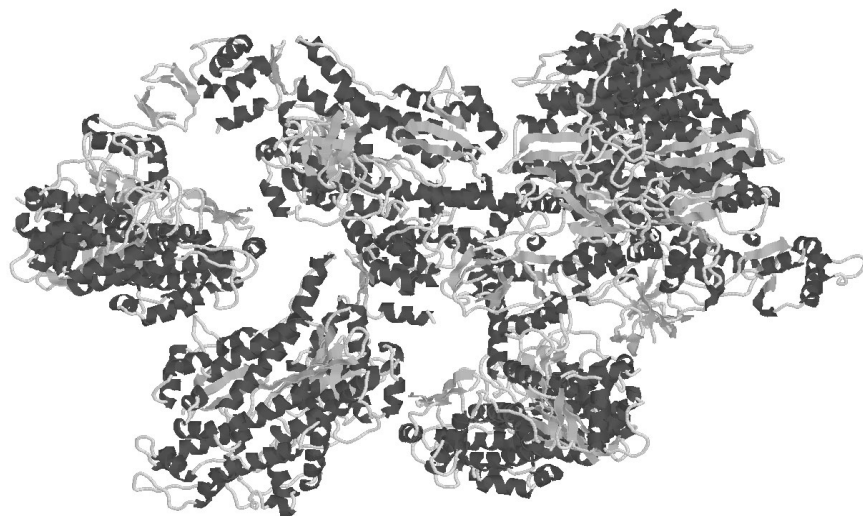


## Part 2: Chemical Evolution



The figure is a cartoon representation of the protein Myosin. Myosin and actin are the two proteins responsible for muscle contraction.

## **Chapter 5: Information & Knowledge before the Genetic Code**

To calculate the information and knowledge for insulin in the last chapter, the genetic code was used to assign a probability to each amino acid arising by chance. For this calculation to be meaningful, both the code and a method to turn the knowledge in DNA into proteins must already exist. So the calculations in chapter 4 assume that life already exists. What about before life exists? How would one calculate the information or knowledge in the very first protein? This is not an easy problem.

Several authors have used thermodynamics, but thermodynamics only applies when the system reaches equilibrium. The relevance of thermodynamic calculations is questionable as amino acids do not polymerize into peptides chains unless external conditions force them away from equilibrium.

This chapter will use information theory to solve the problem. Unlike thermodynamics, information theory can easily deal with non-equilibrium systems.

Information theory cannot normally be used to predict how chemicals will react because some chemicals react with each other readily, and others only react very slowly. Others do not react with each other at all. Thus, the likelihood of two chemicals joining together depends on both the quantity of the chemicals present and their chemical properties. Information theory can easily deal with the effects of quantity, but it has no way to deal with chemical properties.

This chapter will require several assumptions. Without these assumptions information theory cannot be applied to chemical reactions. Fortunately, these assumptions will improve the probability for creating a protein in the primordial soup.

### Assumptions:

- The probability of a peptide bond forming between two amino acids only depends on how many of each amino acid is present in the system.
- The primordial soup only contains amino acids.
- Amino acids do not form non-proteinous bonds with each other. So for example, the carboxylic acid functional groups in aspartate and glutamate do not react with the n-terminus of other amino acids.

The first assumption allows all amino acids to be treated equally. While this assumption ignores the chemical properties of each amino acid, the assumption is not an unreasonable approximation because all amino acids must join together by forming a peptide bond. The second assumption greatly improves the odds of creating a functional protein. By excluding chemicals that react quickly with amino acids, this assumption eliminates chemical reactions that can prematurely terminate a growing peptide chain. It also ensures that the amino acids will be available to interact with each other. The third assumption is not true, but it greatly simplifies the math, and at the same time, it improves the odds of creating a protein in the soup.

With these assumptions, information theory may be applied to the primordial soup. The first step is to estimate the number of each amino acid in the primordial soup. There are two methods. For 50 years, scientists have been trying to find better ways to synthesize amino acids under plausible prebiotic conditions. Many of the 20 amino acids used by life have been synthesized. Because these experiments are riddled with speculation about conditions on the primitive earth and investigator interference, the second method is preferable. This method relies on the amino acids found in meteorites.

