A brief history of the genetic code

The central dogma

In the modern world, we observe a set of basic biochemical principles, and describe them in the textbooks. Most of them we take as granted, we see them already formulated in the textbooks, try our best to conceive them, give them proper description, study their relationships, etc.

The *central dogma* in one of the most fundamental natural laws, which is the cornerstone principle of molecular biology. The central dogma tells us that the information goes from nucleic acids to proteins, and not in the reverse manner. Say, a gene sequence imprints itself into a protein sequence, but a protein sequence does not imprint itself into the structure of genes. To understand how the protein-gene feedback happens, scientists came up with the concept of mutations and evolution of organisms. The power of the central dogma is so significant, that this rule does not have exceptions. People, who are familiar with biochemistry know that almost every rule you formulate in this scientific discipline will at some point have exceptions. Biochemistry is full of exceptions. However, the central dogma does not have exceptions. Proteins will not copy themselves into an RNA or DNA stretch. This is a no go.

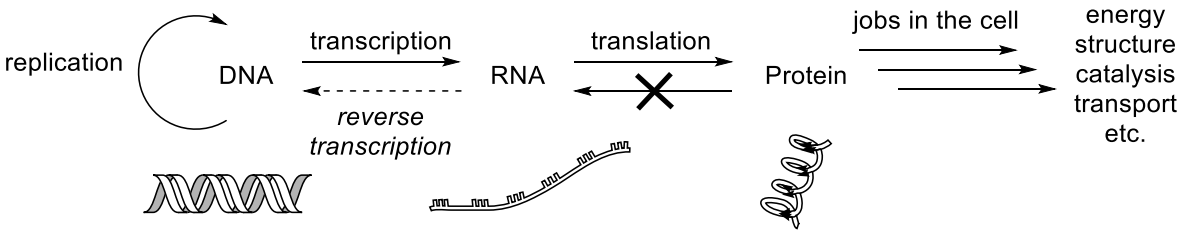

Figure 1. The central dogma

This simple fact shows us, how fundamental the central dogma is. It provides us with the basic principle, which allows us to structure further information regarding biochemical systems and their relationships.

The genetic code

The next fundamental principle which comes directly from the central dogma is the *genetic code*. The genetic code prescribes how the nucleic acid sequence should be interpreted as an amino acid sequence. In the sequence of DNA or messenger RNA, we always find same principles how these should be interpreted. A stretch of a nucleic acid which corresponds to one protein is known as a *gene*. A gene always starts with a start codon, which codes for methionine, ATG. In bacteria, the start codon codes for N-formyl-methionine, however, this does not change the principle: every gene starts with ATG. Then comes the sequence of nucleotides that we break into so-called triplets, each of them will code for one amino acid. For example, a gene sequence that codes for human myoglobin will look like this:

ATGGGGCTCAGCGACGGGGAATGGCAGTTGGTGCTGAACGTCTGGGGGAAGGTGGAGGCTGACATCCCAG
GCCATGGGCAGGAAGTCCTCATCAGGCTCTTTAAGGGTCACCCAGAGACTCTGGAGAAGTTTGACAAGTTC
AAGCACCTGAAGTCAGAGGACGAGATGAAGGCATCTGAGGACTTAAAGAAGCATGGTGCCACTGTGCTCAC
CGCCCTGGGTGGCATCCTTAAGAAGAAGGGGCATCATGAGGCAGAGATTAAGCCCCTGGCACAGTCGCATG
CCACCAAGCACAAGATCCCCGTGAAGTACCTGGAGTTCATCTCGGAATGCATCATCCAGGTTCTGCAGAGC
AAGCATCCCGGGGACTTTGGTGCTGATGCCCAGGGGGCCATGAACAAGGCCCTGGAGCTGTTCCGGAAGG
ACATGGCCTCCAACTACAAGGAGCTGGGCTTCCAGGGCTAG

It starts with ATG, which is methionine/start, then comes GGG, which corresponds to glycine, CTC, which codes for leucine, AGC, which codes for serine, and so on, until we come to one of the stop signals, this will be TAA, TAG or TGA. In the case of myoglobin, the stop signal is TAG. Every triplet of nucleobases codes for either start/methionine, or for one of the 19 amino acids (20 minus methionine), or stop signal (TAA, TAG or TGA). When we translate the abovegiven sequence into the protein sequence we will get the protein sequence of myoglobin:

MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHG ATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKA LELFRKDMASNYKEL

In this sequence, we highlighted methionine (**M**) to show that this amino acid may come at start, but can come again in the middle of the protein sequence. Every protein start with the methionine coding triplet, ATG, but not every methionine coding triplet indicates a start.

Thus, every triplet of nucleobases corresponds to a particular event in protein biosynthesis, therefore we can simply *translate* the gene sequence into a protein sequence. The process of the protein biosynthesis, where a messenger RNA is interpreted into a protein sequence is therefore called *translation*. The triplets that code for an event, start/methionine, amino acid or stop, are also called *codons*. In the latter we will use this word to designate a coding triplet in the sequences of DNA or messenger RNA.

