CHAPTER 8

Theory of Early Molecular Evolution:
Predictions and Confirmations

Edward N. Trifonov*

Abstract

A new theory of early molecular evolution is described, proceeding from original speculations to specific predictions and their confirmations. This classical cycle is then repeated generating the earliest picture of evolving Life. First, a consensus temporal order ("chronology") of appearance of amino acids and their respective codons on evolutionary scene is reconstructed on the basis of 60 different criteria, resulting in the order: G, A, D, V, P, S, E, L, T, R, I, Q, N, K, H, C, F, Y, M, W. It reveals two fundamental features: the amino acids synthesized in experiments imitating primordial conditions appeared first, while the amino acids associated with codon capture events came last. The reconstruction of codon chronology then follows based on the above consensus temporal order, supplemented by the stability and complementarity rules first suggested by M. Eigen and P. Schuster, and on earlier established processivity rule. The derived genealogy of all 64 codons suggests several important predictions that are confirmed: Gradual decay of glycine content in protein evolution; traces of the most ancient 6-residue long gly-rich and ala-rich minigenes in extant sequences; and manifestations of a fundamental binary code of protein sequences.

Introduction

Hot rocks and boiling water—that, presumably, was the "weather" on the planet Earth when 3.9 billion years ago the LIFE started (ref. 1, and references therein). It would not be fair if a skeptical reader had asked: what exactly is life? There are many answers to that question[2] though only one is needed. But it would be equally unfair to claim that the emerging life was as complex and omnipotent as today. It was surely primitive, even, perhaps, trivial, but what was it?

The one who knows what was the most primitive start is Stanley Miller who thought, in 1953, that perhaps in a primordial atmosphere a mere chemistry would take a chance. The imitation experiments[3,4] brought a spectacular result: among many other substances 10 amino acids were synthesized, half of the amino-acid repertoire of modern proteins: alanine, glycine, aspartate, valine, leucine, glutamate, serine, isoleucine, proline, and threonine (A, G, D, V, L, E, S, I, P, T). The earliest attempt of this kind, with the same thought, was the work of Löb[5] in 1913 (see also ref. 6). Analytical chemistry of that time was able to detect only one amino acid in the mixture—glycine.

Those 10 amino acids were not life yet, but a good chemical beginning, on the long way from primitive to simple, and from simple to complex. There are many dramatic stations in this journey: formation of first very small proteins, formation of the membranes and cells,

*Edward N. Trifonov—Genome Diversity Center, Institute of Evolution, University of Haifa. Mount Carmel, Haifa 31905, Israel. Email: trifonov@research.haifa.ac.il.

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber. ©2006 Landes Bioscience and Springer Science+Business Media.

development of replicating molecules and systems, emergence of nucleic acids, invention of the triplet code, formation of the last common ancestor, first bifurcations of the tree of Life. Each step is a mystery, and it is not clear at all what was the sequence of the events.

First Move Towards a New Theory

Let us make a jump straight to the origin of the genetic code, a pretty early stage anyway, not far from the very beginning. Within last few years my colleagues and I were lucky to have asked several very pointed, turned-to-be-right questions, and find tantalizing answers. In particular, given the earliest small proteins and nucleic acids (perhaps, RNA)—what were the very first RNA triplets (codons), and what were the amino acids they encoded? There are many speculations on that matter, including our own attempt.[7] This work, however, was not just yet another one of the speculative kind. It had an element of reconstruction of early biomolecular history, based on specific prediction that was confirmed. Such reconstruction was later expanded and turned, actually, in a vibrant theory of early molecular evolution, that suggested new predictions, followed by confirmations. The development of the theory is described in the following sections in all its logical and some technical details.

The first clue was thrown in by Thomas Bettecken, who in 1996 overviewed a group of so-called triplet expansion diseases. These neurodegenerative diseases are associated with repeating sequences located around certain genes. The repeats are of the type CUGCUGCUG or $(CUG)_n$ where n is the number of the repeats, normally 20 - 50. The repeat number all of a sudden changes to a much larger one, of the order of few hundreds, and that results in a disease. What Thomas noticed is that the most of the observed seemingly different expansions, as documented in literature, actually, correspond to the same structure. E.g., repeats $(CAG)_n$, $(GCU)_n$ and the above $(CUG)_n$, obviously, correspond to the same repeating duplex $(GCU)_n{\cdot}(AGC)_n$. As it turned out, the repeats $(GCU)_n$ and $(GCC)_n$ make a majority of all known triplet expansions. In other words, these two triplets are most expandable, whatever the reason. This is also confirmed in prokaryotic system.[8] This observation per se is not yet enlightening. To make the bell ring one needs another important piece of information of which we were in possession already in 1987,[9] being unaware of its explosive value. This is the $(GNN)_n$ periodicity hidden in all modern protein coding sequences. Later analysis allowed to refine the pattern to $(GCU)_n$.[10] Thus, the hidden mRNA consensus would be $(GCU)_n$, - probably, reflection of an ancient mRNA pattern $(GCU)_n$ (and, perhaps, $(GCC)_n$ as well?).

A thrilling thought then burst in: the $(GCU)_n$ and/or $(GCC)_n$, readily expanding sequences, could be indeed the first coding sequences that later evolved to the modern sequences, where the original pattern is almost lost. An obvious advantage of these sequences at that time was their exceptional ability to expand, i.e., to become longer. The relentless $(GCU)_n$ and $(GCC)_n$ repeats are still in labor—in the modern diseases.

If, thus, the GCU and GCC were the very first codons, they could only code for two amino acids. Several more amino acids should have been accommodated, probably, by single point mutations of the generic GCU and GCC triplets which gives total 15 different triplets (codons).

But which amino acids came first? One only could speculate about it, and we picked up three most natural speculations: (1) The very first amino acids were chemically simplest. (2) They would be expected to appear in the Miller's imitation mix; and (3) They would be likely to have been served by more ancient of known two classes of aminoacyl-tRNA synthetases.[11] By a consensus of these three criteria the amino acids alanine, aspartic acid, glycine, proline, serine and threonine (A, D, G, P, S, T) should have been the very first amino acids, to be served by those first 15 triplets above. Remarkably, 13 of the triplets do, indeed, encode *today* the speculated six earliest amino acids. Correspondence of the 13 predicted earliest codons to 6 predicted earliest amino acids confirms both speculations, and may be considered as a first very promising step in the possible full reconstruction of the origin and evolution of the triplet code. Encouragingly, the match between these two sets is in full agreement with present-day code. That is, being set up once, the code probably never changed.