## **Meteorites**

Some meteorites contain organic carbon, and several of these have been analyzed for amino acids. This analysis has shown that the amino acids, glycine and alanine, are quite common in some meteorites. Most of the other amino acids used by life are rare, but some are present. In addition, many amino acids not used by life are present. More than 50 non-biological amino acids are found in the Murchison meteorite.

Meteorites are easily contaminated by biological amino acids. So samples are always taken from the meteorite interior. Unfortunately, contamination is still a major problem. Nevertheless, several generalizations are possible.

- The biologically relevant amino acids in meteorites are always predominantly glycine and alanine. Sometimes aspartate and glutamate run a close second, but in many cases, this appears to be the result of contamination. Serine and valine are sometimes present. The other amino acids used by life are absent.
- Non-biological amino acids are common in variety and in number. The most common non-biological amino acids are the many isomers of aminobutyric acid. The second most common non-biological amino acids are two forms of alanine that life does not use.

One comparison of four different meteorites that contain amino acids revealed that only 25% of the amino acids are biologically relevant.<sup>4</sup> If the primordial soup has a similar composition, then only 25% of the amino acids in the soup are biologically relevant, and even if a way is found to make the amino acids join together, the odds of a protein emerging are very small.

The average protein in the Swiss Prot database contains 362 amino acids, and most contain more than 150 amino acids. If the composition of amino acids in the soup is similar to that of meteorites, what is the probability of creating a peptide composed of 150 amino acids if all of the amino acids must be biological?

The knowledge required to build this peptide is simply the knowledge required to exclude all amino acids not used by life. Today, random amino acid sequences do not contain such knowledge because the machinery used by life to build proteins ensures that one of the 20 amino acids used by life will always be placed at each position in the growing chain. This is not true in the primordial soup. The term molecular knowledge in this book is reserved for useful information that conveys a selective advantage. A random sequence of biological amino acids that evolves in the soup will not possess molecular knowledge because the sequence will most likely have no function. Thus, the term information is preferred in this case. To avoid any possible confusion with terminology, this type of information will always be referred to as primordial information. Primordial information is the information needed to exclude non-biological chemicals found in the primordial soup from a growing peptide, RNA or DNA molecule. Since primordial information is a form of knowledge, it can be safely related to a probability. Furthermore, this calculation does not rely on human insight. Before self replication, natural selection cannot exist, so all events are guided by chance and chance alone.

Each addition to the growing chain has a 25% chance of being an amino acid used by life. So each amino acid added to the chain has a 1 in 4 chance of being correct. Thus, there are 4 possible outcomes and only one is desirable. Using equation 1 in chapter 1,  $2^{\text{information}} = 4/1$ , and because  $2^2 = 4$ , the information content for each amino acid added is 2 bits. So a random chain of 150 amino acids that emerges from the soup will contain 300 bits of information. The odds of this arising by chance are 1 time in  $2^{300}$  tries or a 1 in  $2 \times 10^{90}$  chance. What do odds like 1 in  $2 \times 10^{90}$  really mean?

The number  $10^{90}$  is so large that naturalistic explanations will always fail to explain any event whose odds are this poor. To understand why, assume that every single star in the universe has one planet composed entirely of amino acids. Further assume that every one of these amino acids exists as a 150 amino acid peptide chain. The highest estimate for the number of stars in the universe currently available is  $7 \times 10^{22}$ . If the planets orbiting these  $7 \times 10^{22}$  stars are about the same size as the earth, then on average each has a mass of  $6 \times 10^{24}$  Kg. A planet with this mass composed entirely of the amino acid glycine will be made from  $5 \times 10^{49}$  glycine molecules. If all the planets have the same number of amino acids, then there will be  $3.5 \times 10^{72}$  amino acids in the universe. Since every amino acid exists in a chain of 150, there will be  $2.3 \times 10^{70}$  peptide chains. The odds that 1 of these chains will contain only biologically relevant amino acids is only 1 in  $8.6 \times 10^{19}$ . So further assume, that all of these peptide chains break down each year only to reform, and that this process has been going on every year for 15 billion years. The odds improve to 1 in 6 billion. So while the odds are not zero, they might as well be. Nature simply cannot accumulate enough tries to overcome the poor odds.

One can certainly speculate that the first proteins used amino acids that are no longer used today or that these proteins were very short. Both assumptions improve the likelihood for evolution. Nevertheless, all readers need to realize that when a scientist in the lab mixes together pure amino acids that are only used by life, the scientist is adding so much information to the system that the experiment can no longer be considered representative of the conditions on the early earth. The starting point for such experiments is not plausible.

