The Codes of Life

BIOSEMIOTICS

VOLUME 1

Aims and Scope of the Series

Combining research approaches from biology, philosophy and linguistics, the emerging field of biosemiotics proposes that animals, plants and single cells all engage in semiosis – the conversion of physical signals into conventional signs. This has important implications and applications for issues ranging from natural selection to animal behaviour and human psychology, leaving biosemiotics at the cutting edge of the research on the fundamentals of life.

The Springer book series *Biosemiotics* draws together contributions from leading players in international biosemiotics, producing an unparalleled series that will appeal to all those interested in the origins and evolution of life, including molecular and evolutionary biologists, ecologists, anthropologists, psychologists, philosophers and historians of science, linguists, semioticians and researchers in artificial life, information theory and communication technology.

The Codes of Life
The Rules of Macroevolution

Edited by

Marcello Barbieri
University of Ferrara
Italy

Springer

Marcello Barbieri
University of Ferrara
Italy

ISBN 978-1-4020-6339-8 e-ISBN 978-1-4020-6340-4

Library of Congress Control Number: 2007939395

Front Cover: The illustrations represent cell samples that have been utilized in signal transduction studies of the kind described in Chapter 9 by Nadir Maraldi. The micrographs show a hippocampal neuron labelled with biocytin (the yellow signal indicates synaptic contacts) at the left, and mouse myoblasts stained for emerin and f-actin, at the right. (By kind permission of the authors of the micrographs, S. Santi and G. Lattanzi, who obtained them in the laboratory directed by Nadir Maraldi).

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Editorial

The New Frontier of Biology

Marcello Barbieri

Codes and conventions are the basis of all cultural phenomena and from time immemorial have divided the world of culture from the world of nature. The rules of grammar, the laws of government, the precepts of religion, the value of money, the cooking recipes, the fairy tales, and the rules of chess are all human conventions that are profoundly different from the laws of physics and chemistry, and this has led to the conclusion that there is an unbridgeable gap between nature and culture. Nature is *governed* by objective immutable laws, whereas culture is *produced* by the mutable conventions of the human mind.

In this century-old framework, the discovery of the genetic code, in the early 1960s, came as a bolt from the blue, but strangely enough it did not bring down the barrier between nature and culture. On the contrary, a "protective belt" was quickly built around the old divide with an argument that effectively emptied the discovery of the genetic code of all its revolutionary potential. The argument is that the genetic code is fundamentally a *metaphor* because it must be reducible, in principle, to physical quantities. It is a secondary structure, like those computer programs that allow us to write our instructions in English, thus saving us the trouble to write them in binary digits. Ultimately, however, there are only binary digits in the machine language of the computer, and in the same way, it is argued, there are only physical quantities at the most fundamental level of Nature.

This conclusion, known as *physicalism*, is based on one fact and one assumption. The fact is that all spontaneous reactions are completely accounted for by the laws of physics and chemistry. The assumption is that it was spontaneous reactions that gave origin to the first cells on the primitive Earth. According to physicalism, in short, genes and proteins are spontaneous molecules that evolved into the first cells by spontaneous processes.

This, however, is precisely the point that molecular biology has proved wrong. Genes and proteins are *not* produced by spontaneous processes in living systems. They are produced by molecular machines which physically stick their subunits together in the order provided by *external* templates. They are assembled by molecular robots on the basis of outside instructions, and this makes them as different from ordinary molecules as *artificial* objects are from *natural* ones. Indeed, if we agree that objects are natural when their structure is determined from within and artificial when it is determined from without, we can truly say that genes and

proteins are *artificial molecules*, that they are *artifacts made by molecular machines*. This in turn implies that all biological objects are artifacts, i.e. that the whole of life is artifact-making.

Spontaneous genes and spontaneous proteins did appear on the primitive Earth but they did not evolve into the first cells, because spontaneous processes do not have biological specificity. They gave origin to *molecular machines* and it was these machines and their products that evolved into the first cells. The simplest molecular machines we can think of are molecules that can join other molecules together by chemical bonds, and for this reason we may call them *bondmakers*. Some could form bonds between amino acids, some between nucleotides, others between sugars, and so on. Among the various types of bondmakers, some developed the ability to join nucleotides together in the order provided by a *template*. Those bondmakers started *making copies* of nucleic acids, so we can call them *copymakers*. The first Major Transition of the history of life (Maynard Smith and Szathmáry, 1995) is generally described as the origin of genes, but it seems more accurate to say that it was the origin of molecular *copying*, or the origin of *copymakers*, the first molecular machines that started multiplying nucleic acids by making copies of them.

Proteins, on the other hand, cannot be made by copying, and yet the information to make them must come from molecules that can be copied, because only those molecules can be inherited. The information for manufacturing proteins, therefore, had to come from genes, so it was necessary to bring together a carrier of genetic information (a messenger RNA), a peptide bondmaker (a piece of ribosomal RNA), and molecules that could carry both nucleotides and amino acids (the transfer RNAs). The first protein-makers, in short, had to bring together three different types of molecules (messenger, ribosomal, and transfer RNAs), and were therefore much more complex than copymakers. The outstanding feature of the protein-makers, however, was not the number of components. It was the ability to ensure a one-to-one correspondence between genes and proteins, because without it there would be no biological specificity, and without specificity there would be no heredity and no reproduction. Life, as we know, it simply would not exist without a one-to-one correspondence between genes and proteins.

