**Abstract**

**Motivation:** Although one would normally expect a given regulatory element to perform best when it fully matches its consensus sequence, this is generally far from being the case. Usually, almost none of the actual sites fits the consensus exactly, and some of those that do fit do not perform well. The main reason for that is the very nature of the sequences and the messages (codes) they contain. Normally, any given stretch of the sequence with one or another regulatory site not only carries this regulatory message, but several more messages of various types as well. These messages overlap with the regulatory element in such a way that the letter (base) which actually appears in any given sequence position simultaneously belongs to one or more additional codes. Apart from numerous individual codes (sequence patterns) specific for a given species or gene, there are many different general (universal) sequence codes all interacting with one another. These are the classical triplet code, DNA shape code, chromatin code, gene splicing code, modulation code and many more, including those that have not yet been discovered. Examples of overlapping of different codes and their interaction are discussed, as well as the role of degeneracy of the codes and the sequence complexity as a function of code density.

**Contact:** E-mail: bptrifoll@weizmann.weizmann.ac.il

**Introduction**

Recognition sites for various transcription factors in the DNA sequences as well as, generally, specific sites of all types appear as islands in the sea of an irrelevant ‘random’ background, different for every individual site. Since many positions in the consensus sequence of the site are frequently degenerate, even the signal sequence itself is variable, the variations being, again, seemingly random. The natural DNA sequences (as well as RNA and protein sequences), however, are far from being random. What was considered in earlier days as the sea of ‘junk’ sequences is today shrinking to remaining lakes and puddles of seemingly featureless and functionless sequences which, most likely, also carry some messages that are just not yet recognized.

As the reader will see below, many of those natural sequences which are found to be functional and recognizable, after further scrutiny, turn out to carry several interfering (overlapping) messages simultaneously. This paper describes various kinds of these interfering contexts and upgrades some of them to the category of ‘codes’ of which the classical triplet code is one veteran representative.

**There is something else in the same sequence**

*Do we get right what the consensus sequences tell?*

Common belief is that having enough versions of a given type of the recognition sequence, one can derive a consensus with a typical or even exemplary performance as the recognition site. Surprisingly, however, there are several hard-to-believe cases when some rather different sequences turn out to perform better than the consensus.

Two such examples emerged from experiments with randomization of the recognition sequences in a search for a better one. In both cases, some of the sequences selected from random sets were found to be superior in some important aspects as compared to well-established consensus sequences. The *tet* promoter of pBR322, TTGACA TTTAAT, with an almost ideal TATA box and ideal -35 sequence, was replaced by the sequence TTTCTT TAAGCT with only marginal resemblance to the wild type; yet 8-fold more mRNA was produced with this artificial ‘promoter’ (Horwitz and Loeb, 1988). Obviously, the promoter consensus reflects not only sequence requirements for good mRNA production, but also some other features that it is important for natural promoters to have. In other words, the consensus tells something else, not just what our limited knowledge about promoter structure would suggest. Another case like that is a study on RNA ligands of T4 DNA polymerase (Tuerk and Gold, 1990). The consensus ligand AAUAACUC was, indeed, fished out from the random sequences as one of the best performers, but not the best. The sequence AGCAACCU showed, actually, even stronger binding, although it was only a 50% match to the natural consensus. Again, it appears that the consensus sequence is not only about the binding. It has to satisfy some other sequence requirements escaping our straightforward understanding.

All this suggests that reading the sequences in only one way, i.e. expecting one message only, is generally wrong. Not
ininfrequently some other messages hide both in the consensus and in the contexts of the individual sites. This is not unlike the following analogy (not to be taken as more than that), where the ‘messages’ have even no connection with their formal consensus. The words COOKY, MANGO, MELON, HONEY, SWEET all suggest something sweet or sweet-sour and could be considered, thus, as recognition sequences for the ‘sweet’ quality. Their consensus sequence, however, conveys a rather different message: MONEY.

