A quote from the book 'information theory, evolution and the origin of life' by Hurbert.p.yocky(page no.93): "many papers have been published with titles indicating that their subject is the origin of the genetic code, but actually the content deals with only with its evolution……we find the origin of the genetic code is unknowable……."

If the above lines are true, the following article is probably the first one to handle this 'unknowable' puzzle with pure logic…..Read on…….

# ORIGIN OF GENETICS CODE: WHAT LOGIC TELLS US

## INTRODUCTION:

In living cells, function of a particular molecule is directed mainly by its structure. Proteins are the primary functional molecule and they are formed by amino acids. On the other hand DNA is made up of nucleotides (A separate chemical entity from amino acid). Though separate in chemical and physical characteristics, DNA acts as a database from which proteins are synthesized via the mechanisms known as 'Transcription' & 'Translation'. The most significant event occurred here is the conversion of language (Language of DNA to Language of proteins), for which the term trna (Translator RNA) has been coined, because its action is very similar to a translator or interpreter. It translates 'codon' [3 letter words of nucleic acid] into 'Amino acid' (Structural unit of protein). Origin of such a translator system & genetic code is hard to explain because such system requires high fidelity & strong evolutionary force in primitive environment. So, in this article we will explore the possible mechanisms along with their advantages & limitations to explain the origin of such a system.

## SIGNIFICANCE OF GENETIC CODE

In a living system (i.e. cell), every molecule is somehow important & necessary. So, the translation machinery & its function certainly carries some significance. A translation system is required only when there exists two different languages [DNA & Protein]. Language of protein is purely functional & it is directly implemented. In contrast, language of DNA is absolutely non-functional in the absence of a translation machinery. This finding suggests that origin of an information storage system (primitive DNA / RNA or any other molecule) must be coupled with origin of translation system & it stores information in a way totally different from proteins. The language of information storage system (which is expressed) must be linear or two dimensional & it is called genetic code. The information storage system must carry some special properties –

1. It can't be 3 dimensional as protein, so that it would contain maximum information in minimum space.
2. It can be copied efficiently. So it can not be a 3 dimensional structure as protein. It must be a linear structure.(it is obvious, because dna is replicated by a mechanism which can not be applied for protein

On the other hand, proteins must be 3 dimensional to carry out its function efficiency, simply be cause the environment in which it works, is 3 dimensional.

So, this translation actually helps the molecule to change its dimension (From 2D to 3D). We can call it a dimensional conversion.

## WHICH DIMENSION CAME FIRST ?

Let us imagine the primitive environment on earth. There were thousands of non functional molecules, interacting with each other without any foresight or intention. No doubt, those non-living molecules were the precursor of a first living system. But how such conversion could occur? Here we can imagine a car (representing molecules / system) running from place A to Place B (From non-living or disordered state to a relatively ordered / living state). Now, the pathway of this journey is not important. There are thousands of such pathways, which are imaginable. The more fundamental problems is, what was the fuel of that car ? What could be the evolutionary force in absence of replication (& thus Darwinism) ? We will be able to think of an evolutionary force if we expand the Darwinism a bit, from its original statement.

Now, if we look carefully, to Darwinism, we will see that the key process of Darwinism is ‘natural selection’ that favours certain properties in living systems.

| Components of living system | | Properties favoured by natural selection |
|---|---|---|
| A single molecule :- | → | Stability |
| | → | Structural flexibility |
| | → | Catalytic efficiency |
| | → | Information storage |
| A multi-molecular system:- | → | Replication |
| | → | Energy utilization |
| A group of Bio-system | → | Co-operatively among systems |

Therefore, in the first step of Life formation, the possible evolutionary force could be – Natural selection based on stability yet structural flexibility among the molecules.This could be achieved by structural & functional properties of those molecules.

In the absence of replication & translation the stability with a degree of flexibility in structure becomes the major property for natural selection. This two

properties must be implemented through a direct & independent manner. So they must have had a 3 dimensional working structure. So if logically traced, we see that 3-D structure must came prior to 2-D structure.

[Note: What I mean by 2-D structure is that to serve its function the molecule may not have diverse / variable 3 dimensional involvement, though no molecule in the world is literally 2 dimensional (according to some school of thought language of DNA is 1 dimensional. But here I am using 2 dimensional for the same purpose with the same meaning)].

