REACTION ENUMERATION & CONDENSATION OF DOMAIN-LEVEL STRAND DISPLACEMENT SYSTEMS

Stefan Badelt

DNA and Natural Algorithms (DNA) Group, Caltech
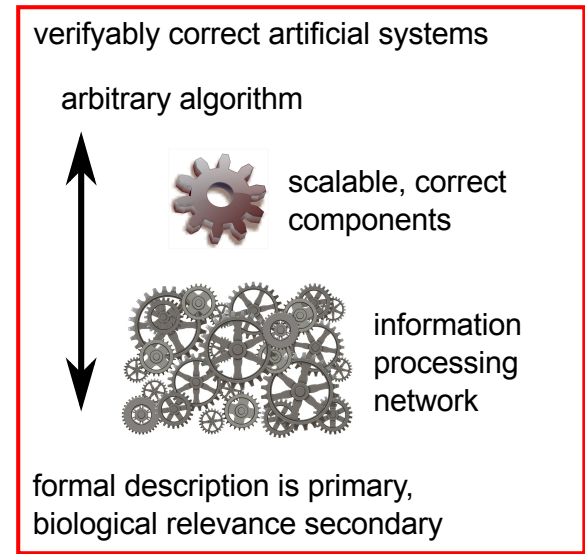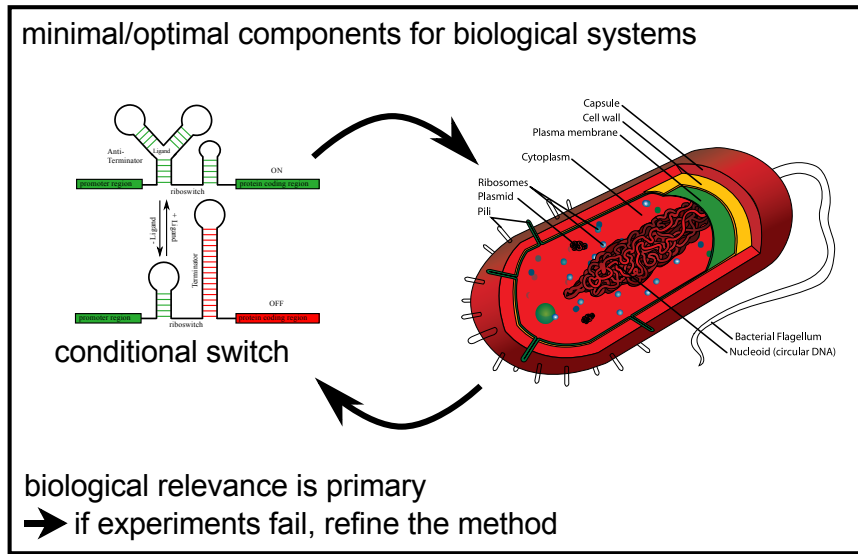
Feb 14th, 2018

33rd TBI Winterseminar, Bled, Slovenia

Grun, Badelt, Sarma, Shin, Wolfe, and Winfree (manuscript in preparation)
http://www.github.com/DNA-and-Natural-Algorithms-Group/peppercornenumerator
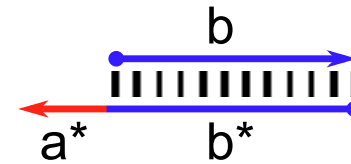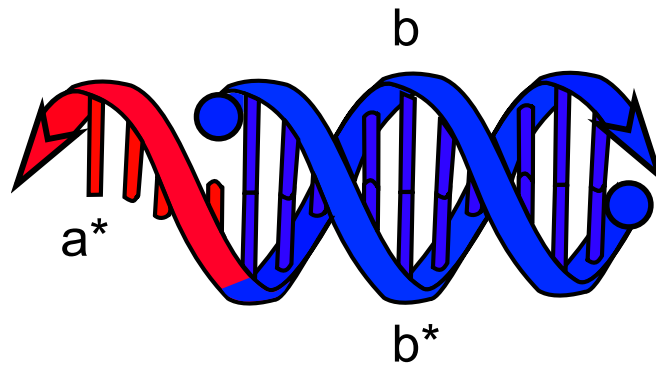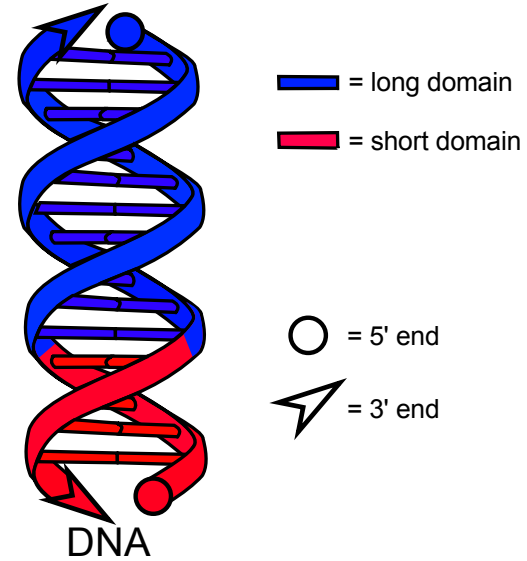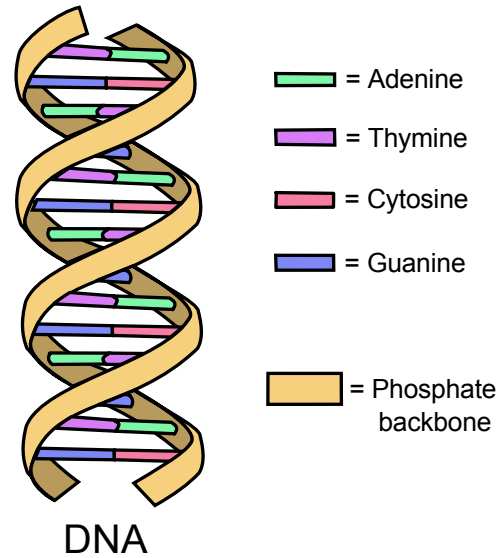
MOLECULAR PROGRAMMING

(in terms of the nuskell compiler project)

nucleic acids are architecture to implement algorithms
chemical reaction networks are a programming language
formal/experimental verification of correct implementation

minimal/optimal components for biological systems

conditional switch

biological relevance is primary
➡ if experiments fail, refine the method

verifyably correct artificial systems

arbitrary algorithm

scalable, correct components

information processing network

formal description is primary,
biological relevance secondary

DNA STRAND DISPLACEMENT

= Adenine

= Thymine

= Cytosine

= Guanine

= Phosphate backbone

DNA

= long domain

= short domain

= 5' end

= 3' end

DNA

b

a*

b*

b

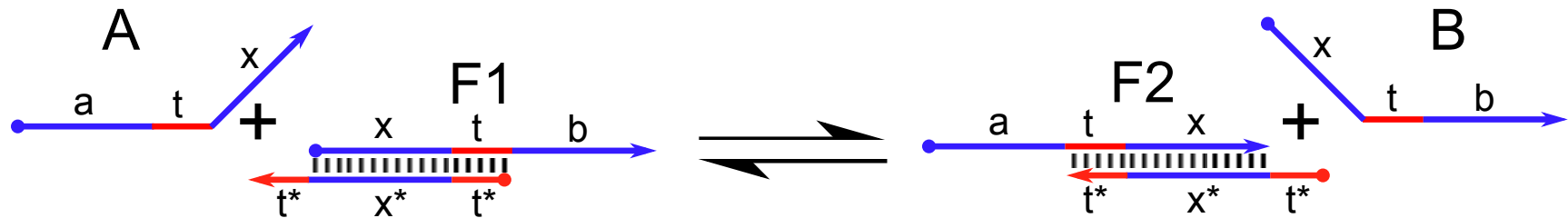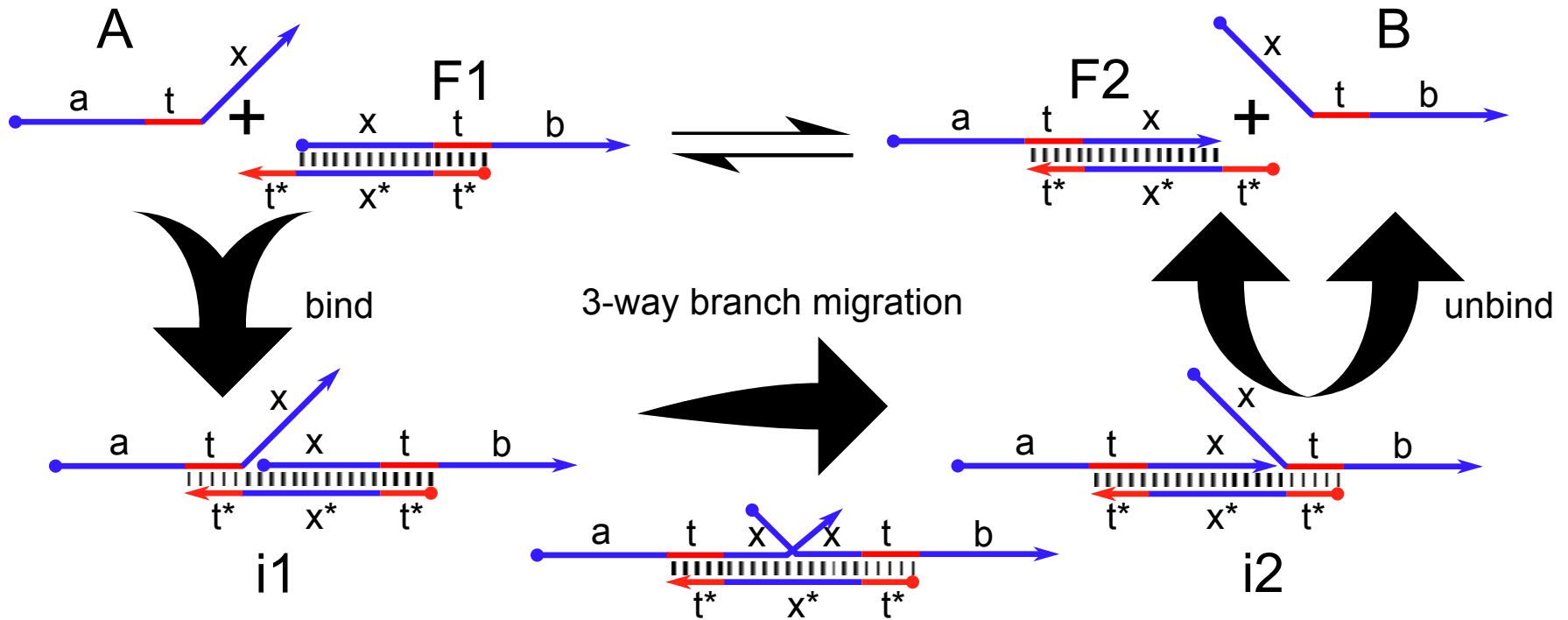a*   b*