Every codon means something. In the DNA sequence, the codons are made out of four nucleobases, A, T, G and C, while in messenger RNA, there is also for nucleobases, A, U, G and C. When we combine these into codons (triplets), we find that there is 64 unique combination of the nucleobases. One of them means start/methionine or methionine (AUG), three mean stop (UAA, UAG and UGA), the rest 60 code for 19 amino acids. We see that there is actually more codons than amino acids, therefore, some amino acids are coded by several codons. Let's list them:

- The least number of codons have two amino acids, methionine (AUG) and tryptophan (UGG), only one each;
- 9 amino acids, phenylalanine (UUU, UUC), tyrosine (UAU, UAC), cysteine (UGU, UGC), histidine (CAU, CAC), glutamine (CAA, CAG), asparagine (AAU, AAC), glutamic acid (GAA, GAG), aspartic acid (GAU, GAC) and lysine (AAA, AAG) are coded by two codons each;
- One amino acid, isoleucine, is coded by three codons (AUU, AUC, AUA);
- 5 amino acids are coded by four codons, valine (GU**X**), proline (CC**X**), alanine (GC**X**), glycine (GG**X**) and threonine (AC**X**);
- 3 amino acids are coded by six codons each, arginine (CG**X**, AGA, AGG), serine (UC**X**, AGU, AGC) and leucine (CU**X**, UUA, UUG).

(here **X** indicates any nulceobase)

The fact that several codons can have the same meaning is often called *degeneracy of the genetic code*. The degeneracy is an important feature of the genetic code, and it has several important functions in the biochemical systems. We will not explain them in details here, but to briefly list a few: 1) it helps to avoid misinterpretation of codons through wobbling, 2) ensures that some synonymous mutations will not have detrimental effects, 3) helps to regulate the folding of messenger RNA, form riboswitches and other structures in mRNA, 4) helps to regulate codon-dependent speed of translation and associated co-translational folding process.

For example, even when codons are coding for the same amino acids, they are usually translated with a different speed. By placing slow or fast codon in the same sequence, an organism can regulate the speed at which the protein is formed. This is important to keep in mind, because a protein needs some time to fold into a correct 3D structure. Some difficult protein sequences which need time to fold, assemble or brought to a membrane co-translationally, can benefit from slowing down their biosynthesis speed at some points of the sequence. This can be quite nicely regulated with exchange of the codons, without the change of the actual protein sequence.

Genetic code is universal (on Earth, at least)

Finally, one critically important feature of the genetic code is its *universality*. The universality of the genetic code says that the same codon will have same amino acid meaning in any organism on Earth. For example, the gene quoted above can be taken across from human to another organism, a pig, worm, mushroom or a bacterium, and everywhere the amino acid sequence coded by this gene will be same. This gene will lead to myoglobin no matter in which host organism it will be translated. As any biochemical rule, this one has a few exceptions. Some codons can have their meanings re-assigned in specific organisms, which usually reside in unique biological niches. However, these exceptions are relatively rare.

The universality of the genetic code enables for the genetic cross talk between organisms. For example, when we try to eradicate pathogenic bacteria with antibiotics, some of them can develop a special protein that will help them to survive, the resistance protein. This protein will be synthetized according its gene, the resistance gene. This resistance gene can be shared between other bacteria, and there this gene will be translated to the same protein, hereby the other bacteria will also get the resistance. According to the genetic cross talk, some biochemical innovations can be easily shared between species. All because the amino acid meaning of a gene will be same in all organisms: CCA will always mean proline, AAU will always mean asparagine, UAC will always be translated to tyrosine, and so on.

Where are proteins coming from?

This has been a puzzling issue for quite some time, how the central dogma with the universal genetic code all got established back in the days. Why is this question so important? Well, it's because we would like to track the basic life principles back to the moment when they just appeared and in this way try to realize whether they appeared through a chance or a necessity. In the modern organisms we cannot avoid these principles, of course, but was it always like this? We would like to address this question in the subsequent sections. Here we will summarize our current view on the topic. We would like to note that the tracking of the history of biochemistry is a tremendously complex task, and many relevant concepts have been expressed by a number of excellent scientists, and we would like to thank to all for them for their efforts.

Our summary will start with the reference to the central dogma. Yes, we cannot avoid this basic principle in our understanding of life, including understanding of the life history. The central dogma suggests that RNA will come to protein but a protein will not come to RNA. Therefore, we think that the RNA should come first, and the protein biosynthesis appeared from a place, where RNA were doing all biochemical jobs. For example, they were accompanying all the catalytic processes (the RNA molecules doing catalysis are called ribozymes) in the early primitive stage of life development. This phase is called the *RNA World*, thus the biochemical world where RNA were doing all the jobs. The polypeptide synthesis was not an intended process, but started as an accident. In the next step, the newly synthetized peptide sequences turned out to be beneficial for the RNA functions, and the peptides started to accumulate, while the

repertoire of the amino acids was expanding on this stage. This scheme of development from RNA doing biochemical jobs to proteins doing biochemical jobs, is what is called the *RNA World hypothesis*.

Let's try to think what was the driving force of this whole process of transformation from one world to another. This is a principle of chemical diversity and versatility. Chemical composition of biochemical systems makes them versatile and able to fulfil a number of tasks in the cells and outside of them. There is many chemical functions needed for survival of organisms, including some trivial, like positive or negative charges, to very specific, such as binding of specific metals, utilization of special elements (e.g. selenium, iodine), fancy functional groups, cofactors etc. All these chemical diversity of modern organisms was accumulating through acquisition of what we call chemical innovations. Every new chemical function is a chemical innovation, because it allows an organism to do new jobs, or perhaps, do same jobs as before, but more efficiently. In the RNA World phase the major chemical processes were done by RNA-based structures, but then as proteins started to accumulate they started to fulfil biochemical tasks, and their repertoire of functional was expanding, from a few amino acids at the begin, to the set of 20 amino acids as we know them. Then the protein biosynthesis machinery stopped to acquire new chemical elements. This does not mean however, that proteins stopped to diversify their chemical compositions, just the opposite. From this stage, the chemical modifications in proteins were introduced not through incorporation of novel amino acids to protein sequences, but through the modification of the ready proteins, so-called post-translational modifications. Post-translational modification machinery was already based on protein chemistry, in this way, protein started to diversify themselves. There is a huge number of modifications in nature, there is literally hundreds of them, therefore, on this stage incorporation of novel amino acids through their addition to the set of 20 coded amino acids was not any more necessary. Just the opposite, it was beneficial to keep the existent repertoire as a reliable basis set of chemical functions.

