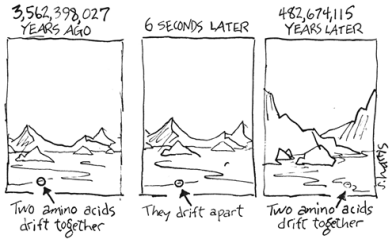
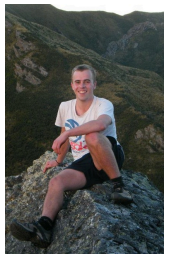
Autocatalytic sets and models of early life



Wim Hordijk



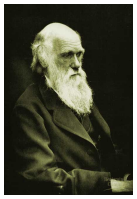
Joshua Smith

- Mike Steel

Joint work with…

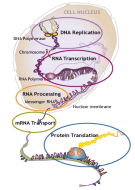
Elchanan Mossel

Stuart Kauffman

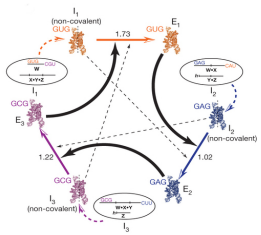


It is often said that all the conditions for the first production of a living organism are now present, which could ever have been present.— But if (& oh what a big if) we could conceive **in some warm little pond** with all sorts of ammonia & phosphoric salts,—light, heat, electricity &c present, that a protein compound was chemically formed, ready to undergo still more complex changes, at the present day such matter would be instantly devoured, or absorbed, which would not have been the case before living creatures were formed. Letter to J. D. Hooker, 1 Feb [1871]

- Many ideas/theories re. origin of life ('RNA world', genetic first/vs metabolism first, hydrothermal vents etc).
- Current DNA/RNA/protein molecular machinery too complex to have arisen spontaneously all at once.



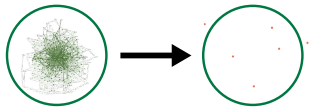
- Key early steps require the emergence (and evolution) of **self-sustaining** and **autocatalytic** networks of reactions.



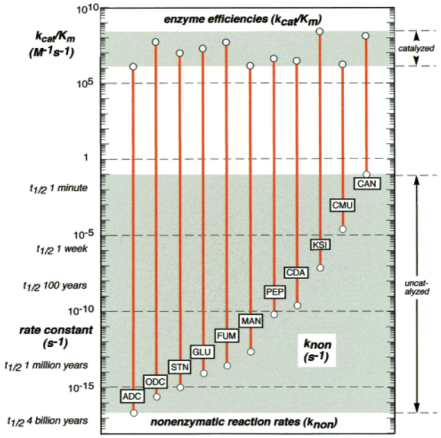
Vaidya et al., *Nature*, 2012

Two features of catalysis

Accelerates the production of molecules in the network so they accumulate spatially in concentrations sufficient to sustain further reactions and fight diffusion.



Not only much faster rates, but also tightly 'coordinated'



Wolfenden, Snider, Acc. Chem. Res, 2001

Catalytic Reaction System (CRS)

$Q = (X, R_0, C, F)$

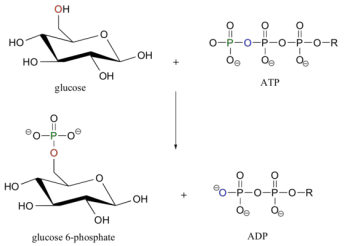
Molecule types X

Reactions $R_0 \subseteq 2^X \times 2^X$

Catalysis $C \subseteq X \times R$ (x,r)

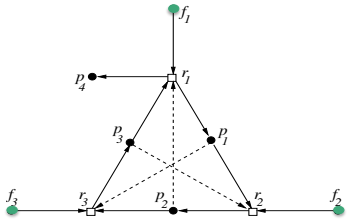
"Food" set $F \subseteq X$

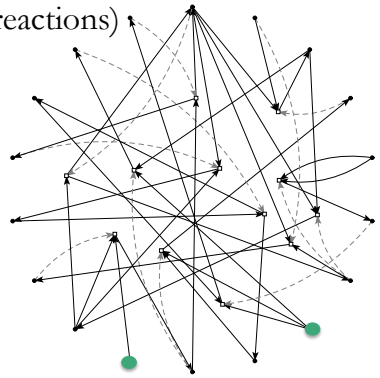
$r = \{\text{reactants}, \text{products}\} = (\rho(r), \pi(r))$



Another view of a CRS Q

A directed (and bipartite) graph with
two types of vertices (molecules, reactions)
and
two types of arrows
(reactants + products, catalysis).