Such a correspondence would be automatically ensured if the bridge between genes and proteins could have been determined by *stereochemistry*, as one of the earliest models suggested, but that is not what happened. The bridge was provided by molecules called *adaptors* (transfer RNAs) that have two recognition sites: one for a group of three nucleotides (a *codon*) and another for an amino acid. The crucial point is that the two recognition sites are physically separated and chemically independent. There is no deterministic link between codons and amino acids, and a one-to-one correspondence between them could only be the result of conventional rules. Only a real code, in short, could guarantee biological specificity, and this means that the evolution of the translation apparatus had to be coupled with the evolution of the genetic code.

Protein synthesis arose, therefore, from the integration of two different processes, and the final machine was a *code-and-template-dependent-peptide-maker*, or, more simply, a *codemaker*. The second Major Transition of the history of life is

generally described as the origin of proteins, but it would be more accurate to say that it was the origin of *codemaking*, or the origin of *codemakers*, the first molecular machines that discovered molecular coding and started populating the Earth with codified proteins.

But what happened afterwards? According to modern biology, the genetic code is the only organic code that exists in the living world, whereas the world of culture has a virtually unlimited number of codes. We know, furthermore, that the genetic code came into being at the origin of life, whereas the cultural codes arrived almost 4 billion years later. This appears to suggest that evolution went on for almost the entire history of life on Earth without producing any other organic code after the first one. According to modern biology, in short, the genetic code was a single extraordinary exception, and if nature has only one exceptional code whereas culture contains an unlimited number of them, the real world of codes is culture and the barrier between the two worlds remains intact.

At a closer inspection, however, we realize that the existence of other organic codes cannot be ruled out, and that we can actually test it. Any organic code is a set of rules of correspondence between two independent worlds, and this requires molecular structures that act like *adaptors*, i.e. that perform two independent recognition processes. The adaptors are required because there is no necessary link between the two worlds, and a set of rules is required in order to guarantee the specificity of the correspondence. The adaptors, in short, are necessary in all organic codes. They are the molecular *fingerprints* of the codes, and their presence in a biological process is a sure sign that the process is based on a code. In splicing and signal transduction, for example, it has been shown that there are true adaptors at work, and that allows us to conclude that those processes are based on *splicing codes* and on *signal transduction codes* (Barbieri, 1998, 2003). In a similar way, the presence of adaptors has suggested the existence of *cytoskeleton codes* and of *compartment codes* (Barbieri, 2003).

Many other organic codes have been discovered with different theoretical and experimental criteria. Among them, are the *Sequence Codes* (Trifonov, 1989, 1996, 1999), the *Adhesive Code* (Redies and Takeichi, 1996; Shapiro and Colman, 1999), the *Sugar Code* (Gabius, 2000; Gabius et al., 2002), and the *Histone Code* (Strahl and Allis, 2000; Jenuwein and Allis, 2001; Turner, 2000, 2002; Gamble and Freedman, 2002; Richards and Elgin, 2002). Even if the definition of code has been somewhat different from case to case, these findings tell us that the living world is teeming with organic codes.

What is most important is the realization that organic codes appeared throughout the history of life and that their appearance marked the origin of great biological innovations. When the splicing codes appeared, for example, the *temporal* separation between transcription and translation could be transformed into a *spatial* separation between nucleus and cytoplasm. That set the stage for the origin of the nucleus, and we can say, therefore, that the splicing codes created the conditions for the origin of the eukaryotes. In a similar way, we can associate the origin of multicellular organisms with membrane codes (*cell adhesion codes*), and the origin of animals with embryonic codes (in particular, with the *codes of pattern*).

We realize in this way that there is a deep link between codes and macroevolution, and that opens up an entirely new scenario. The appearance of new organic codes went on throughout the history of life and was responsible for most of its major transitions, from the origin of protein synthesis, with the genetic code, all the way up to the origin of culture with the codes of language. That is the new frontier of biology.

This book is addressed to students, researchers, and academics who are interested in *all* codes of life, from the genetic code to the codes of language. The aim of the book is to show not only that a *plurality* of organic codes exists in nature, but also that codes exist at all levels of life, from the molecular world to the world of language. They exist in cells, embryos, nervous systems, minds, and cultures, and are fundamental components of those systems. In order to illustrate this plurality of codes at a plurality of levels, the book has been divided into four parts: (1) codes and evolution; (2) the genetic code; (3) protein, lipid, and sugar codes; and (4) neural, mental, and cultural codes.

Part 1 describes the first organic codes that have been recognized at the molecular level in addition to the classical triplet code of protein synthesis, and discusses the evolutionary consequences of that extraordinary fact. They range from the idea that the Major Transitions of the history of life were associated with the origin of new organic codes, to the concept that natural selection and natural conventions are two distinct mechanisms of evolution.

Part 2 is dedicated to the genetic code and describes the results of five distinct lines of investigation. More precisely, three chapters illustrate three different biological models on the origin of the genetic code and two chapters argue for the existence of mathematical features in it. This makes us realize that the study of the genetic code is still a very active field of research and, above all, that we have barely scratched its surface. Contrary to a widespread impression, the genetic code has not been explained by modern biology, and this should convince us that the problem of coding is a very central issue in biology.

Part 3 presents a collection of independent research projects that have revealed the existence of organic codes in the worlds of proteins, lipids, and sugars. Each of these cases has individual characteristics and represents a field of research in its own right. This may give an impression of fragmentation, at first, but in reality it is precisely such a *diversity* that makes us realize that codes are *normal* components of living systems. Diversity is the quintessence of the living world and the presence of codes at all levels of diversity is clearly a sign that coding is a basic tool of life.