DOMAIN-LEVEL STRAND DISPLACEMENT

long (branch-migration) domain: binds irreversibly
short (toehold) domain: binds reversibly

DOMAIN-LEVEL STRAND DISPLACEMENT

long (branch-migration) domain: binds irreversibly
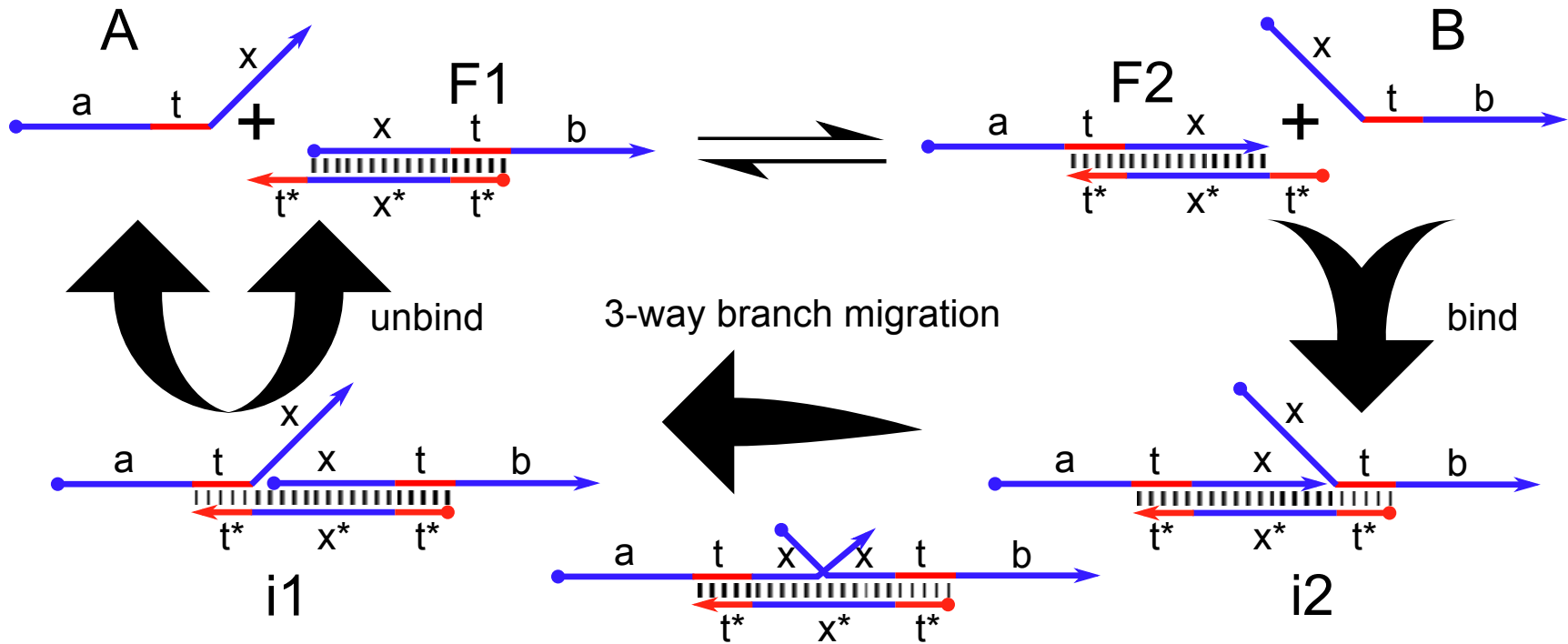short (toehold) domain: binds reversibly

detailed network
condensed network

A

B

F1

F2

+

+

bind

3-way branch migration

unbind

i1

i2

DOMAIN-LEVEL STRAND DISPLACEMENT

long (branch-migration) domain: binds irreversibly

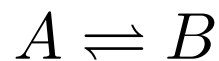short (toehold) domain: binds reversibly

detailed network

condensed network

A

B

F1

F2

+

+

unbind

3-way branch migration

bind

i1

i2

DOMAIN-LEVEL STRAND DISPLACEMENT

long (branch-migration) domain: binds irreversibly
short (toehold) domain: binds reversibly

formal CRN

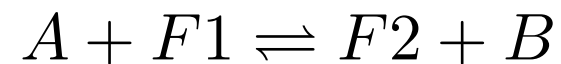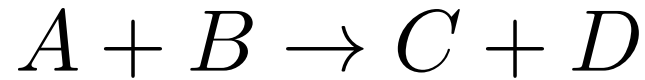$$A \rightleftharpoons B$$
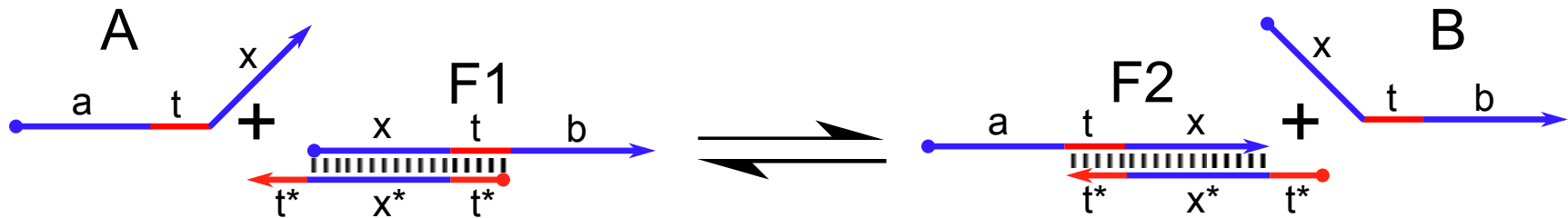
formal species: {A, B}
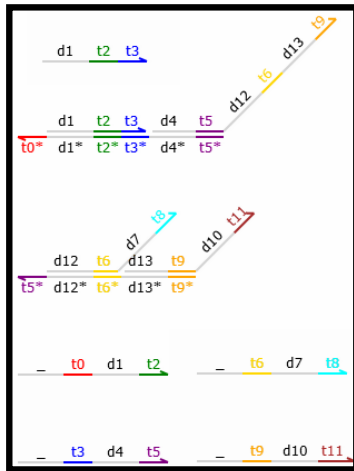
DSD sytem specification

$$A + F1 \rightleftharpoons F2 + B$$

signal species (low concentation): {A, B}

fuel species (high concentration): {F1, F2}

FROM CRN TO DSD SYSTEMS

$$A + B \rightarrow C + D$$
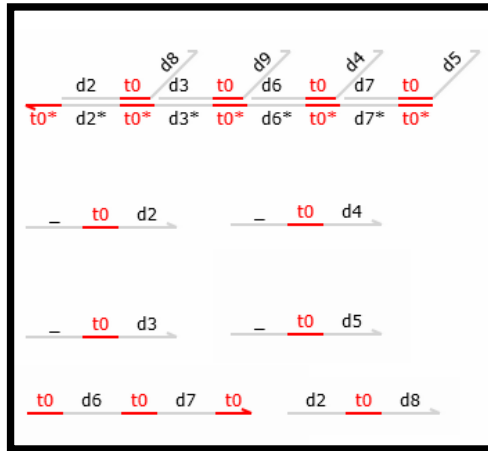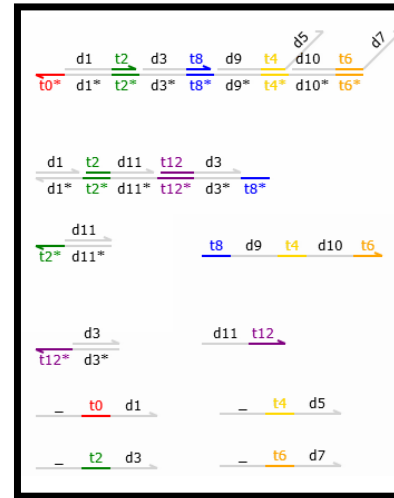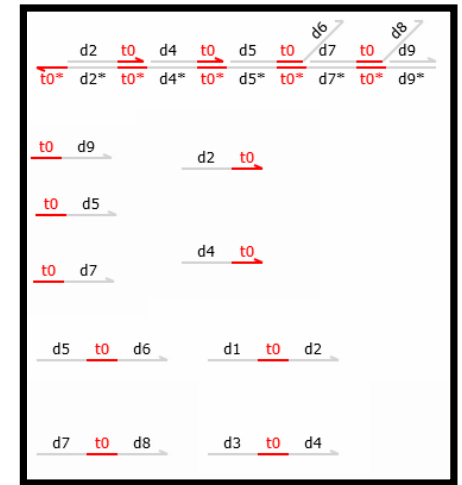

Soloveichik et al. (2010)

Lakin et al. (2012)

Cardelli (2011)

Qian et al. (2011)

Chen et al. (2012), Cardelli (2013), Srinivas (2015), Lakin et al. (2016), ...

