Molecular Programming

Luca Cardelli
Microsoft Research

Seven Keys to the Digital Future, Edinburgh 2010-09-30
http://lucacardelli.name


<10 iterations of Moore's Law left! The race is on for molecular scale integrated circuits.
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.*
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 each other.

• At the molecular scale many such materials exist; let’s pick one...
DNA

GC Base Pair
Guanine–Cytosine

TA Base Pair
Thymine–Adenine

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

Sequence of Base Pairs (GACT alphabet)
Robust, and Long

- **DNA in each human cell:****
  - 3 billion base pairs
  - 2 meters long, 2nm thick
  - folded into a 6µm ball
  - 750 MegaBytes

- **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
Zipping Along

- 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
Nanoscale Engineering

- Sensing
  - Reacting to forces
  - Binding to molecules
- Actuating
  - Releasing molecules
  - Producing forces
- Constructing
  - Chassis
  - Growth
- Computing
  - Signal Processing
  - Decision Making

Nucleic Acids can do all this. And interface to biology.
Hybridization

- Strands with **opposite orientation and complementary base pairs** stick to each other (Watson–Crick duality).
- This is all we are going to use
  - We are not going to exploit DNA replication, transcription, translation, restriction and ligation enzymes, etc., which enable other classes of tricks.
Hybridization Tricks
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

36 nt, 12.6 nm
2D DNA Lattices

Chengde Mao
Purdue University, USA

N–point Stars
3D DNA Structures

Ned Seeman
NYU

Andrew Tuberfield
Oxford

3D Crystal

Tetrahedron
S.M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf and W. M. Shih
Self-assembly of DNA into nanoscale three-dimensional shapes, Nature (2009)
DNA Origami

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


Black: long viral strand
Color: short staple strands
DNA Origami

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

This means we can already self-assemble meso-scale structures.

Paul W K Rothemund
California Institute of Technology
DNA Circuit Boards

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

Greg Wallraff, IBM

Sensing
Aptamers

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

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

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

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

A Synthetic DNA Walker for Molecular Transport
Jong-Shik Shin* and Niles A. Pierce*†
Department of Bioengineering and Applied Mathematics, California Institute of Technology, Pasadena, California 91125
Hybridization Chain Reaction

Stable mixture of two hairpins

Initiator

1. N by hybridization

Robert M. Dirks and Niles A. Piercel-5
Polymerization Motor

An autonomous polymerization motor powered by DNA hybridization

SUVR VENKATARAMAN*, ROBERT M. DIRKS†, PAUL W. K. ROTHMUND†, ERIK WINFREE† and NILES A. PIERCE†∗

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

ROBERT A. YERKES, STANLEY F. BATES, MARCEL G. FECHAT, and TED RICKRATH
Computing

Basic Notions
Compositionality

• Sensors and Actuators at the 'edge' of the system
  o They can use disparate kinds of inputs (sensors) and outputs (actuators)

• The 'kernel' of the system computes
  o Must use uniform inputs and outputs

• Compositionality in the kernel
  o Supporting 'arbitrary' computing complexity
  o The output of each computing components must be the same kind of 'signal' as the input

  sdf
  o If the inputs are voltages, the outputs must be voltages
  o If the inputs are DNA, the outputs must be DNA

• Central design question
  o What should our signals (not components!) be?
  o Then design components that manipulate those signals.
What does DNA Compute?

• Electronics has *electrons*
  o All electrons are the same: you can only count them
  o *Few* electrons = *False*; *lots* of electrons = *True*
  o But *Boolean Logic* is only a necessary evil to build symbolic computation

• DNA computing has *symbols* (DNA words)
  o DNA words are not all the same
  o *Symbolic computation on abstract signals* can be done *directly*
  o Signals are presented *concurrently* (in a soup)
  o No requirement to do Boolean Logic

• Then, what are our ‘*gates*’ (if not Boolean?)
  o Theory of Concurrency
  o Process Algebra as the “Boolean Algebra” of DNA Computing
Why Compute with DNA?