Part 4 is dedicated to the codes that are associated with the nervous system, the animal mind, and finally language and culture. Here we have the impression to be in a more familiar territory, but in reality the mystery is possibly even deeper. The origin of language does not look any simpler than the origin of life, and some believe that it is probably the hardest problem of all. The point made here is that there has been a long stream of biological codes before language and that it was those codes that created the conditions for the origin of language. The great obstacle to our understanding of language, in short, is the present paradigm where biological codes are regarded as unlikely accidents or extraordinary exceptions rather than normal components of nature.

Acknowledgements

This is the first collective book on the codes of life and it has become a reality because its authors have accepted to take part in a project that goes beyond their individual fields. Each of them has contributed a vital piece to the new mosaic of life that we now have before us, and I wish I could adequately thank them for joining the enterprise. I am also profoundly grateful to the publishing manager, Catherine Cotton, who has created the Springer Book Series in Biosemiotics with the specific purpose to promote the study of the codes of life and to turn it into a new academic discipline. This book is the beginning of that project and I wish to acknowledge, with thanks, that Catherine has been the driving force behind it.

References

Barbieri M (1998) The organic codes. The basic mechanism of macroevolution. Rivista di Biologia-Biology Forum 91:481–514

Barbieri M (2003) The Organic Codes. An Introduction to Semantic Biology. Cambridge University Press, Cambridge

Gabius H-J (2000) Biological information transfer beyond the genetic code: the sugar code. Naturwissenschaften 87:108–121

Gabius H-J, André S, Kaltner H, Siebert H-C (2002) The sugar code: functional lectinomics. Biochimica et Biophysica Acta 1572:165–177

Gamble MJ, Freedman LP (2002) A coactivator code for transcription. Trends Biochem Sci 27(4):165–167

Jenuwein T, Allis D (2001) Translating the histone code. Science 293:1074–1080

Maynard Smith J, Szathmáry E (1995) The Major Transitions in Evolution. Oxford University Press, Oxford

Readies C, Takeichi M (1996) Cadherine in the developing central nervous system: an adhesive code for segmental and functional subdivisions. Develop Biol 180:413–423

Richards EJ, Elgin SCR (2002) Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. Cell 108:489–500

Shapiro L, Colman DR (1999) The diversity of cadherins and implications for a synaptic adhesive code in the CNS. Neuron 23:427–430

Strahl BD, Allis D (2000) The language of covalent histone modifications. Nature 403:41–45

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432

Trifonov EN (1996) Interfering contexts of regulatory sequence elements. Cabios 12:423–429

Trifonov EN (1999) Elucidating sequence codes: three codes for evolution. Annal NY Acad Sci 870:330–338

Turner BM (2000) Histone acetylation and an epigenetic code. BioEssay 22:836–845

Turner BM (2002) Cellular memory and the histone code. Cell 111:285–291

Contents

Part 1
Codes and Evolution

Chapter 1

Codes of Biosequences

Edward N. Trifonov

Abstract Contrary to common belief that the nucleotide sequences only encode proteins, there are numerous additional codes, each of a different nature. The codes, at DNA, RNA, and protein sequence levels, are superposed, i.e. the same nucleotide in a given sequence may be simultaneously involved in several different encoded functions, at different levels. Such coexistence is possible due to degeneracy of the messages present in the sequence. Protein sequences are degenerate as well: involved not only in the functions related to the protein, but also adjusting to sequence requirements at the DNA level.

1 Introduction

All manifestations of life, from elementary biomolecular interactions to human behavior, are tightly associated with, if not in full command of, sequence-specific interactions. Nucleic acid or protein sequence patterns involved in the molecular or higher-level functions stand for the sequence codes of the functions. The genome that carries or encodes all these sequence patterns is, thus, a compact, intricately organized, informational depot. To single out all major sequence codes and trace them in action may be viewed as the major challenge of modern molecular biology, sequence biology.

The nucleotide sequences, thus, not only encode proteins, as an inexperienced reader would think. Various sequence instructions are read from the DNA, RNA, or protein molecule each in its own way, via one or another specific molecular interaction or a whole network of interactions. In the triplet code the reading device is the ribosome. In gene splicing the sequence signals are recognized by the spliceosome. There are also numerous relatively simple sequence-specific DNA–protein and RNA–protein interactions, where the respective sequences are read by a single protein.

After the triplet code was spectacularly cracked (Ochoa et al. 1963; Khorana et al., 1966; Nirenberg et al., 1966), the impact of this event was such that

Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel, e-mail: trifonov@research.haifa.ac.il

M. Barbieri (ed.), *The Codes of Life: The Rules of Macroevolution*.

nobody could even think of other possible codes. The triplet code was even called "genetic code," in other words *the only* code, not leaving any room for doubts. All early history of bioinformatics revolved around this single code (Trifonov, 2000a). Yet, already in 1968, R. Holliday noted almost en passant that, perhaps, recombination signals in yeast might reside on the same sequence that encodes proteins. This remark not only introduced the notion of other possible codes, but also the overlapping of different codes on the same sequence. The existence of codes, other than the classical translation triplet code, is already suggested by degeneracy of the triplet code (Schaap, 1971). Freedom in the choice of codons allows significant changes in the nucleotide sequence without changing the encoded protein sequence. This makes it possible, in principle, to utilize the interchangeable bases of the mRNA sequence for some additional, different codes. In this case, the codes would coexist in interspersed form as mosaics of two or more "colors." It is known today that a more general and widespread case is when the codes literally overlap so that some letters in specific positions of a given sequence (nucleotides or amino acids) are simultaneously involved in two or more different codes (sequence patterns). Such is the case with the coexisting triplet code and chromatin code – sequence instructions for nucleosome positioning (Trifonov, 1980; Mengeritsky and Trifonov, 1983). This was the first demonstration of the actual existence (Trifonov, 1981) of the hypothetical overlapping codes. Sequences that do not encode proteins, despite their traditional classification as noncoding, carry some important messages (codes) as well. Especially striking are the cases of sequence conservation in the noncoding regions (Koop and Hood, 1994), suggesting that the so-called noncoding sequences are associated with some function.

