Today in Astronomy 106: polymers to life

- Translation: the current fashion in protein manufacture.
- The chicken-egg problem
  - Protein-based primitive life?
  - RNA World
- Emergence of the genetic code.
- How long does all this take? The importance of baby steps, and the heroism of time.

Primordial soup (Nature)
Replication of proteins

This has been envisaged in several different ways:

- Long chains of amino acids can polymerize in solution without much help.

- Some configurations might be self-organizing: certain sequences surely have higher binding energies than others and will form preferentially.
  - Would it help if all were L or all were D isomers?

- There will be lots of copies of these for this reason.

- Alternative mass production or replication: perhaps formation on a sequenced structure, like a crystal or another molecule, that attracts the monomers preferentially in a certain order?
Translation

In modern lifeforms, this is quite complicated, and involves nucleic acids intimately, following a transcription.

- Section of DNA or RNA transcribed onto a short length of RNA, called messenger RNA (mRNA).

- mRNA transported to ribosome, a collection of proteins and nucleic acids that will H-bond onto a certain codon of the mRNA.

- The ribosome has shorter stretches of RNA around, called transfer RNA (tRNA), that have exposed codons and, bonded elsewhere, an amino acid captured from the solution that is unique to the exposed codon.
tRNA

From Neal Evans’s AST 390L lecture notes.
Replication of proteins (continued)

- The tRNA complementary to the first codon on the mRNA H-bonds to it, and the AA is the first element of the protein. The next element is brought in with the next tRNA, and so on.

- The ribosome’s (tRNA’s) job is called translation. Animation of translation in modern lifeforms: go to

[Link to animation](http://vcell.ndsu.nodak.edu/animations/translation/movie.htm)

Again from Phillip McClean and Christina Johnson (UND).
The original chicken and egg problem

So, in modern Earthly life, the replication or mass production of nucleic acids and protein are interdependent: each needs certain of the others in order to be produced in large quantities.

- Which came first? Was the most primitive life originally just protein or nucleic acid?
- How did they come to be interdependent?
- How did useful proteins and nucleic acids develop?
Protein-based primitive life?

There is no experimental evidence that protein-only life could exist, but there are theories and candidate substances.

- Closest to modern life: peptide nucleic acids, which are proteins, but which have nucleobases that allow replication and information storage à la RNA.

- So far peptide nucleic acids are a lab curiosity; no evidence in substructure of cells or chemistry of primitive monomers that it is a viable life form.

  - They can hydrogen-bind DNA, though.

Peptide nucleic acid, with a N-(2-aminoethyl) glycine backbone (Wikimedia Commons).
Protein-based primitive life? (continued)

- Prions don’t count: they reproduce by folding existing proteins into weird shapes, and in general seem to be things that developed after life was already on its way with both nucleic acids and proteins.

- Before the genetic code was traced to DNA, the originator of the idea behind Miller-Urey, Alexander Oparin (1924), noted the tendency of amino acids to combine, and fatty acids (lipids) to form, in solution, and postulated that this would lead to spherical coacervate droplets of these compounds in the “primordial pond.”

  - He thought this would lead to cells: membranes made from lipids, most of the rest proteins.

- Similarly Fox later proposed proteinoid droplets.
Protein-based primitive life? (continued)

Freeman Dyson (1982, 1999), a famous physicist turned evolutionary biologist, picked up on Oparin’s scheme and conjured a model in which metabolism and proteins were primary, with nucleic acids and the genetic code developing much later.

- “Cells” first, enzymes next, nucleic acids much later.
- Cells provide sites for adsorption of monomers.
  - He had acervates/proteinoids and amino acids in mind, but noted that the reasoning would work on crystals and nucleotides as well.
- Repeatable assembly of complex polymers through “order-disorder” transition of adsorbed molecules.
Protein-based primitive life? (continued)

- Dyson’s model predicts that polymers with 100-200 monomers can be made repeatably.
  - About the minimum required for reproducing system, according to experiments on viruses (and thus RNA, not proteins).
- Works best if there are 8-10 different monomers; does not work at all if there are 3-4, so the mechanism prefers amino acids to nucleotides.
- Also achieves an error rate of about 1 in 4 if the total number of monomers present is of order 10000 (fewer errors if more monomers).
- Hard to extend to nucleic acids, though, so this model hasn’t found much favor.
Nucleic-acid based primitive life?

