Telling Molecules What To Do

Luca Cardelli

Microsoft Research &
University of Oxford

Edinburgh, 2017-05-31, Milner Lecture
Objectives

• The promises of Molecular Programming
  • In Science & Medicine
  • In Engineering
  • In Computing

• The current practice of Molecular Programming
  • DNA technology
  • Molecular languages and tools
  • Example of a molecular algorithm
Molecular Programming: The Hardware Aspect

Smaller and smaller things can be built
Smaller and Smaller

First working transistor
John Bardeen and Walter Brattain, Dec. 23, 1947

First integrated circuit

50 years later

25nm NAND flash
Intel&Micron, Jan. 2010. ~50 atoms

Single molecule transistor
Observation of molecular orbital gating
*Nature*, 2009; 462 (7276): 1039

Molecules on a chip

~10 Moore’s Law cycles left!

Scanning tunneling microscope image of a silicon surface showing 10nm is ~20 atoms across

Race to the Bottom

Moore’s Law is approaching the single-molecule limit

Carlson’s Curve is the new exponential growth in technology

In both cases, we are now down to *molecules*
Building the *Smallest* Things

- How do we build structures that are by definition smaller than your tools?
- Basic answer: you can’t. Structures (and tools) should build themselves!
- By *programmed self-assembly*

www.youtube.com/watch?v=Ey7Emmdlf7Y
Molecular IKEA

- Nature can self-assemble. Can we?

- “Dear IKEA, please send me a chest of drawers that assembles itself.”

- We need a magical material where the pieces are pre-programmed to fit into to each other.

- At the molecular scale many such materials exist...

Programmed Self-Assembly

Proteins

Membranes

DNA/RNA
Molecular Programming: The Software Aspect

Smaller and smaller things can be programmed
We can program...

- Information
- Completely!
We can program...

- Forces
  - Completely!  
    (Modulo sensors/actuators)
We can program...

- **Matter**
  - Completely and directly! By self-assembly.
  - Currently: only DNA/RNA.

- But DNA is an amazing *material*

*It's like a 3D printer without the printer!*
[Andrew Hellington]
DNA

G-C Base Pair
Guanine-Cytosine

T-A Base Pair
Thymine-Adenine

Sequence of Base Pairs (GACT alphabet)

Interactive DNA Tutorial
(https://www.biosciences.bham.ac.uk/labs/minchin/tutorials/dna.html)
Structure

- DNA in each human cell:
  - 3 billion base pairs
  - 2 meters long, 2nm thick
  - 750 megabytes
  - folded into a 6µm ball,
    140 exabytes (million terabytes)/mm³

- A huge amount for a cell
  - Every time a cell replicates it has to copy 2 meters of DNA reliably.
  - To get a feeling for the scale disparity, compute:

- DNA in human body
  - 10 trillion cells
  - 133 Astronomical Units long
  - 7.5 octabytes

- DNA in human population
  - 20 million light years long
Function

- DNA can support structural and computational complexity.

DNA replication in real time
- In Humans: 50 nucleotides/second
  Whole genome in a few hours (with parallel processing)
- In Bacteria: 1000 nucleotides/second
  (higher error rate)

DNA transcription in real time
- RNA polymerase II: 15-30 base/second

Drew Berry
http://www.wehi.edu.au/wehi-tv
What is special about DNA?

- There are many, many nanofabrication techniques and materials

- But only DNA (and RNA) can:
  - Organize ANY other matter [caveats apply]
  - Execute ANY kinetics [caveats: up to time scaling]
  - Assemble Nano-Control Devices
  - Interface to Biology

Organizing Any Matter

- Use one kind of programmable matter (e.g. DNA).
- To organize (almost) ANY matter through it.

"What we are really making are tiny DNA circuit boards that will be used to assemble other components."

Greg Wallraff, IBM

6 nm grid of individually addressable DNA pixels

Executing Any Kinetics

- The kinetics of any finite network of chemical reactions, can be implemented (physically) with especially programmed DNA molecules.

