Today in Astronomy 106: the important polymers and from polymers to life

- Replication or mass production of nucleic acids and proteins
- Interdependence: which came first, protein mass production or nucleic-acid replication
- Translation: the current fashion in protein manufacture.
- The chicken-egg problem
  - Protein-based primitive life?
  - RNA World
- Emergence of the genetic code.
- Baby steps and the heroism of time.

Primordial soup (Nature)

What is a zwitterion?

A. A molecule with positive charges on each end.
B. A molecule with negative charges on each end.
C. A molecule with positive and negative charges on opposite ends.
D. A molecule with a net positive or negative charge.

This is an example of a zwitterion:

A. Any amino acid in water.
B. Any amino acid in crystalline form.
C. Any acid in water.
D. Any acid in crystalline form.
E. Any organic acid or alcohol in water.

What do zwitterions have to do with polymerization or amino acids?

Replication of nucleic acid

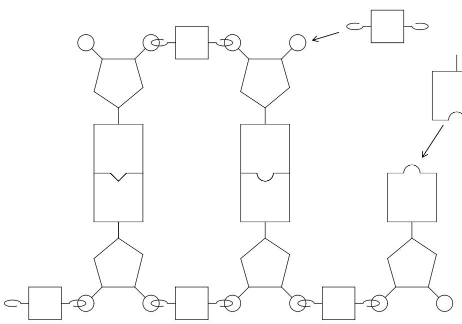
Requirement (#3) of life is replication. Can proteins and nucleic acids replicate, or at least mass-produce?

- If they can, this could explain the emergence of classes of proteins and nucleic acids as components, the assembly of which could serve more complex, life-like roles.

Perhaps easiest to envisiage for short strands of RNA. Suppose a short RNA lies in a solution containing nucleobases, ribose and phosphoric acid.

- A **complementary** polymer can form by hydrogen-bonding nucleotides onto the RNA's nucleobases, hooking up the phosphates and sugars, and then severing the hydrogen bonds.

Replication of RNA



Replication of nucleic acid (continued)

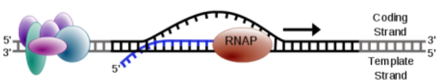
- The second polymer can do the same, thus replicating the first.
- Hydrogen bonds are much weaker than covalent bonds. The "copies" can be stripped off without harm to the original chain.
- Other polymers are capable of "unzipping" the H bonds in DNA. Once unzipped, both sides can bond to nucleotides in the solution, and thus the DNA replicates: bases in same order as original.
- Long chains of DNA can be partially unzipped, capture a complementary chain of nucleotides, have this new chain of RNA zip off, and have the DNA zip back up. This form of partial replication is called **transcription**.

Replication of nucleic acid (continued)

Unzipping does not happen spontaneously very frequently, except with small increases in solution temperature.

- Certain polymers with the right molecules in the right spacing can do it though.
- Among Earth life forms, the "unzipping" and "rezipping" involved in copying of DNA is facilitated by a protein we call **RNA polymerase**.
- This is an example of an **enzyme**: proteins that catalyze chemical reactions which means they don't get chemically changed in the process, like the dust grains that catalyze the formation of molecular hydrogen.
- So replication of DNA and RNA, and transcription, require – in the current age – a special class of proteins.

Replication of nucleic acid (continued)



Animation of transcription in modern lifeforms: go to

http://vcell.ndsu.nodak.edu/animations/transcription/movie.htm

From Phillip McClean and Christina Johnson, the Virtual Cell Animation collection, Molecular and Cellular Biology Learning Center, University of North Dakota.

Image from Wikimedia Commons.

Codons

Three-nucleotide sequences have a special significance, and thus their own name: **codon**.

- Because there are four different nucleobases used in either nucleic acid, there are

$$4 \times 4 \times 4 = 64$$

  different codons.
- As we will see, this has significance for the encoding of information used by modern organisms to build proteins: there are, for example, 22 amino acids used in human proteins, and base sequences that indicate the beginnings and ends of codon sequences that specify the amino acids in certain proteins. 24 < 64, so this works.

Suppose Earthly organisms used only 10 different amino acids. How many nucleotides could be used in a codon?

A. 1 B. 2 C. 3 D. 4 E. Any number greater than 1.

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?

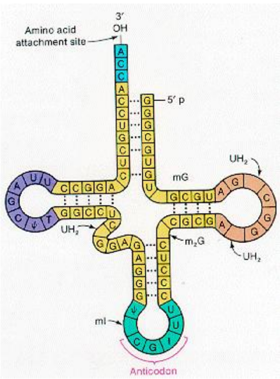
Replication of proteins (continued)

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 an amino acid captured from the solution that is unique to the exposed codon.

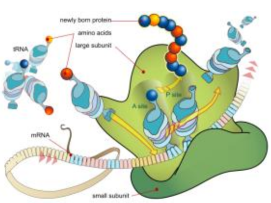
tRNA

From Neal Evans's AST 390L lecture notes.



Replication of proteins (continued)

- The tRNA complementary to the first codon on the mRNA H-bond 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.



Wikimedia Commons

Animation of translation in modern lifeforms: go to http://vcell.ndsu.nodak.edu/animations/translation/movie.htm

Again from Phillip McClean and Christina Johnson (UND).