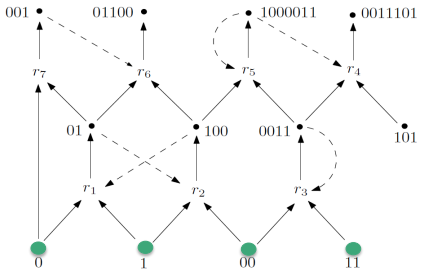


Simple example: **Polymer Model**

A set of molecules represented by strings over an alphabet (e.g. 0, 1) up to length n, with food molecules up to length t (with $t << n$).

A set of reactions of two types:
ligation: 000+111→000111
cleavage: 0101010→0101+010

Randomly assigned catalysis:
Pr[x catalyzes r] = $p(x,r)$
'vanilla model' $p(x,r)=p$

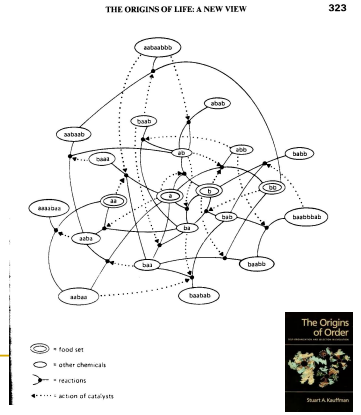


Early claim:

"The formation of autocatalytic sets of polypeptide catalysts is an expected emergent collective property of sufficiently complex sets of polypeptides, amino acids, and other small molecules."

(Kauffman, 1986)

Basic idea: Given a fixed probability of catalysis p and increasing n, at some point there is a phase transition where the entire reaction graph becomes an autocatalytic set, similar to giant connected components appearing in random graphs.



Main Criticisms

- Argument requires an exponential growth rate in level of catalysis (Lifson, 1996).
- Autocatalytic sets lack evolvability (Vasas, Szathmáry & Santos, 2010).
- Binary polymer model is not realistic enough (Wills & Henderson, 1997).

We will consider all these issues….

Our approach

Use mathematics (and simulations) to study the polymer model and its extension.

First we need to formalize some notions….

Definitions: *Closure*

- Given any subset R of R_0, the ***closure of F*** (relative to *R)*

$$\mathrm{cl}_R(F)$$

is the set of molecules that can be constructed from F by applying just reactions from R (whether they are catalysed or not).

- Formally, $\mathrm{cl}_R(F)$ is the unique (minimal) subset W of X that contains F and satisfies:

$$\rho(r) \subseteq W \Rightarrow \pi(r) \subseteq W$$

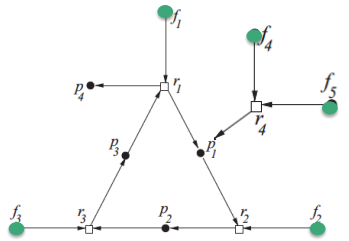
- $\mathrm{cl}_R(F)$ is computable in polynomial time in $|Q|$

Definition: *F-generated*

R is *F - generated* if $cl_R(F)$ contains every reactant of every reaction in *R*.

$\Rightarrow$ each reactant of any reaction in *R* is either in *F*
or is a product of some other reaction in *R*

$\Leftarrow$?



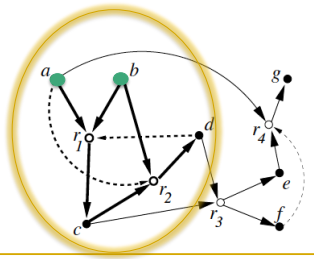
Definition: *RAF* (Reflexively Autocatalytic network over F)

A subset R of R_0 is an RAF if

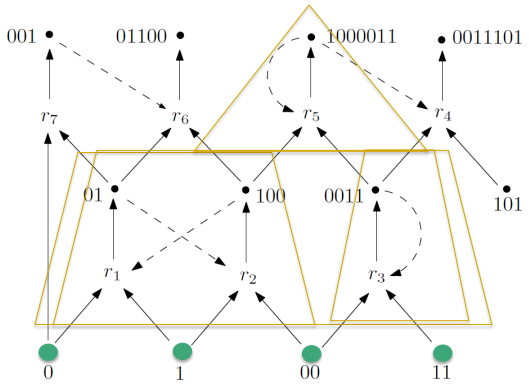
$R \neq \emptyset$, and it satisfies the two properties:

(RA): each reaction r in R, is catalysed by a product of some other reaction (or by an element of F),