The third positions do have something to say

The third, degenerate, positions of the codons in the mRNA are normally given an unimpressive role of nearly neutral drifting. Indeed, if two very closely related gene sequences are taken, the difference between them resides almost exclusively in the third positions, leaving the encoded amino acid sequences unchanged. However, there are several observations indicating that the third position is not quite neutral. For example, a certain selection pressure in favor of local G + C composition is well documented (D’Onofrio and Bernardi, 1992). Analysis of the occurrences of four bases in the three codon positions of mRNA also indicates that the distribution of bases in the almost ‘silent’ third position displays a significant degree of species specificity (Zhang and Chou, 1994). Yet another bias is the frequent preference for U in the third positions, a reflection of the hidden mRNA consensus (GCU)n, presumably responsible for keeping the ribosome in the correct reading frame (Lagunez-Otero and Trifonov, 1992). In other words, the third positions, indeed, seem to carry some functional load or rather several different loads, not necessarily one at a time. The degeneracy of the triplet code (alternative bases in the third positions and alternative triplets for leucine, serine and arginine) is not wasted. Below, some particularly striking examples of the utilization of the degeneracy are described.

The largest subunit of RNA polymerase II contains a C-terminal domain with tandem repeats of the highly conserved seven amino acid sequence YSPTSPS. In the corresponding cDNA, all six possible codons for serine are used (Corden et al., 1985). One would expect that if the selection pressure were only exerted on the amino acid sequence, thus allowing mutations of the genomic sequence to all possible triplets for serine. Close inspection of the distribution of the AGY serine triplets along the cDNA reveals, however, that the most frequent distances between the AGY triplets are not 6, 9, 12, 15 or 21 bases as the amino acid sequence would suggest but, very selectively, 21 (also 42 and 63) residues. In other words, the AGY triplets appear periodically, with the period 21 bases. Obviously, this has nothing to do with the selection pressure on the amino acid sequence. The AGY periodicity is needed for whatever purpose by the DNA (or RNA) sequence. Perhaps, this is some message carried by DNA for itself, since the period 21 bases corresponds closely to two helical repeats of DNA.

A similar example, but apparently with a different type of DNA level message, is provided by the sequence of one of the MHC genes (Levi-Strauss et al., 1988). Here as well the region 725–880 of the sequence codes for a repeating amino acid sequence with strictly alternating positively and negatively charged amino acids, R, K and E, D, respectively. Four different triplets for arginine are present in this sequence: CGA, CGG, AGA and AGG; and, again, the degeneracy of these triplets and of the third positions of other triplets is highly non-random. The positions of purine bases strictly follow a consensus (GA)n, such that all AG dinucleotides occupy only odd positions in the sequence (727 and other odd coordinates), while the GA dinucleotides are placed in even positions (728 and other even coordinates). Perhaps at some point in evolution this sequence was just an alternation of A and G, and this alternation is still kept wherever the degeneracy of the triplet code allows this. Why the two-base periodical alternation of A and G is almost as important as six-base periodical alternation of triplets for positively and negatively charged amino acids remains to be seen. Interestingly, the presumed ancestral repeating sequence (AGAGAG)n coding for the repeat (RE)n would ideally satisfy both alternations. Two other charged amino acids, K and D, apparently, had to be introduced, compromising the original simple pattern, but perhaps bringing in some important advantage.

Another very interesting example is hypervariability of the V3 region of the HIV-1 envelope gene. Here the changes occur in the first two positions of the codons, while the third positions are conserved (!): of 11 documented point changes in this region, only one hits the third position (J.Hirshon and J.Goudsmit, personal communication). This conservation, obviously, is unrelated to the encoded protein and keeps intact some additional message. It remains to be found what that message means.

Codes other than triplet code

The code is any pattern or bias in the sequence which corresponds to one or another specific biological (biomolecular) function or interaction (Trifonov, 1989). The codes could be specific or general, depending on the respective functions. The sequence of a regulatory site operating with only a limited group of genes or common to only a few species would be an example of a specific code. Many transcription factor binding sites are of this category. The universal triplet code and, say, signals for transcription initiation and termination are examples of general codes common either to all species or at least to large kingdoms. This section is a brief overview of current knowledge about various general codes starting with those known sequence patterns which are
normally not considered as codes, although they are according to the above definition. The following codes are very briefly described below: transcription codes, gene splicing code, translation pausing code, DNA structure codes, chromatin code, translation framing code, modulation (adaptation) code and gene segmentation code.

