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

Shore D, Langowski J, Baldwin RL (1981) DNA flexibility studied by covalent closure of short fragments into circles. Proc Natl Acad Sci USA 78:4833–4838
Shpigelman ES, Trifonov EN, Bolshoy A (1993) CURVATURE: software for the analysis of curved DNA. CABIOS 9:435–440
Sobolevsky Y, Trifonov EN (2005) Conserved sequences of prokaryotic proteomes and their compositional age. J Mol Evol 61:591–596
Sobolevsky Y, Trifonov EN (2006) Protein modules conserved since LUCA. J Mol Evol 63:622–634
Tompa P, Scasz C, Buday L (2005) Structural disorder throws new light on moonlighting. Trends Bioch Sci 30:484–489
Trifonov EN (1980) Sequence-dependent deformational anisotropy of chromatin DNA. Nucl Acids Res 8:4041–4053
Trifonov EN (1981) Structure of DNA in chromatin. In: Schweiger H (ed) International Cell Biology 1980–1981. Springer-Verlag, Berlin
Trifonov EN (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. J Mol Biol 194:643–652
Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432
Trifonov EN (1991) DNA in profile. Trends Biochem Sci 16:467–470
Trifonov EN (1990) Making sense of the human genome. In: Sarma RH, Sarma MH (eds) Structure and Methods, vol. 1, Human Genome Initiative and DNA Recombination. Adenine Press, New York
Trifonov EN (1995) Segmented structure of protein sequences and early evolution of genome by combinatorial fusion of DNA elements. J Mol Evol 40:337–342
Trifonov EN (1996) Interfering contexts of regulatory sequence elements. CABIOS 12:423–429
Trifonov EN (1997) Genetic sequences as product of compression by inclusive superposition of many codes. Mol Biol 31:759–767
Trifonov EN (1999) Elucidating sequence codes: three codes for evolution. Annals NY Acad Sci 870:330–338
Trifonov EN (2000a) Earliest pages of bioinformatics. Bioinformatics 16:5–9
Trifonov EN (2000b) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261:139–151
Trifonov EN (2002) Segmented genome: elementary units of genome structure. Russian J Genet. 38:659–663
Trifonov EN (2004) The triplet code from first principles. J Biomol Struct Dyn 22:1–11
Trifonov EN (2006) Theory of early molecular evolution: predictions and confirmations. In: Eisenhaber F (ed) Discovering Biomolecular Mechanisms with Computational Biology. Landes Bioscience, Georgetown
Trifonov EN, Berezovsky IN (2003) Evolutionary aspects of protein structure and folding, Curr Opinion Struct Biol 13:110–114
Trifonov EN, Bettecken T (1997) Sequence fossils, triplet expansion, and reconstruction of earliest codons. Gene 205:1–6
Trifonov EN, Sussman JL (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA 77:3816–3820
Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN (2001) Distinct stages of protein evolution as suggested by protein sequence analysis. J Mol Evol 53:394–401
Trifonov EN, Kogan S, Cohanim AB (2006a) Latest on the nucleosome positioning sequence patterns. In: Kiyama R, Shimizu M (eds) DNA Structure, Chromatin and Gene Expression. Transworld Research Network. Trivandrum, India
Trifonov EN, Gabdank I, Barash D, Sobolevsky Y (2006b) *Primordia vita*. Deconvolution from modern sequences. Origin Life Evol Biosph 36(5–6):559–565
Ulanovsky LE, Trifonov EN (1986) A different view point on the chromatin higher order structure: steric exclusion effects. In: Sarma RH, Sarma MH (eds) Biomolecular stereodynamics III. Adenine Press, New York