Amongst known general sequence codes, other than the triplet code, are transcription signals (*transcription code*) in promoters such as TATAAA box in eukaryotes, and TATAAT and TTGACA boxes in bacteria coding for initiation of transcription. Another broadly known sequence code is the *gene splicing code*, the GT–AG rule (Breathnach and Chambon, 1981) and some sequence preferences around the intron–exon junctions. A complex set of sequence rules describes details of DNA shape important for DNA–protein interactions and DNA folding in the cell.

At the level of amino acid sequences, the most important is the *protein folding code*, which is not yet described as a sequence pattern. One can single out the modular component of the folding code – organization of the globular proteins as linear succession of the modules in the form of loops of 25–30 residues closed at the ends by interactions between hydrophobic residues (Berezovsky et al., 2000; Berezovsky and Trifonov, 2002). The 3D structure of proteins appears to be encoded largely by a *binary code* (Trifonov et al., 2001; Trifonov, 2006; Gabdank et al., 2006) that, essentially, reduces the 20-letter alphabet to only two letters, for nonpolar and polar residues (more accurately, residues encoded by codons with pyrimidine or purine in the middle). The binary code also suggests the ancestral form for any given sequence.

As the carriers of instructions, biological sequences may be considered a language. Indeed, according to an appealing definition of Russian philosopher V. Nalimov (1981), language is a communication tool to carry instructions to the operator at the receiving end. Such languages as computer programs (frequently called "codes" as well) and written (spoken) human languages convey instructions expressed in the form of one code, for one reading device that takes consecutively letter by letter, word by word, until the transmitted command is fully uttered. As mentioned above, a unique property of the biological sequences is the superposition of the codes they carry. That is, the same sequence is meant to be read by several reading devices, each geared to its own specific code. Many cases of such overlapping are known (Trifonov, 1981; Normark et al., 1983). The overlapping is possible due to degeneracy of the codes. There is, of course, an informational limit for such superposition, when the freedom of degeneracy becomes insufficient to accommodate additional messages without loss of quality of many or all other messages present.

2 Hierarchy of the Codes

The commonly considered information flow from DNA to RNA and to protein is accompanied by massive loss of the sequences involved. Indeed, neither all DNA is transcribed, nor is the whole mass of RNA transcripts translated. This is especially obvious in eukaryotic genomes that contain large intergenic regions, and large intervening sequences that are passed from DNA to pre-mRNA. Is that loss of sequences also a loss of information? The multiplicity of the codes and their superposition suggest that some information is lost, indeed, together with those sequences that are not transcribed and not translated. In other words, DNA carries the sequence codes, serving at the DNA level, of which some are transferred to pre-mRNA. The sequences of the transcripts carry codes serving at RNA level, of which some are passed to the protein sequences, via mRNA. One, thus, has to consider the codes characteristic for the three sequence levels, hierarchically.

One could think of yet higher-level codes, beyond the purely molecular level. Among them would be organ/tissue-specific codes, i.e. genomic sequence features characteristic for one or another physiological function. These could be specifically placed tandem repeats, dispersed repeats, amplified genes, or whole groups of genes. One could also imagine "personal code(s)" – various sequence details responsible for individual traits, such as distinct facial features (Fondon and Garner, 2004) and mimic (Peleg et al., 2006) body set, favorite postures and gestures, and, perhaps, personal behavioral traits. Well-documented existence of population-specific genetic diseases and disorders indicates that there are also sequence features responsible for ethnicity traits. These may include specific sequence polymorphisms and, perhaps, some "guest" sequences present in one ethnical group and absent in others. The higher-level codes are likely to become a major focus of molecular

medicine in coming decades. In the mean time the sequence codes of molecular levels are still struggling to make it from singular to plural.

2.1 DNA Level Codes

The DNA structure is not monotonously uniform. It is modulated by the sequence-dependent local deviations from standard geometry, which may accumulate, for example, to a net DNA curvature (Trifonov and Sussman, 1980). Geometry of every base-pair step in the simple wedge model is described by three angles – wedge roll, wedge tilt, and twist. By following the sequence and deflecting the DNA axis at every step, according to the wedge and twist angles from the table of the dinucleotide codons (Bolshoy et al., 1991; Trifonov, 1991), one can calculate the predicted path of DNA axis – its local shape for any given sequence (Shpigelman et al., 1993). Hence, *DNA shape code*.

