Prebiotic evolution: replicator systems and information threshold
Life as evolved (evolving) complex multilevel information processing system

how to model it?
(1) how did it get started?/ bootstrap itself
Origin of life

studied by
• phylogenetic reconstruction (LUCA)
  quite complex - what before that?....
• Sufficient chemistry / environment
• experimental studies of minimal 'living' systems
  (re)constructing/engineering/evolving) such systems
• modeling studies of minimal 'living' systems

different approaches (focus) dependent on:
what is life?
  including alternative forms (e.g. extra terrestrial / lab.)
Life is ....
Unique properties of life not shared by technological systems

'In stark contrast with current computer technology, biological cells compute in construction using molecular and spatial information, in order to delimit, organize, power, sustain, repair, move, communicate, reproduce, protect and evolve themselves robustly from simple and scarce material and energy resources in their complex environments'
J. McCaskill and S. Rasmussen EU report (2012)

not good starting point...
Hypothesized Environments of Prebiotic 'life' (metabolism)

Hydrothermal vents: (black smokers) energy/energy gradients for free compartments (concentration of ingredients) catalysis by metal sulphides; acetyl-coA pathway abiogenic aminoacid synthesis

Menez et al PNAS 2018

OR

Origin of first cells at terrestrial, anoxic geothermal fields Because of 'open' cell environment should match internal cell composition

“shallow ponds of condensed and cooled geothermal vapor that were lined with porous silicate minerals mixed with metal (primarily Zn) sulfides and enriched in K+, Zn2+, and phosphorous compounds.”

Armen Y........ Eugene V. Koonin, 2012 PNAS
“The individual taxonomic units evolve and go extinct, yet the core machines survive surprisingly unperturbed.”

PG Falkowski et al, Science 2008
conserved metabolic pathway: glycolysis/gluconeogenesis

WHY??
“unique?” , “optimal?”
contingency?, (evolvability?)

Court, Waclaw & Allen 2015
Mapping all possible trunc pathways
Glyceraaldehyde 3 phosphate to pyruvate

All (~1500) unbranched aliphatic CHOPN upto 4 carbon, negatively charged free energy for formation

All possible reaction

All possible pathways which produce (at least) 2 ATP length 4, 5 or 6

1787 glycolysis pathways

6445 gluconeogenic pathways

<table>
<thead>
<tr>
<th>EC class</th>
<th>Oxidation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.1</td>
<td>phosphorylation</td>
</tr>
<tr>
<td>1.2.1</td>
<td>Deamination</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Transamination</td>
</tr>
<tr>
<td>1.4.1</td>
<td>Phosphate transfer</td>
</tr>
<tr>
<td>2.6.1</td>
<td>Hydrolysis</td>
</tr>
<tr>
<td>2.7.1</td>
<td>Decarboxylation</td>
</tr>
<tr>
<td>2.7.2</td>
<td>Dehydration</td>
</tr>
<tr>
<td>2.7.9</td>
<td>Ammonia-lyase</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Isomerization</td>
</tr>
<tr>
<td>3.5.1</td>
<td>ATP-driven amine ligase</td>
</tr>
<tr>
<td>3.6.1</td>
<td>ATP-driven carboxylation</td>
</tr>
</tbody>
</table>
Sample 10000 conditions of 11 external metabolites + G3P and Pyruvate (log sampling around typical existing levels)

Limit internal metabolite conc. 0.1-100mM

average relative flux in all samples
different optima for different conditions

alternative 'bests'

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td>[source]</td>
<td>1μM to 1 mM</td>
</tr>
<tr>
<td>[source]/[product]</td>
<td>0.01 to 100</td>
</tr>
<tr>
<td>[ATP]/[ADP]</td>
<td>0.1 to 100</td>
</tr>
<tr>
<td>[NAD]/[NADH]</td>
<td>0.1 to 1000</td>
</tr>
<tr>
<td>[AMP]</td>
<td>0.01 to 0.1 mM</td>
</tr>
<tr>
<td>[Pi]</td>
<td>1.0 to 100 mM</td>
</tr>
<tr>
<td>[PPI]</td>
<td>0.1 to 10 mM</td>
</tr>
<tr>
<td>[CO₂]</td>
<td>1.0 μM to 0.1 mM</td>
</tr>
<tr>
<td>[NH₃]</td>
<td>1.0 μM to 0.1 mM</td>
</tr>
<tr>
<td>[GLUT]</td>
<td>1.0 to 100 mM</td>
</tr>
<tr>
<td>[2-OXO]</td>
<td>0.1 to 10 mM</td>
</tr>
</tbody>
</table>

sampling space
“Biological systems are distinguishable from chemical systems because they contain components that have many potential alternative compositions but adopt a particular composition based on the history of the system. In this sense biological systems have a molecular memory (genotype), which is shaped by experience (selection) and maintained by self-reproduction”

