Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Review

Thirty Years of Multiple Sequence Codes

Edward N. Trifonov[1,2*]

[1]*Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel;*
[2]*Division of Functional Genomics and Proteomics, Faculty of Science, Masaryk University, Brno 61137, Czech Republic.*

Genomics Proteomics Bioinformatics 2011 Apr; 9(1-2): 1-6 DOI: 10.1016/S1672-0229(11)60001-6

Received: Sep 15, 2010; Accepted: Dec 09, 2010

Abstract

An overview is presented on the status of studies on multiple codes in genetic sequences. Indirectly, the existence of multiple codes is recognized in the form of several rediscoveries of Second Genetic Code that is different each time. A due credit is given to earlier seminal work related to the codes often neglected in literature. The latest developments in the field of chromatin code are discussed, as well as perspectives of single-base resolution studies of nucleosome positioning, including rotational setting of DNA on the surface of the histone octamers.

Key words: genomic code, chromatin code, second genetic code, code interaction, code degeneracy

Introduction

Discovery of the triplet code 40 years ago has been an event of such intellectual power that nobody, even with very open mind, could then imagine that there could be something else in the sequences, despite the irritating fact that the triplet code turned out to be degenerate. That is, the code allows for some DNA sequence variations, while the protein sequence stays unchanged. The hypnotic effect of the discovery lasted very long, to such a degree that even 20 years after, when it became clear that the genomes are full of sequences not encoding proteins, some researchers (*1, 2*) gave the bulk of genomic sequences the role of "selfish" DNA, just a minimal role, not to dismiss the "non-coding" sequences completely. That further fueled the already firmly rooted belief in the almost divine singularity of the triplet code. The first very much unnoticed voice that sounded like heresy was the speculation by Holliday that, perhaps, protein-coding sequences may simultaneously harbor here and there another message—recombination signal (*3*). Later suggestions, on the special role of G+C composition (genomic code) (*4, 5*) and on the chromatin code overlapping with the triplet code (*6, 7*), also have been barely afloat for many years. The announcement of the second genetic code (*8, 9*) was a first crack in the ice of dogma. The very thought that there could be one more code, was a revelation. In the meantime it became clear that the sequences that do not code for proteins—non-coding sequences—actually do code for numerous small functional RNAs, the first of which, U-RNA, was discovered in 1979 (*10*). Since then a swarm of other small non-coding RNAs with various largely regulatory functions have been discovered (*11*), making more visible the coding functions of the "non-coding" DNA sequences.

Thus, apparently, there are some other codes in the sequences, in addition to the triplet code. But to say that there are many codes would still sound too unorthodox. Here we present an overview on the status of

*Corresponding author.
E-mail: trifonov@research.haifa.ac.il

studies on multiple codes in genetic sequences.

It should be noted that the term "code" is often used in literature in a very broad sense, as a biosemiotic term (*12*). In this overview we concentrate specifically on the sequence codes, as distinct patterns carried by DNA, RNA and protein sequences.

How Many "Second Genetic Codes"?

According to the media sympathetic to science and enthusiastic about sensational discoveries, the "Second Genetic Code" as it was called by *New York Times* (*8*) was discovered by Ya-Ming Hou and Paul Schimmel and published in *Nature* in 1988 (*9*). It was about recognition of tRNAs by respective aminoacyl-tRNA synthetases. Thirteen years later *New Scientist* announced the second Second Genetic Code (*13*), discovered by Jenuwein and Allis (*14*) and published in *Science*. This time it was about histone modifications. Five years later, *New York Times*, again, reported about "a second code in DNA in addition to the genetic code" (*15*). This was already the third Second Genetic Code, discovered by Segal *et al* (*16*), suggesting now nucleosome positioning rules. One, surely, would raise eyebrows having learned that there is also the fourth Second Genetic Code (*17*)—on interaction specificities between proteins and DNA, and the fifth Second Genetic Code, the name given by *Nature* magazine (*18*) to the set of rules governing gene splicing (*19*). Bewildered reader, naturally, would say "I'm done with seconds, can I have a third?" (*20*)

The conclusion from the above is obvious: one has to admit that the genetic sequences carry many different codes. If we are to know what the sequences are about, we have to detect and decipher these codes. The times of surrender to "junk" and "selfish DNA" are over, and "non-coding" becomes a misnomer.