- Chemical reactions as an executable programming language for dynamical systems!
Building Nano-Control Devices

- All the components of nanocontrollers can already be built entirely and solely with DNA, and interfaced to the environment

- DNA Aptamers
- Sensing
- DNA Logical Gates
- Computing
- Constructing
- Self-assembling DNA Tiles
- Actuating
- DNA Walkers & Tweezers
Constructing
Crosslinking
Crosslinking
Crosslinking
Crosslinking
Crosslinking

In nature, crosslinking is deadly (blocks DNA replication).

In engineering, crosslinking is the key to using DNA as a construction material.
DNA Tiling

4 sticky ends

crosslinking

Construction and manipulation of DNA tiles in free space

Pakhtazi
2D DNA Lattices

Chengde Mao
Purdue University, USA

N-point Stars
3D DNA Structures

Ned Seeman
NYU

Andrew Tuberfield
Oxford

3D Crystal

Tetrahedron

3D Tetratetrad 

GAGCAGCC
CQTCCG

ACATGCC

ACATCA

TGTAGCT

TGTAGCT
DNA Origami

*Folding* long (7000bp) naturally occurring (viral) ssDNA
By lots of short ‘staple’ strands that constrain it

Paul W K Rothemund
California Institute of Technology


Black/gray: 1 long viral strand (natural)
Color: many short staple strands (synthetic)

Paul Rothemund’s “Disc with three holes” (2006)
DNA Circuit Boards

- DNA origami are arrays of uniquely-addressable locations
  - Each staple is different and binds to a unique location on the origami
  - It can be extended with a unique sequence so that something else will attach uniquely to it.

- More generally, we can bind “DNA gates” to specific locations
  - And so connect them into “DNA circuits” on a grid
  - Only neighboring gates will interact

Some staples are attached to “green blobs” (as part of their synthesis) Other staples aren’t

Dalchau, Chandran, Gopalkrishnan, Reif, Phillips. 2014
DNA has a data density of 140 exabytes ($1.4 \times 10^{20}$ bytes) per $mm^3$ compared to state-of-the-art storage media that reaches ~500 megabytes ($5 \times 10^8$ bytes) per $mm^3$. DNA has been shown to be stable for millions of years.

We have machines that can read (sequence) and write (synthesize) DNA. The **Carslon Curve** of “productivity” is growing much faster than Moore’s Law.

Cost of sequencing is decreasing rapidly ($1000$ whole human genome), while cost of synthesis is decreasing very slowly. [Rob Carlson, www.synthesis.cc]
Sensing
Aptamers

Artificially evolved DNA molecules that stick to anything you like highly selectively.
Pathogen Spotlights

• DNA aptamer binds to:
  - A) a pathogen
  - B) a molecule our immune system (when allergic) hates and immediately removes (eats) along with anything attached to it!

• Result: instant immunity
  - Mice poisoned with Anthrax plus aptamer (100% survival)
  - Mice poisoned with Anthrax (not so good)

Kary Mullis (incidentally, also Nobel prize for inventing the Polymerase Chain Reaction)
Actuating
DNA Tweezers
DNA Walkers
Polymerization Motor

Rickettsia (spotted fever)

An autonomous polymerization motor powered by DNA hybridization

Triggered amplification by hybridization chain reaction

Robert M. Dicks and Niles A. Pierce

Developed Actin Polymerization Associated with Spotted Fever Group Rickettsia Infection of Vero Cells

Robert M. Dicks, Niles A. Pierce

AMERICAN JOURNAL OF PHYSIOLOGY, Vol. 257, No. 1, Supplement 1, January 1989
Curing
Computational Drugs

- An automaton sequentially reading the string PPAP2B, GSTP1, PIM1, HPS (known cancer indicators) and sequentially cutting the DNA hairpin until a ssDNA drug (Vitravene) is released.
Interfacing to Biology

- A doctor in each cell

*Fig. 1 Medicine in 2050: “Doctor in a Cell”*
Molecular Programming: The Biological Aspect

Biological systems are already ‘molecularly programmed’
Abstract Machines of Biology

Gene Machine
Nucleotides

Protein Machine
Aminoacids

Membrane Machine
Phospholipids

Regulation

Make proteins

Send signals

Direct construction and regulators

Hold receptors, host reactions

Enact fusion, fission

Confinement, Storage
Bulk Transport

Metabolism, Propulsion
Signaling, Transport

Surface and Extracellular Features

Biological Languages

Molecular Interaction Maps

Gene Networks

Transport Networks
But ...