Images drawn using VisualDSD, Lakin et al. (2012)

FROM A DIGITAL CIRCUIT TO DSD

$$y_2 y_1 = \left\lfloor \sqrt{x_4 x_3 x_2 x_1} \right\rfloor$$

Qian et al. (2011)

fanout:

$$x_{on} \rightarrow i_{on} + j_{on} + k_{on}$$
$$x_{off} \rightarrow i_{off} + j_{off} + k_{off}$$

AND gate:

$$i_{off} + j_{off} \rightarrow y_{off}$$
$$i_{off} + j_{on} \rightarrow y_{off}$$
$$i_{on} + j_{off} \rightarrow y_{off}$$
$$i_{on} + j_{on} \rightarrow y_{on}$$

```
# The CRN computes the floor of the squareroot of a four-bit binary number.

# Fanout X3 -> F5 + F6 + F7
X3_OFF -> F5_OFF + F6_OFF + F7_OFF          # G16b = F5 AND F8
X3_ON  -> F5_ON  + F6_ON  + F7_ON           F5_OFF + F8_OFF -> G16b_OFF
# Fanout X4 -> F8 + F9 + F10                 F5_OFF + F8_ON -> G16b_OFF
X4_OFF -> F8_OFF + F9_OFF + F10_OFF          F5_ON  + F8_OFF -> G16b_OFF
X4_ON  -> F8_ON  + F9_ON  + F10_ON           F5_ON  + F8_ON  -> G16b_ON
# G11 = NOT(X1 OR X2) + Fanout G11 -> F14 + F15   # G16 = NOT(F14 AND G16b)
X1_OFF + X2_OFF -> F14_ON  + F15_ON          F14_OFF + G16b_OFF -> G16_ON
X1_OFF + X2_ON  -> F14_OFF + F15_OFF         F14_OFF + G16b_ON -> G16_ON
X1_ON  + X2_OFF -> F14_OFF + F15_OFF         F14_ON  + G16b_OFF -> G16_ON
X1_ON  + X2_ON  -> F14_OFF + F15_OFF         F14_ON  + G16b_ON -> G16_OFF
# G12 = F6 AND (NOT F9)                       # G17 = F15 OR G12
F6_OFF + F9_OFF -> G12_OFF                   F15_OFF + G12_OFF -> G17_OFF
F6_OFF + F9_ON -> G12_OFF                    F15_OFF + G12_ON -> G17_ON
F6_ON  + F9_OFF -> G12_ON                    F15_ON  + G12_OFF -> G17_ON
F6_ON  + F9_ON -> G12_OFF                    F15_ON  + G12_ON -> G17_ON
# Y2 = F7 OR F10                              # Y1 = NOT(G16 AND G17)
F7_OFF + F10_OFF -> Y2_OFF                   G16_OFF + G17_OFF -> Y1_ON
F7_OFF + F10_ON -> Y2_ON                     G16_OFF + G17_ON -> Y1_ON
F7_ON  + F10_OFF -> Y2_ON                    G16_ON  + G17_OFF -> Y1_ON
F7_ON  + F10_ON -> Y2_ON                     G16_ON  + G17_ON -> Y1_OFF
```

Input for the nuskell compiler: **32** formal reactions.

soloveichik2010.ts: **52** signal species, **92** fuel species, **172** intermediate species, **180** reactions.
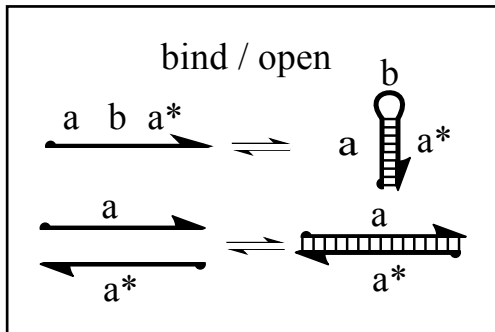
verifies as **correct** according to the **pathway decomposition** and **CRN bisimulation** equivalence

Badelt, Johnson, Dong, Shin, Thachuk and Winfree: A general-purpose CRN-to-DSD compiler with formal verification, optimization, and simulation capabilities. LNCS (2017)
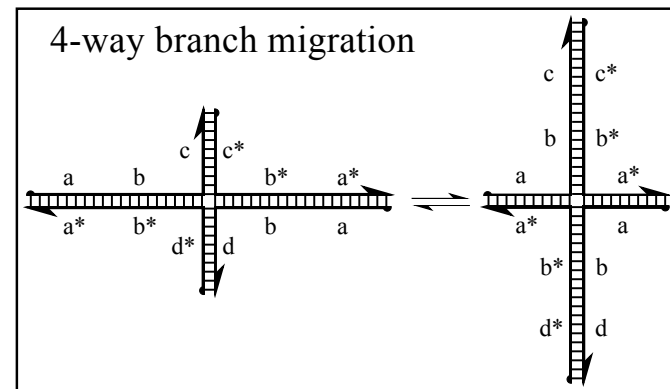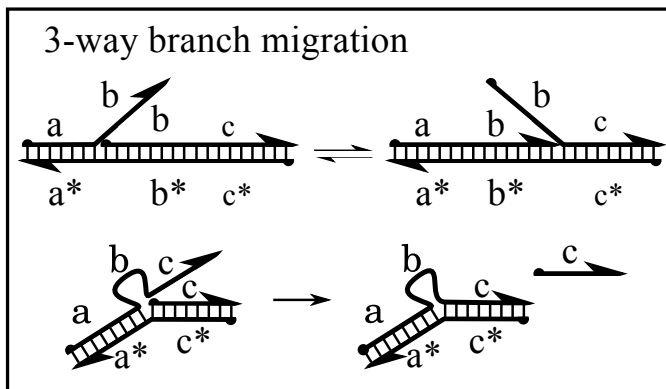
REACTION TYPES

bind / open

3-way branch migration
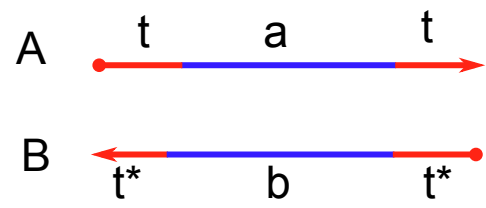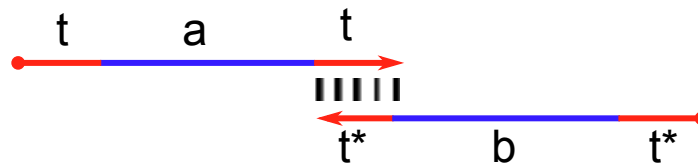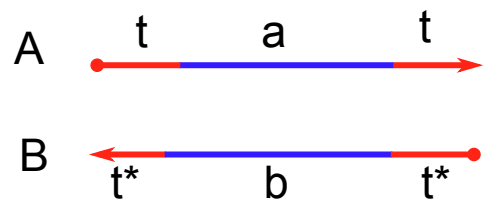
4-way branch migration
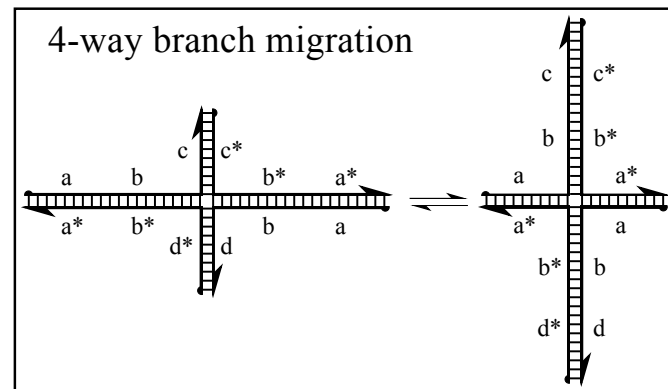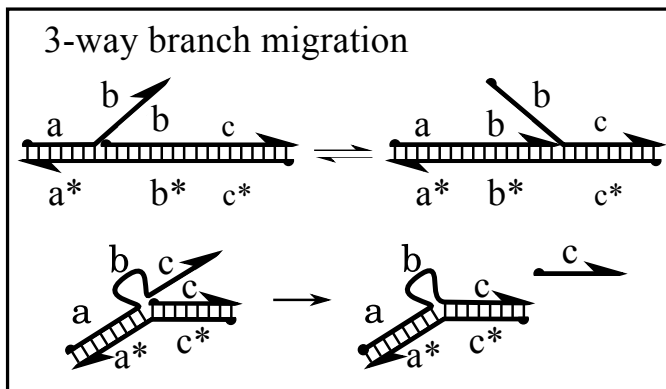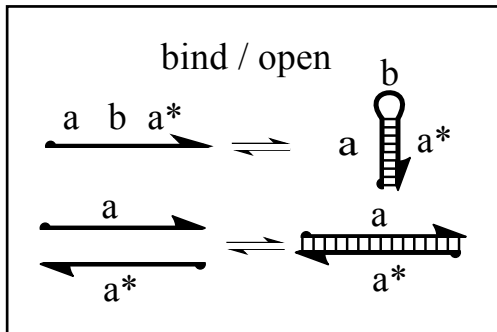
REACTION TYPES

bind / open

allows all secondary structures (pseudoknots excluded)
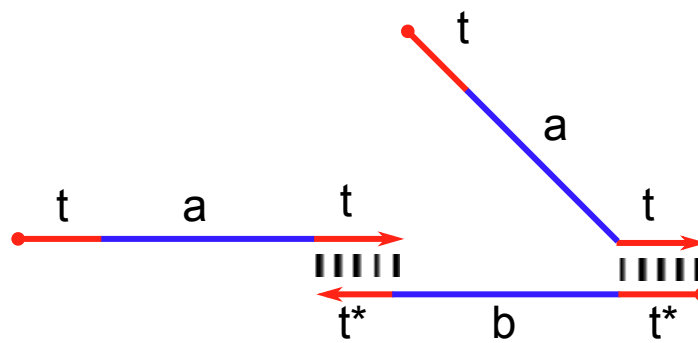open reactions of domains with length > L are forbidden

open & branch migration reactions are always
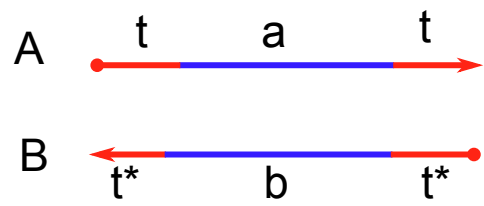unimolecular, but may lead to dissociation.

bind reactions are the only valid bimolecular reactions

3-way branch migration

4-way branch migration

A   t   a   t

B   t*   b   t*

A   t   a   t

B   t*   b   t*

t   a   t

t*   b   t*

A   t ——— a ——— t →

B   ← t* ——— b ——— t*

?

t a t + t* b t* + t a t + t* b t*
. . (+ (. . +) . (+) .)