In the world made by proteins, their identities were imprinted in the mRNA sequence, their synthesis was done according to mRNA templates by the ribosomal RNA molecules. However, the biochemical jobs, such as making catalysis were gradually taken over by proteins, until proteins toll almost all of them. As the result we have the biochemical world, where absolute majority of the cellular tasks are done by proteins, and this is called the *Protein World*. Thereby, by following the principle of chemical diversification, we draw the concept of the *RNA to Protein World transition* directly from the central dogma.

Why are proteins better at doing catalysis than RNAs? The answer this question, let's use one analogy. Try to pick a sheet of paper using fists only. Now try to do it using your fingers. What is simpler? Probably, fingers, right? Smaller instruments are better at doing more fine precise and efficient jobs. Protein backbone and side-chains are much smaller compared to those in the structure of RNA. Therefore, proteins are doing the biochemical jobs much better than the RNAs. Like fingers and fists. It should not be surprising that fine molecular tools took over the functions from the cumbersome RNA molecules.

Why should there be amino acids in the RNA World?

Lets' imagine, we are in the RNA World. The protein biosynthesis did not start yet, and RNA molecules are doing all the jobs. Where would then amino acids come from and in which form? Even if we assume that the protein biosynthesis started by an accident, we should assume that the precursors were readily available. In the translational machinery, the amino acids are coming as attached at the 3'-end of a transfer RNA by an ester bond. With this attachment, the amino group of an amino acid remains freely

accessible, while the carboxylic group is attached to the tRNA. This will be an important note for our considerations.

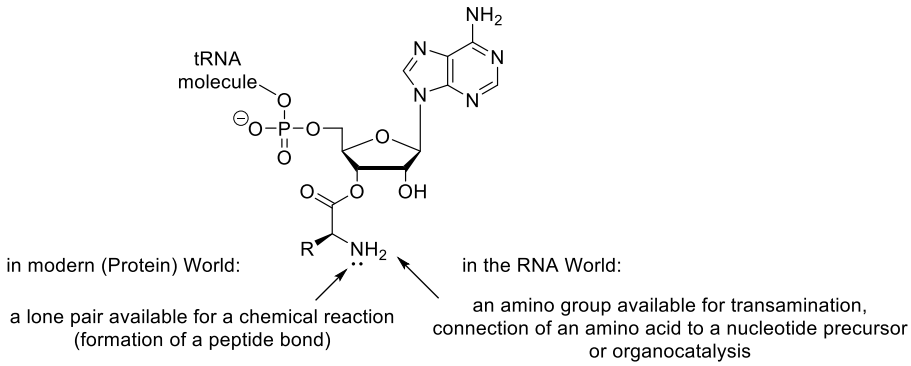

Figure 2. Attachment of an amino acid to an RNA adaptor molecule (tRNA) is made through an ester bond.

Now let's think: why the amino acids should be available in the RNA World? We will need just a few simple assumptions to answer this question. First, we should assume that there was a biosynthetic production of the RNA molecules, in order to live the RNA should reproduce. For this reason, there should be a biosynthetic way to produce the basic elements such as sugars (at least, ribose) and nucleobases (purines and pyrimidines). Second assumption, we will assume that there were already some basic biochemical pathways in this phase of life. Most important are citric acid cycle and C1-metabolism, which are certainly the most central basic and primitive. As the result, we should have nucleobase, sugar synthesis, citric acid and C1 metabolic cycles.

How do we know, that these basic metabolic pathways were going in the same or at least similar manner, as we have them now? This is an important and not trivial question. In the RNA World hypothesis, we assume that the catalytic transformations between molecules were at fist done by RNA molecules, but then these were taken over by more efficient protein catalysts, enzymes. If some biochemical pathway (for example, purine biosynthesis) was done by one type of molecules, and then at the next stage, it was done by other, the process should not necessarily be same, tight? When John makes a task, and then we fire John, and hire Steve to this job, will Steve do the task in exact same way, as it was done by John before? Not necessarily. So, how can we project the existent biochemical pathways into the RNA World phase?

To make this projection, we will need another simple, but quite logical assumption. Let's assume that the process, in which the biosynthetic pathways were taken over from RNA to protein, were proceeding not in a single discrete event but gradually. Remember, that the processes we are talking about are the core biochemical processes, there cannot be any interruption in them. We cannot just fire John, the advertise the open position, collect applications from those who are interested in the job, and then make a decision. This is how employers act on a job market. But when the job cannot be interrupted by a single minute, we need to make this substitution differently. While John is still working, we already take some other people for probation period, and as soon as we see that one of them actually performs better than John, we will offer them the job. We need to make absolutely sure, the new employee is better at this very job and ready to take it over immediately, since the processes are at the core of the biochemistry, and they are essential for survival. Therefore, we can assume that while RNA molecules were still doing the catalysis, more efficient protein catalysts started to make same processes, just more efficiently, and eventually, they prevailed.

In this way, we will assume that main scheme of the core biochemical pathways were similar in the RNA World, and we can actually project them from the existent pathways. If we agree in this, let's proceed. If we have a brief look at the de novo synthesis of purines, we immediately find amino acids! There are at least few of them, although, they have different functions. To build up one purine core we absolutely require one glycine (which is going to make part of the heterocycle), two glutamines and one aspartate as the sourced of the amino groups, another aspartate will be needed to convert inosine into an adenosine, otherwise inosine will be transformed into a xanthosine, and the latter will be transformed into a guanosine with the help of a glutamine. We will also find amino acids in the biosynthesis of pyrimidines. One aspartate molecule will be at the core of the newly formed heterocycle, and some amino groups will be taken across from glutamine.