Transcription codes include promoters and terminators, and are rather universal though different in prokaryotes and in eukaryotes. Most of the promoter sequences include the TATA box or a close version of it (Pribnow, 1975; Goldberg, and are rather universal though different in prokaryotes and in eukaryotes. Most of the promoter sequences include the TATA box or a close version of it (Pribnow, 1975; Goldberg, 1979). Prokaryotic promoter sequences are rather thoroughly characterized and the algorithms are available to recognize them in the sequences (e.g. Alexandrov and Mironov, 1990; Demeler and Zhou, 1991). Prokaryotic terminators are characterized as well (Brendel et al., 1986; d’Aubenton Carafa et al., 1990; Yager and von Hippel, 1991). Eukaryotic promoters are well reviewed (Bucher and Trifonov, 1986; Bucher, 1996) and several algorithms to locate them in the sequences have been attempted [Prestige (1995) and references therein].

The gene splicing code for the processing of nuclear pre-mRNA is largely deciphered. Its main components are obligatory GU- and AG-ends of introns (Breathnach and Chambon, 1981), as well as rather conserved consensus sequences and other sequence features around the ends (Mount, 1982; Soloviev et al., 1994). This information, unfortunately, is insufficient to locate the splice sites in the genomic sequences unequivocally. Apparent high precision of the gene splicing would suggest that there are some other important features of the excised and spliced sequences, which remain to be revealed in future studies.

Translation pausing important for the regulation of translation is encoded by clusters of rare triplets for which the respective aminoacyl-tRNAs are in limited supply (Varenne and Lazdunski, 1986; Kurland, 1991). A higher potential of hairpin formation immediately downstream from the rare triplets (Shpaer, 1985) also contributes to the translation pausing pattern, providing a transient obstacle for the ribosome. Since the rare codons are also sites of more frequent translational errors, the translation pausing code may also be called a ‘fidelity code’ (Shpaer, 1985).

The last two decades of studies on DNA structure revealed that DNA is not a monotonous double helix all the same for whatever sequence of base pairs is hidden in its interior. Rather, it is a spectrum of local variations of the shape dictated by the nucleotide sequence. In particular, the molecule is frequently curved, which was originally revealed from the DNA sequence analysis showing some dinucleotides reappearing with the DNA helical repeat (Trifonov, 1980; Trifonov and Sussman, 1980). The sequence-dependent local shape of DNA is a crucial component of the protein–DNA recognition. The corresponding sequence code (DNA shape code) is universal and can be expressed in the form of a set of 26 angular parameters characterizing geometries of all 10 possible base pair stacks in DNA. The angles—wedge roll, wedge tilt, and twist—have been estimated by computational fitting to results of solution experiments with many DNA fragments of different nucleotide sequences (Kabsch et al., 1982; Bolshoy et al., 1991). A computer algorithm exists which calculates the DNA shape from its nucleotide sequence based on the above 26 angles (Shipgelman et al., 1993). There are also several alternative models describing the sequence-dependent DNA shape [Dickerson et al. (1996) and references therein]. These models, however, predict DNA curvatures with average misfit >2 SD away from the experimental estimates (Bolshoy et al., 1991).

A gross change in DNA helical shape is caused by transition from one DNA form to another. Each form can be characterized by A-, B- or Z-propensity energies calculated for all trinucleotides (Basham et al., 1995). This constitutes DNA form codes for all three basic DNA helical structures.