17

multistranded pseudoknot

A   t   a   t
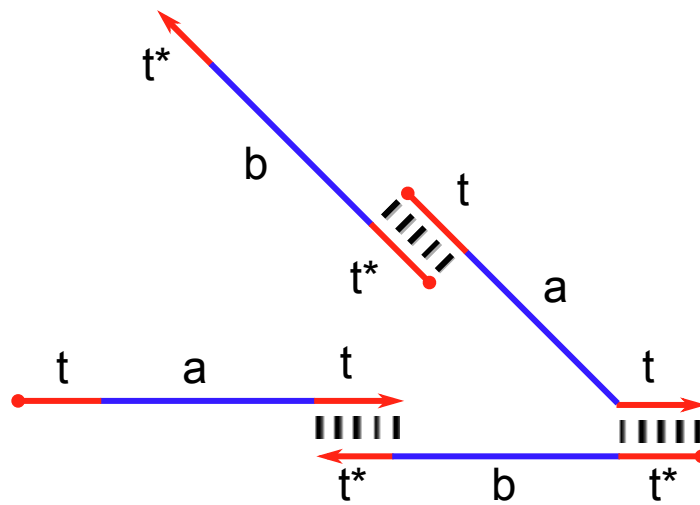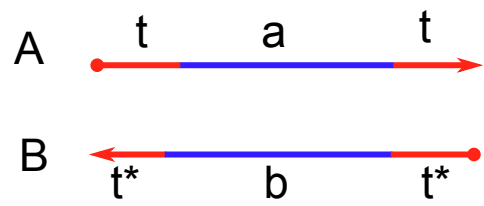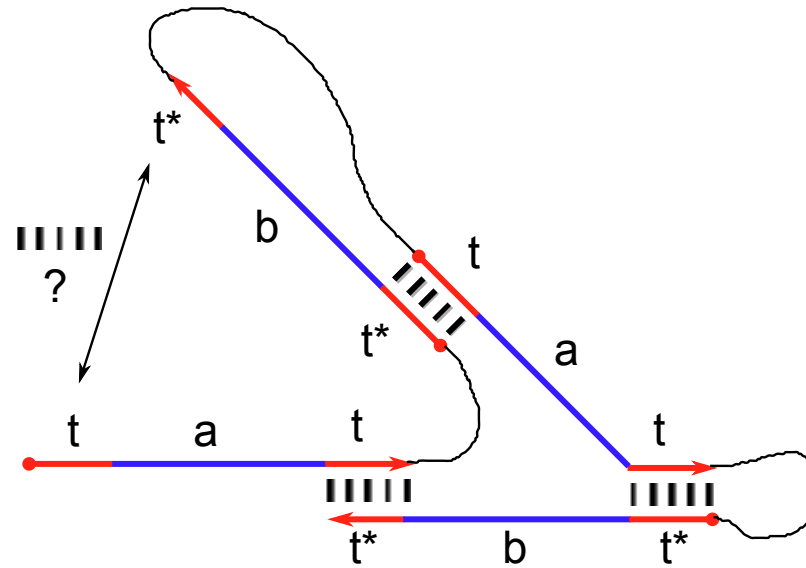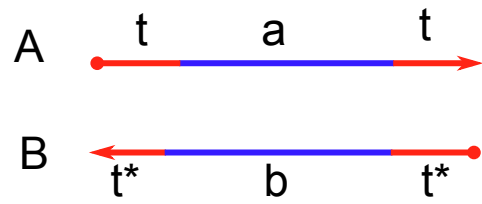
B   t*   b   t*
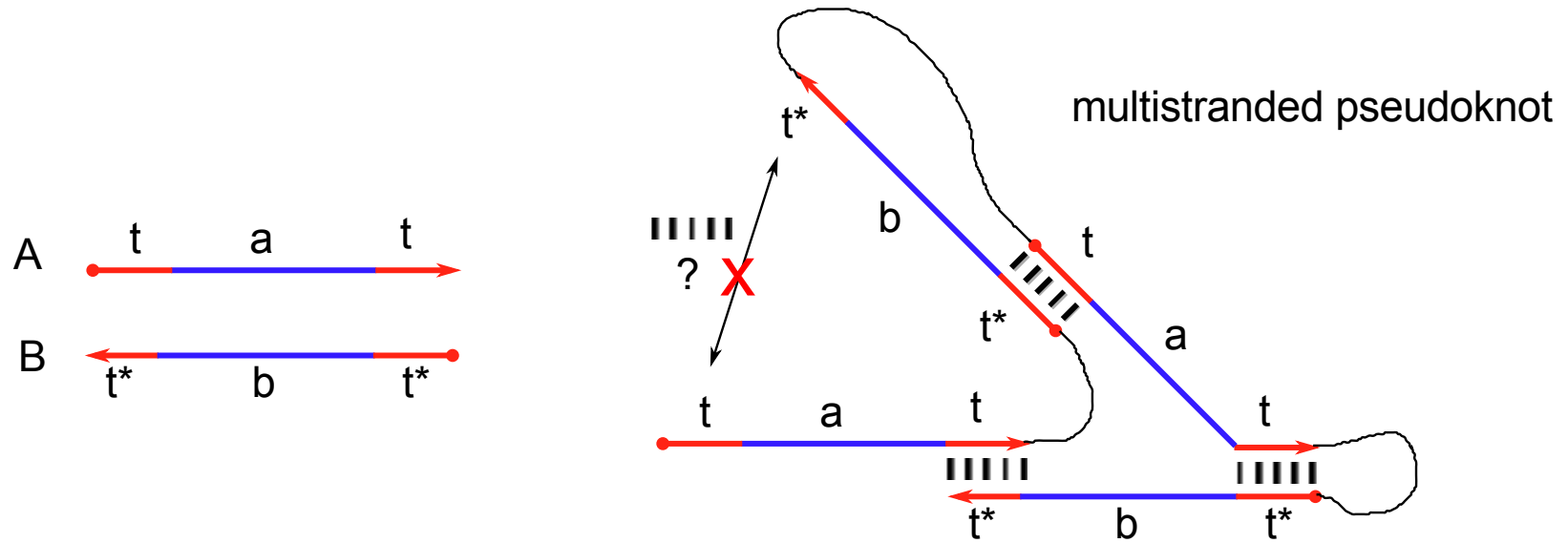
t a t + t* b t* + t a t + t* b t*
. . (+ (. . +) . (+) .)

SEPARATION OF TIMESCALES

unimolecular reactions are fast
bimolecular reactions are slow

resting complexes   transient complex

$$\{X \xrightarrow{k_\alpha} A + B; \; A + B \xrightarrow{k_\beta} X\}$$
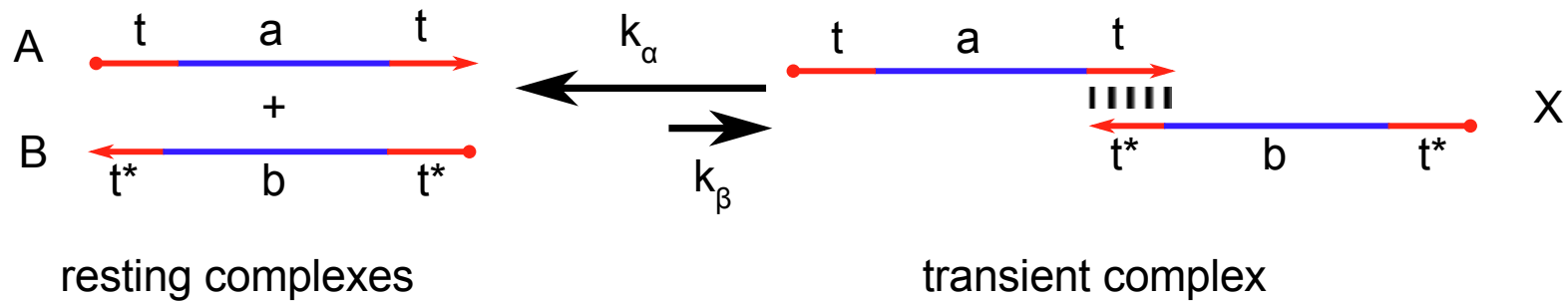
at low concentrations:

$$k_\beta[A][B] << k_\alpha[X]$$

MODEL PARAMETERS

rate-independent model

open reactions where domain-length $> L$ are negligible

unimolecular reactions are fast

bimolecular reactions are slow

rate-dependent model

assume typical rate constant for every reaction:

k = rate(rtype, dlength)

unimolecular reactions with $k < k_{\text{slow}}$ are negligible

unimolecular reactions with $k < k_{\text{fast}}$ are slow

unimolecular reactions with $k \geq k_{\text{fast}}$ are fast

bimolecular reactions are slow

REACTION ENUMERATION

- every complex has all valid fast reactions enumerated
- transient complexes have no slow reactions enumerated
- resting complexes have all valid slow reactions enumerated
- all initial complexes are included

**valid** according to enumeration semantics:

- all valid, except open $> L$
- max-helix semantics: reaction types are greedy
- probability threshold for reactants of bimolecular reactions.
- probability threshold for products of unimolecular reactions.