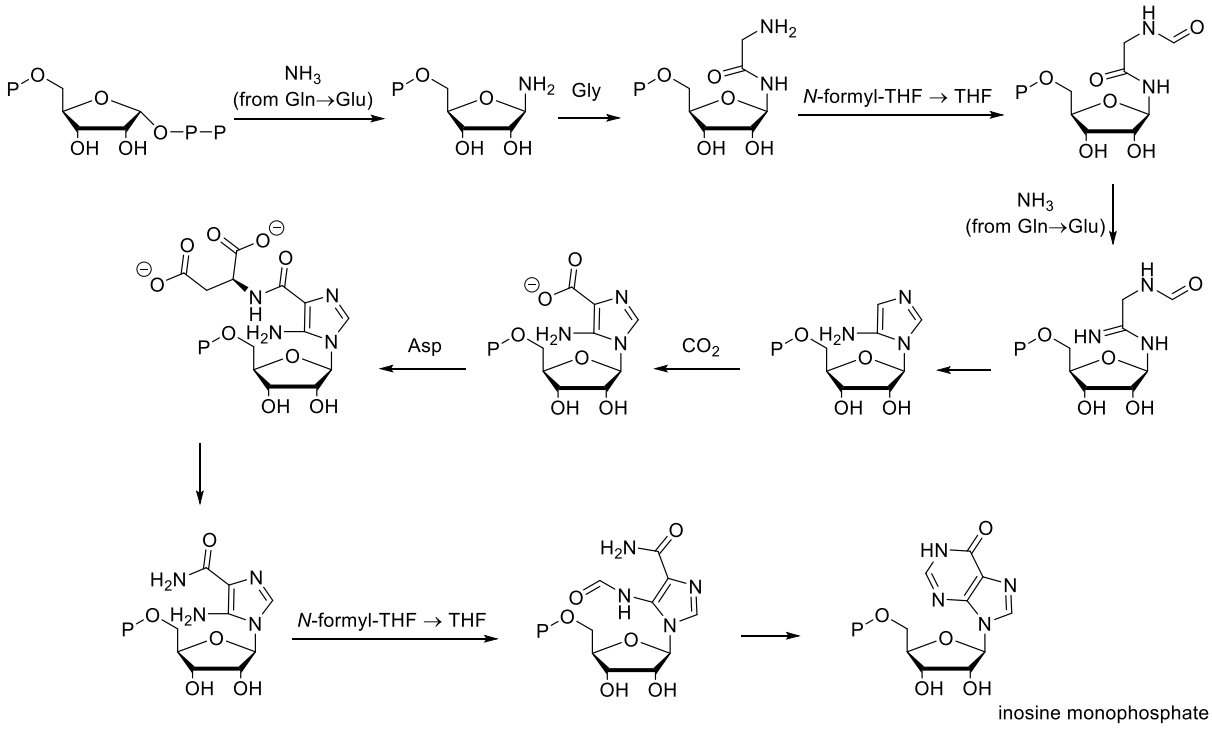

Figure 3. De novo purine biosynthesis in modern biochemistry.

What do these pathways show us? There are three important things:

1. In order to construct a nucleobase, amino acids are absolutely essential;
2. While some of them (glutamine, aspartate) are acting as simple donors of amino groups, others (glycine, aspartate) are making the core of the nucleobases;
3. The biosynthesis involves sequences of steps, where amino acid components come alongside with other important cofactors, mainly, NAD and THF based.

The amino acids aren't coming from the air, they should be produced too. Glycine is usually made out of serine in a process involving THF, which is a part of the C1 metabolism. Aspartate is derived from the citric acid cycle in one simple step (transamination of oxalacetate), glutamine in made out of glutamate, which is again coming in one step from the citric acid cycle (transamination of α-ketoglutarate). Let's add one important note. What we know now as the amino groups donors (glutamine, aspartate, glutamate, etc.)

are actually swappable units. Every amino acid, which can be transaminated into an α-ketoacid, can donor an amino group. The simplest α-ketoacid is pyruvic acid, and corresponding aminated counterpart is known as alanine. Therefore, in the primitive world, the simplest amino-group carrier could be alanine. This conclusion can be supported by the fact that pyruvate is one of the most central metabolites now and (most probably) then.

Okay, we already see that some amino acids are required for essential nucleobase producing processes. This is not surprising since all nucleobsases contain nitrogen atoms, same as amino acids. Where are the sugars coming from though? This question is really important, because we need to have sugars with a correct structure and stereochemistry to replicate RNA. The stereochemically correct synthesis of sugars will necessarily require an asymmetric catalysis. To make a projection on how the sugars were made in the RNA World phase we will rely on the fact that amino acids can actually catalyze the processes in which sugar-like molecules are formed (aldol condensation and other condensation type reaction). There is one amino acid, which is exceptionally good at this, it is proline. When proline has a free amino-group, this molecule can perform condensation chemistry in a quite efficient and stereoselective manner. In organic chemistry, proline is one of the classical organocatalysts. And if we can do it in organic chemistry, the RNA World could also take advantage of this structure. How is proline made metabolically? Either from glutamate or from ornithine. Ornithine is a precursor of arginine.

In this way we come up with a set of amino acids, which we expect to be there in the RNA World.