If the soup existed, then the first proteins evolved in a soup that contained many amino acids not used by life. The soup also contained a host of other chemicals like aldehydes that react readily with amino acids. These undesirable side reactions make the evolution of information in the primordial soup very difficult to explain. When a scientific experiment models evolution by excluding these other chemicals, the experiment no longer models the origin of life. Such experiments only model evolution in a test tube.

## **The Evolution of Primordial Knowledge**

The odds of a random amino acid chain evolving in the soup are quite poor, but what about a protein? A protein is not a sequence of random amino acids. The order and type of amino acids in a protein determine how it folds, how it behaves, and its biological function. The sequences are not random. They contain knowledge. The odds of a functional protein evolving are certainly expected to be much less than that of a random sequence.

### **How Many Solutions?**

One of the more important experiments concerning the origin of life was performed by Keefe and Szostak.<sup>1</sup> The authors of this paper in Nature searched six trillion random peptides each composed of 80 amino acids. They were looking for a sequence that could bind the chemical, ATP. They found four sequences in this large pool with ATP binding activity.

This allows for a direct computation of the molecular knowledge required for ATP binding. Using equation 2 in chapter 1, molecular knowledge =  $3.32 \times \log(6 \text{ trillion}/4) = 40$  bits. Notice, that this is not 40 bits of information because the proteins that were selected only possessed minimal functionality. These proteins were subjected to several rounds of selection greatly improving their affinity for ATP.

This experiment provides a direct measurement of molecular knowledge. It also shows that there are very few solutions.

Binding a chemical like ATP is one of the functions that many enzymes possess. So while Keefe and Szostak did not actually find a useful enzyme, they did find a function that many enzymes require. The 40 bits calculated above are for evolution in a test tube. How many bits are required for evolution in the primordial soup?

The minimum possible primordial information in a random sequence of 80 peptides is 160 bits (2 bits per amino acid). The odds of such a peptide evolving are one in  $1.5 \times 10^{48}$ . Given that the odds that a random sequence of 80 peptides will bind ATP are only 4 in 6 trillion, the odds of finding a primitive peptide on the earth that can bind ATP are simply the product of the two numbers or one chance in  $2.2 \times 10^{60}$ . Alternatively, the 160 bits needed to construct an 80 amino acid in the soup may be added to the 40 bits calculated above. The total knowledge is thus 200 bits, and the odds of this happening are 1 in  $2^{200}$  or 1 in  $2.2 \times 10^{60}$ . In this calculation, the total knowledge required is simply the sum of the molecular knowledge and the primordial information. After life exists, primordial information always equals zero, and molecular knowledge always equals total knowledge.

Binding ATP is a simple function. Clay, a simple mineral, binds ATP. Furthermore, the function by itself does not confer a selective advantage. Thus, ATP binding is below the threshold of molecular knowledge. This function must be combined with another function before natural selection will preserve it. To create a functional enzyme that can be preserved by natural selection quite a bit more knowledge is required. Since it takes 200 bits to bind ATP, assume that it also takes 200 bits to bind another molecule. Thus, 400 bits is a more reasonable approximation for a functional enzyme, and the odds for such evolution are given by 1 time in  $2^{400}$  tries or a 1 in  $2.5 \times 10^{120}$  chance.

The origin of the first enzyme just cannot be explained in this way. The odds are too poor.

## **Molecular Knowledge Before Life**

This section will investigate how the composition of the soup influences knowledge. If the soup contains mostly glycine and alanine along with a host of other amino acids not used by life, then the probability of a useful protein emerging from it must be very low. Chapter 4 calculated the molecular knowledge for the protein insulin. This chapter will repeat this procedure assuming that insulin emerged in the soup before life. By this calculation this chapter does not mean to suggest that insulin originated in the soup. The calculation is for comparison only. Remember insulin was only chosen because it does not contain many amino acids, and this makes the calculations easier.

## **The Composition of the Soup**

If meteorites are used to reconstruct the composition of the soup, then 14 of the 20 amino acids used by life will be absent. Only glycine, alanine, valine, serine, aspartate and glutamate would be available in the soup. The proteins used by life today require more than 6 amino acids. While this prediction of the soup's composition is probably the most accurate, it is an undesirable composition. So this chapter will assume a much more favorable composition.