CRN CONDENSATION

Goal: represent CRN in terms of overall slow reactions

properties / requirements:

- all fast reactions are unimolecular

- reactions have arity (n,m)
  with n > 0 and m > 0

- reactants of slow reactions
  must be resting states

- reactants and products of
  fast (1-2) reactions are in
  different SCCs (mass conservation)

- ● microstate (complex)
- → fast (1,1) reaction
- ⇢ slow (1,1) reaction
- ⅄ fast (1,2) reaction
- ⇣ slow (2,1) reaction
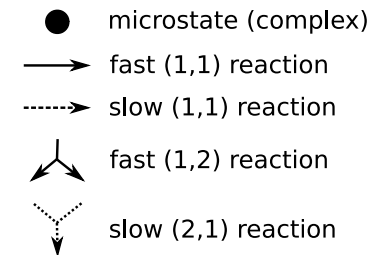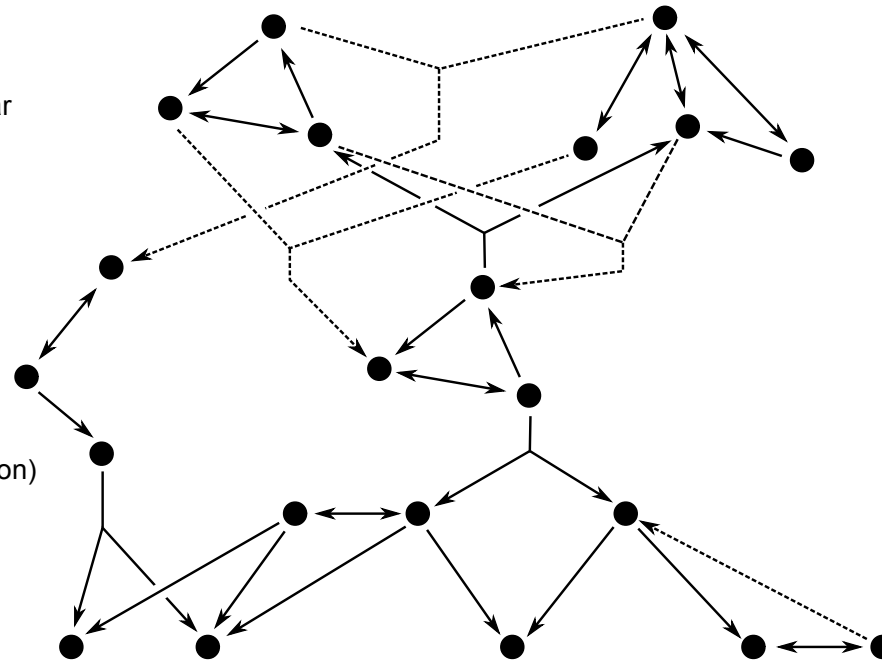
CRN CONDENSATION

Step 1: Make a graph that contains only fast (1,1) reactions

microstate (complex)
fast (1,1) reaction
slow (1,1) reaction
fast (1,2) reaction
slow (2,1) reaction
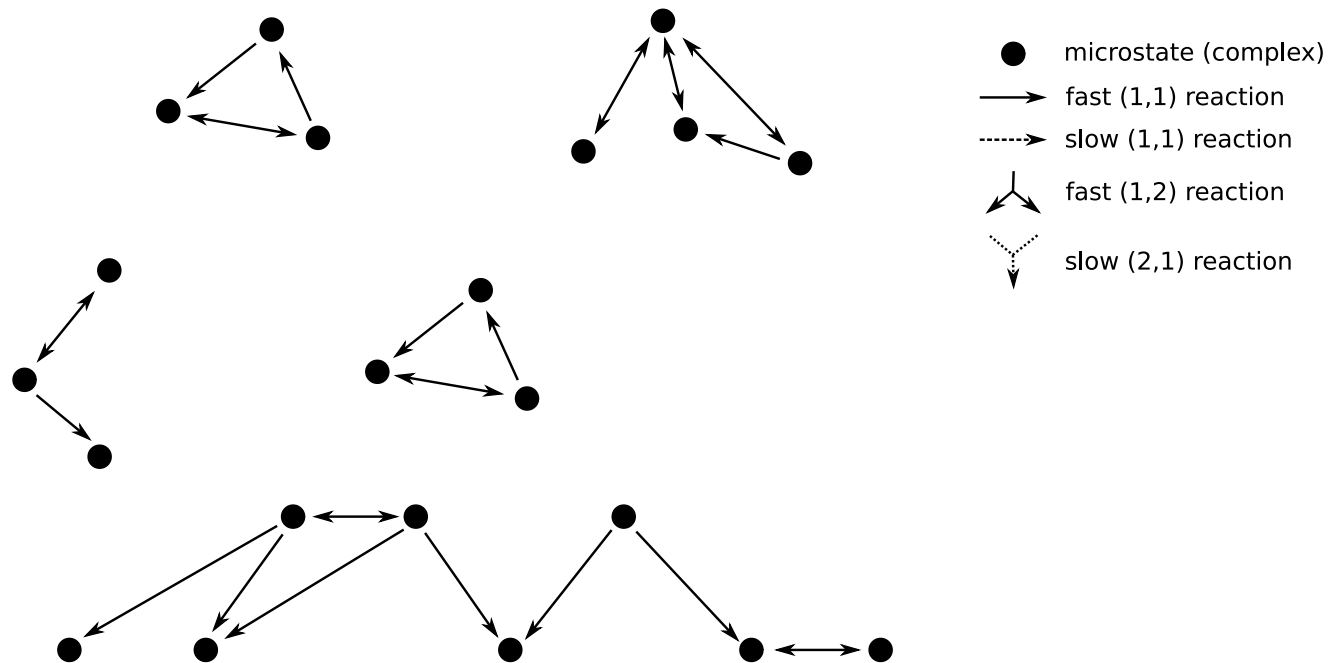
CRN CONDENSATION

Step 2: Identify strongly connected components (SCCs)