Joyce (2012) Bit by Bit: The Darwinian Basis of Life:

“How many heritable bits are involved, and where did they come from
Evolution-first scenario of the origin of life

RNA world

The RNA world hypothesis:

The worst theory of early evolution of life (except for all the others)
(Harold S Bernhardt, Biology Direct 2012)

- RNA for information storage and amplification
- (template and catalyst)
- potential 'generic' replication
- RNA in core processes of current biological systems
- New (old?) catalytic functions easily evolvable

HOWEVER

Many chemical caveats raised, and partially solved...

BUT

“The presumed RNA world should be viewed as a milestone, a plateau in the early history of life on Earth. So, too, the concept of an RNA world has been a milestone in the scientific study of life’s origins. Although this concept does not fully explain how life originated, it has helped to guide scientific thinking and has served to focus experimental efforts”

Protocells and RNA Self-Replication
Gerald F. Joyce1 and Jack W. Szostak 2018
review of and towards RNA world from chemical point of view.
First polynucleotide synthesis without prior metabolism?

wet-dry cycling
as in hydrothermic ponds
enough energy for
RNA polymerization?

Visualizing RNA polymers
produced by hot dry-wet cycling
(Hassenkam & Deamer 2022
Scientific reports
RNA dependent RNA polymerase evolved from a ligase (Bartel & Szostak 1993), and improved by design and evolution to current form:

Replicated RNA's

Works also Reversed transcriptase!
Current state after 56 rounds (26 mutations):
can replicate own ancestor
Portillo ...Joyce 2021 elife

Evolution of the novel pseudoknot structure.
RNA world hypothesis

here assumed as starting point for developing

bioinformatic theory prebiotic evolution

focusing on informatic rather than chemical constraints
AND as starting point for
modeling biotic systems as
evolving mutilevel information processing systems

Informatic potential and limitation of RNA world hypothesis
limited evolvability?
minimal evolution system, 
replicator equation

• 'generic replicators'
• independent synthesis and decay
• mutation
• competition

Eigen: Replicator Equation (in chemostat)
\[
dX_i/dt = A_i Q_i X_i - d_i X_i + \sum w_{ij} X_j - \Omega_i
\]

\[
\Omega_i = \left(\frac{X_i}{\sum X_j}\right) \sum (A_j - d_j) X_j \quad \text{(constant population size)}
\]

\[
w_{i,i} = A_i Q_i - d_i
\]

\[
w_{ij} = \mu_{ij} A_j (1 - Q_j)
\]

Converges to eigenvector belonging to largest eigenvalue of W

\[==\text{Quasispecies (wildtype)}\]

growthrate Quasispecies: largest eigenvalue
HOW DOES COMPLEXITY ARISE THROUGH DARWINIAN EVOLUTION?

Survival of the fittest does not imply increase of complexity.

On the contrary:

(If we take genome size as measure of complexity)

more complexity tends to decrease replication rate

MOREOVER

Classical caveat for the evolvability of complexity

**Error catastrophe (Information threshold)** (Eigen 1971)

*limited replicator complexity*
Error catastrophe
Survival of the fittest (quasi)species?

Assume: $w_{ii} >> w_{jj} = C$ for all $j$
Simplify: lump all $j$ into one equation (Y)

$$
dX/dt = a_1Q_1X - d_1X - X((a_1 - d_1)X + (a_2 - d_2)Y)
dY/dt = a_2Y - d_2Y + a_1(1 - Q_1)X - Y((a_1 - d_1)X + (a_2 - d_2)Y)
$$

Is the fittest ($X$) selected? Does it “survive”

Can the fittest invade?