- Biology is programmable, but (mostly) not by us!

- Still work in progress:
  - Gene networks are being programmed in synthetic biology, but using existing ‘parts’
  - Protein networks are a good candidate, but we cannot yet effectively design proteins
  - Transport networks are being investigated for programming microfluidic devices that manipulate vesicles
Molecular Languages

... that we can execute
Our Programming Language: Chemistry

- A Lingua Franca between Biology, Dynamical Systems, and Concurrent Languages

- Chemical Reaction Networks
  - $A + B \rightarrow C + D$ (the program)

- Ordinary Differential Equations
  - $d[A]/dt = -r[A][B]$ ... (the behavior)

- Rich analytical techniques based on Calculus

- But prone to combinatorial explosion
  - E.g., due to the peculiarities of protein interactions
Chemical Programming Examples

**specification**

\[
Y := \min(X1, X2)
\]

**program**

\[
X1 + X2 -> Y
\]

\[
Y := \max(X1, X2)
\]

\[
X1 -> L1 + Y \\
X2 -> L2 + Y \\
L1 + L2 -> K \\
Y + K -> 0
\]

\[
\max(X1, X2) = (X1+X2) - \min(X1, X2)
\]

(but is not computed “sequentially”; it is a form of concurrent computation)

**chemical reaction network**
How do we “run” Chemistry?

- Chemistry is not easily executable
  - “Please Mr Chemist, execute me this bunch of reactions that I just made up”

- Most molecular languages are not executable
  - They are *descriptive* (modeling) languages

- How can we *execute* molecular languages?
  - With real molecules?
  - That we can design ourselves?
  - And that we can buy on the web?
Action Plan

- Building a full software/hardware pipeline for a new fundamental technology
  - Mathematical Foundations [~ concurrency theory in the 80’s]
  - Programming Languages [~ software engineering in the 70’s]
  - Analytical Methods and Tools [~ formal methods in the 90’s]
  - Device Architecture and Manufacturing [~ electronics in the 60’s]

- To realize the potential of Molecular Programming

- “With no alien technology” [David Soloveichik]

- This is largely a ‘software problem’ even when working on device design
Molecular Compilation

- Programs
  - "High-Level" Language
  - Intermediate Language
  - Architecture
  - Molecules
- Gates
- Sequences
- Devices

- Boolean Networks
- Petri Nets
- Chemical Reaction Networks
  - DSD
  - ...
  - 4-domain Signals
  - 3-domain Signals
  - 2-domain Signals

- Petri Nets
- Boolean Networks
- Chemical Reaction Networks

- Intermediate Language
- Architecture
- Molecules

- "High-Level" Language
- Petri Nets
- Chemical Reaction Networks

- Programs
- Sequences
- Devices

- Molecules
- 4-domain Signals
- 3-domain Signals
- 2-domain Signals

- Devices Sequences Gates Programs
- Molecules

- Petri Nets
- Chemical Reaction Networks

- Petri Nets
- Chemical Reaction Networks
Towards High(er)-Level Languages

- Gene Networks
  - Synchronous Boolean networks
    - Stewart Kauffman, etc.
  - Asynchronous Boolean networks
    - René Thomas, etc.

- Protein Networks
  - Process Algebra (stochastic π-calculus etc.)
    - Priami, Regev-Shapiro, etc.
  - Graph Rewriting (kappa, BioNetGen etc.)
    - Danos-Laneve, Fontana & al., etc.