Legend:
- ● microstate (complex)
- → fast (1,1) reaction
- ┄→ slow (1,1) reaction
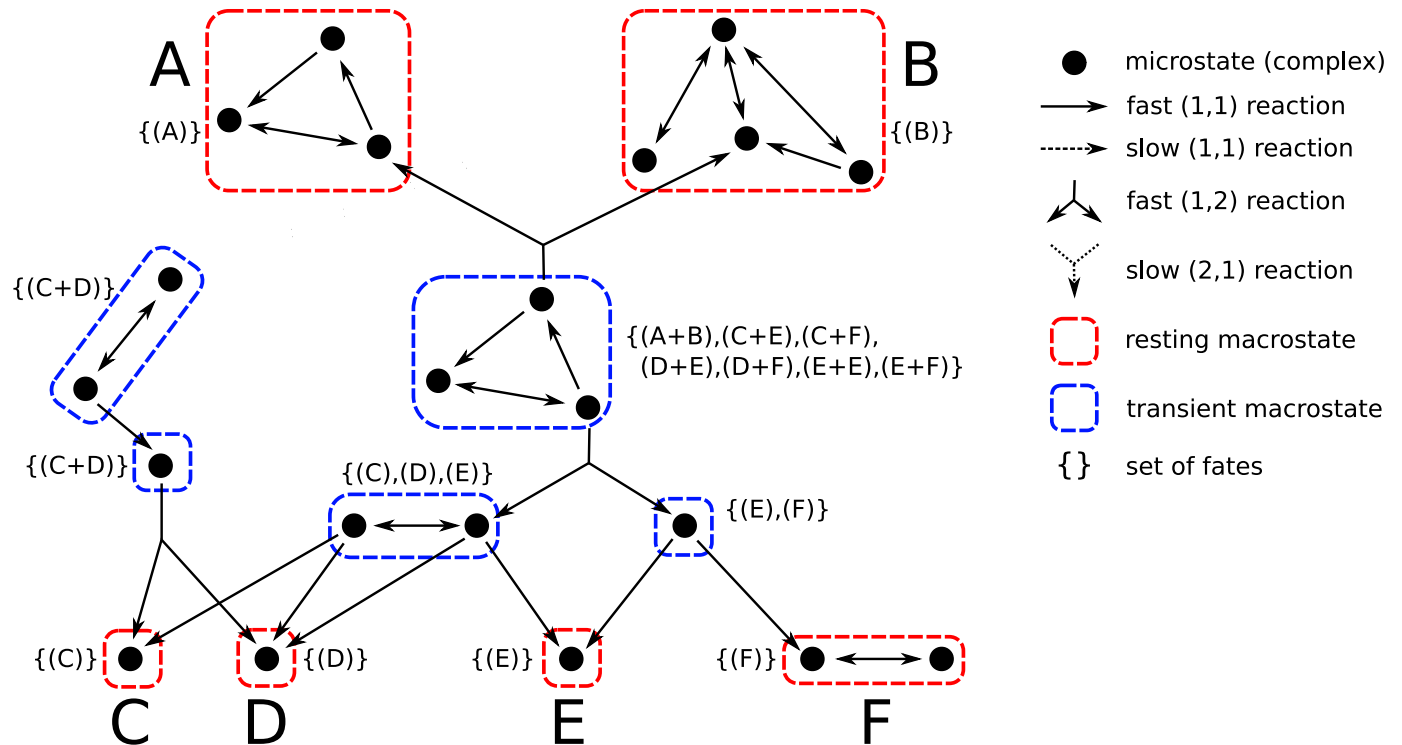- ⅄ fast (1,2) reaction
- ⅄ slow (2,1) reaction
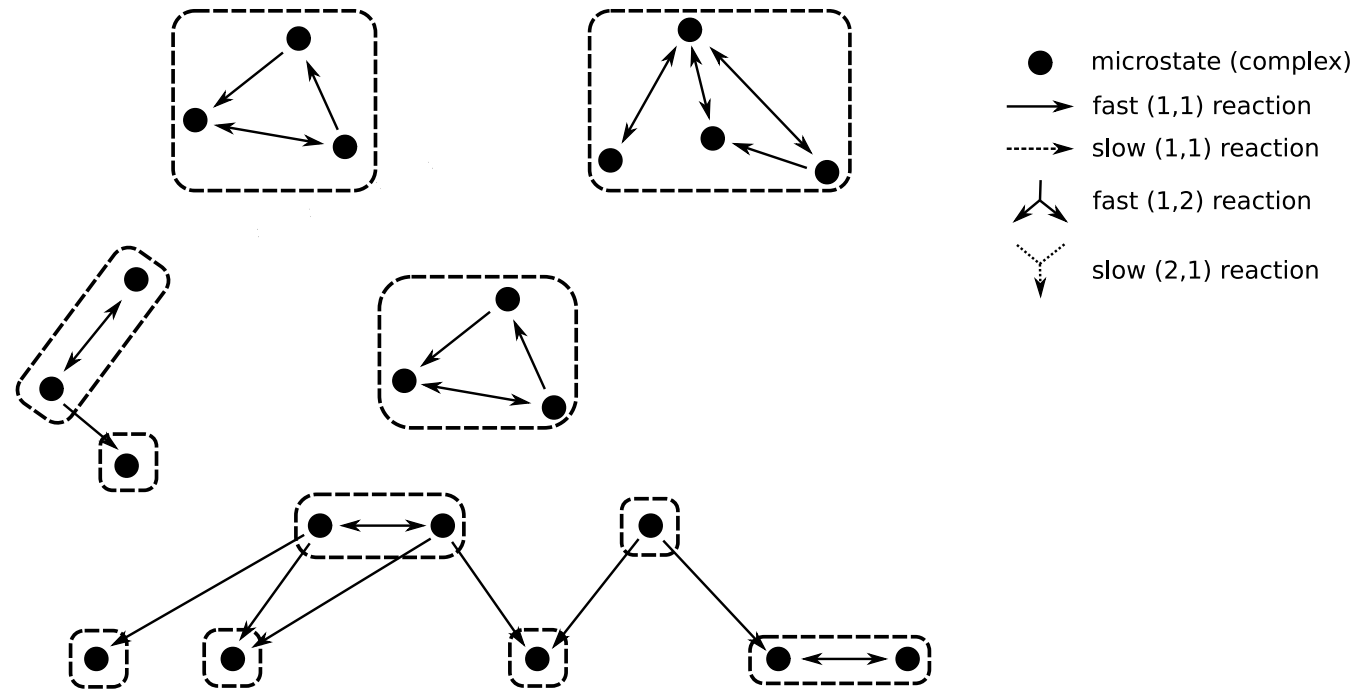
CRN CONDENSATION

Step 3: Define transient and resting macrostates
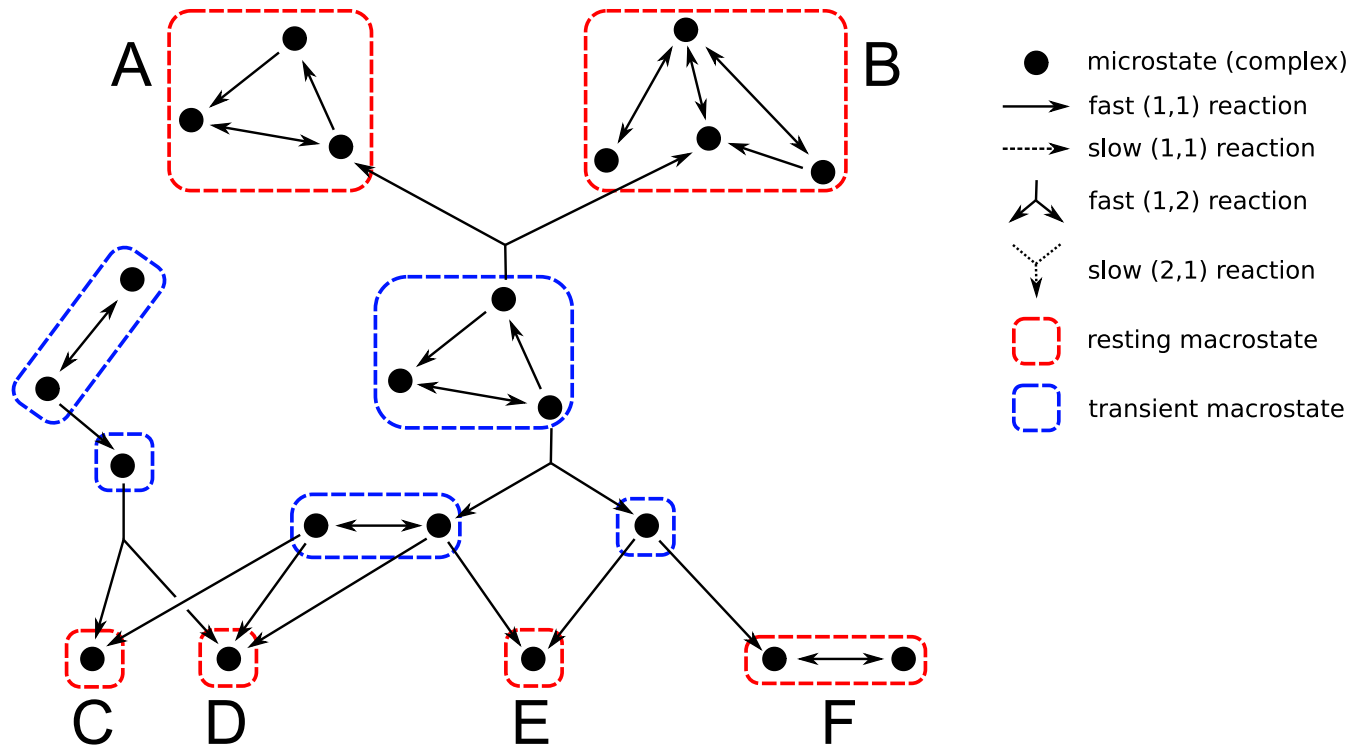
CRN CONDENSATION
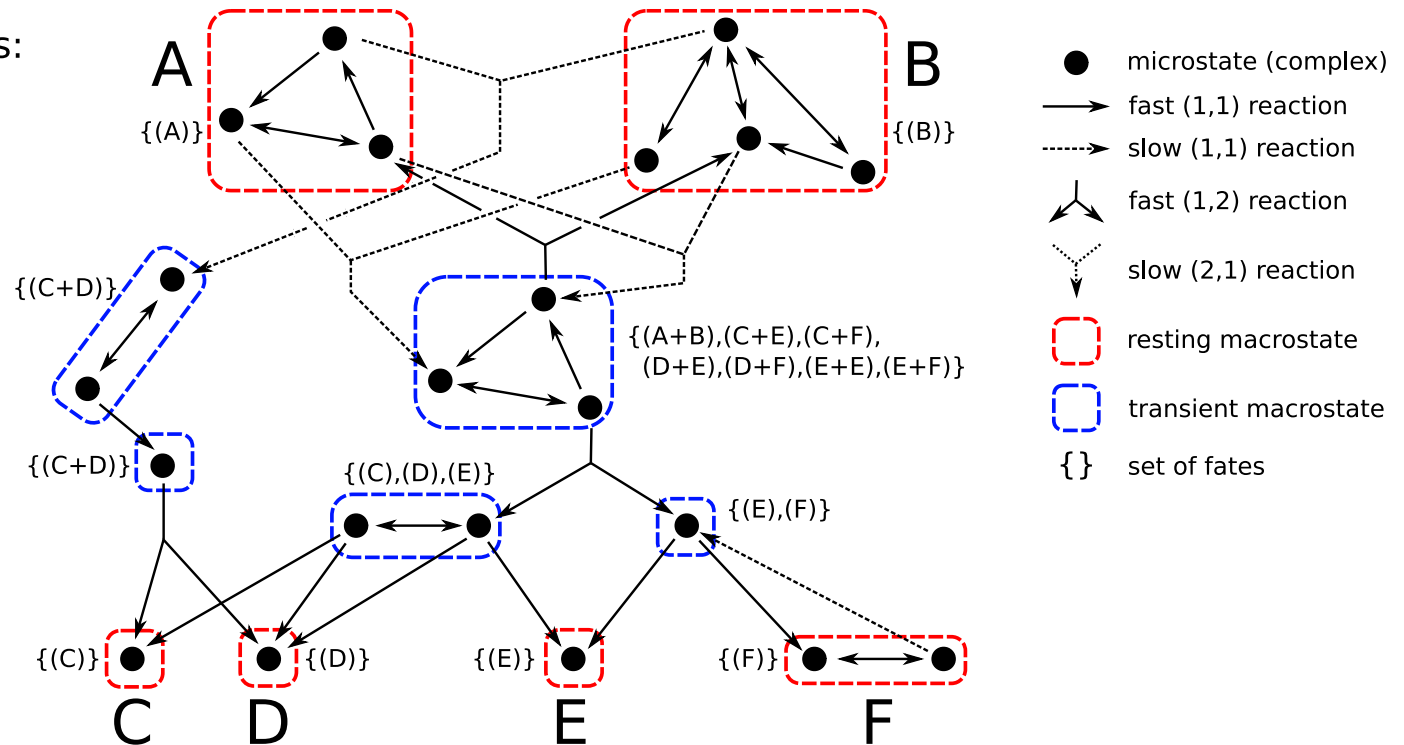
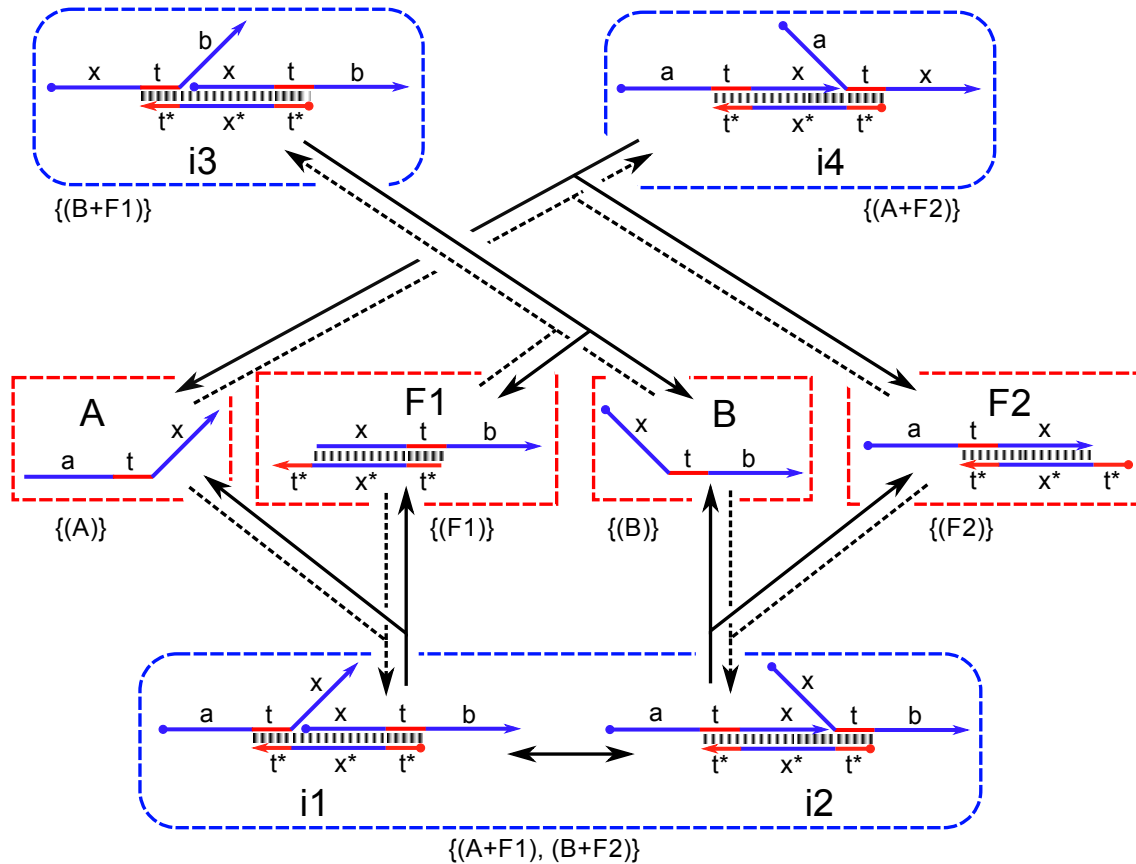## Step 4: Assign fates to complexes (or macrostates)