Domestication of amino acids

We hope, you're now convinced that the amino acids are very important molecules, and these couldn't be avoided in the RNA world phase. The main function of amino acids is 1) to be core building blocks for de novo nucleobase synthesis, 2) to donate an amino group when needed, 3) to perform organocatalysis with the free amino group.

However, now we run into a big trouble. Remember our size issue from above, when we were comparing fingers and fists? Here it comes again. If we think that the RNA molecules were utilizing amino acids in their life, how could such large molecules (ribozymes) operate with such small molecules (amino acids)? This question is not difficult to answer. You only need to link (anchor) the small molecule substrate to a larger molecule, which we could call an adaptor RNA molecule. When an amino acid is linked to a corresponding adaptor, the ribozymes should not any more operate with amino acids, they should operate with adaptors. Plain and simple. This is how ribosomal synthesis of proteins work too. Instead of operating with amino acids directly, the ribosome operates with tRNA (adaptor) molecules with the amino acids attached to them at the 3'-end (as esters).

In the RNA World, though, the amino acids were involved in the metabolism. Here, we will use same principle. In order to utilize an amino acid in a metabolic production of a nucleoside, we attach it to a corresponding adaptor RNA molecule, and now operate with the adaptors. In this way, the RNA World domesticates the amino acids to make use of them. Fingers become attached to the fists.

Now, imagine that we need to construct a metabolic cycle, we will need to attach the amino acids to corresponding adaptor molecules, and then put them in some sort of a sequence. For example, we will start to assemble a purine nucleoside with a sugar-building cluster (prolines), then comes an anono-group donor (for example, alanine), then glycine, then, *N*-formyl-THF, then another ramino group donor, and so

on. The best way is if we place these adaptors on some sort of a template matrix, where the template sequence will correspond to a particular metabolite production, or even complete RNA stretch synthesis.

Let's note that in this phase, the amino acids were not meant to react with one another, they should simply stay attached to the adaptors that are based on a messenger matrix. Other than amino acids, the sequence of the attached molecules should involve THF, NAD, FAD and other notable cofactors. Interestingly, all these cofactors are derivatives of either nucleotides or short RNAs. For instance, both, NAD and FAD are dinucleotides, and this shows that perhaps, earlier, they were simply attached to a particular oligonucleotide adaptor, which was later simplified in the course of evolution.

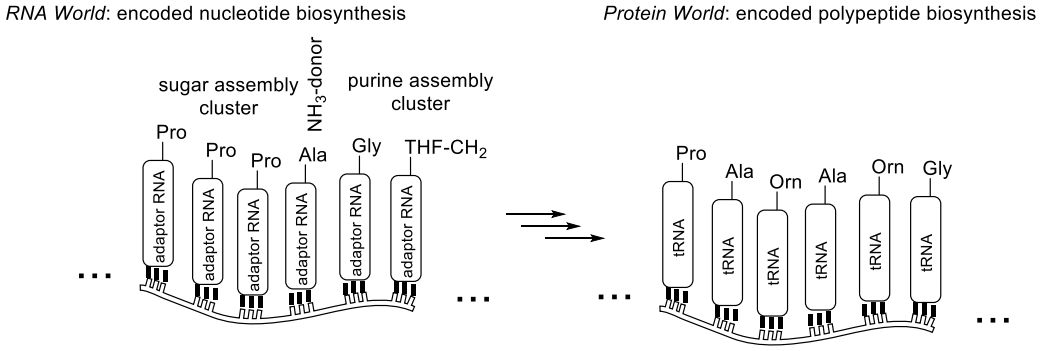

Figure 4. The sequential arrangement of the small molecules on a messenger RNA matrix: then and now.

Now we have one step to the polypeptide synthesis. Once we have placed the amino esters on a matrix in some proximity to each other, it will be the matter of time, when the chemistry between them will happen. This chemical step is called peptide bond formation. At first, it was most probably a side reaction, or in other words, the peptides were undesired side products. However, we can assume that the peptide stretches were accumulated around the RNA. At some point, they RNA functions started to benefit from their presence, and then there were accumulated more and more, and in this way, an evolution happened towards the Protein World. Eventually, the matrix RNA stretches, which were previously coding for metabolism, were coding for a polypeptide synthesis, have become messenger RNA.

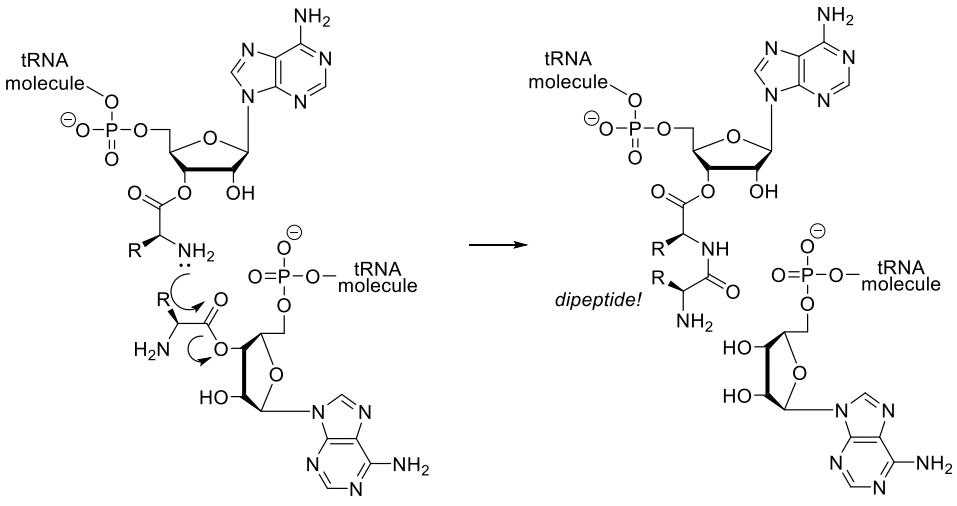

Figure 5. A peptide bond formation between two RNA amino esters.