- Not to solve NP-complete problems.
- Not to put Intel out of business.
- Not to orchestrate protein production.
- To precisely control the organization and dynamics of matter and information at the molecular level.
  - The use of DNA is “accidental”.
  - No genes involved.
  - In fact, no material of biological origin.
Rules of the Game

- Short complementary segments hybridize reversibly

- Long complementary segments hybridize irreversibly
DNA Strand Displacement

- Short strand (*toehold*): reversible binding
- Long strand (*body*): irreversible binding

Random collision/breakup  Random walk  Entropy gain
Failed Strand Displacement

• What if the input does not match the gate?
Failed Strand Displacement
Failed Strand Displacement
Failed Strand Displacement
Failed Strand Displacement
Failed Strand Displacement
Failed Strand Displacement

- **Hence an incorrect binding will undo**
  - That’s why toeholds must bind reversibly

- **Matching depends on the long segment only**
  - Strand displacement succeeds iff the whole long segment matches
  - The address space is determined by the size of the long segment, which is unbounded (not by the size of the toehold)
  - The toehold is just a ‘cache’ of the address
Computing

Implementing “Arbitrary” Computing Functions
Signals

• A signal is the representation of an abstract event
  o E.g. generated by a sensor
  o E.g. accepted by an effector
  o We are not limited to true/false

• 3-domain signals
  o $x_h$: history (ignore)
  o $x_t$: toehold (binding)
  o $x_b$: body (recognition)

• Signals (single stranded DNA) are prepared by (artificial) DNA synthesis
Gates

- Double-stranded structures with free toeholds

Gates are prepared by self-assembly from single-stranded DNA that is synthesized
Fork Gate

- \( x \rightarrow y + z \)
- \( x \rightarrow y + 0 \) transform \( x \) to \( y \) (transducer)
- \( x \rightarrow x + y \) linear production of \( y \) (catalyst)
- \( x \rightarrow x + x \) exponential production of \( x \) (amplifier)
This is the Fork Gate structure
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate
Fork Gate

This is Waste
Join Gate

- $x + y \rightarrow z$
Join Gate

This is the Join Gate structure.
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate
Join Gate

This is Waste
General n-Join/m-Fork Gate

Garbage collection
Gate Design Verification

• Active garbage
  o The active join residuals slow down the performance of following joins.
  o → Add a garbage collector to remove the active residuals.

• Interference between gates
  o The join garbage collector interferes with the fork gate.
  o → Modify the fork gate to remove the interference.

• What else could go wrong?
  o Endless possibilities.
  o → Prove that the fork/join gate structures correctly implement fork/join in all larger circuits.
Strand Algebra

\[ x_1 \mid \ldots \mid x_n \mid [x_1, \ldots, x_n] \cdot [y_1, \ldots, y_m] \rightarrow y_1 \mid \ldots \mid y_m \]

- Join + Fork + Populations = (Stochastic) Petri Nets
Curing
A Doctor in Each Cell

Fig. 1 Medicine in 2050: “Doctor in a Cell”
Tools
So we can in principle work at this level.
Visual DSD
A Strand Displacement Simulator

Matthew Lakin, Simon Youssef, Andrew Phillips

http://lepton.research.microsoft.com/webdna/
A programming language for composable DNA circuits

Andrew Phillips* and Luca Cardelli

A. Syntax of DNA molecules $D$

Upper strand with sequence complementary to $S$

\[ \text{S} \]

\[ \langle S \rangle \]

Molecule with segments $G_1, \ldots, G_k$

\[ G_1 : G_2 : \ldots : G_k \]

Parallel molecules $D_1, \ldots, D_k$

\[ D_1 \mid D_2 \mid \ldots \mid D_k \]

Molecules $D$ with private domains $N_1, \ldots, N_k$

\[ (N_1, \ldots, N_k) \]

\[ \text{new } (N_1, \ldots, N_k) \]

B. Syntax of DNA segments $G$

Lower strand with toehold $N^c$

\[ N^c \]

\[ N^c \]

Double strand with sequence $S$ and overhangs $L, R$

\[ <L> [S] <R> \]

C. Syntax of DNA sequences $S, L, R$

Sequence of domains $O_1, \ldots, O_k$

\[ O_1 \quad O_2 \quad \ldots \quad O_k \]

\[ 01 \quad 02 \quad \ldots \quad 0k \]
Dynamics

1. Toehold binding and unbinding
   \[ L N R \leftrightarrow G_1 N^c G_2 \]

2. Strand displacement to the right
   \[ L_1 L_2 S_2 R_2 \rightarrow L_1 S_1 S_2 R_1 \]
   \[ L_1[S_1][S_2]R_1 \leftrightarrow L_2[S_2]R_2 \]