Step 5: Insert slow reactions & derive condensed reactions

condensed reactions:

```
A+B -> A+B
A+B -> C+D
A+B -> C+E
A+B -> C+F
A+B -> D+E
A+B -> D+F
A+B -> E+E
A+B -> E+F
F   -> F
F   -> E
```

A {(A)}

B {(B)}

{(C+D)}

{(A+B),(C+E),(C+F),
(D+E),(D+F),(E+E),(E+F)}

{(C+D)}

{(C),(D),(E)}

{(E),(F)}

{(C)} C

{(D)} D

{(E)} E

{(F)} F

● microstate (complex)
→ fast (1,1) reaction
⤑ slow (1,1) reaction
fast (1,2) reaction
slow (2,1) reaction
resting macrostate
transient macrostate
{} set of fates

27

DSD CONDENSATION

detailed reactions:

```
A + F1 -> i1
i1 -> i2
i2 -> B + F2
B + F2 -> i2
i2 -> i1
i1 -> A + F1
A + F2 -> i4
i4 -> A + F2
B + F1 -> i3
i3 -> B + F1
```

condensed reactions:

```
A + F1 -> B + F2
B + F2 -> A + F1
```

REACTION RATE CONDENSATION

Consider a condensed reaction:
$$P + Q \rightarrow K + L + M$$

It is composed of all detailed slow reactions:
$$p + q \rightarrow I$$
weighted by the decay probability over all pathways:
$$I \rightarrow \cdots \rightarrow k + l + m$$

where $p \in P, q \in Q, k \in K, l \in L, m \in M$
and I is a multiset of intermediate species

REACTION RATE CONDENSATION

Notation:

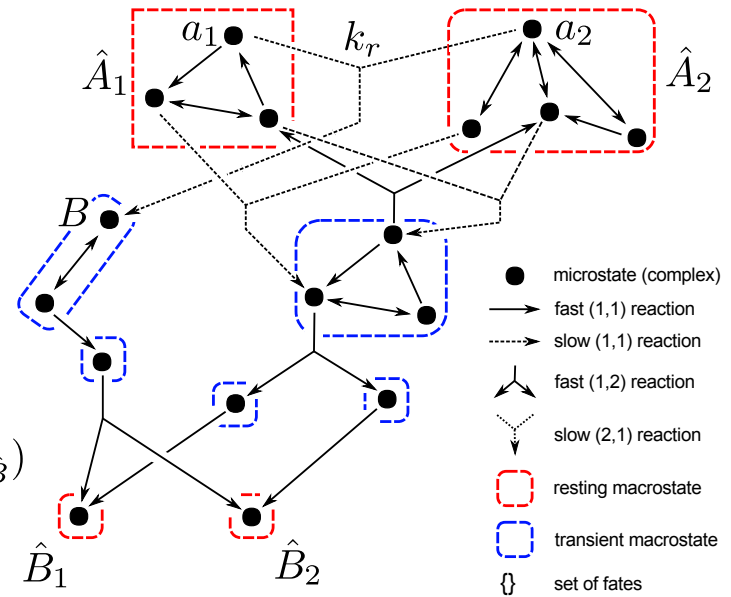detailed reaction: $\quad r = (A, B) \qquad A = \{|a_i|\}$

condensed reaction: $\quad \hat{r} = (\hat{A}, \hat{B}) \qquad \hat{A} = \{|\hat{A}_i|\}$

given: $\quad \hat{A} = (\hat{A}_1, \hat{A}_2) \quad \hat{B} = (\hat{B}_1, \hat{B}_2)$

define: $\quad R_{\hat{A}} = \{r = ((a_1, a_2), B) : a_1 \in \hat{A}_1, a_2 \in \hat{A}_2\}$

then the condensed rate is:

$$k_{\hat{r}} = \sum_{r = ((a_1, a_2), B) \in R_{\hat{A}}} P(a_1|\hat{A}_1) \cdot P(a_2|\hat{A}_2) \cdot k_r \cdot P(T_{B \to \hat{B}})$$

- ● microstate (complex)
- → fast (1,1) reaction
- ------→ slow (1,1) reaction
- fast (1,2) reaction
- slow (2,1) reaction
- resting macrostate
- transient macrostate
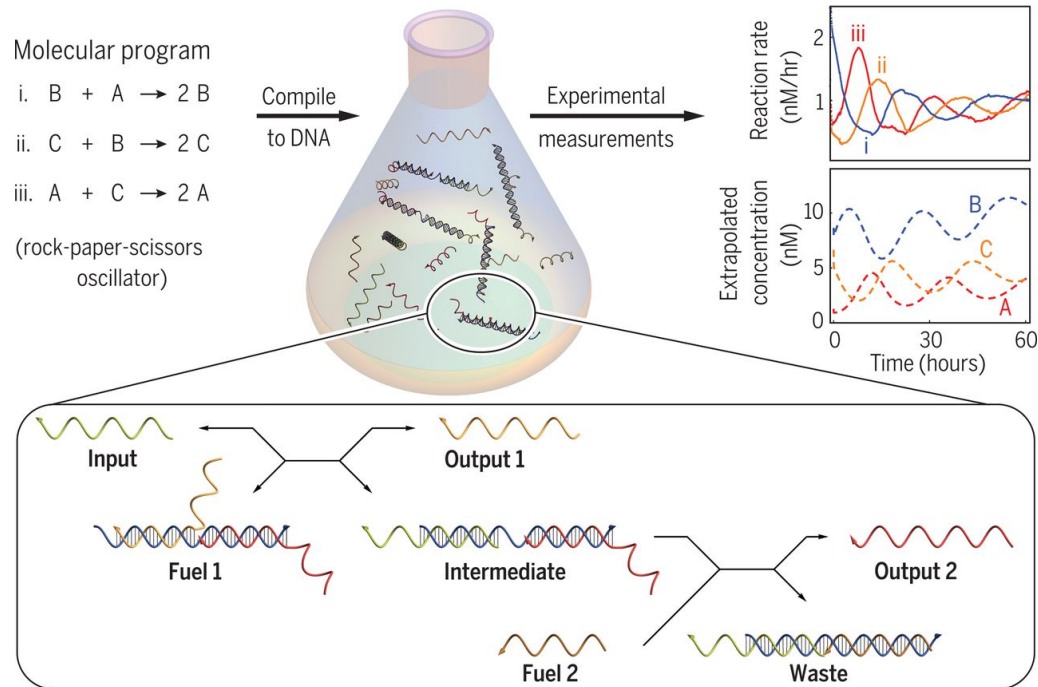- {} set of fates