What was the first amino acid set?

In the modern world we have the set of 20 amino acids, which all have different functions and properties in proteins. How was this set formed? We will again lay down an assumption. Let's assume that the set of 20 did not come at once, but it was created gradually, most probably through expansion. At first, just few amino acids constituted primitive polypeptides, then a few more were added, then a few more, and so on, until we have them 20. Later, the chemical diversification of proteins was taken over by post-translational modification machinery.

There are few lines of evidences, which suggested that the primitive RNA code for the first polypeptides was based on just two letters, G and C. When we now look at the genetic code, with these two we can code four amino acids: glycine (GG**X**), alanine (GC**X**), proline (CC**X**) and arginine (CG**X**). We already discussed what was the possible function of the first three in this set, glycine (purine biosynthesis), alanine (amino group carrier) and proline (sugar assembly), but what is the role of arginine? We should admit that arginine is an absolutely key amino acid in the GC-coded set, because it was making the first polypeptides cationic. Without a cationic residue, the polypeptides will simply swim away, and they won't adhere to the polyanionic RNA molecules. Therefore, if the first polypeptide sequences were not cationic, they could not create a feedback for evolution and this makes them irrelevant. Even if there were neutral or anionic peptide stretches formed in the RNA World phase, we can now nothing about them. Conversely, cationic peptides could be evolved. Therefore, there should be at least one cationic amino acid in the first set. Now, this amino acid is arginine, but if we simplify the side-chain group to a simple ammonium group, resulting amino acid, ornithine, is a known precursor for proline. If proline was utilized in the RNA World phase, then ornithine was utilized for the synthesis of proline.

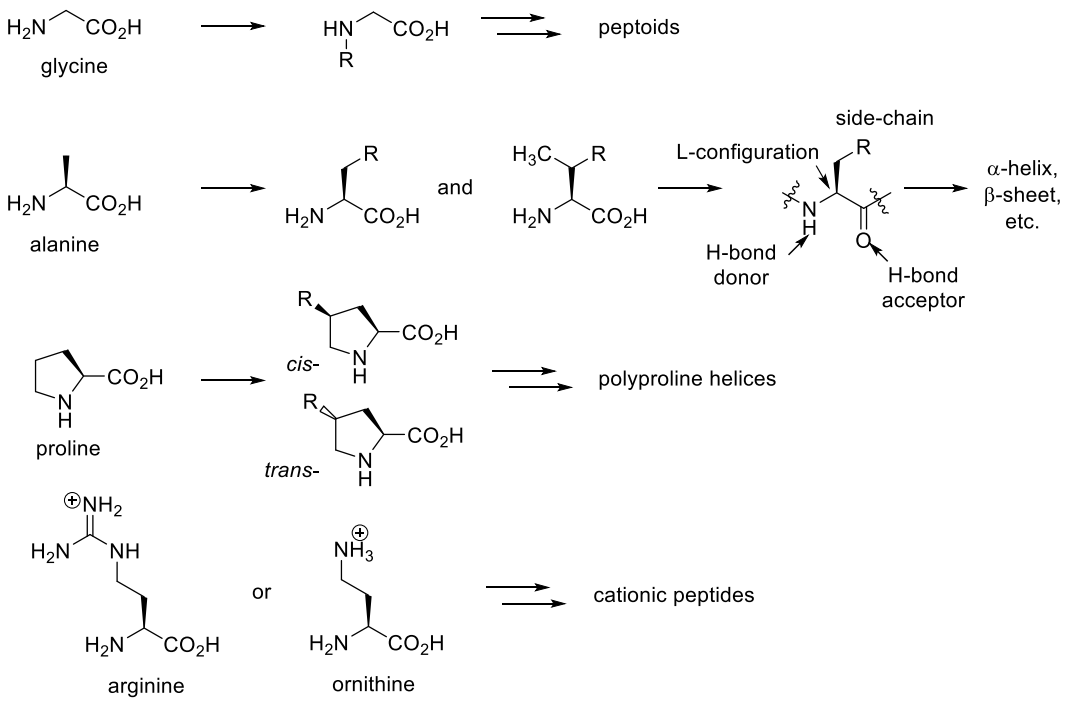

Figure 6. Analysis of the first set of the amino acid from the GC-code phase.

The protein structure

Let's pose one interesting question. If you have the GC-coded peptides, what could you do out of them? If you'd show this set of amino acids to a peptide chemists, they will say that this set is an absolute horror. Indeed, two amino acids from this set (glycine and proline) are known as the structure breakers, and homooligomeric sequences for each of the four mainly create a fancy set of structures known as extended structures. There could be perhaps, only one good use for the primitive peptides. Given that they're positively charged, they could bind to the surfaces of large RNA bodies, and in this way help two polyanionic bodies interact, form an assembly, or something alike. At the same time, these peptides, will most probably not form a set of structures capably of independent functioning. To form defined structures the first set had to be expanded.

The expansion happened with the addition of letter A to the messenger code. This is known as the GCA code or GCA stage of the genetic code developments. In the GCA code, there is a whole new set of amino acids. Let's look at them closer. These will be: aspartic and glutamic acids, asparagine and glutamine, serine, threonine, lysine, histidine. These amino acids are relatively easy to produce metabolically. Remember, that we had already aspartic acid in the pyrimidine de novo synthesis, and serine as the precursor for glycine. Other amino acids should also be easily accessible, as they are just few steps away from the core metabolism. For example, histidine is made out of ATP by reconstruction of the heterocydle in just a few steps.