Life uses 20 amino acids. Seventeen of these have been synthesized in the lab under conditions that might be similar to the conditions found on earth 4 billion years ago. Some amino acids are quite easy to synthesize, and others are very difficult. The amino acids that are easy to synthesize invariably are the primary product of these experiments. The other amino acids occur in various concentrations depending on the conditions chosen to carry out the experiment. Three amino acids, histidine, arginine, and lysine, have not been synthesized under plausible conditions.<sup>2</sup>

Because no single experiment has generated all of the amino acids, if the soup's composition is taken from the results of a single prebiotic experiment, then the composition will be unfavorable for protein evolution. Most proteins need 18 or 19 different amino acids to function. To construct a favorable composition for protein evolution, it is either necessary to combine many different prebiotic experiments or to just assume that the absent amino acids are present. This section will take the latter approach.

On page 87 of his book, Miller lists the results from one of the most successful prebiotic experiments.<sup>3</sup> The yields of ten amino acids are listed in this table.

As a reasonable starting point, assume the abundance of the amino acids in the primordial soup tracks Miller's table. Ten amino acids are not found in Miller's table. Seven of these have been synthesized under plausible prebiotic conditions. Assume that these seven are as abundant as threonine. Threonine is the least common amino acid listed in Miller's table. Three amino acids have not been synthesized in the lab. Assume that these are found in the soup at 1/10 the concentration of threonine. Finally, assume that the 20 amino acids that life uses comprise 1/4 of all amino acids present in the soup. Thus, the soup ratio of biological to non-biological amino acids is similar to the ratio found in meteorites.

These assumptions improve the odds that a protein will emerge in the soup. For example, one could easily assume that the ten proteins not found in Miller's table were also absent from the soup. With this single assumption, the information and molecular knowledge found in most proteins becomes infinite. Furthermore, the assumption to exclude chemicals like aldehydes and formic acid greatly improves the likelihood for protein evolution.

With these assumptions in place, labeling wooden blocks according to amino acid abundance yields table 5.1. The number of blocks in the second column are taken from Miller's table. The right column is based on what might have been given the constraints of the favorable assumptions discussed above.

Table 5.1: Wooden Blocks Used to represent Chemicals in the Soup

amino acid	number of blocks	amino acid	numbers of blocks
glycine*	440,000	tryptophan	400
alanine	395,000	tyrosine	400
valine	9,750	histidine	40
leucine	5,650	lysine	40
isoleucine	2,400	cysteine	400
proline	750	methionine	400
aspartate	17,000	phenylalanine	400
glutamate	3,850	arginine	40
serine	2500	asparagine	400
threonine	400	glutamine	400
Total number of blocks labeled with amino acids used by life		880220 (sum of column 2 and 4)	
Total number of blocks		4 x 880220 = 3,520,880	

\* Most amino acids exist in two forms. The forms are mirror images of each other. Life only uses one image. Glycine is the only amino acid that does not have a mirror image. Thus, the number reported for glycine in table 5.1 corresponds to the concentration reported in Miller's table. The numbers associated with all other amino acids in the left column are ½ the value reported in Miller's table.

## The Evolution of a Functional Protein in the Primordial Soup

Because three million blocks cannot fit in a basket, the trapped scientist is now given a truck (figure 5.1). The blocks in the truck are determined by table 5.1. The scientist can draw blocks from a tube that connects his room to the back of the truck. How much information would insulin contain, if it evolves given these constraints.