But RNA can

- replicate, as we’ve seen (transcription).
- catalyze the formation of proteins (tRNA +RNA) and nucleic acids.
- store information: the genetic code, so that specific proteins can be mass produced.
- transmit information: mRNA copies of short sections of a master RNA (translation).

This is the most consistent, and therefore popular, story of emergence of biology from chemistry: an RNA-dominated primitive form of life that preceded proteins and DNA.

- This stage of evolution is generally called RNA World.
RNA World

Scenario, due mainly to Eigen, Orgel, and Cech:

- In the beginning there were short strands of bare, self replicating RNA (Eigen et al. 1981). Eigen calls the varieties of the strands quasispecies.

- Most numerous and robust quasispecies predominate (appearance of natural selection).

- Quasispecies mutate, diversify as time goes on.

- Truncated stretches of these RNAs catalyze reactions of other RNAs (Cech and Bass 1986).
  
  - Example: 192-base RNA that catalyzes the replication of an RNA as long as itself with 98.9% accuracy.
  
  - D-sugars “chosen” by now.
RNA World (continued)

- These RNA catalysts, called **ribozymes**, presage the function of modern ribosomes.
  - **1989 Nobel Prize** to Cech and Altman for discovering ribozymes.
- Soon the ribozymes begin to catalyze protein formation.
  - L-amino acids “chosen” at this point.
- In turn some of the proteins thus formed begin to influence the local RNA, catalyzing RNA cleaving and hydrolysis; these are the first **enzymes**.
- This interchange of RNA and protein catalysts is termed by Eigen a **hypercycle**.
RNA World (continued)

- Simulations and theories show that proteins made from other quasispecies would be likely to perform novel catalytic functions on a given quasispecies.

- RNA can hydrogen-bond to complementary DNA. Eventually DNA took over as the more robust and stable master copy of the base sequences that encode the formation of proteins, but RNA retained its role in transcription and translation.

Mid-lecture Break.

- Homework #3 is due a week from yesterday, at 7PM.
The genetic code

One of the most striking adaptations of ribozymes is the development of tRNA: rather long molecules, each with a three-base sequence on one end uniquely matched to an amino acid on the other.

- As we have mentioned, life on Earth has 20** amino acids in use, which could not be specified by two bases (16 combinations) but can with three (64 combinations).

- This code must have been established very early in the development of life, as even cell-organelle DNA and RNA differ very little from nuclear DNA in the code they use.
The genetic code (continued)

<table>
<thead>
<tr>
<th>Nucleotide 2</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>CUC</td>
<td>AUC</td>
<td>GUC</td>
</tr>
<tr>
<td></td>
<td>Phenylalanine</td>
<td>Serine</td>
<td>Tyrosine</td>
<td>Cysteine</td>
</tr>
<tr>
<td>U</td>
<td>UUC</td>
<td>UCC</td>
<td>UAC</td>
<td>UGC</td>
</tr>
<tr>
<td></td>
<td>Phenylalanine</td>
<td>Serine</td>
<td>Tyrosine</td>
<td>Cysteine</td>
</tr>
<tr>
<td>U</td>
<td>UUA</td>
<td>UCA</td>
<td>UAA</td>
<td>UGA</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Ochre (Stop)</td>
<td>Opal (Stop)</td>
</tr>
<tr>
<td>U</td>
<td>UUG</td>
<td>UCG</td>
<td>UAG</td>
<td>UGG</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Amber (Stop)</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>CUC</td>
<td>CUA</td>
<td>CUG</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
</tr>
<tr>
<td>C</td>
<td>CUC</td>
<td>CCC</td>
<td>CCA</td>
<td>CCG</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
<td>Leucine</td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>AUC</td>
<td>ACA</td>
<td>AGA</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Isoleucine</td>
<td>Threonine</td>
<td>Lysine</td>
<td>Arginine</td>
</tr>
<tr>
<td>A</td>
<td>AUC</td>
<td>ACC</td>
<td>AAA</td>
<td>AGA</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Isoleucine</td>
<td>Threonine</td>
<td>Lysine</td>
<td>Arginine</td>
</tr>
<tr>
<td>A</td>
<td>AUG</td>
<td>ACA</td>
<td>AAG</td>
<td>AGG</td>
</tr>
<tr>
<td>Methionine</td>
<td>Methionine</td>
<td>Threonine</td>
<td>Lysine</td>
<td>Arginine</td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>GCU</td>
<td>GAU</td>
<td>GGU</td>
</tr>
<tr>
<td>Valine</td>
<td>Valine</td>
<td>Alanine</td>
<td>Aspartic acid</td>
<td>Glycine</td>
</tr>
<tr>
<td>G</td>
<td>GUC</td>
<td>GCC</td>
<td>GAC</td>
<td>GGC</td>
</tr>
<tr>
<td>Valine</td>
<td>Valine</td>
<td>Alanine</td>
<td>Aspartic acid</td>
<td>Glycine</td>
</tr>
<tr>
<td>G</td>
<td>GUA</td>
<td>GCA</td>
<td>GAA</td>
<td>GGA</td>
</tr>
<tr>
<td>Valine</td>
<td>Valine</td>
<td>Alanine</td>
<td>Glutamic acid</td>
<td>Glycine</td>
</tr>
<tr>
<td>G</td>
<td>GUG</td>
<td>GCG</td>
<td>GAG</td>
<td>GGG</td>
</tr>
<tr>
<td>Valine</td>
<td>Valine</td>
<td>Alanine</td>
<td>Glutamic acid</td>
<td>Glycine</td>
</tr>
</tbody>
</table>
The genetic code (continued)