$X > 0$ iff $dX/dt > 0$ close to $X = 0$

$$a_1Q - d_1 > a_2 - d_2$$

$Q > a_2/a_1 = 1/\sigma$ (assuming $d_1 = d_2$)

*Darwinian evolution only when mutation rate small enough*

(survival of the fittest NOT a tautology)

Similar for full model: N species with back mutation.
From error threshold to information threshold

If replicator is a information carrying polymer (e.g. RNA or DNA)

Mutations can happen at each position (nucleotide) (with rate 1-q)

\[ Q = q^L = e^{-L(1-q)} \]

\[ L < \frac{\ln \sigma}{(1 - q)} \]

Given a mutation rate, given selction coefficient
Length of sequence limited -- >

Only limited information accumulation possible
this is 'best' case scenario

- infinite population size
  - always to 'best' quasispecies
  - no stochastic population dynamics
  - no extinction (everybody viable-replicatable)

- strong selection, single peak landscape
  - therefore sharp transition (threshold)
  - delocalization vs threshold

- fixed length - no other constraints
  - NO negative selection on length (rate, energy)
Population just before the information threshold
“sequence logo”, consensus sequence

Note: different contribution to fitness of different positions
information threshold, further characterisation
Before the error threshold common ancestor is master sequence beyond the error threshold NOT

Common Ancestor: D to master seq.

Delocalization but no threshold for exponential fitness landscape

Takeuchi & Hogeweg (2007, BMC-evol)
However, if also lethal mutations - there is a sharp threshold
Common ancestor in finite population

Common Ancestor: D to master seq.

Information threshold - any observational evidence?

*Drake’s rule:* 
*constant (BUT LOW!) per genome mutation rate*

Mutation rate “evolved” property 
(cf Sulfolobus in very harsh environment)

Sniegowski “ Evolution: constantly avoiding mutation” current biology 2001
retaining low mutation rates impossible in eukaryotes because of small population sizes? ("above" error threshold?)

Lynch 2010 TIG
Error threshold and antiviral strategies

error threshold and vs extinction threshold and/or new mutants

Perales C, Agudo R, Domingo E. PLOS-one 2009

WT extinction by mutagenesis

mutant resistant to mutagen (mutation in RNAdep RNA pol.)

Bull et al 2005 Plos comp biol
for more info we need better replication
for better replication we need more info

artefact of too simple replicator model?
Error Catastrophe Can Be Avoided by Proofreading Innate to Template-Directed Polymerization

elongation-rate: \( r(l, s, m, S) = \beta(l, s, S) \nu(l + 1, m, S)f(S) \).

\( \beta \) binding to template; \( \nu \) elongation rate; \( f(S) \) ”fitness”

\[
\dot{x}_{l,s} = x_{l-1,s'} \sum_{S \in \{0,1\}^L} r(l-1, s', m, S)x_{L,S} - x_{l,s} \sum_{S \in \{0,1\}^L} (r(l, s, 0, S) + r(l, s, 1, S))x_{L,S} - \phi x_{l,s},
\]

population dyn: \( \text{TIMESCALES!} \)
**innate proofreading increases information threshold and more efficient for longer sequences**

if influx of primers (= efflux) **slow** replicator model
if influx of primers (= efflux) **fast** rate of polymerization matters

iff fast:

**innate proofreading**

lower binding + lower elongation rate $\rightarrow$ slower replication $\rightarrow$ more efflux of mutated sequences before acting as template

NB "attofox problem": blacklines: allmost only free primers
How otherwise to 'solve' or 'circumvent' information threshold?

Did we ask the wrong question?
Did we use the wrong model?
Only little information needed for higher quality replication?

2(3) main directions to try to circumvent problem
“more replicators” “more RNA in replicators”

BOTH 2 & 3

FIRST
more replicators: ecosystem based solution
  Hypercycles (Eigen’s original solution)

Emergence of higher levels of selection
First attempt to circumvent information threshold: Hypercycles, Eigen and Schuster

If one replicator has too little information - use many. However beyond the many of the quasispecies: evolved and coordinately optimized.

*Specific catalysis of reactions*

\[ \frac{dX_i}{dt} = a_i X_i + b_i X_i X_j - \Omega_i \]

• (no mutations): look at 'ecosystem'
• ONLY stable topology: cycle
Hypercycle properties

- Selection LOCAL on amount of catalysis received
- growth and contraction of cycles

**HOWEVER**

- Once only selection/survival of the first
- NO selection for GIVING catalysis: Parasites