The *chromatin code* is a set of rules directing sequence-specific positioning of the nucleosomes. Sequence-dependent deformational anisotropy (bendability) of DNA appears to be an underlying principle of the nucleosome sequence specificity (Trifonov, 1980). As the strands of the nucleosome DNA follow the path of the deformed DNA duplex, they pass through inner contact points with histones (interface positions) and outward points (exposed to nucleoplasm). Various sequence elements that prefer the inner or outward positions would thus, ideally, reappear in the sequence at the distances that are multiples of nucleosome DNA. Indeed, the sequence periodicity is the most conspicuous feature of the nucleosome DNA sequences (Trifonov and Sussman, 1980). According to the latest updates (Cohanim et al., 2005, 2006a; Kogan et al., 2006; Trifonov et al., 2006a), there are at least three major periodical patterns in the nucleosome DNA: counter-phase AA/TT pattern, counter-phase GG/CC pattern (both combined in RR/YY pattern), and in-phase AA/TT pattern. Several other possible patterns are discussed in literature (reviewed in Kiyama and Trifonov, 2002; Segal et al., 2006).

An important issue in the elucidation of the chromatin sequence code is mandatory weakness of the nucleosome positioning sequence signal. This is required by the necessity of unfolding the nucleosomes during template processes. That is, the DNA complexes with the histone cores in the nucleosomes should be of marginal stability only. Accordingly, the sequence elements associated with the DNA bendability should be rather scarce in the nucleosome DNA sequence, especially those elements that are strong contributors to the bendability. Regrettably, it makes the deciphering of the nucleosome positioning code quite a challenge.

One of the factors influencing the nucleosome positioning is sterical exclusion of the nucleosomes by other nucleosomes, neighbors in 3D space (Ulanovsky and Trifonov, 1986). The most obvious sterical rule is the rule of linkers, first formulated and experimentally observed by Noll et al. (1980). Since every extra base pair in the linker causes rotation of the nucleosome around the axis of the linker by ~34°, the rotation may result in a sterical clash between the nucleosomes connected

by the common linker. This effect, indeed, is observed at short linkers. It is expressed in preferential appearance of the linkers of lengths about 5–11, 16–21, and 26–31 bases (Noll et al., 1980; Mengeritsky and Trifonov, 1983; Ulanovsky and Trifonov, 1986; Cohanim et al., 2006a). Intermediate linker lengths are forbidden due to the sterical clashes ("interpenetration" of the nucleosomes). The rule of linkers, thus, is an important part of the chromatin code.

2.2 RNA Level Codes

Those messages contained in the transcribed DNA are passed to RNA. The transcribed DNA, thus, contains overlapping messages of both DNA and RNA levels. The major mRNA level message is the classical triplet code – *RNA-to-protein translation code*. The chapters about this code appear in every textbook on molecular biology, and it will not be described here.

Eukaryotic transcripts also carry the *RNA splicing code*. This code is only poorly described (Breathnach and Chambon, 1981; Mount, 1982), so that existing sequence-based algorithms are not sufficient for detection of the splice sites in the sequences with as high a precision as in natural splicing process.

Overlapping with the protein-coding message, sequence of codons-triplets, is the universal 3-base periodicity with the consensus $(\text{G-nonG-N})_n$ (Trifonov, 1987) or, more accurately, $(\text{GCU})_n$ (Lagunez-Otero and Trifonov, 1992). Since the mRNA binding sites in the ribosome possess a complementary periodicity $(\text{xxC})_n$, with obligatory cytosines complementary to the frequent guanines of the first codon positions in mRNA, these 3-base periodicities have been interpreted as a device to maintain correct reading frame during translation of mRNA – the *framing code* (Trifonov, 1987). As described below, the periodical pattern $(\text{GCU})_n$ in mRNA appears to be a fossil of very ancient organization of codons (Trifonov and Bettecken, 1997).

The usage of codons corresponding to the same amino acid is known to be different for different organisms and even different genes. Among the alternative codons, the rare codons are of special interest. Their occurrence along the mRNA sequence is not random. It is shown, for example, that clusters of infrequently used codons in prokaryotic mRNA often follow at a distance about 150 triplets from one another. This is interpreted as *translation pausing code*, to slow down the translation after a protein domain (fold) is synthesized: to give the newly synthesized chain sufficient time for its proper folding (Makhoul and Trifonov, 2002).

2.3 Codes of Protein Sequences

According to common belief, the protein sequence carries instructions on how the polypeptide chain folds, for the reliable performance of respective function of the protein, encoded in the sequence as well. At the same time, it is well known that

proteins with the same fold and the same function may have rather different sequences. As in the case of the triplet code, this degeneracy of the protein sequence may allow incorporation in the same sequence of some additional messages.

The *protein folding code* is a major challenge for the protein structure community. There are plenty of sophisticated approaches offering partial solutions of the problem, but the conclusive sequence rules for protein folding are still to be found.

An apparent major obstacle is estimated colossal time required for the unfolded polypeptide chain to go through all intermediate states until the final native fold structure is reached – the so-called Levinthal paradox. By some trick of nature, a special sequence organization should be there, in the protein sequences, to ensure the folding in realistic time of milliseconds to seconds. One possible way out is suggested by modular organization of the protein folds (Berezovsky and Trifonov, 2002). Indeed, if the chain length of the module is 20–30 amino acid residues, the time required for its folding fits well to the realistic limits. And, as numerous recent studies demonstrate, globular proteins *are* built of such modules of standard size 25–30 residues in form of closed loops (Berezovsky et al., 2000; Trifonov and Berezovsky, 2003; Berezovsky et al., 2003a, b; Aharonovsky and Trifonov, 2005; Sobolevsky and Trifonov, 2006).