Figure 5.1: Trapped Scientist with a Truck

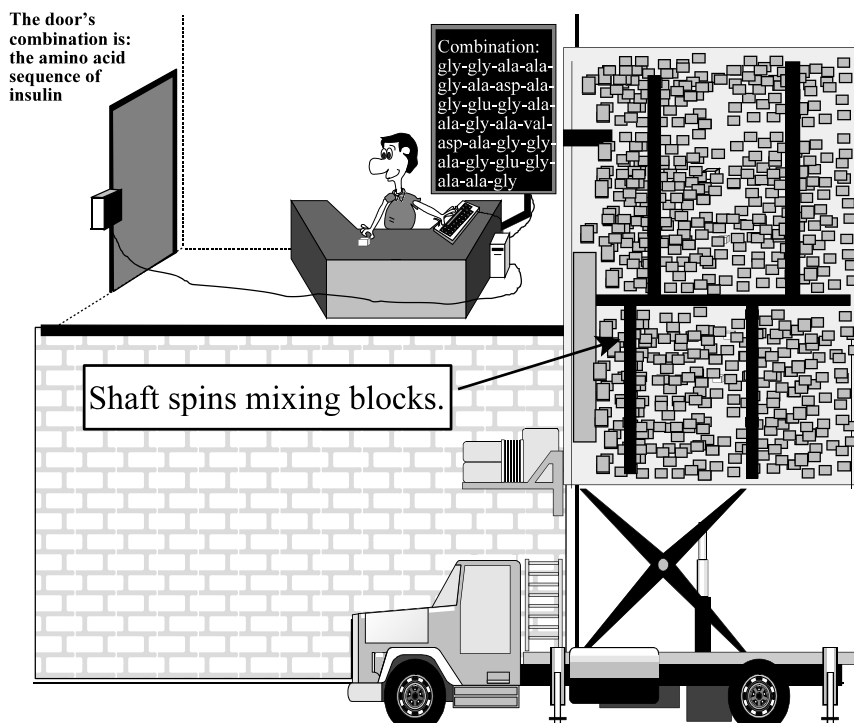


Table 5.2: Information in Insulin B Chain (Primordial Evolution)

pos	allowed amino acids	bits	pos	allowed amino acids	bits
2	phe, ala, leu, val	3.1	17	phe, tyr	12.1
3	val, ala, pro	3.1	18	leu	9.3
4	pro, lys, asn	11.5	19	val, ile	8.2
5	gln	13.1	20	cys	13.1
6	his, arg	15.4	21	gly	3.0
7	leu	9.3	22	asp, glu	7.4
8	cys	13.1	23	arg	16.4
9	gly	3.0	24	gly	3.0
10	ala, pro, ser	3.1	25	phe	13.1
11	his	16.4	26	phe, tyr	12.1
12	leu	9.3	27	tyr	13.1
13	val	8.5	28	thr, ser, asn	10.1
14	glu, asp	7.4	29	pro	12.2
15	ala	3.2	30	lys, arg	15.4
16	leu	9.3	31	ala, arg, thr, ser	3.1

**Total bits = 280.5 bits**

**Example calculation:** Phenylalanine, alanine, leucine and valine are possible at position one. There are 5,650 blocks labeled leu, 395,000 labeled alanine, 9,750 labeled valine, and 400 labeled phenylalanine in the truck. The total number of blocks is 3,520,880. So the probability that the scientist will pull a leucine, alanine, phenylalanine, or valine is  $(395,000+9,750+5,650+400) = 410,800$  times in 3,520,880 tries. Thus the information at position one is calculated as follows:  $\text{information} = 3.32 \times \log(3,520,880/410,800) = 3.1$  bits.

The total number of bits is 280.5. In chapter 4, the total for the B chain was only 108 bits. Intuitively, this is obvious because any proteins that emerge in the primordial soup will be composed of mostly alanine and glycine. Since real proteins do not follow this pattern, they are less likely to evolve in the primordial soup. The conclusion is that it is much harder for information and knowledge to evolve in the primordial soup.

The B chain of insulin contains 30 amino acids. So the average information contributed by each amino acid is equal to the total information divided by 30.

Information before life =  $280.5 / 30 = 9.35$  bits per amino acid

Information with the genetic code =  $108/30 = 3.6$  bits per amino acid

Because knowledge is defined in terms of information, it too must increase.

## Molecular Knowledge in The Primordial Soup

Table 5.3 calculates the knowledge in the B chain of insulin assuming that the protein evolved in the primordial soup.

Table 5.3: Molecular Knowledge in Insulin B Chain

pos	allowed amino acids	bits	pos	allowed amino acids	bits
2	phe, ala, leu, val	2.0	17	phe, tyr ,(trp)	2.0
3	val, ala, pro	2.0	18	leu, (ile),(val), (ala), (met)	3.1
4	pro, lys, asn	2.0	19	val, ile, (ala), (leu), (met)	3.1
5	gln, (asn)	12.1	20	cys	13.1
6	his, arg, (lys)	14.8	21	gly	3.0
7	leu, (ile), (leu), (val), (met)	3.1	22	asp, glu	7.4
8	cys	13.1	23	arg, (lys), (his)	14.8
9	gly	3.0	24	gly	3.0
10	ala, pro, ser	2.0	25	phe, (tyr), (trp)	11.5
11	his, (lys), (arg)	14.8	26	phe, tyr, (trp)	11.5
12	leu, (ile), (val), (ala), (met)	3.1	27	tyr, (phe), (trp)	11.5
13	val, (ile), (leu), (ala), (met)	3.1	28	thr, ser, asn	2.0
14	glu, asp	7.4	29	pro	12.2
15	ala, (leu), (ile),(val), (met)	3.1	30	lys, arg, (his)	14.8
16	leu, (ala), (val),(Ile), (met)	3.1	31	ala, thr, ser, -	0*

Total = 211 bits.