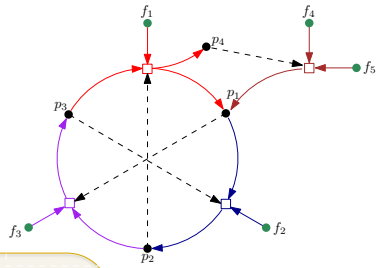
(F): R is F-generated



Earlier example



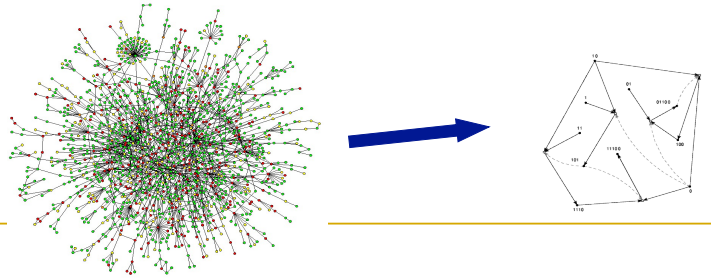
Equivalent definition



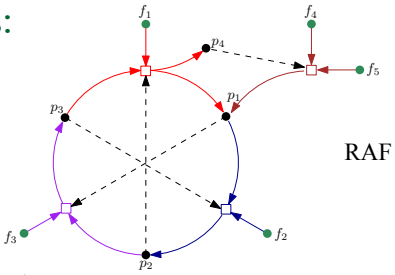
A subset R of R_0 is an RAF if
$R \neq \emptyset$, and for each reaction r in R,
all of the reactants and at least one catalyst of r
are present in $cl_R(F)$.

Two nice properties of RAFs

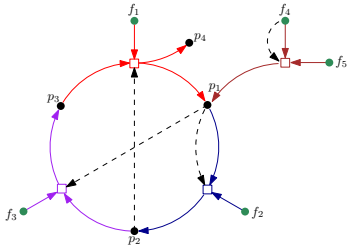
- The union of any collections of RAFs is itself an RAF
  - So if Q has an RAF then it contains a unique maximal RAF.
  - Denote this by **maxRAF(Q)**
- There is a simple algorithm to determine whether or not Q has an RAF, and if so to compute maxRAF(Q) (polynomial time in |Q|).



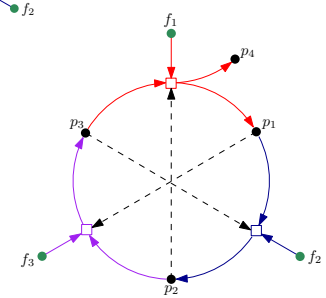
Related notions:



RAF



CAF (Constructively Autocatalytic network over F)



pseudo-RAF

RAF Algorithm

- $R_0, R_1, \ldots$. (nested decreasing sequence) with limit R
- (*) R_{i+1}= reactions in R_i that have *all* their reactants and *at least one* catalyst in $cl_{R_i}(F)$
- If R is the empty set, R_0 has no RAF, otherwise R=maxRAF(R_0).

More efficient:

- i odd: perform step (*) above
- i even R_{i+1}=max-pseudoRAF(R_i) [~ HORN-SAT, linear time]

Quantities of Interest

$Q = (X, R_0, C, F)$ full binary polymer model (on all sequences of length up to n). $|X| \sim 2^{n+1}$; $|R_0| \sim n2^{n+1}$.

- Average number of reactions catalyzed by any molecule type: $\mu = p \cdot |R_0|$
- Probability P_n = Pr(R_0 contains an RAF) that an instance of the binary polymer model contains an RAF set.

Early results

- p constant (i.e. independent of n):

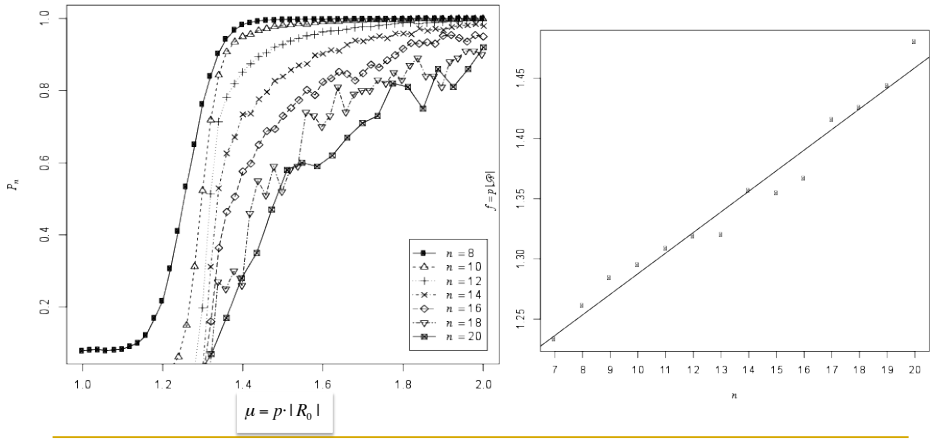
[Kauffman, '86; '93] $P_n \to 1$, as $n \to \infty$