There is one important note with regards to the GCA phase. The newly added amino acids in this phase are all polar amino acids, and they are all structural derivatives of alanine, and share same basic architecture as alanine. They share L-stereochemistry (not like in glycine, which is non-chiral), have a side chain on a one- or more-carbon linker attached to the backbone carbon, amino group remains unmodified (not like proline with a secondary amino group). These set of structural features enabled one important structural feature, the α-helix. Ever since this phase, the world has become the world of the α-helix. β-sheets are enabled by this set too. These structures are known as the secondary structure, as they represent the second step in the hierarchy of the protein folding.

If the amino acid set would develop in different direction, the secondary structure phase would also look differently. For example, if the side-chain functions would be based on the proline structure, we would have so-called polyproline helix dominating the world. Otherwise, if the side-chain would not be placed on the backbone carbon, but on backbone nitrogen, there will be structures called peptoids with special and interesting structural features. There won't be a world dominated by the α-helices and β-sheets, that's for sure. Neither oligoprolines nor peptoids are able to sustain these structures. The reason why the α-helix have become so dominant is the fact that at this decisive step, at the GCA code development phase, the structures recruited to the polypeptide synthesis were based on alanine. Alanine basic architecture is ideal for the α-helix, as this amino acids has a max propensity α-helical propensity according to quantification scales.

Having the secondary structure set, in particular, the α-helix, in not enough to build proteins. There should be a tertiary fold two, and for this reason, we need what is called *motifs*. Motifs rely on a particular patter of side chains on a secondary structure carries, which makes defined interacting interfaces. For example, there are electrostatic interactions. One could expect that the primitive helices could start to interact with each other through the charge moieties, for instance, glutamate on one side (negatively charged) and arginine – on another (positively charged). This scenario would be problematic though. There is

unfortunately lots of charged and polycharged species in the RNA World, most typically, polyanionic RNA bodies. The RNAs would interact with positively charged counterparts quite well, and in this way they would disturb the intraprotein interactions based on charges.

There was a need to make an entirely new motif, which could allow proteins to assemble in an independent manned, and this motif was found. It is called hydrophobic motif. For this reason, another new letter was added to the code, the U-letter. In the GCAU phase, the code acquired additional amino acids, leucine, isoleucine, valine, phenylalanine, tyrosine, tryptophan, methionine. All hydrophobic amino acids were recruited on this stage of development. Their recruitment made it possible to make use hydrophobic motif for assembling principle for the tertiary fold of proteins. Moreover, this was the fist point in the history of molecular evolution, when the hydrophobic motif was recruited by natural biopolymers. Notably, in contract to GCA, the amino acids from the GCAU phase are relatively difficult to synthetize, as they requite quite complex and sophisticated metabolism, which was not available at the earlier stages.

At this phase, the protein fold hierarchy was completed, and the Protein World was settled. One important amino acid, cysteine, was also added to the repertoire on this stage. Cysteine is now known for forming disulfide bridges that stabilize tertiary and quaternary fold through covalent bonds. In the GCAU phase, though, when the conditions were reducing, the cross-linking through cysteine were probably involving metal chelation. Therefore, a bridge would probably be not S-S bridge, but a S-M-S bridge, where M is a metal ion.

One critically important innovation, which was enabled at the GCAU phase: the addition of the hydrophobic amino acid set allowed for the interaction with the membrane and formation of membrane proteins. Without the interaction with membranes, there would be no cells. From this point, proteins started to mediate relationships between RNAs (negatively charged) and lipids (negatively charged) that are not able to interact directly.

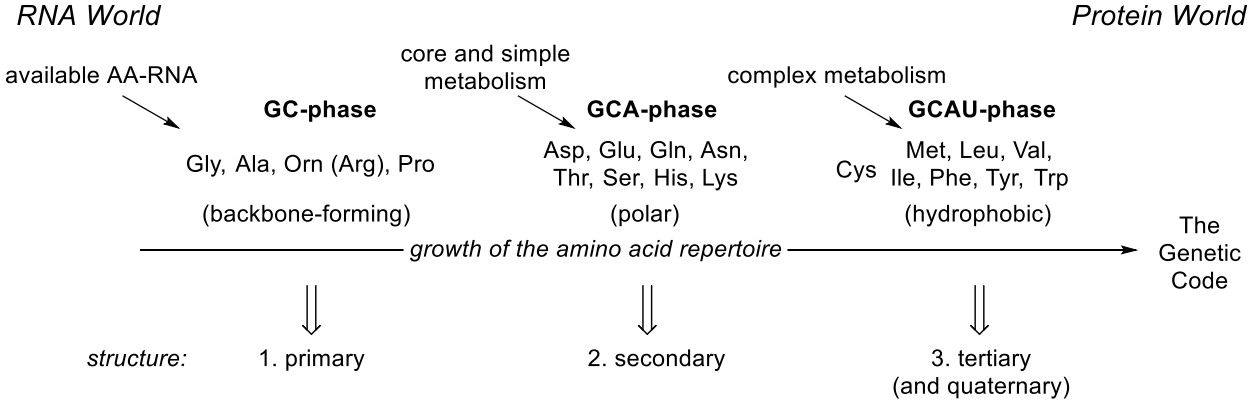

Figure 7. The GC-GCA-GCAU scheme of the genetic code development reflects the hierarchy of protein folding and complexity of the amino acid biosynthesis.