The original chicken and egg problem

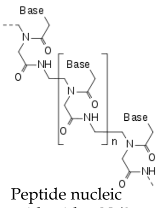
So, in modern Earthly life, the replication or mass production of nucleic acids and proteins are interdependent: each needs certain of the others in order to be produced in large quantities.

- 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 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.



Peptide nucleic acid, with a *N*-(2-aminoethyl) glycine backbone (Wikimedia Commons).

Protein-based 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 life? (continued)

Freeman Dyson (1982, 1995), 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.

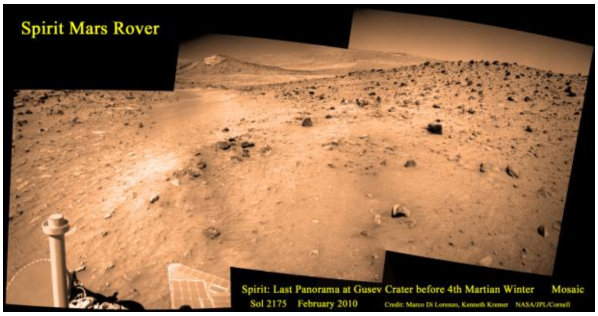
- "Cells" first, enzymes next, nucleic acids much later.
- Cells provide sites for adsorption of monomers.
  - He had acervates/proteinoids and amino acids in mind, but noted that the reasoning would work on crystals and nucleotides as well.
- Repeatable assembly of complex polymers through "order-disorder"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

Protein-based life? (continued)

- 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.

Mid-Lecture Break

- Homework #3 due tomorrow by midnight.



RNA World

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.

RNA World (continued)

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

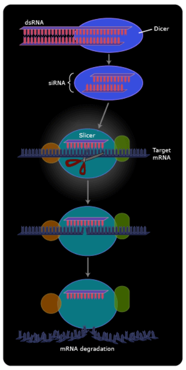
- In the beginning there were short strands of bare, self replicating RNA (Eigen et al. 1981). Eigen calls the varieties of the strands **quasispecies**.
- Most numerous and robust quasispecies predominate (appearance of natural selection).
- Quasispecies mutate, diversify as time goes on.
- Truncated 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 disovering 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)

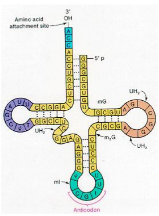
- Simulations and theories show that proteins made from other quasispecies would be likely to perform novel catalytic functions on a given quasispecies.
- RNA can hydrogen-bond to complementary DNA. Eventually DNA took over as the more robust and stable master copy of the base sequences that encode the formation of proteins, but RNA retained its role in transcription and translation.



Ribozymes in action on RNA (Cech 2004).

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)

Nucleotide 1 (5')	Nucleotide 2: U	C	A	G
U	UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine
	UUC Phenylalanine	UCC Serine	UAC Tyrosine	UGC Cysteine
	UUA Leucine	UCA Serine	UAA Ochre (*Stop*)	UGA Opal (*Stop*)
	UUG Leucine	UCG Serine	UAG Amber (*Stop*)	UGG Tryptophan
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine
	CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine
	CUA Leucine	CCA Proline	CAA Glutamine	CGA Arginine
	CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine
A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine
	AUC Isoleucine	ACC Threonine	AAC Asparagine	AGC Serine
	AUA Isoleucine	ACA Threonine	AAA Lysine	AGA Arginine
	AUG Methionine	ACG Threonine	AAG Lysine	AGG Arginine
G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine
	GUC Valine	GCC Alanine	GAC Aspartic acid	GGC Glycine
	GUA Valine	GCA Alanine	GAA Glutamic acid	GGA Glycine
	GUG Valine	GCG Alanine	GAG Glutamic acid	GGG Glycine

The genetic code (continued)

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

Small differences from this occur in mitochondria and bacteria.

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

$$n = \underbrace{4\times4\times4\times\ldots\times4}_{\text{192 factors of 4}} = 4^{192} \cong 4\times10^{115}$$

If they form randomly, we may have to try 4×10^{115} times for a good chance of getting the right one.

Quick education in probability. Flip a coin. What is the probability of it coming up heads?

A. 1/2 B. 1/20 C. 1 D. 2/3

What is the probability of the coin turning up heads on ten consecutive tosses?

A. $\frac{1}{2}$ B. $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$ C. $\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$ D. $\left(\frac{1}{2}\right)^{100} = 8\times10^{-31}$

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. 4 C. 1024 D. 1.3×10^{30} E. Can't estimate from information given

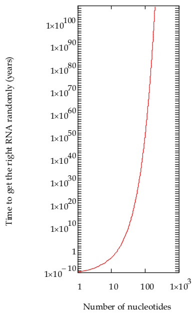
So how many coin tosses total would that be?

A. 20 B. 40 C. 10240 D. 1.3×10^{31} E. Can't estimate from information given

Baby steps (continued)

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 $4^{24}\times0.04$ sec = 3.6×10^{5} years, while it takes $4^{192}\times0.32$ sec = 4×10^{107} 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

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. 16 C. 64 D. 128 E. 256

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?

Enter a *number* this time.

Baby steps (continued)

- 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 $24^{(192/24)}\times 1.6$ sec = 5000 years to have a good chance of making a *specific* 192-long RNA, like our replicator 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.

Time is still the hero.

Thus a sophisticated biochemical function like the replicator 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.