Resetting Clocks

The repeated announcements of every new code as the second one also demonstrate sort of a collective amnesia, like in a chain car accident. The actual succession of events is scrambled, and some are, indeed, forgotten.

Historically, the first genetic code discovered and described is, of course, the classical triplet code of translation of RNA sequence into protein sequence. The first non-triplet code, in the form of significant biases in G+C compositions of large sections in eukaryotic genomes (isochores), was described in 1973 by Bernardi's group (*4*). Later, with accumulation of data on biological involvement of the isochores, this was called a genomic code (*5*). The code exerts compositional pressure on all types of sequences and on all three codon positions of the protein-coding sequences. As this is a sequence bias reflecting biologically relevant features, it fits well to the definition of the sequence code.

The chromatin code—sequence periodicity responsible for nucleosome positioning—was discovered in 1980 by Trifonov and Sussman, and was first described as a code by Trifonov (*6, 7*). The first description of the main sequence features of the gene splicing code was published by Breathnach and Chambon (*21*). They were later called "Chambon rules".

These are the codes most frequently discussed in literature and continuously explored since they have been discovered. Two of them, the chromatin code and the gene splicing code, received the same honorary title of "Second Genetic Code", under fresh authorships. The term "genomic code" has been borrowed for a different meaning (*16*). The Chambon rules did not appear in the complex description of the fifth Second Genetic Code (*19*).

In the chronology of events, the notion of multiple overlapping codes has been known since 1980, when it was understood that the chromatin code and the triplet code coexist in the same sequences so that many bases of the sequences simultaneously serve at least two different codes (*6, 7*). The multiplicity and overlapping (interaction) of the sequence codes were later addressed in 1989-1996. A sequence code was defined as "any pattern or bias in the sequence which corresponds to one or another specific biological (biomolecular) function or interaction" (*22, 23*).

The sequence codes can be naturally divided into three groups: (1) the codes related to DNA functions and structures, such as genomic code (isochores), DNA shape (curvature) code, and chromatin code; (2) codes of RNA level—triplet code, gene splicing code;

and (3) protein level codes—folding rules, N-end rule and others (*22-24*).

The true interest of scientific community to the variety of the sequence codes and their interactions is, surely, coming and new large-scale efforts are anticipated to further unravel the multiplicity of the codes, carried by the genomic sequences.

Latest Developments on Chromatin Code

An apparent finale

One of the most significant recent breakthroughs in the genetic cryptology is the final establishment of the universal nucleosome positioning pattern (chromatin code), derived by three independent approaches—analysis of nucleosome DNA sequences (*25*), deformational properties of the dinucleotide stacks of nucleosome DNA (*26*), and Shannon N-gram extention for genomic sequences (*27, 28*).

The pattern derived from the nucleosome DNA sequence database of *C. elegans* (*29*), CGRAAATTTYCG (*25*), is identical to what physics of DNA deformation suggests, and very close to CRAAAATTTTYG, derived by Shannon extension.

The chromatin code is well hidden in the sequences. Essentially, the nucleosomes should not be very strong, not to hinder too much the template processes. On the other hand, the code has to be rather degenerate, not to interfere with other messages carried by the sequences. As a result, good half of numerous attempts, since 1980, to derive the pattern, with rather diverse suggested solutions, failed to make it to finale (*22*). The survivors are patterns with alternating dinucleotides AA/TT (*6, 30*), RR/YY (*31*), WW/SS (*32*) (here W stands for A and T, while S stands for C and G) and CC & GG (*33*). They all reflect various sequence features of the final pattern. The nucleosome mapping motif suggested as chromatin code in the above-mentioned publication of Segal *et al* (*16*) is, essentially, the WW/SS motif (*32*).

The cracking of the chromatin code became possible when a large database of the nucleosome DNA sequences appeared in 2006 (*29*). The mere size of it substantially improved statistics of the signal, so that it became, finally, fully visible (*25*).

