Molecular codes in biology: the genetic code and beyond

Annick Lesne

Molecular codes in biology: the genetic code and beyond

Annick Lesne

Sorbonne Université, CNRS, Laboratoire de Physique Théorique de la Matière Condensée, LPTMC, F- 75252, Paris, France & IGMM, UMR5535, University of Montpellier, CNRS, F-34293, Montpellier, France

(January 30, 2011)

Keywords: adaptor, allostery, genetic code, gratuity
Abbreviations: AARS, aminoacyl-tRNA synthetase; mRNA, messenger RNA; PTM, post-translational modification; TF, transcription factor; tRNA, transfer RNA.

Abstract

The trend to call a code any correspondence between molecular entities requires clarification. Taking as a basis the reference biological code, the genetic code, we argue that codes are defined by the existence of adaptors that recognize separately symbols, also termed *codewords,* and objects to be encoded. The genetic code actually involves two nested adaptors: the enzymes aminoacyl-tRNA synthetases bridge amino acids and tRNAs into aminoacyl-tRNAs, which in turn bridge codons and amino acids, thus defining the genetic code. Natural and synthetic variants demonstrate that today the genetic code is arbitrary. In contrast, the correspondence between codons and anticodons is a mere physicochemical pairing rule. Among proposed molecular codes (sequence codes, the histone code, the transcriptional regulatory code) few if any qualify as a code. Calling codes simple pairing rules, contingency tables, or puzzles is misleading. Faced to a potential code, the agenda is to investigate whether coevolved adaptors establishing an arbitrary correspondence between objects and codewords can be evidenced.

Introduction

Since the discovery of the genetic code in the 60's [1-4], a series of papers have introduced other codes in a genomic context: sequence codes [5, 6], nucleosome codes [7, 8], or a splicing code [9-12]. These codes are claimed to encode in genomic sequences more than protein sequences. Other proposed codes include histone codes [13-17], transcriptional regulatory codes [18-21], epigenetic codes [22-24], chromatin codes [19, 25-27] and several other ones. For these codes, codewords (the entities encoding the biological objects of interest) are post-transcriptional modifications of histones and transcription factors (TFs). This profusion has led to a questioning, if not a confusion, about what a biological code is. Are the associated mappings all codes, or just rules, e.g. pairing rules? Is the use of the word 'code' a mere metaphor or does it involve a meaningful analogy with the notion of code in either semiotics or information theory?

In the plain language, the word has several meanings. The word originates from the latin *codex*, a book as opposed to a papyrus or parchment roll, typically used to gather a set of laws. This use of the word is still encountered today in law for instance as a penal code or a highway code. We also speak of a computer code (by which the computer is ascribed to make a computation), of a zip code (designating a label), of the Morse code (a conventional representation of a spoken message devised for easy communication by telegraphy), or of the once popular Da Vinci code (subtending the idea of a cryptic knowledge to be unraveled). The use of the term in biological literature [28] may allude to any of these meanings. The word code is currently used in journals or article titles to attract attention without any reference to the notion of code that has been developed in semiotics and information theory. It is often used to designate just a rule or even a puzzle.

In order to clarify the notion of *biological code*, I will first show that the universally accepted one, the

genetic code, is indeed a code according to semiotics, while this is not necessarily the case for tables of correspondence or association rules between two sets of molecular objects. Within this framework, I will review the wealth of molecular codes encountered in the literature and show that this attractive designation is often inappropriate, and can therefore be misleading as to the real nature of the underlying biological processes.

I will restrict this study to *molecular codes* without entering the realm of codes in neuroscience, namely neural codes, retinal codes, or odor codes [29-32], which encode input signals in the features of neuronal spike trains (the codewords) rather than in molecular entities.

A set of binary relationships can be termed a code when mediated by an adaptor

In semiotics, a code is defined by i) a discrete set of *objects* (the *signified)* to be encoded, ii) a set of symbols, termed the *signifiers* or simply the *codewords* and iii) the *conventional* mapping according to which codewords represent objects. In a code, distinct mappings can be imagined without any particular one being imposed by the laws of nature. Semiotics explicitly underlines that the choice of one mapping among the possible ones endows the code with a semantic content. These basic conceptual developments date back to Peirce, considering a code as the triplet of a sign, and object and an interpretant [33]. In a biological context, these notions have been adapted into biosemiotics, emphasizing the role of organic codes in establishing conventional rules between different worlds [34-37]. This notion of code meets that of *gratuity* introduced by Monod [38] as an essential feature of many relationships within living systems, the two exemplary instances being the genetic code and allosteric enzymes [39-41].

