reachability for Chemical Reaction Networks.** (QEST'16)
- **Efficient Syntax-Driven Lumping of Differential Equations** (TACAS'16)
- **Symbolic Computation of Differential Equivalences** (POPL'16)

http://lucacardelli.name/