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codons</th>
<th>Amino acid</th>
<th>Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>GCU, GCC, GCA, GCG</td>
<td>Leucine</td>
<td>UUA, UUG, CUU, CUC, CUA, CUG</td>
</tr>
<tr>
<td>Arginine</td>
<td>CGU, CGC, CGA, CGG, AGA, AGG</td>
<td>Lysine</td>
<td>AAA, AAG</td>
</tr>
<tr>
<td>Asparagine</td>
<td>AAU, AAC</td>
<td>Methionine</td>
<td>AUG</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>GAU, GAC</td>
<td>Phenylalanine</td>
<td>UUU, UUC</td>
</tr>
<tr>
<td>Cysteine</td>
<td>UGU, UGC</td>
<td>Proline</td>
<td>CCU, CCC, CCA, CCG</td>
</tr>
<tr>
<td>Glutamine</td>
<td>CAA, CAG</td>
<td>Serine</td>
<td>UCU, UCC, UCA, UCG, AGU, AGC</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>GAA, GAG</td>
<td>Threonine</td>
<td>ACU, ACC, ACA, ACG</td>
</tr>
<tr>
<td>Glycine</td>
<td>GGU, GGC, GGA, GGG</td>
<td>Tryptophan</td>
<td>UGG</td>
</tr>
<tr>
<td>Histidine</td>
<td>CAU, CAC</td>
<td>Tyrosine</td>
<td>UAU, UAC</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>AUU, AUC, AUA</td>
<td>Valine</td>
<td>GUU, GUC, GUA, GUG</td>
</tr>
<tr>
<td>Start</td>
<td>AUG</td>
<td>Stop</td>
<td>UAA, UGA, UAG</td>
</tr>
</tbody>
</table>

Small differences from this occur in mitochondria and bacteria.

Also used to code selenocysteine.
Baby steps

Every step in the development of life has involved the development, not just of molecules of a certain size and class, but molecules that do something specific. That’s a big restriction.

- Take the minimum-length RNA-replication ribozyme mentioned above, for example: 192 nucleotides.

- There are four different nucleotides, so the number of combinations of 192 nucleotides is

\[
    n = \underbrace{4 \times 4 \times 4 \times \ldots \times 4}_{192 \text{ factors of } 4} = 4^{192} \approx 4 \times 10^{115}
\]

If they form randomly, we may have to try \(4 \times 10^{115}\) times for a good chance of getting the right one. (Why? Well…)
Quick education in probability. Flip a coin. What is the probability of it coming up heads?

A. $\frac{1}{2}$
B. $\frac{1}{20}$
C. 1
D. $\frac{2}{3}$
E. Can’t tell until you flip the coin.
What is the probability of the coin turning up heads on ten consecutive tosses?

A. $\frac{1}{2} = 0.5$
B. $\frac{1}{10} = 0.1$
C. $\left(\frac{1}{2}\right)^{10} = 9.8 \times 10^{-4}$
D. $\left(\frac{1}{2}\right)^{1/10} = 0.933$
E. $(1/10)^2 = 0.01$
F. $(1/10)^{1/2} = 0.316$
So how many trials of ten coin tosses would you think you’d have to make, to have a good chance of ten consecutive “heads” at some point?