Chromatin code describes those sequence features that direct the histone octamer’s binding to DNA and formation of the nucleosomes. The pattern reflects deformatonal anisotropy of DNA (Trifonov, 1980), and consists primarily of periodically alternating AA and TT dinucleotides with the period ~10.4 bases (Mengeritsky and Trifonov, 1983; Ioshikhes et al., 1992). The latest analysis (Ioshikhes et al., 1996) of compiled experimentally mapped nucleosome DNA sequences (Ioshikhes and Trifonov, 1993) confirmed earlier estimates of the phase shift between AA and TT dinucleotides in the nucleosomal pattern (~5 bases), and allowed the location of the AA dinucleotides preferentially on the histone octamer surface (Ioshikhes et al., 1996). This places the complementary TT dinucleotides on the outer face of the nucleosomal DNA, in accordance with UV-irradiation experiments (Gale et al., 1987; Pehrson, 1989). An alternative nucleosome DNA pattern is also suggested where the periodical AA and TT dinucleotides are in-phase (Satchwell et al., 1986). An important part of the chromatin code is information about orientations of neighboring nucleosomes in space indispensable for the description of local architecture of the chromatin fiber. This information is, presumably, expressed in the lengths of the linkers connecting the nucleosomes (Noll et al., 1980; Ulanovsky and Trifonov, 1986).

Overlapping with the triplet code is the translation framing code (Trifonov, 1987). The correct reading frame during translation is, apparently, secured by complementarity of the ‘consensus’ mRNA pattern, (Gcu)n (Trifonov, 1987; Laguna-Otero and Trifonov, 1992), with the (Cxx)n mRNA contact sites in the small subunit ribosomal RNA (Trifonov, 1987; McCarthy and Brimacombe, 1994). The frame monitoring by this scheme is reflected in several known cases of translational frameshifting where the change in the (Gcu)n frame, apparently, causes the jump of the ribosome to the new frame.
(Trifonov, 1987). The presence of the G-periodical pattern in the protein-coding sequences should be important, especially for long polypeptide chains, to make the probability of the ribosome jumping sufficiently low so that the translation would faithfully proceed to the completion of the synthesis of the entire long chain.

The most conspicuous of all sequence types are tandemly repeating sequences, especially simple repeats. At first sight, the tandem repeats seem to be meaningless and dispensable since the copy numbers of the repeats are highly variable. Analysis of the literature on the occurrences and functional involvement of the repeats leads to the conclusion that the very copy number matters no less than the repeat sequence itself (Trifonov, 1989, 1990). It is suggested that the variable copy number of a repeat serves as an adjustable variable to modulate expression of the nearby gene or other biological function, i.e. the genome units or the fusion sites are remains to be found.

Undoubtedly, there are many more codes hidden in the sequences. Some of them could only be guessed, some already disclose themselves by one or another sequence feature. For example, the alcoholic syndrome in mice is accompanied by a change in the copy number of the (RY) repeat sequence in the first intron of the alcohol dehydrogenase gene (Zhang et al., 1987). It is speculated that the tuning mechanism—fast changes in the copy numbers of the tuning tandem repeats—also participates in the fast adaptation to a changing environment (Trifonov, 1989, 1990). Indeed, the unequal crossing-over and replication slippage on the repeating sequences, both causing changes in the copy numbers, are rather frequent events compared to rare point mutations. The given gene's activity could be changed much faster by the tuning mechanism than by the point mutations. The modulation code as described above could thus also be considered as an adaptation code.

One of the emerging new codes is the genome segmentation code, i.e. the genomes appear to be built of rather standard size units of an unknown number of types fused together in various orders, very much like shuffled packs of cards. The first evidence comes from the size distributions of the natural polypeptide chains. Eukaryotic proteins, for example, show preference for the size 123 amino acid residues and multiples of this unit size (Berman et al., 1994). Such a distribution could result from many acts of combinatorial fusion of different small genes, all of the same unit size, at some early stage of evolution (Trifonov, 1995). One prediction would be that at the points of fusion, i.e. every 120–125 residues, there would be some excess of methionines, formerly initiation residues of the fused small genes. This expectation is well met, indeed (Kolker and Trifonov, 1995), although most of these boundary methionines disappeared, apparently, due to mutations. The corresponding DNA unit size, ~370 bp, is also observed in the size distributions of DNA mobile elements (E.N.Trifonov, in preparation), in particular, in the sizes of extrachromosomal circular DNA (Gaubatz, 1990). What the sequence features of the genome units or the fusion sites are remains to be found.