- But this requires μ to grow exponentially with n which is biochemically unrealistic (Lifson '96)
- What if μ grows more slowly?
  - [S: 2000]

$$\text{If } \mu < \frac{1}{3}e^{-1} \text{ then } P_n \to 0, \text{ as } n \to \infty$$
$$\text{If } \mu > cn^2 \text{ then } P_n \to 1, \text{ as } n \to \infty$$

[Conjecture: sub-quadratic]

Probability of RAF Sets



[Hordijk+S, 2004]

Main theoretical results I (Mossel+S, 2005)

- Theorem 1: Linear transition for RAFs

  If $\mu \propto n^{1+\delta}$ then $P_n \to 1$, as $n \to \infty$

  $\mu \propto n^{1-\delta}$ then $P_n \to 0$, as $n \to \infty$

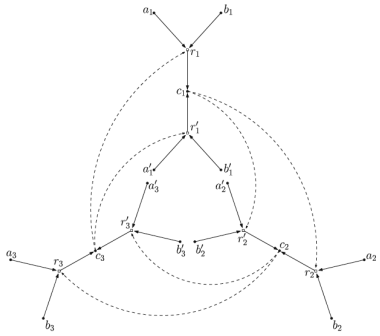
- Theorem 2: Exponential transition for CAFs

  If $\mu \propto (2-\delta)^n$ then $P_n \to 0$, as $n \to \infty$

  $\mu \propto (2+\delta)^n$ then $P_n \to 1$, as $n \to \infty$

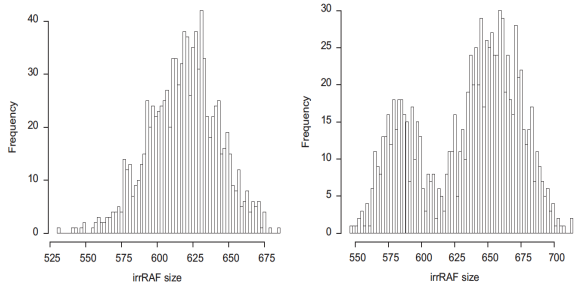
Small RAFs

- Finding a smallest RAF is NP-hard.
- **Definition:** An RAF of Q is **irreducible** if it contains no proper sub-RAF
- Finding an irrRAFs is easy…
- But there can be exponentially many of them



Small RAFs

- irrRAFs can be of different sizes:



- **Computational question:** Is it hard to find a largest irrRAF?

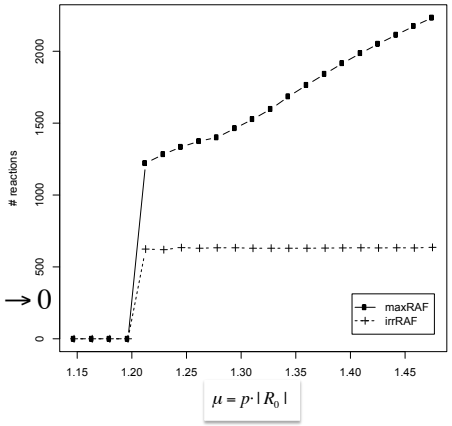
Main theoretical results II (S+Hordijk+Smith 2013)

Theorem 3

There are no small RAFs when they first appear

If $\mu = n^{1+\delta}$ then as $n \to \infty$:

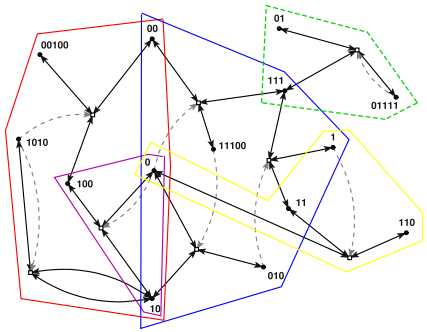
$\Pr(R_0$ contains an RAF of size $< 2^{cn}) \to 0$



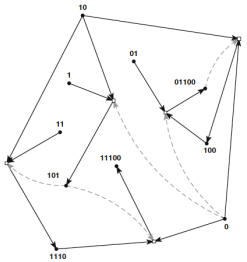
- But when first cycles first appear in a random directed graph, they are of all sizes! (Bollobás and Rasmussen, 1989)