In this way, the hierarchy of the protein folding directly reflects the development scheme of the genetic code. At first (GC phase) there is just the primary structure (sequence) and the fact that the structure is coded, thereby the mRNA represent information about it. In the second (GCA phase), the amino acids are

recruited, which form the set of the dominant secondary structures. In the third (GCAU phase) the hydrophobic motifs are added, and onbuilt on the existent set of the secondary structures. In this way, the tertiary fold (globular proteins) and transmembrane proteins were formed. The world came to the Protein World phase, which dominates the biochemistry until now.

Can we change the genetic code?

In the narrative above, we were trying to map the history of the genetic code development as follows from the central dogma. Nonetheless, we should put an important disclaimer here. We do not know how the evolution life was exactly happening, and whether it was indeed happening the way we proposed. This does not mean that the history development mapped above is wrong though. It simply means that the history of life is a complex topic, and tracking of the exact sequence of events now, 4 billion years later, is most probably impossible.

The map above is proposed to show how the RNA to Protein World transformation is suggested from the central dogma, and how can we detalize this transit by using a few simple assumptions. As the result, we have a mind game, which helps us understand basic principles of life, and things like the structure of the genetic code. When we look at the genetic code now, we find that all typical hydrophobic amino acids (phenylalanine, leucine, isoleucine, valine and methionine) have U at the codon central letter. Wow! This must mean something, we say. This means that the nature ensures that even if there is a misinterpretation of a codon, a hydrophobic is replaced by a hydrophobic, so the structure does not suffer.

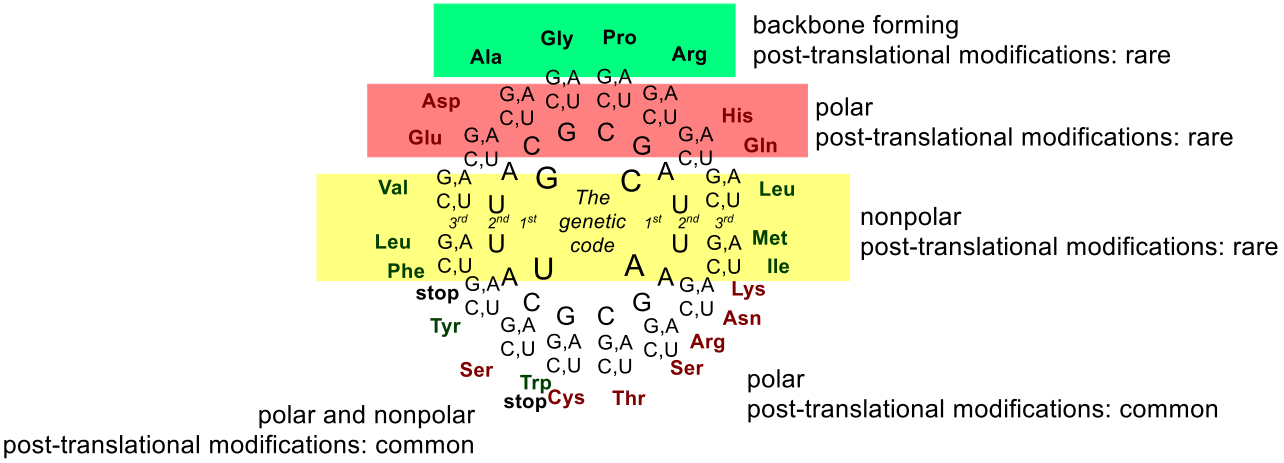

Figure 8. Some patterns behind the genetic code structure.

This is how we look at it now, and this is how we argue that the codon choice is clever and helps to maintain some existent functions. However, is there is a biochemical benefit from a certain biochemical relationship, this benefit was not necessarily the reason why the relationship was established in the first place. In the above, we showed that the U letter was added to the coding space at the very late stages of the genetic code development, and recruitment of hydrophobic motif in the formed phases was not possible due to biochemical unavailability and complexity of hydrophobic amino acids. This was the actual reason. We cannot assume that there was a clever engineer, who designed the genetic code in a way to

avoid detrimental mutations and help maintain its efficiency. We do not need such an assumption for our narrative. We simply go from basic principles and with a few logical steps, get the existent code as it is.

In fact, the genetic code is in many ways imperfect. For example, there is wobbling, and both the tRNA adaptors and the messenger RNA contain multiple post-transcriptional modifications, with the aim to rescue low fidelity and speeds of the translation process, which are otherwise quite poor. However, since the existent genetic code was around for 4 billion years, there are so many molecular mechanisms related to it, that changing some genetic code components becomes a tremendously difficult task. Imagine, replacing a cornerstone in a basement of a building. When the basement is still not complete and the construction is still in progress, we can easily do it. However, when the house is complete, changing one cornerstone can cause a complete collapse of the construction. We say: this cornerstone is very important, do not replace it. This does not mean, however, that this cornerstone is not replicable in principle.

The same with the genetic code. The map above shows that the amino acid set for the primitive code development was dictated by their metabolic availability. What goes around comes around. This is a very optimistic view, actually. It shows that if we succeed to sufficiently degrade the complexity of the cellular mechanisms responsible for chemical identity, and if we will offer other set of amino acids, we could expect an alternative code development, and other repertoire as the result. For example, modern synthetic chemistry can offer a large set of god amino acid building blocks that are not available metabolically otherwise, for example, with fluorine-containing side-chains. If we do it in a clever way, and ensure that the newly added amino acid will not repeat exact same functions as already existent, but add some new chemistry to the cells, the cells might be interested to acquire and keep them permanently.

Imagine making cells, which recruit different set of amino acid building blocks, some of which are exclusively provided by humans? An addicted cell. An alien cell. A new being with an alien chemistry. An alien in a test tube. This is the eventual goal of our efforts in the field of genetic code engineering.