No code at all?

The perpetuating failures of describing the complete nucleosome positioning sequence pattern have lead, naturally, to the development of views on the positioning that, essentially, does not depend on the sequences. A physically attractive model has been suggested by Kornberg and Stryer (*34*) according to which mere presence of some barriers for free nucleosome sliding would cause preferential nucleosome positioning close to the barriers. Several papers are devoted to this model (*35*). This component of the nucleosome positioning should be best seen in those cases where the sequence does not provide sufficiently strong signal. For example, human and mouse genomes display only very weak sequence periodicity that actually is invisible in case of mouse (*36*). The positioning in these cases would be expected to be influenced by occasional strong nucleosomes ("barriers"). Alu-sequence containing nucleosomes may serve as such chromatin organizing "barrier" nucleosomes (*37*).

It is about physics

The CGRAAATTTYCG pattern reflects deformational properties of DNA and indicates where various dinucleotide elements should be positioned within the period of DNA wrapped around histones, to make the bending energetically less expensive (*26*). In particular, the central dinucleotide AT should be positioned at the dyad axis of the structural DNA repeat, in the minor groove oriented outwards. Respectively, the CG elements, five bases apart from AT, should be positioned at minor grooves contacting the surface of the histone octamer. The presence in the DNA sequence of the repeating pattern in its complete form would be, of course, very good for the bending of DNA. That would make an ideally deformable DNA, but then this unique DNA sequence with no degeneracy would serve only one purpose—to make a very stable nucleosome. Reality of multiple overlapping degenerate codes in the same sequence allows in every nucleosome DNA sequence only few favorably posi-

tioned (oriented) dinucleotides, just about enough to suggest where along the molecule it would be easier to form the nucleosome. Actual nucleosome DNA sequences are very far from being a full match to the ideal repeating bendability pattern. It is only a resemblance (usually rather weak) of a given sequence to the standard positioning pattern, which causes the preferential binding of the histone octamer to the sequence. The physics of DNA, thus, is reflected (encoded) in the sequence. Comparing (aligning) real sequence with the "nucleosome probe", periodically repeating motif $(GRAAATTTYC)_n$, one would find those few dinucleotides that match, and whether their amount is higher than anywhere in close vicinity. Displacement of the sequence by only one base would result in rotation by 34° (the structural period of the nucleosome DNA is 10.4 base pairs) of all those favorably oriented base pair stacks, away from the optimum. That, of course, would result in corresponding loss of the DNA bendability. This is a physical basis of high-accuracy mapping of the nucleosome sequence positions (*38, 39*).

The nucleosome DNA bendability pattern CGRAAATTTYCG is derived from very basic deformational properties of DNA (*26*), without detailed energy calculations. Such calculations are highly desirable as they may eventually be able to evaluate directly the deformation energy of DNA for any sequence of interest. That could provide energy-based estimates of the amplitudes in the calculated nucleosome maps, perhaps, more accurate than current indirect procedure, via sequence similarity to the pattern outlined above. One promising computational approach dealing with the deformational details of all base pair stacks is the one pursued by Zhurkin's group (*40-42*). This approach has good chances to arrive to yet another, independent, formulation of the pattern, with single-base accuracy mapping capability. It is likely to be very similar to the one suggested above as an ultimate solution. Indeed, the detailed deformational analysis of YR steps in the nucleosome DNA demonstrated as well that these dinucleotides are located preferentially in minor grooves facing the histone octamer (*42*). Interestingly, energy minimization calculations (*43*) for DNA curvature motifs arrive to the same pattern CGAAAATTTTCG, for which an appreciable curvature is also experimentally observed (*44*). The maximal curvature sequence known, however, is different—CGGCAAAAAACG (*45, 46*). The DNA bendability pattern and intrinsic DNA curvature pattern are two different sequence codes, based on different physics.