- Membrane Networks
  - Membrane Computing
    - Gheorghe Păun, etc.
  - Brane Calculi
    - Luca Cardelli, etc.

- Waiting for an architecture to run on...
Molecular Programming with DNA

Building the cores of programmable molecular controllers
The role of DNA Computing

• Non-goals
  • Not to solve NP-complete problems with large vats of DNA
  • Not to replace silicon

• Bootstrapping a carbon-based technology
  • To precisely control the organization and dynamics of matter and information at the molecular level
  • DNA is our engineering material
    • Its biological origin is “accidental” (but convenient)
    • It is an information-bearing programmable material
    • Other such materials will be (are being) developed
Domains

- Subsequences on a DNA strand are called **domains**
  - provided they are “independent” of each other

- Differently named domains must not **hybridize**
  - With each other, with each other’s complement, with subsequences of each other, with concatenations of other domains (or their complements), etc.
Short Domains

Reversible Hybridization

DNA double strand
Long Domains

Irreversible Hybridization
Strand Displacement
Strand Displacement

“Toehold Mediated”
Strand Displacement

Toehold Binding
Strand Displacement

Branch Migration
Strand Displacement

Displacement
Strand Displacement

Irreversible release
Bad Match
Bad Match
Bad Match
Bad Match

Cannot proceed
Hence will undo
Two-Domain Architecture

- Signals: 1 toehold + 1 recognition region

- Gates: “top-nicked double strands” with open toeholds

Garbage collection “built into” the gate operation
Transducer
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$

**Input**

$tx$

$ta$

$xtata$

Built by self-assembly!

**ta** is a *private* signal (a different ‘a’ for each xy pair)
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$

Active waste
Transducer $x \rightarrow y$
Transducer $x\rightarrow y$

So far, a **tx signal** has produced an **at cosignal**. But we want signals as output, not cosignals.
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Here is our output \textit{ty signal}.
But we are not done yet:
1) We need to make the output irreversible.
2) We need to remove the garbage.
We can use (2) to achieve (1).
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$
Transducer $x \rightarrow y$

Done.

N.B. the gate is consumed: it is the energy source
(no proteins, no enzymes, no heat-cycling, etc.; just DNA in salty water)
Transducer $x \rightarrow y$
Join $x + y \rightarrow z$
Strand Algebra

- Simple exercise, but forced focus on garbage collection
  - Otherwise algebraic equalities would not hold
  - Led to the two-domain strategy

- Previous proposals for strand displacement
  - Emulate chemical reactions by “overhang” structures. Three-domain:
    - They require fewer steps, but garbage collection is more complex than in two-domain
    - Moreover, the overhang strands can be experimentally problematics

- Accidentally enabled a new implementation technology
  - Plasmid-produced DNA gates (as opposed to synthetic DNA gates)
  - Because 2-domain structures are “flat”

$$P ::= x \mid [x_1, \ldots, x_n], [y_1, \ldots, y_m] \mid 0 \mid P_1 | P_2 | P^* \quad n \geq 1, m \geq 0$$
Plasmidic Gate Technology

- Synthetic DNA is length-limited
  - Finite error probability at each nucleotide addition, hence ~ 200nt max
- Bacteria can replicate plasmids for us
  - Loops of DNA 1000’s nt, with extremely high fidelity
  - Practically no structural limitations on gate fan-in/fan-out

Only possible with two-domain architecture
Correctness

- Eventually the two-domain implementation of chemical reactions (after fixing some bugs) was proven correct.

Modular verification of chemical reaction network encodings via serializability analysis

Matthew R. Lakin\textsuperscript{a,*}, Darko Stefanovic\textsuperscript{a,b}, Andrew Phillips\textsuperscript{c,*}

\textsuperscript{a}Department of Computer Science, University of New Mexico, Albuquerque, NM, USA
\textsuperscript{b}Center for Biomedical Engineering, University of New Mexico, Albuquerque, NM, USA
\textsuperscript{c}Microsoft Research, Cambridge, UK

- This involves more than simple algebra: it’s a rather sophisticated proof of equivalence of concurrent systems.
Large-scale Circuits (so far...)