## FIRST STEP TOWARDS LIVINGNESS

The above-mentioned two properties can be achieved only when a molecule can facilitates its own formation along with some altered structure from the original. In the absence of replication, this can be achieved by two means –

1. Auto catalysis
2. Hyper cycle

Autocatalysis is actually a hyper cycle, which contains only 1 member (In hyper cycle number of members are many more). So how many members should be there to establish an efficient hyper cycle ?.

As we see the key pattern of life is the central dogma. From amoeba to human beings, this basic pattern is the basis of everything. As it is the central event of all biological events, it must had come in a very early time. Which is to be noted at this point, that central dogma itself is a hyper cycle of 2 members (DNA & Protein), though involvement of other molecule is also seen. This finding suggests

that in nature, two membered cycle (now we don't need to call it a 'hyper cycle'), are suited best in bio-systems (as only 2 different kind of languages is necessary to store information and for functional activity). So next, we will imagine the origin & possible fate of such a two membered cycle in primitive earth because of natural selection.(two membered means presence of minimum types of languages it contains. involvement of accessory molecules are ignored).

## ORIGIN OF TRANSLATION – BASIC CONCEPT

Modern proteins carry out different functions. Those functions are highly precise & unlikely to be present in the most primitive form of protein. The most primitive function of protein appears to be its enzymatic / catalytic property. So, Let's imagine the first functional molecule (precursor of protein) to have the catalytic property. Now suppose it has two functional domains I & J. This molecule suppose, is able to catalyze two separate reactions forming 2 separate molecule A & B. Now if formation of A & B would impossible without the molecule M, we can actually say – I codes for A & J codes for B. This event must be the first step towards the origin of 'Language of molecules' providing those molecules A & B somehow helps M to be stable or Reformed. So, suppose M molecule catalyzes the following event with domain D.

$$X \xrightarrow{D\ (M)} Y$$

If we compare this event with central dogma we will see following things.

| Component | | Precursor |
|---|---|---|
| DNA / mRNA | → | D domain |
| trnA | → | X |

Protein / Amino acids ⟶ Y

Now we need to see, how a simple system could be converted into complex translation machinery.

## EVOLUTION OF EACH COMPONENTS

Evolution of the above system would be possible if –

1. All the molecules are to be stable / formation of a ‘Re forming cycle’
2. Scope of structural alteration
3. Every single step will carry some functional advantages

1. *Evolution of ‘D’ domain* – Autocatalytic / hypercyclical feedback could create lot of functional domain / sub-domains in M or similar molecule. They carry the information for synthesis of other molecule. Advantage of such a molecule like M is that if it could be copied, formation of thousands of molecules can be facilitated (they are formed by reactions catalyzed by M) & need not to be copied separately. But for this event to occur, the M is to be converted into 2 dimensional molecule, slowly & it will require other catalysts to execute its function. This dimensional conversion will be considered later in details.

2. *Evolution of X* – It should be logically co-evolved with ‘D’ domain to give it functional support.

3. Evolution of Y – As original functional molecule, it forms different pathways (all catalysed by M family) to form different functional molecules. Now they also catalyze reactions, form hypercycle etc to form new molecules to support evolution of ‘D’ domain. But their catalytic diversity became more prominent, and their structural basic hinders them from being 2D structure. Therefore, they become protein.

## DIMENSIONAL CONVERSION OF ‘D’ DOMAIN

So, the key process in the origin of genetic code appears to be the conversion of M from a 3 dimensial to 2 dimensional structures. Once it possible, it is copied (origin of replication & Darwinism) and from then on, the domain will be read using code (origin of genetic code). As the 2-dimensional structure (information storage molecule) is made up of nucleotides today, we should think of a similar origin. It suggests that from beginning, the M molecule is structurally similar to a nucleic acid molecule and as it is catalyzing the formation of Y (precursor of protein), we should think of a nucleic acid that can synthesized protein. Only one kind of such molecule is known, that is peptidyl transferase. Basically it joins amino acids together & the chain grows. Now the way it acts is found to have a special advantage.