REACTION RATE CONDENSATION

general form:

$$k_{\hat{r}} = \sum_{r=(A,B)\in R_{\hat{A}}} k_r \cdot \mathbb{P}[T_{B\to\hat{B}}] \cdot \prod_{a_i\in A} \mathbb{P}[a_i : \hat{A}_i]$$

where

$$\mathbb{P}[a_i : \hat{A}_i] = \text{stationary distribution}$$
$$\mathbb{P}[T_{B\to\hat{B}}] = \text{reaction decay probability}$$

A DNA OSCIALLATOR

Molecular program

i. $B + A \rightarrow 2B$

ii. $C + B \rightarrow 2C$

iii. $A + C \rightarrow 2A$

(rock-paper-scissors oscillator)

Compile to DNA

Experimental measurements

Input Output 1
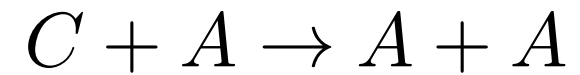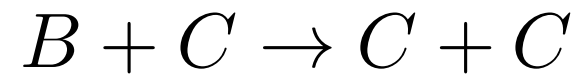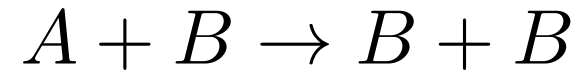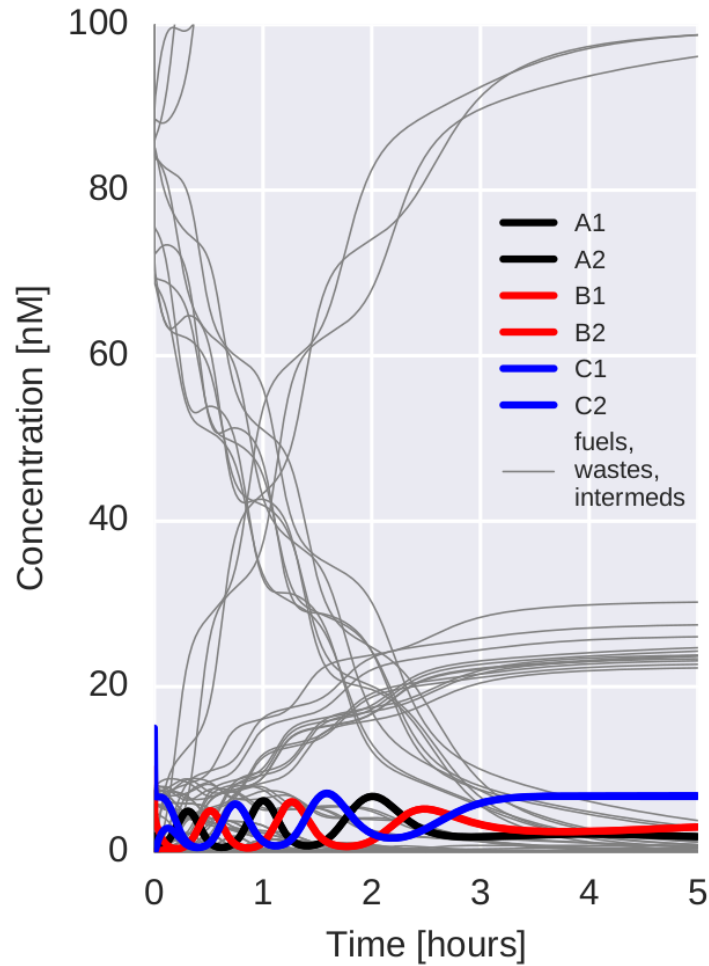Fuel 1 Intermediate Output 2
Fuel 2 Waste

Srinivas, Parkin, Seelig, Winfree, Soloveichik:
Enzyme-free nucleic acid dynamical systems. Science (2017)

A DNA OSCIALLATOR
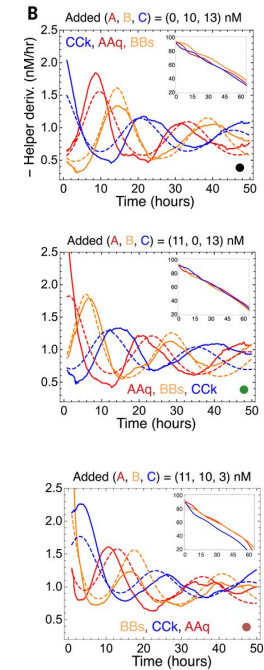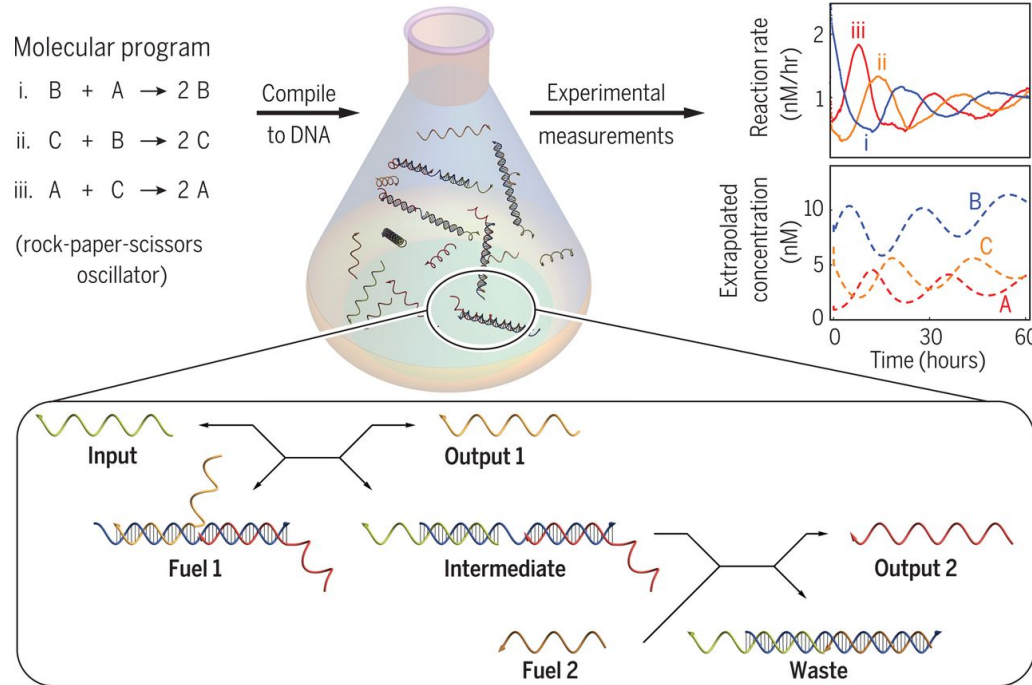
Srinivas, Parkin, Seelig, Winfree, Soloveichik:
Enzyme-free nucleic acid dynamical systems. Science (2017)

DETAILED VS. CONDENSED SIMULATION

$$A + B \rightarrow B + B$$
$$B + C \rightarrow C + C$$
$$C + A \rightarrow A + A$$
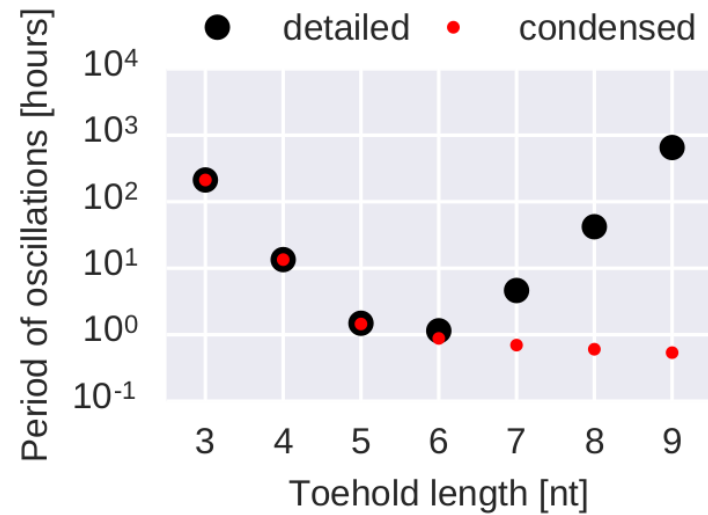
translation scheme: srinivas2017.ts

REACTION ENUMERATOR

model limitations

- no multistranded pseudoknots
- assumption of low concentrations
    - assumption of "typical" reaction rate constants

model parameters

- multiple layers of reaction-semantics
    - reaction types
    - max-helix notion (representation-independent)
    - reaction rate dependent enumeration

What the domain level can do:

- enumerate intended reaction pathways
- detect unintended reaction pathways
- **very fast** assessment of overall dynamics
- define a CRN for sequence-level simulations

What the domain level cannot do:

- include sequence-level variations within the domains

What the domain level could do:

- detect and quantify particular leak reactions
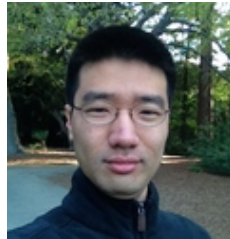- provide a coarse-graining for stochastic simulations

THANKS TO

Erik Winfree   Casey Grun   Karthik Sarma

Seung Woo Shin   you   Brian Wolfe

http://www.github.com/DNA-and-Natural-Algorithms-Group/peppercornenumerator