3. Strand displacement to the left
   \[ L_1 L_2 S_2 R_2 \rightarrow L_2 S_1 S_2 R_1 \]
   \[ L_1[S_1][S_2]R_1 \leftrightarrow L_2[S_2]R_2 \]

4. Branch migration
   \[ L_1 L_2 S_2 S_3 R_2 \leftrightarrow L_1 S_1 S_2 S_3 S_2 R_2 \]
   \[ L_1[S_1][S_2]R_1 \leftrightarrow L_2[S_2][S_3][S_2]R_2 \]
Initial Species
Simulation
### DNA Sequences

<table>
<thead>
<tr>
<th>Code</th>
<th>DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check sequences</td>
<td>Reset</td>
</tr>
</tbody>
</table>

#### TOEHOLES SEQUENCES
- TATTCC
- GCTA
- GTCA
- TACCA
- CTTGC
- ACTACAC
- CTCAG
- CTGAATC
- CCTAGG
- TCTCAG
- TCTCAG
- CCACT
- TAGC
- ACCT
- TAGGCA
- CACACA
- AGAC

#### SPECIFICITY SEQUENCES
- CCCAAAACAACACACAAACACAA
- CCCCCCAATAACACAAAATACACAA
- CCCCCCACTATATCACAAACAA
- CCCCCCACTATATACAAACAAA
- CCCCCCACTATATACAAACAAA
- CCCCCCACTATATACAAACAAA

<table>
<thead>
<tr>
<th>Spec</th>
<th>Reactions</th>
<th>Graph</th>
<th>Text</th>
<th>Data</th>
<th>Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>TATTCC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GCTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CCCCTTACATTACATAACAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CCCAAAAACAAAAACAAAAACAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CCCCCCTTAACATAACAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CCCCCATCATATCAATACAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Zoom:**
- 0
Experiments
How are they Actually Done?

Postdoc

Faculty
Sequences to DNA

Oligonucleotide Synthesis

Gene Synthesis
- Synthesize gene at $0.39/bp (until 3/31/2010)
- Guaranteed 100% sequence fidelity
- CloneEZ® seamless cloning technology

Struggling with cloning? Try Gene-on-Demand® Service!
- Fastest turnaround time for less money!
- Standard gene synthesis from 0.36 €/bp in just 8 days

Find Out G-Reward
- Earn rewards for every purchase!

Your Chance to Win a Nintendo Wii

SameDay® Oligo Service

Only £0.57 GBP / Base!

<table>
<thead>
<tr>
<th>Synthesis Scale</th>
<th>Base Pricing</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 nmole DNA Oligo</td>
<td>£0.25 GBP / Base</td>
<td></td>
</tr>
<tr>
<td>100 nmole DNA oligo</td>
<td>£0.45 GBP / Base</td>
<td></td>
</tr>
<tr>
<td>250 nmole DNA oligo</td>
<td>£0.80 GBP / Base</td>
<td></td>
</tr>
<tr>
<td>1 µmole DNA oligo</td>
<td>£1.60 GBP / Base</td>
<td></td>
</tr>
<tr>
<td>5 µmole DNA oligo</td>
<td>£7.50 GBP / Base</td>
<td>Order</td>
</tr>
<tr>
<td>10 µmole DNA oligo</td>
<td>£14.50 GBP / Base</td>
<td>Order</td>
</tr>
</tbody>
</table>

Please inquire for larger quantities.
Next-Day Oligos!

SameDay® Oligo Service

Only $0.60 USD / Base!
The Current Time is 8:12PM ET

Order on or before 3:00pm SameDay, priority shipping for delivery by 10:30am on second business day
2-OD minimum guarantee (sufficient for > 250 PCR reactions)
15-45 bases
Shipped lyophilized in tubes
Differentiated & desalted
Available within the U.S. and Canada
Unmodified

(25% shipping and handling fee for expedited services on SameDay® oligos within the U.S.)
(25% within Canada.)