Scaling Up Digital Circuit Computation with DNA Strand Displacement Cascades
Lulu Qian and Erik Winfree

Neural network computation with DNA strand displacement cascades
Lulu Qian, Erik Winfree, and Joshua Bruck
A Molecular Algorithm

Running something interesting with DNA
Approximate Majority Algorithm

- Given two populations of agents (or molecules)
  - Randomly communicating by radio (or by collisions)
  - Reach an agreement about which population is in majority
  - By converting all the minority to the majority
  [Angluin et al., Distributed Computing, 2007]

- 3 rules of agent (or molecule) interaction
  - \(X + Y \rightarrow B + B\)
  - \(B + X \rightarrow X + X\)
  - \(B + Y \rightarrow Y + Y\)

“our program”
Surprisingly good (in fact, optimal)

- Fast: reaches agreement in $O(\log n)$ time w.h.p.
  - $O(n \log n)$ communications/collisions
  - Even when initially $#X = #Y!$ (stochastic symmetry breaking)

- Robust: true majority wins w.h.p.
  - If initial majority exceeds minority by $\omega(\sqrt{n \log n})$
  - Hence the agreement state is stable

Stochastic simulation of worst-case scenario with initially $#X = #Y$
Circuit component $X + Y \rightarrow 2B$
DNA Implementation, at U.W.

- Programmable chemical controllers made from DNA

[Yuan-Jyue Chen, Neil Dalchau, Niranjan Srinivas, Andrew Phillips, Luca Cardelli, David Soloveichik and Georg Seelig]
Execution

A wetlab pipeline for Molecular Programming
Output of Design Process

- Domain structures
  - (DNA sequences to be determined)

“Ok, how do I run this for real”
From Structures to Sequences

DSD Structure → “Dot-Paren” representation

Output Sequences

Thermodynamic Synthesis

“Ok, where do I buy these?”

www.nupack.org
“DNA Synthesis”

DNA synthesis is a process by which DNA is created. It is commonly referred to as DNA replication, DNA biosynthesis, or in vivo DNA amplification. It is a fundamental process in biology, as it is the basis for hereditary inheritance. DNA synthesis can be performed in vitro using enzymes such as polymerase chain reaction or enzymatic DNA synthesis.
From Sequences to Molecules

• Copy&Paste from nupack
“Ok, how do I run these?”
Add Water
Execute (finally!)

- Fluorescence is your one-bit ‘print’ statement
Output
Debugging

- A core dump

DNA strand length

Various processing stages

Calibration scale
Delivery!

Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA
David Yu Zhang, et al.
Science 318, 1121 (2007);
DOI: 10.1126/science.1148532

A

B

C

D

E

Fluorescence (arbitrary units)
0 0.25 0.5 0.75 1
0 0.5 1 1.5 2
0 0.25 0.5 0.75 1
0 0.25 0.5 0.75 1
0 0.25 0.5 0.75 1
0 0.25 0.5 0.75 1
Final Remarks
A Brief History of DNA

Turing Machine, 1936

Transistor, 1947

Computer programming

DNA, -3,800,000,000

Systematic manipulation of information

20th century

DNA Algorithm, 1994

Systematic manipulation of matter

21st century

Structural DNA Nonotech, 1982

Molecular programming
Acknowledgments

- Microsoft Research
  - Andrew Phillips, Biological Computation Group

- Caltech
  - Winfree Lab

- U.Washington
  - Seelig Lab
Resources

• Biological Computation Group at MSR
  https://www.microsoft.com/en-us/research/group/biological-computation/

• Molecular Programming Project at Caltech
  http://molecular-programming.org/

• Georg Seelig’s DNA Nanotech Lab at U.W. CS&E
  http://homes.cs.washington.edu/~seelig/

• “DNA Computing and Molecular Programming” Conference Proceedings
  http://www.dna-computing.org/