Superposition and interaction of the codes

One of the most remarkable properties of genetic sequences is that the codes they carry overlap, i.e. a given base in a given nucleotide sequence may well belong to two or more different codes simultaneously. This phenomenon, unique for the genetic sequences, was predicted in 1968 by Holliday who suggested, in particular, that the signals responsible for recombination may well reside within protein-coding sequences. The general idea was developed further in later works (Schaap, 1971; Zuckerkanldl, 1976a,b; Trifonov, 1981) and soon confirmed by numerous experimental data, as reviewed by Normark et al. (1983). The idea on the multiplicity of the codes and their overlapping also appeared in later independent studies (Caporale, 1984; Kypr, 1986; Konings et al., 1987).

Essentially, the genetic sequences appear as texts designed for several different reading devices, each one seeing in it a message of its own kind. This is possible, of course, primarily due to degeneracy of the superimposed codes which allows a given letter (base, amino acid) serving a given message to be replaced by another letter. As a result, some other message is also satisfied, written by a different code. Such a design is most natural in terms of selection: if, for example, in a given sequence after some mutational changes a certain pattern emerged which causes some useful interaction (protein–DNA,
with the whole sequence context with the patterns selected earlier. Or if the sequence responsible for one function happened to be good for something else, it would serve both functions simultaneously, and there is no reason why it should not. A good descriptive term for such sequence behavior is ‘molecular opportunism’, used by Doolittle (1988) for one of the most striking examples of overlapping codes: crystallins, eye lens proteins, serve as very specific enzymes in different tissues of the same species, which have nothing to do with their physical and optical properties suitable for the eye lens.

A theoretical treatment of the multicode information systems with overlapping is not available. Such systems have no precedent either in information theory or in theories on codes. The main issue here is, perhaps, the relationship between the number of codes present in the sequence and the code degeneracies. The overlapping codes have to be degenerate to allow the adjustment of the superimposed sequence patterns to a common sequence which, thus, represents all the interacting codes simultaneously. The number of different codes overlapping in the given sequence should be a function of their degeneracy: fewer codes of low degeneracy and more codes of high degeneracy. Low-degeneracy codes, obviously, would leave only a little possibility for compromise, and the sequence would not, thus, accommodate much of other codes. The superposition of the codes should enrich the vocabulary of the sequences, i.e. to make a broader spectrum of the oligonucleotides (oligopeptides) present. Indeed, the combined vocabularies of sufficiently different overlapping codes are certainly richer than the component vocabularies. On the other hand, the condition of compromise would force the interacting codes to use all the wealth of their ‘linguistic’ arsenals, again making the final vocabulary rich. The richness of the vocabulary can be expressed as the sequence complexity, which can be defined in several different ways (Konopka, 1990; Trifonov, 1990). Contrary to the complexity approach, classical measures of information based on statistical non-uniformity of the sequences are not applicable to the multicode systems. Indeed, a highly complex sequence which carries many overlapping codes is as far off the statistical mean as a simple repeating sequence with only one code in it.

The linguistic consideration finds some confirmation in the analysis of the complexity of protein-coding and non-coding sequences (Konopka, 1990; Trifonov, 1990), and in complexity comparisons of protein sequences with human texts (Popov et al., 1996). In particular, the prokaryotic sequences, basically open for all kinds of interaction at all three levels (DNA, RNA and protein), would, perhaps, carry many overlapping sequence patterns all over their genomes, including the rather conserved protein-coding regions [as originally suggested by Holliday (1968)]. These regions would be closer to the exhaustion of their capacity to accommodate additional codes and should, therefore, have higher linguistic complexity as compared to adjacent ‘non-coding’ sequences. Indeed, this is found to be the case (Konopka, 1990; Trifonov, 1990). The human texts are relevant in the context of sequence complexity because they provide the only example available of the sequences carrying only one code (read in only one way, consecutively letter by letter). Such an example cannot be taken from biological sequences, since they all are likely to carry many overlapping messages. Comparison of the sequence complexities of the human texts and of natural protein sequences, both in 20-letter alphabets, shows that as one would expect human texts are simpler (Popov et al., 1996). This, however, can only be used as an illustration of the point on higher complexity of multicode sequences, rather than proof, because the relative simplicity of the human texts may also reflect imperfectness of human languages at this stage of their evolution.