The modular structure of proteins suggests a principally new, compressed way of presentation of amino acid sequences rather as, sequences of the modules, descendants of the early sequence/structure/function prototypes (Berezovsky et al., 2003a, b), in a new alphabet of the prototypes. This would represent the *proteomic code* contained in the amino acid sequences. The prototype modules, then, would appear as the codons of the proteomic code.

2.4 Fast Adaptation Code

This code resides and functions in all three types of genetic sequences. It is believed to be responsible for special type of quick, significant changes in the sequences, apparently, in response to environmental changes. It involves the most variable sequences – simple tandem repeats of the structure $(AB\ldots MN)_n$. Remarkably, the information carried in the sequences resides not as much in the sequence AB…MN of the repeating unit, as rather in the copy number *n* of the repeats (Trifonov, 1989, 2004). Indeed, after the spontaneous change in the repeating sequence, its extension or shortening, the sequence in brackets stays intact while the copy number *n* becomes larger or smaller, respectively. Since the repeats are involved in gene expression in one or another way, the change of *n* results in the modulation of gene activities, as a response to environmental challenges, and thus in fast adaptation (Trifonov, 1989, 1990, 1999, 2004; Holliday, 1991; King, 1994; Künzler et al., 1995, King et al., 1997). An important faculty of this mechanism is an apparent directionality of the mutational changes of this type (Trifonov, 2004). Indeed, small variations in the *n* values corresponding to repeats serving genes *irrelevant* to a

given environmental stress do not change the expression patterns of these genes. On the contrary, if *relevant* responsive genes are involved, the copy numbers of the respective repeats become subject of systematic selection towards better repeat copy number (better gene expression) patterns. The relevant genes (but only relevant ones) become, thus, retuned (King et al., 1997; Trifonov, 1999).

2.5 The Codes of Evolutionary Past

Every sequence has its evolutionary history, and those sequences or sequence fragments, that have been successful in the earliest times of molecular evolution, are, perhaps, still around in hidden form or even unchanged since those times. The proteomic code described above is an example of such code of evolutionary record. The modern sequence modules are not the same as their ancestral prototypes, but a certain degree of resemblance to the ancestors is conserved allowing classification of present-day modules.

The earliest traced sequence elements go back to the very first codons, which are described as the triplets GGU, GCC, and their point mutational versions (Trifonov and Bettecken, 1997). More detailed reconstruction confirmed this conclusion (Trifonov, 2000b, 2004). According to the reconstruction of the earliest stages of molecular evolution, the very first "genes" had a duplex structure with complementary sequences $(GGC)_n$ and $(GCC)_n$, encoding, Gly_n and Ala_n, respectively. Thus, the mRNA consensus $(GCU)_n$ and the consensus $(xxC)_n$ of the mRNA binding sites in the ribosome are both fossils of the earliest mRNA sequences (Trifonov, 1987; Lagunez-Otero and Trifonov, 1992; Trifonov and Bettecken, 1997).

The size of the earliest minigenes, as it turns out, can be estimated by distance analysis of modern mRNA sequences (Trifonov et al., 2001). For this purpose the sequences were first rewritten in binary form, in an alphabet of two letters, *G* and *A*, for Gly series of amino acids and codons and Ala series (see above). Respective codons contain in their middle positions either purines (in *G*) or pyrimidines (in *A*). From the reconstructed chart of evolution of the codons (Trifonov, 2000b, 2004), it follows that all codons of *G*-series are descendants of the GGC codon, with purine in the middle, while codons of *A*-series originate from GCC codon, with pyrimidine in the middle. If the products of very first genes had the structures either G_n or A_n, of a certain size *n*, then after fusion of the minigenes the alternating patterns $G_nA_nG_nA_n$ … may have been formed. Later mutations could, of course, have completely destroyed this pattern, but they did not. Analysis of large ensembles of the mRNA sequences showed that the pattern did survive, though in rather hidden form (Berezovsky and Trifonov, 2001; Trifonov et al., 2001) so that the estimation of the very first gene size became possible, 6–7 codons encoding hexa- and heptapeptides. This estimate is strongly supported by independent calculation of the sizes of the most ancient mRNA hairpins that arrived at the same minigene size (Gabdank et al., 2006; Trifonov et al., 2006b). Moreover, most conserved oligopeptide

sequences, present in every prokaryotic proteome, also have the size of 6–9 amino acids (Sobolevsky and Trifonov, 2005, Sobolevsky et al., submitted).

The ancient conservation of the middle purines and pyrimidines in the codons during the evolution of the codon table, actually, has very much survived till now. This is confirmed by an analysis of amino acid substitutions in modern proteins (Trifonov, 2006; Gabdank et al., 2006). Every modern protein sequence, thus, can be written in the *A* and *G* alphabet. Such presentations of modern sequences in the *binary code* would suggest the most ancient version of the sequences.

The binary code, the mosaic of *A*- and *G*-minigenes, and the proteomic code describe various stages of protein evolution, from simple to more complex. Today one can also detect the next stage – combining the closed loop modules in the protein folds, domains.

First, the next level is seen already in protein sizes, which appear to be multiples of 120–150 amino acid units (Berman et al., 1994; Kolker et al., 2002). This size is a good match to the optimal DNA ring closure size, about 400 base pairs (Shore et al., 1981). This attractive numerology may well reflect original formation of modern genes and genomes by fusion of individual DNA circles (genome units) of this standard size (Trifonov, 1995, 2002). This would constitute the *genome segmentation code*. How this code is expressed in the sequence form is not yet specified, except for preferential appearance of methionines (former translation starts) at genome unit size distances (Kolker and Trifonov, 1995).