Place an Order Now

© Copyright 2010 Integrated DNA Technologies, Inc.
Wait 24 Hours
DNA by Mail
Spec Sheet
Add Water
Put DNA into Gel

- Polyacrylamide gel electrophoresis (PAGE)
- Sorts DNA strands by length
Wait 6 Hours
Get DNA out of Gel

• Find DNA with ultraviolet light. Cut it out.
Wait 12 Hours
Mix DNA Up

• Screaming for robotic automation
Spectrofluorometer

- Fluorescence is your ‘print’ statement
Go To Lunch
Execution Trace
Core Dump
RINSE & Repeat
Publish!

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

• Don’t try this at home
  o (Although you could)

• Latex gloves, UV glasses
  o Fear the Gel (acrylamide)
  o Fear the Light (UV)

• Otherwise safe
  o No smells
  o No fires
  o No biohazards
  o No life forms

• Most complex machines:
  o Gel machine
  o Fluorometer
  o Atomic force microscope

• Most dangerous activity:
  o Replacing the light bulb in the fluorometer (hot; may explode)
DNA Compilation
Compilers

Monolithic Compilers

Language Design #1
Boolean Networks
Language Implementation #1

Language Design #2
Petri Nets
Language Implementation #2

Language Design #3
... Language Implementation #3
Intermediate Languages

Front End

Intermediate Language

Back End

Strand Algebra

Boolean Networks

Petri Nets

The algebra of fork and join gates
Front Ends

Circuit Design

Intermediate Language

Boolean Networks

Petri Nets

Strand Algebra

Intermediate Language #2
Back Ends

Intermediary Language

Gate Design

Structural Language

Device Design

4-domain Signals

3-domain Signals

2-domain Signals

Strand Algebra
Compiling Abstract Machines
Boolean Networks

This encoding is *compositional*, and can encode *any* Boolean network:
- multi-stage networks can be assembled (*combinatorial logic*)
- network loops are allowed (*sequential logic*)
Petri Nets

Petri Nets to Strand Algebra

Transitions as Gates
Place markings as Signals

\[
\begin{array}{c}
\text{p}_1 \quad \text{p}_2 \\
\text{p}_3 \quad \text{p}_4
\end{array}
\quad
([\text{p}_1, \text{p}_2].[\text{p}_3, \text{p}_4])^* \\
\text{p}_1 \mid \text{p}_1 \mid \text{p}_4
\]
Chemical Reaction Networks

Implementing an arbitrary finite chemical system in DNA with asymptotically correct kinetics
Soloveichick & al. DNA 15

Species become signals
Reactions become gates

\[ A + B \rightarrow C + D \quad \Rightarrow \quad \lbrack A,B \rbrack \cdot \lbrack C,D \rbrack \]
Interacting Automata

This is a uniform population of identical automata, but heterogeneous populations of interacting automata can be similarly handled.

([A,B].[B,B])\*  |  ([B,C].[C,C])\*  |  ([C,A].[A,A])\*  |
A | A | B | C
This is a uniform population of identical automata, but heterogeneous populations of interacting automata can be similarly handled.
Interacting Automata

This is a uniform population of identical automata, but heterogeneous populations of interacting automata can be similarly handled.
Interacting Automata

This is a uniform population of identical automata, but heterogeneous populations of interacting automata can be similarly handled.
And finally...
Workflow

- Abstract Machines to Strand Algebra
  - Or other intermediate language
- Strand Algebra to DSD
  - Or other structural language
- Simulation, analysis, etc.
  - Design iteration
- DSD to Sequences
  - E.g. NuPack, or pre-build strand libraries
- Sequences to DNA
  - Web order
- DNA experiments
  - Fairly basic wet lab
- Deployable Nanotech
Conclusions

• Programmable Matter
  o Nucleic acids

• Molecular Computation
  o DNA strand displacement

• Molecular Compilation
  o From programming abstractions (Petri Nets, Process Algebra, etc.), through intermediate language (Strand Algebra) to molecule synthesis (DNA).

• Correctness
  o Ensuring molecular programs work as intended
  o Through thermodynamic analysis, simulation, formal verification.
Acknowledgments

• Illustrations
  o John Reif, Duke
  o Ned Seeman, NYU
  o Erik Winfree, Caltech
  o Bernard Yurke, Boise State
  o Molecular movies by Drew Berry
  o Wikipedia, YouTube

• David Soloveichik