Suppose, the ‘D’ domain catalyzes protein synthesis in different ways. Maximum variation of protein could be achieved using minimum number of domain if domains function as amino acid ligator (joining the units together). So by single peptide bond forming action it could give rise to enormous variety of proteins (the number depends on several environmental factors like presence of amino acids and other chemicals, pH, Temperature etc). If some of those proteins

serve back i.e. help the M protein to maintain its stability & function the whole system survives.

Now, obviously, only very few proteins will do that job. Let's think that they are the members of a protein group named 'P'. Now M proteins will survive more if they can generate more 'P' proteins. Now how could that be done ?

Now, let us think of a selective mechanism that could control the protein formation. This mechanism can act by –

1. Selecting only those amino acids those are useful to form 'P' group of proteins.
2. Arranging those amino acids in a meaningful way.

This selective mechanism can not be formed in one day. It should take time but it is not impossible as Darwinism already start functioning. To form the 'P' proteins more efficiently, it requires a database to dictate correct amino acid towards 'D' domain at the right time. As the database is made up of nucleic acid, it requires an adaptor molecule that will bind both to the database and the amino acid. And this adaptor molecule subsequently would become trna.

## FORMATION OF A DATABASE OUT OF NOTHING

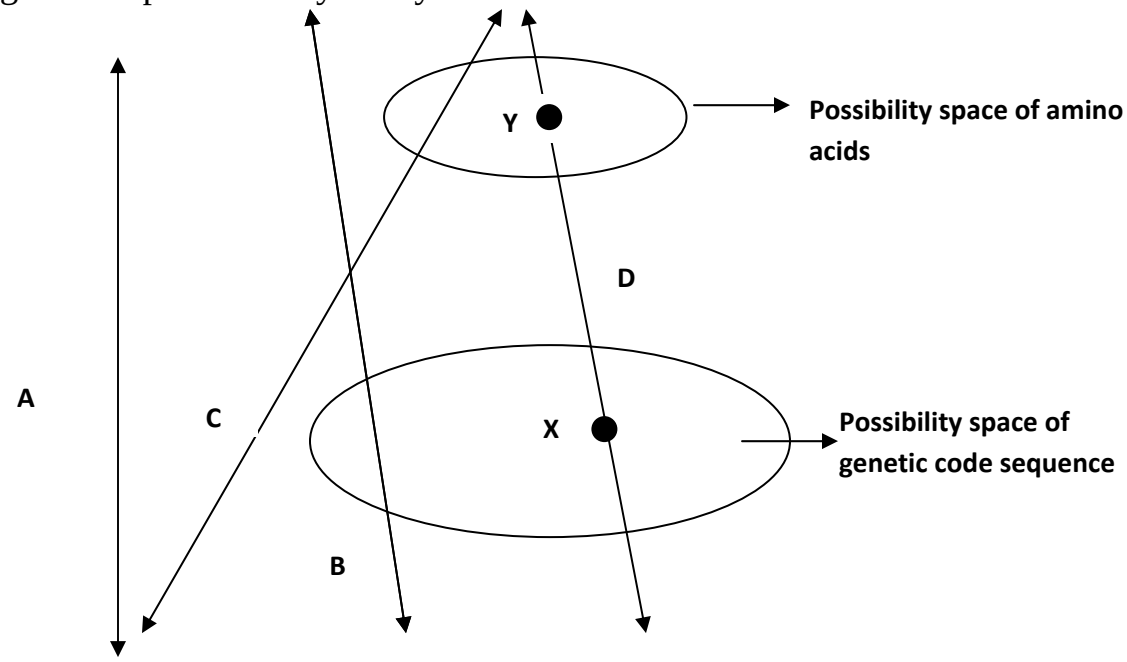
But how a database could be formed specifically for the p protein? Actually, it is really hard to imagine today because the optimum language (genetic code and amino acids) has already been invented. But it is relatively easier in a language less era.

Let's imagine a particular nucleic acid sequence (N) in a molecule M. Now suppose there is a molecule 'A' that is able to bind both N & a particular amino acid 'V' (it will dictate more 'V' in the formed proteins). If presence of V amino acid makes the 'P' proteins more stable or active, this property would be selected & presence & formation of 'N' & 'A' will be favoured together. This 'N' sequence is a step towards the database formation & in presence of 'A' the fact that 'N' codes for 'V' would be a step towards language or genetic code formation.