3 Superposition of the Codes and Interactions Between Them

As most of the codes described above are degenerate, allowing alternative or sometimes even wrong letters here and there, they may coexist as a superposition of several codes, on the same sequence (reviewed in Normark et al., 1983; Trifonov, 1981, 1989, 1996, 1997). The most spectacular case is the overlapping of the chromatin code (nucleosome positioning) with protein coding and gene splicing. Indeed, the alternating AA/TT nucleosome pattern is demonstrated to be located largely, if not fully, on those sections of the protein-coding regions that correspond to amphipathic α-helices (Cohanim et al., 2006a,b). The third positions of the codons within the region occupied by the nucleosome are responsible as well for the creation of the periodical AA/TT pattern. Moreover, even the encoded amino acid sequence is also biased to a certain degree to contribute to the nucleosome sequence pattern (Cohanim et al., 2006b). In addition, the nucleosomes are preferentially centered at the splice junctions, apparently for their protection (Denisov et al., 1997; Kogan and Trifonov, 2005). Since the coding sequences also carry at least one more message – translation framing, the nucleosome sequences display superposition of at least four different codes, on the same sequence.

The adjustment of the protein sequence, to contribute to the DNA sequence periodicity, both in prokaryotes and in eukaryotes (Cohanim et al., 2006b), is an interesting case. Apparently, on one hand, the 10–11 base DNA sequence

periodicity is of no less importance for the cell than the proteins encoded in the DNA sequence. On the other hand, this example of interactions between the codes shows that the DNA sequence level message is projected all the way through mRNA to the protein sequence level. The latter one, thus, carries (reflects) the sequence patterns of the whole hierarchy – of DNA, RNA, and protein levels.

A neat example of the overlapping at the level of protein sequences is the "moonlighting" of intrinsically unfolded proteins (IUPs) (Tompa et al., 2005). That is, the same molecule of the IUP, the same sequence, can be involved in more than one function, thus, carrying different superimposed messages. Structural and functional promiscuity of the IUPs is carried through, perhaps, since the earliest times of molecular evolution. Highly structured functionally specialized proteins were not yet around, and the multi-functionality of simpler IUP molecules was of an obvious advantage for survival.

4 Is That All?

There are still many nondeciphered codes around. Nature would utilize every useful combination of letters. This is because of eternal molecular opportunism (Doolittle, 1988) that drives the molecules of life towards better and more diverse performance in the challenging conditions of the changing environment. In this struggle for survival (natural selection) and for better well-being (opportunism), living matter developed intricate levels of complexity, including sequence complexity. It would be naive to say that all the codes are already known, as it was, indeed, naive to content oneself with the single "genetic code" 30 years ago.

On the one hand, there are sequence biases and patterns that are still not fully explained, such as species-specific G + C content of genomes – genomic code (D'Onofrio and Bernardi, 1992), and general avoidance of the CG dinucleotides. On the other hand, many of the known molecular functions still do not have explicit sequence descriptions, such as RNA interference (Fire et al., 1998) or RNA editing (Gott and Emeson, 2000). The so-called noncoding sequences have the provocative property of being rather dispensable, though they do carry some of the codes described in the review (chromatin code, fast adaptation code). The famous case of the Fugu-fish genome, with the reduced amount of noncoding sequences in it (Aparicio et al., 2002), is often taken as an example of a seemingly insignificant role the noncoding sequences play. Yet, it is known that the noncoding sequences harbor various repeats, of dispersed type (transposons), and tandem repeats. It is also known that transposable elements play an important role in evolution and adaptation (Reanney, 1976). The tandem repeats serve as tuners of gene expression (Trifonov, 1989, 2004; King et al., 1997; Fondon and Garner, 2004) (see *Fast adaptation code*, above). Could it be that the Fugu-fish is in an evolutionary steady state, with virtually no need for adaptive sequence changes? That could be only if there are no environmental challenges for this species. Indeed, the small-genome Fugu-fish has a narrow habitat (Hinegardner, 1976), living only in coral reefs with

well-defined fauna, around the islands of Japan. Thus, even dispensable sequences deserve respect, as they seem to code for the vital ability for adaptation.

The conspicuously primitive simple tandem repeats are the best advocates in favor of all sequences, no matter how nonsensical, primitive, or even dispensable they appear. In a recent study (Bacolla et al., 2006), the pure purine or pyrimidine repeats are shown to be the only difference between human and chimpanzee sequences (over 800 large segments studied). The repeats are also the same, but the copy numbers of the repeat units (total lengths of the repeat regions) are different in these two species. Referring to the fast adaptation code (above), one would think that humans and chimpanzees are nearly the same species, only well adapted to completely different living conditions. So much for even the primitive sequences.

The answer to the question in the title of this section, thus, is a firm "No."