A. 2
B. 10
C. 100
D. $2^{10} = 1024$
E. $2^{100} = 1.3 \times 10^{30}$
How many coin tosses would that be, in total?

A. 20
B. 40
C. **10240**
D. **1.3×10^3**
E. Can’t estimate from information given.
If one had to try that many times to have a reasonable chance of getting the right molecule, making them out of single nucleotides one by one, it would take much longer than the age of the Universe.

But it takes much less time to make shorter RNAs. And that, plus natural selection, is the key: make short RNAs first, then make longer RNAs out of these.
Baby steps (continued)

Consider, for example, a two-step process.

- Start with water at $T = 300K$ that has the Earth-crust abundance of carbon, 10% of it in nucleotides.
  - Then it takes about 0.04 seconds to form an RNA with 24 nucleotides, and 0.32 seconds to form one with 192, adding nucleotides one by one.
  - And to have a good chance to get a specific RNA with 24 nucleotides by this means, it takes $4^{24} \times 0.04 \text{ sec} = 3.6 \times 10^5 \text{ years}$, while it takes $4^{192} \times 0.32 \text{ sec} = 4 \times 10^{107} \text{ years}$ to get a specific 192-long RNA.

- Clearly 24-nucleotide RNAs are going to form the slow way much faster than the 192s; also at least some of them can self-replicate, as we’ve discussed.
How many different RNA molecules, four nucleotides long, are there?

A. 4  
B. 8  
C. 16  
D. 32  
E. 64  
F. 128  
G. 256
About how many times would you have to construct 4-base RNAs at random to have a good chance of getting a specific one?

A. 4  
B. 8  
C. 16  
D. 32  
E. 64  
F. 128  
G. 256

[Bar chart showing percentages for each option]
In the worst case you would have to do this serially, one RNA at a time. If it takes 0.05 sec to synthesize a four-base RNA, how long does it take to have a good chance of synthesizing a specific one?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.8</td>
</tr>
<tr>
<td>2</td>
<td>1.28</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Values: 2
Value Matches: 62%

$t = 12.8$ sec
So suppose that once certain 24-long RNAs formed, self-replication led quickly (<10^5 years) to the incorporation of all the nucleotides around into chains of 24.

- It takes about 1.6 sec to make a chain of 192 out of chains of 24, given the conditions above.

Furthermore let’s assume that there are 24 different ones that have risen to prominence by replication.

Then it will take 24^{(192/24)} \times 1.6 \text{ sec} = 5000 \text{ years} to have a good chance of making a specific 192-long RNA, like our “RNA replicase” ribozyme. The total time is therefore less than a million years, instead of greater than 10^{107} years.

- And thus much less than the age of the Solar system.
- Goes even faster with more, shorter steps, as we’ll see.
Baby steps (continued)

It would be nice to start with primordial soup and see the evolution of ribozymes produced in the lab in this manner, but unfortunately there are few graduate students willing to wait 5000 years to get their Ph.D.s.

- Existing ribozymes have been “intelligently designed” by the likes of Cech and Altman, rather than evolved from a mixture of prebiotic chemicals.

Meanwhile there are computer simulations based on lab data:

- Ma et al. (2007, 2010a): starting with a prebiotic mixture and allowing the basic sequences to be 5-10 nucleotides long (instead of the 24 we used a minute ago), “RNA replicase” is generated in only 52 years.
Baby steps (continued)

- Also in the simulations: once the first kind of RNA replicase is created, more efficient, *mutated* replicases begin to arise.

- Ma et al. (2010b) have also shown in similar computer simulations that ribozymes provided with a protective cell-like membrane soon (within only a few years!) produce ribozymes which can replicate membrane-component molecules.
  
  - Emergence of *cell* self-replication?

Computer simulation of RNA-replicase evolution, using 10-nucleotide (upper) and 6-nucleotide (lower) basic sequences. A “Monte Carlo step” is approximately 0.38 hours. From Ma et al. 2007.
Time is still the hero.

Thus a sophisticated biochemical function like the “RNA replicase” ribozyme takes less than a million years to produce from scratch, via chemistry, and biological natural selection.

- There are of course lots of functions to evolve, but…

- All the functions of the molecules of life didn’t need to develop serially.

- So thousands of millions of years (i.e. Gyr), very roughly, seem reasonable for making many functions of this level of sophistication. Complex simulations bear this out.

- Still a long time, but much shorter than the time available, and consistent with the fossil record of primitive life on Earth.