The interaction between the codes can be studied by singling out two interacting codes of interest and analyzing the changes in the corresponding sequence patterns which are caused by the superposition. One example of such a study is known: analysis of the interaction of the gene splicing code with the triplet code (Gelfand, 1992). It is found that the codons flanking the splice junctions are biased to satisfy the sequence requirements of the splicing signals. Interestingly, the bias is different depending on the frame of the interruption by the intervening sequence.

One more study case is interaction of the gene splicing code with the chromatin code. The interaction is not yet described at the sequence level, but well demonstrated by other means. The connection chromatin–gene splicing has been originally suggested as the very essence of the phenomenon of gene splicing (Zuckerkandl, 1981). In the terminology of the interacting codes, the suggestion was that the intervening sequences had to be introduced to separate two otherwise interfering messages: triplet code and chromatin code. The exons were given primarily the protein-coding function, while the introns were loaded by the sequence requirements of the chromatin structure. The connection splicing–chromatin soon found numerous supporting evidence. The typical size of the intron–exon pairs was found to be ~205 bp which is close to the nucleosome ‘repeat’ size (Beckmann and Trifonov, 1991). The linguistic sequence complexity of exons turned out to be lower than that for prokaryotic protein-coding sequences (Konopka and Owens, 1990; Trifonov, 1991), as one would expect, indeed, in the case of spatial separation of the interacting messages in eukaryotes. The introns show higher affinity to the nucleosomes (Soloviev and Kolchanov, 1986; D. Denisov et al., in preparation) and the splice junctions are preferentially located in the central positions of the nucleosomes (D. Denisov et al., in preparation). Finally, experimental studies clearly demonstrate
that the chromatin reconstituted on cDNA (without introns) has no unique structure, contrary to the chromatin reconstituted on the genomic DNA (with introns) (Liu et al., 1995).

Yet another case of apparent interaction of different sequence patterns is periodic positional distribution of transcription factor binding sites CCAAT and GGCGGG in eukaryotic promoters (P. Bucher and E.N. Trifonov, in preparation). Since the period observed, 10.2 ± 0.2 bases, is close to the sequence period of the nucleosome DNA (Ioshikhes et al., 1992; Ioshikhes et al., 1996), the periodicity, quite likely, reflects the presence of the nucleosome immediately upstream from the TATA box. The binding sites, then, would be preferentially oriented in a specific way relative to the surface of the histone octamer, e.g. to favor (or disfavor) accessibility of the site for interaction with the transcription factor.

Concluding remarks

The regulatory sequence elements and transcription signals, in particular, should not be torn out of their contexts. The sequence they are located in carries numerous messages both related and unrelated to the transcription function. The choice of the sequence elements belonging to a given binding site is a compromise between several overlapping codes specific to this sequence region. The consensus sequence of the transcription signal is only a first approximation in the description of the signal. A more appropriate, although still approximate, description would be a matrix presentation in which preferences of various signal elements (mono-, di-, and trinucleotides) for different positions along the sequence are described (Trifonov, 1983). In most cases, none of the elements individually is necessary or sufficient for the recognition, very much like individual voters in elections. The recognition is distributional (Trifonov, 1983), such that the decision is made after the weights of different signal elements present in a given site are somehow combined. The distribution of various signal components along the sequence and the use of alternative components at the same position reflects the multiplicity of the codes overlapping with the given recognition sequence and the compromise, yielding to one another in order to co-exist, as much as their degeneracies allow.

References