\* Any position with a gap does not need an amino acid and therefore the knowledge is set to 0 bits.

**Example calculation:** At position 7, only leucine is found in the alignment. Nevertheless, the technique to calculate knowledge assumes that the other amino acids in this group are allowed. The number of blocks labeled with the 5 amino acids belonging to group 1 in the truck (figure 5.1) is 413,200. There are 3,520,880 total blocks. So the knowledge is  $3.32 \times \log(3,520,880/413,200) = 3.1$  bits.

Notice that no position can ever contribute less than 2 bits. If all 20 amino acids are found at a particular position, the position still contributes 2 bits. This accounts for the amino acids not used by life found in the soup.

The average knowledge per amino acid in the soup is calculated as follows: knowledge = 211 total bits / 30 amino acids = 7 bits per amino acid. The average knowledge per amino acid with the genetic code is only 76 total bits / 30 amino acids = 2.5 bits per amino acid. (Refer to pg. 76, table 4.4 for number of bits using the code).

Because of the nature of logarithms, the implications are dramatic. Suppose that one of the first proteins to evolve contains 100 amino acids, and that 30% of this protein shows a conservation pattern similar to insulin.

Knowledge today = molecular knowledge =  
 100 amino acid x 2.5 bits per amino acid x 30% = 75 bits

Odds of evolving today are 1 time in  $2^{75}$  tries or 1 time in  $4 \times 10^{22}$  tries. This could happen with enough tries.

Knowledge soup = molecular knowledge + primordial information

Knowledge soup =  

$$\frac{100 \times 7 \text{ bits per amino acid} \times 30\% + 100 \times 2 \text{ bits per amino acid} \times 70\%}{350 \text{ bits}}$$

Odds of evolving in the primordial soup are 1 time in  $2^{350}$  tries or 1 time in  $2.2 \times 10^{105}$  tries. This can never happen.

Before life exists, chance will require an incredible number of tries to create knowledge, and the vastness of space, the number of atoms in the universe, and the incredible age of the universe do not make a dent in the problem. Nature simply cannot accumulate enough tries to overcome the poor odds.

Finally, this chapter had to make quite a few assumptions. Some readers may be concerned about these assumptions, but realize that almost every assumption was for the benefit of protein evolution. For example, this chapter assumed that primordial soup did not contain aldehydes, carboxylic acids, and amines. This assumption is obviously false, but it greatly improves the chance for a protein evolving because it eliminates many side reactions. Also the amino acids not listed in Miller's table (because they are not present in significant quantities), are assumed to be in the soup at a very generous level. Allowing every star in the universe to have one planet is certainly a generous assumption, but perhaps the most generous assumption is to allow every single one of these planets to be composed entirely of peptide chains each containing 150 amino acids. This assumption is only rivaled by the next one that allows these peptide chains to break down and reform every year, and even with all of these generous assumptions, the probabilities do not budge from zero.

Any scientist who believes that nature can create molecular knowledge before life exists is relying on faith to justify his opinion because the math just does not support this belief. No matter how favorable the assumptions, the results are always the same.

#### References:

- 1) Keefe and Szostak, "Functional Proteins from a Random Sequence Library," *Letters to Nature*, 410: 715-718, 2000.
- 2) Miller, Which Organic Compounds Could have Occurred on the Prebiotic Earth?, *Cold Spring Harbor Symposium of Quantitative Biology Volume L11*, 17-25, 1987.
- 3) Miller, Orgel, *The Origins of Life on Earth*, Prentice Hall, 1974
- 4) Ehrenfreund, Glavin, Botta, Cooper, Bada, "Extraterrestrial amino acids in Orgueil and Ivuna: Tracing the parent body of CI type carbonaceous chondrites," *PNAS*, 98: 2138-2141, 2001.