When a genetic code is to be originated first, it need not to be perfect because it would be so, the chance of formation would become so less it would be practically zero. Let's imagine 3 different possibility space for N, A and V. The more they will be perfect with respect to each other, less will be their chance of superposition and actually, the language could not be formed. But if they would

not be that perfect, the possibility spaces will be increased and when they superimposes, genetic code is invented. Once it is invented, it will be evolved towards perfection more and more.

It is to be noted that the basic feature of such an 'AVN' system is that the formed 'P' proteins will facilitate the system to survive and diverge (mutate) for e.g. one 'P protein' may catalyse to form more A molecule etc.




A = adaptor molecule with no functional relationship with any of the possibility spaces.

B&C = Adaptor molecule with functional relationship with only one possibility space.

D = It is the adaptor molecule which is functionally related with both of the possibility spaces. It crosses the spaces at point X & Y. It means that in the presence of D adaptor molecule X sequence codes for Y amino acid. If Y amino acid increases the activity of P protein, this code (X for Y) will be established.

Today in living organisms 64 such lines are observed which are optimum for functioning. Nevertheless once upon a time a first line (D) was formed, that may not be included in today's 64 lines.

If we imagine ancient human beings, suddenly started speaking in modern English, probability is zero. But they can use simple language that could change into modern language through the course of time.

## DISCUSSION

Though the evolution of genetic code is a very active field of research and amount of literature is being increased day by day, the origin of such a code out of nothing is also a similar and even a more fundamental mystery. But this problem has been attempt less in literature. In this article it is proposed that the gene without genetic code and translation system is meaningless, so gene translation machinery and genetic code was originated simultaneously and randomly and then evolved. As an adaptor molecule (A) is present from beginning, there is no need for physico-chemical similarities between the database sequence N (precursor of gene / genetic code) and amino acids. The whole process also supports the RNA world hypothesis and seek to explain how genetic code could have originated in a primitive RNA world.

A single RNA molecule is synthesized randomly

| Ability faboured by natural selection | | Why it was favoured ? |
|---|---|---|
| Synthesize molecule that serves back that RNA | Step 1 | Serving molecules are advantages for it |
| The catalytic process is like joining the units together | Step 2 | Gives rise to a huge number of molecule so increasing the chance of serving molecule formation |
| Formation of a selective mechanism (database & an adaptor) that could dictate the catalytic process | Step 3 | More formation of serving molecules in a directional manner |

RNA molecule becomes an enzyme, which could also carry information This molecule eventually survives & gives rise to a RNA world

Now an interesting shift is not worthy. As we have mentioned, the central dogma consists of 3 things – a database (DNA/RNA), a translator (translation machinery that can read genetic code) and functional proteins. In this context, we can see that the D-X-Y system has shifted towards N-A-V system. Though the N-A-V type system represent a more recent precursor of today's central dogma, D-X-Y system must came first to facilitate the formation of N-A-V system subsequently D-X-Y became secondary. D-domain only functions as peptidyl transferase. This model implies origin of peptidyl transferase kind of Ribozyme at the first stage favoured by hypercycles. After that came the trna type of activity. The whole process is considered in the context of RNA world hypothesis (the most accepted hypothesis in the origin of life).

## FROM RNA WORLD PERSPECTIVE

Until now, we have discussed the model taking amino acids as a prototype of 3-D structure. However, it needs not to be like that. It can be any molecule

which is functionally active, even can be RNA with enzymatic property (In the context of RNA world hypothesis) pre-RNA molecule with similar properties. Now, let us see, if we assume RNA instead of protein / amino acid at the beginning, how it can affect the sequel. In that case, in our D-X-Y system all the molecules would be RNA.

If X or Y is RNA, it has to be converted into amino acid. That was certainly possible because it would be favored by natural selection (as proteins are way more efficacious than RNA to perform catalytic function and hypercycle had already established Darwinism).

Logic says that both 2D and 3D type of molecule can not be formed simultaneously without a coding mechanism. Logic also says that it is easier to frame a mechanism with similar M, X and Y molecule.

Now, proteins can’t store information. But RNA can do both. So it is more likely that origin of genetic code took place in a RNA world.