Towards highest resolution studies

Scientific community is, thus, in possession of a reliable tool, subject of small future modifications, sequence-directed mapping of the nucleosomes, based on the pattern of DNA bendability. Publicly available server (http://www.cs.bgu.ac.il/~nucleom/) (*39*) allows determination of nucleosome positions in any given sequence in few seconds, with uncertainty of only one base. Considering the time/resource loads of experimental techniques, it is a definite technical breakthrough. Note that the most popular technique today, MNase digestion and mapping by sequencing, has accuracy of 10-15 bases (*29*) and, thus, does not allow to determine rotational setting of the located nucleosomes, providing the "occupancy" profiles only.

One example of application of the new nucleosome mapping technique is the discovery of protective positioning of gene splice junctions within the nucleosomes (*47*). Not only the junctions are more often positioned within the nucleosomes, but the GT and AG dinucleotides at the ends of the introns are preferentially oriented in such a way that C8 and N9 atoms of guanine residues are located closest to the histone octamers, minimizing chance of attack by free radicals and aggressive metabolites. More such fine structure analyses of chromatin are in progress.

Chromatin research today is on unprecedented rise. Especially involved is the part of scientific community for which the sequencing and high-throughput analyses are an everyday routine. A credit for the recent boost of the broad interest to chromatin studies should be given, in particular, to the above-mentioned paper by Segal *et al* (*16*). Regretfully, however, the computer partners of the efforts serve the low-resolution data, without even trying to step on high-resolution path. Various genomes and even sequence types may well have their specific biased versions of the basic DNA bendability pattern. Development of many

appropriate species-specific sequence-directed nucleosome mapping routines would be then a must, to the benefit of high quality and resolution chromatin research.

The triplet code has been for many years one of the major topics in life sciences. Today the chromatin code takes its share. And since it is already clear that the genetic sequences carry many more messages, the other codes are at the doors.

Acknowledgements

This work was supported by Grant 222/09 of Israel Science Foundation, and by Fellowship of SoMoPro (South Moravian Program, Czech Republic) with financial contribution of European Union within the seventh framework program (FP/2007-2013, Grant No. 229603).

References

1 Doolittle, W.F. and Sapienza, C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601-603.
2 Orgel, L.E. and Crick, F.H. 1980. Selfish DNA: the ultimate parasite. *Nature* 284: 604-607.
3 Holliday, R. 1968. Genetic recombination in fungi. In *Replication and Recombination of Genetic Material* (eds. Peacock, W.J. and Brock, R.D.), pp. 157-174. Australian Academy of Science, Canberra, Australia.
4 Filipski, J., *et al.* 1973. An analysis of the bovine genome by Cs_2SO_4-Ag^+ density gradient centrifugation. *J. Mol. Biol.* 80: 177-197.
5 Bernardi, G. 1990. Le génomes des vertébrés: organization, fonction et evolution. *Biofutur* 94: 43-46.
6 Trifonov, E.N. 1980. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.* 8: 4041-4053.
7 Trifonov, E.N. 1981. Structure of DNA in chromatin. In *International Cell Biology 1980-1981* (ed. Schweiger, H.), pp. 128-138. Springer, Berlin, Germany.
8 Kolata, G. 1988. Second genetic code deciphered, solving a protein synthesis puzzle. *New York Times* May 13.
9 Hou, Y.M. and Schimmel, P. 1988. A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* 333: 140-145.
10 Lerner, M.R., *et al.* 1980. Are snRNPs involved in splicing? *Nature* 283: 220-224.
11 Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* 2:919-929.
12 Barbieri, M. 2003. *The Organic Codes. An Introduction to Semantic Biology.* Cambridge University Press, Cambridge, UK.
13 Young, E. 2001. Packaging proteins may be second genetic code. *New Scientist* August 9.
14 Jenuwein, T. and Allis, C.D. 2001. Translating the histone code. *Science* 293: 1074-1080.
15 Wade, N. 2006. Scientists say they've found a code beyond genetics in DNA. *New York Times* July 25.
16 Segal, E., *et al.* 2006. A genomic code for nucleosome positioning. *Nature* 442: 772-778.
17 Hughes, T. 2008. Cracking the second genetic code. *FASEB J.* 22: 262.2.
18 Tejedor, J.R. and Valcárcel, J. 2010. Gene regulation: breaking the second genetic code. *Nature* 465: 45-46.
19 Barash, Y., *et al.* 2010. Deciphering the splicing code. *Nature* 465: 53-59.
20 Smith, T. 2010. FinchTalk. *Journal Club* May 11.
21 Breathnach, R. and Chambon, P. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* 50: 349-383.
22 Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. *Bull. Math. Biol.* 51: 417-432.
23 Trifonov, E.N. 1996. Interfering contexts of regulatory sequence elements. *Comput. Appl. Biosci.* 12: 423-429.
24 Trifonov, E. N. 1999. Sequence codes. In *Encyclopedia of Molecular Biology* (ed. Creighton, T.E.), pp. 2324-2326. John Wiley & Sons, New York, USA.
25 Gabdank, I., *et al.* 2009. Nucleosome DNA bendability matrix (*C. elegans*). *J. Biomol. Struct. Dyn.* 26: 403-411.
26 Trifonov, E.N. 2010. Base pair stacking in nucleosome DNA and bendability sequence pattern. *J. Theor. Biol.* 263: 337-339.
27 Trifonov, E.N. 2010. Nucleosome positioning by sequence, state of the art and apparent finale. *J. Biomol. Struct. Dyn.* 27: 741-746.
28 Rapoport, A.E., *et al.* 2011. Nucleosome positioning pattern derived from oligonucleotide compositions of genomic sequences. *J. Biomol. Struct. Dyn.* 28: 567-574.
29 Johnson, S.M., *et al.* 2006. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* 16: 1505-1516.
30 Trifonov, E.N. and Sussman, J.L. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA* 77: 3816-3820.
31 Mengeritsky, G. and Trifonov, E.N. 1983. Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Res.* 11: 3833-3851.
32 Satchwell, S.C., *et al.* 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191: 659-675.
33 Bolshoy, A. 1995. CC dinucleotides contribute to the bending of DNA in chromatin. *Nat. Struct. Biol.* 2: 446-448.