References

Aharonovsky E, Trifonov EN (2005) Protein sequence modules. J Biomol Str Dyn 23:237–242

Aparicio S, *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Bacolla A, Collins JR, Gold B et al. (2006) Long homopurine* homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucl Acids Res 34:2663–2675

Berezovsky IN, Trifonov EN (2001) Evolutionary aspects of protein structure and folding. Mol Biol 35:233–239

Berezovsky IN, Trifonov EN (2002) Loop fold structure of proteins: resolution of Levinthal's paradox. J Biomolec Str Dyn 20:5–6

Berezovsky IN, Grosberg AY, Trifonov EN (2000) Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 466:283–286

Berezovsky IN, Kirzhner VM, Kirzhner A et al. (2003a) Protein sequences yield a proteomic code. J Biomol Struct Dyn 21:317–325

Berezovsky IN, Kirzhner A, Kirzhner VM, Trifonov EN (2003b) Spelling protein structure. J Biomol Struct Dyn 21:327–339

Berman AL, Kolker E, Trifonov EN (1994) Underlying order in protein sequence organization. Proc Natl Acad Sci USA 91:4044–4047

Bolshoy A, McNamara P, Harrington RE, Trifonov EN (1991) Curved DNA without AA: experimental estimation of all 16 wedge angles. Proc Natl Acad Sci USA 88:2312–2316

Breathnach R, Chambon P (1981) Organization and expression of eukaryotic split genes coding for proteins. Ann Rev Bioch 50:349–383

Cohanim AB, Kashi Y, Trifonov EN (2005) Yeast nucleosome DNA pattern: deconvolution from genome sequences *of S. cerevisiae*. J Biomol Str Dyn 22:687–694

Cohanim AB, Kashi Y, Trifonov EN (2006a) Three sequence rules for chromatin. J Biomol Struct Dyn 23:559–566

Cohanim AB, Trifonov EN, Kashi Y (2006b) Specific selection pressure on the third codon positions: contribution to 10 - 11 base periodicity in prokaryotic genomes. J Molec Evol (in press)

Denisov DA, Shpigelman ES, Trifonov EN (1997) Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 205:145–149

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110:81–88

Doolittle RF (1988) More molecular opportunism. Nature 336:18

Fire A, Xu S, Montgomery MK et al. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811

Fondon JW, Garner HR (2004) Molecular origin of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 101:18058–18063
Gabdank I, Barash D, Trifonov EN (2006) Tracing ancient mRNA hairpins. J Biomol Str Dyn 24:163–170
Gott JM, Emeson RB (2000) Functions and mechanisms of RNA editing. Ann Rev Genet 34:499–531
Hinegardner R (1976) Evolution of genome size. In: Ayala FJ (ed) Molecular Evolution. Sinauer Association, Sunderland
Holliday R (1968) Genetic recombination in fungi. In: Peacock WJ, Brock RD (eds) Replication and Recombination of Genetic Material. Australian Academy of Science, Canberra, Australia
Holliday R (1991) Quantitative genetic variation and developmental clocks. J Theor Biol 151:351–358
Khorana HG, Büchi H, Ghosh H et al. (1966) Polynucleotide synthesis and the genetic code. Cold Spring Harb Symp Quant Biol 31:39–49
King DG (1994) Triple repeat DNA as a highly mutable regulatory mechanism. Science 263:595–596
King DG, Soller M, Kashi Y (1997) Evolutionary tuning knobs. Endeavor 21:36–40
Kiyama R, Trifonov EN (2002) What positiones nucleosomes? A model. FEBS Lett 523:7–11
Kogan S, Trifonov EN (2005) Gene splice sites correlate with nucleosome positions. Gene 352:57–62
Kogan SB, Kato M, Kiyama R, Trifonov EN (2006) Sequence structure of human nucleosome DNA. J Biomol Struct Dyn 24:43–48
Kolker E, Trifonov EN (1995) Periodic recurrence of methionines: Fossil of gene fusion? Proc Natl Acad Sci USA 92:557–560
Kolker E, Tjaden BC, Hubley R et al. (2002) Spectral analysis of distributions: finding periodic components in eukaryotic enzyme length data. OMICS: J Integr Biol 6:123–130
Koop BF, Hood L (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. Nature Genet 7:48–53
Künzler P, Matsuo K, Schaffner W (1995) Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. Biol Chem Hoppe-Seyler 376:201–211
Lagunez-Otero J, Trifonov EN (1992) mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. J Biomol Struct Dyn 10:455–464
Makhoul CH, Trifonov EN (2002) Distribution of rare triplets along mRNA and their relation to protein folding. J Biomol Struct Dyn 20:413–420
Mengeritsky G, Trifonov EN (1983) Nucleotide sequence-directed mapping of the nucleosomes. Nucl Acids Res 11:3833–3851
Mount SM (1982) A catalogue of splice junction sequences. Nucl Acids Res 10:459–472
Nalimov VV (1981) In the labyrinths of language: A Mathematician's Journey. ISI Press, Philadelphia, USA
Nirenberg M, Caskey T, Marshall R et al. (1966) The RNA code and protein synthesis. Cold Spring Harb Symp Quant Biol 31:11–24
Noll M, Zimmer S, Engel A, Dubochet J (1980) Self-assembly of single and closely spaced nucleosome core particles. Nucl Acids Res 8:21–42
Normark S, Bergstrom S, Edlund T et al (1983) Overlapping genes. Ann Rev Genet 17:499-525
Ochoa S (1963) Synthetic polynucleotides and the amino acid code. Cold Spring Harb Symp Quant Biol 28:559–567
Peleg G, Katzir G, Peleg O et al. (2006) Hereditary family signature of facial expression. Proc Natl Acad Sci USA 103:15921–15926
Reanney DC (1976) Extrachromosomal elements as possible agents of adaption and development. Bact Rev 40:552–590
Schaap T (1971) Dual information in DNA and the evolution of the genetic code. J Theor Biol 32:293–298
Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006) A genome code for nucleosome positioning. Nature 442:772–778