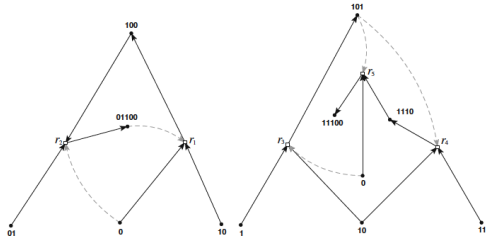
Structure of RAFs

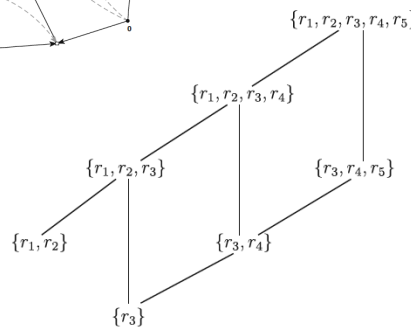
- In general maxRAF(Q) may contain (many) other sub-RAFs.



Poset of RAFs







$r_1 : 10 + 0 \xrightarrow{01100} 100$

$r_2 : 01 + 100 \xrightarrow{0} 01100$

$r_3 : 10 + 1 \xrightarrow{0} 101$

$r_4 : 11 + 10 \xrightarrow{101} 1110$

$r_5 : 1110 + 0 \xrightarrow{101} 11100$

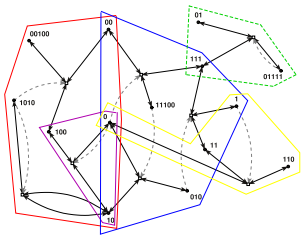
Computing: It's easy to….

Find *all* the maximal proper subRAFs of the maxRAF (polynomial time in $|Q|$).

Construct the poset P of subRAFs of Q.
(poly. in $|P| \times |Q|$).

Some other things that are easy to compute (in poly time in $|Q|$)

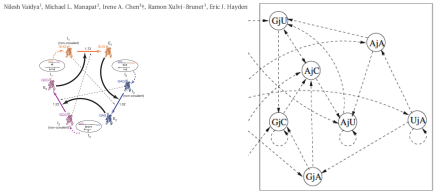
- Determining if:
  - any given reaction (or set of reactions) is in **every** RAF
  - a given set of irrRAFs contains all the irrRAFs
  - if a given R is a 'co-RAF' (i.e. Is $R=R_1-R_2$ for nested $R_1 \subset R_2$ RAFs).

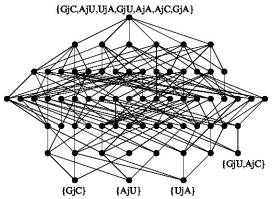


Application to a real experimental system

ARTICLE

Spontaneous network formation among cooperative RNA replicators





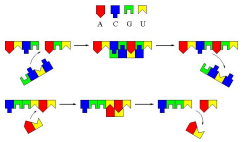
RNA ribozyme replicator system: 16 reactions, 18 molecules, $|F|=2$, 64 catalysation pairs (x,r).

Forms an RAF. Contains many subRAFs.

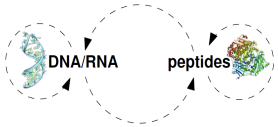
+Dynamics of subRAFs via Gillespie algorithm

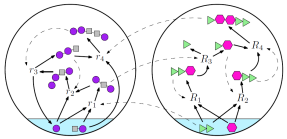
Extensions beyond the 'vanilla' model

- Allowing $p(x,r)$ to vary
  - Template-based catalysis



  - $p(x,r)$ depends on length of x
  - Partitioned polymers system





- Extension of transition theory to non-polymer systems

[n replaced by ratio of # reactions to # molecules]

Questions and future work:

- Analysis of real biochemical networks (in progress)
- Impact of inhibition: (existence of RAFs now NP-
- RAFs that satisfy rate constraints: (also NP-hard)

Further details:

- Hordijk, W. and Steel, M. (2013). A formal model of autocatalytic sets emerging in an RNA replicator system. *J. Systems Chemistry* 4:3.
- Steel, M., Hordijk, W., and Smith, J. (2013). Minimal autocatalytic networks. *Journal of Theoretical Biology* 332: 96-107.
- Hordijk, W., Wills, P. and Steel, M. (2013). Autocatalytic Sets and Biological Specificity. *Bulletin of Mathematical Biology* 76(1): 201--224.
- Hordijk, W., Steel, M. and Kauffman, S. (2012). The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence. *Acta Biotheoretica*, 60: 379-392.
- Hordijk, W. and Steel, M. (2012). Predicting template-based catalysis rates in a simple catalytic reaction model. *Journal of Theoretical Biology* 295: 132-138.