34 Kornberg, R.D. and Stryer, L. 1988. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* 16: 6677-6690.

35 Vaillant, C., *et al.* 2007. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.* 99: 218103.

36 Bettecken, T. and Trifonov, E.N. 2009. Repertoires of the nucleosome-positioning dinucleotides. *PLoS One* 4: e7654.

37 Salih, F., *et al.* 2008. Epigenetic nucleosomes: Alu sequences and CG as nucleosome positioning element. *J. Biomol. Struct. Dyn.* 26: 9-16.

38 Gabdank, I., *et al.* 2010. Single-base resolution nucleosome mapping on DNA sequences. *J. Biomol. Struct. Dyn.* 28: 107-122.

39 Gabdank, I., *et al.* 2010. FineStr: a web server for single-base-resolution nucleosome positioning. *Bioinformatics* 26: 845-846.

40 Tolstorukov, M.Y., *et al.* 2007. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.* 371: 725-738.

41 Cui, F. and Zhurkin, V.B. 2010. Structure-based analysis of DNA sequence patterns guiding nucleosome positioning *in vitro*. *J. Biomol. Struct. Dyn.* 27: 821-841.

42 Wang, D., *et al.* 2010. Sequence-dependent Kink-and-Slide deformations of nucleosomal DNA facilitated by histone arginines bound in the minor groove. *J. Biomol. Struct. Dyn.* 27: 843-859.

43 Lafontaine, I. and Lavery, R. 2000. Optimization of nucleic acid sequences. *Biophys. J.* 79: 680-685.

44 Hagerman, P.J. 1986. Sequence-directed curvature of DNA. *Nature* 321: 449-450.

45 Koo, H.S. and Crothers, D.M. 1988. Calibration of DNA curvature and a unified description of sequence-directed bending. *Proc. Natl. Acad. Sci. USA* 85: 1763-1767.

46 Bolshoy, A., *et al.* 1991. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA* 88: 2312-2316.

47 Hapala, J. and Trifonov, E.N. 2010. Protective rotational positioning of splice junctions in nucleosomes. In Abstracts of *Alternative Splicing—Special Interest Group Meeting*, pp. 51. Boston, USA.