Following the terminology introduced by Crick [42] for aminoacyl-tRNAs (see below), I propose to call *adaptor* a macromolecule capable of recognizing both the codeword and the encoded object *at two different sites* (Fig. 1A). By its very existence, the adaptor bridges two molecular pathways or entities, circumventing the need of direct physical, chemical or stereochemical interactions between them. A code is therefore arbitrary in that the design of adaptors is a product of molecular coevolution.

A major consequence of code arbitrariness is that the mapping between the codewords and the encoded objects can be modified *a posteriori* by modifying the adaptor (Fig. 1B, C). We cannot say that we have a code if a change in the laws of physics is required to change the mapping. For instance, it is tempting to say that the atomic number *Z* of an element encodes its chemical properties. But all the features of the element can be (in principle) predicted from the laws of quantum chemistry from the knowledge of *Z*, with no way to be tuned. The mapping can be summarized in an efficient way (the Mendeleev periodic classification) but it is not a code. Similarly, despite their central role in the replication and expression of genetic information, Watson-Crick pairing rules between the nucleotides composing each single strand of DNA (i.e. adenine A pairs with thymine T and guanine G pairs with cytosine C) are not codes since they follow from physicochemical laws in a computable way. Another example illustrating the evolvability of a true code, as prescribed by an adaptor (Fig. 1B, C), is given by labels and barcodes**.** The barcode is a mediator allowing changing the price of an object without manipulating it. This structure, where the label on the object indicates the barcode is to be contrasted to the case where the label indicates directly the price. The latter is not a code, whereas the barcode indeed deserves to be termed a code.

Figure 1 --- General scheme of a code: the notion of adaptor.
A code is defined as a conventional correspondence between a set of objects and another set of entities, termed symbols or more generally codewords. ***(A)*** *The arbitrariness of the code lies in the independence of the codewords from the objects to be encoded. The correspondence is established by means of an adaptor recognizing both the object and the codeword at different sites (respectively blue and red); in this respect the adaptor is an allosteric entity. Evolving the adaptor allows* ***(B)*** *accommodating a change in the object without modifying the codeword, or* ***(C)*** *establishing variants of the code, where objects are encoded with other codewords.*

The genetic code provides a rich reference case

In a biological context, the genetic code is emblematic of what a code is. Its discovery began with the experimental unraveling of an association between each of the 20 amino acids encountered in natural proteins and one or several 3-nucleotide sequences termed codons. This association is now called the *codon table* [4]. In the fifties, most scientists were assuming direct recognition and docking of amino acids onto the messenger RNA template [43, 44]. The first step towards demonstrating the existence of a code was the proposition by Crick et al. that an intermediary entity, and not the amino acid, fitted onto the nucleotidic template [45]. They envisioned a code in which constraints of unambiguity of the reading frame would have reduced the number of acceptable codons to exactly 20. Despite this remarkable numerical coincidence, experiments lead to reject their hypothesis: 61 among the 64 codons are actually associated with an amino acid (the three remaining ones are stop codons). However, Crick followed up on the vision of an adaptor, making possible to escape a direct and obligatory physicochemical interaction between the codon and the amino acid. He proposed that each amino acid is first attached to its own specific piece of nucleic acid (now known as transfer RNA, tRNA), in an enzyme-catalyzed reaction [42]. The adaptor hypothesis was rapidly validated experimentally [46].

With respect to the definition of a code, amino acids are the objects and codons are the codewords. In addition, the genetic code has two combinatorial aspects, since i) codons are 3-letter words constructed from an alphabet composed of the four nucleotides A, U, G, C (uracil U replacing thymine T in RNAs) and ii) these words are then assembled into a linear messenger RNA (mRNA). The genetic code is degenerate: most amino acids are associated with several codons, up to six for serine, leucine and arginine. Very rare cases of ambiguous decoding are known [47] e.g. ambiguity between serine and leucine in some *Candida* species [48, 49]. The degeneracy of the genetic code can be described quantitatively in the framework of information theory [50].

Figure 2 --- The genetic code and its adaptor: aminoacyl-tRNA.
The figure sketches the molecular parts involved in the correspondence between a codon (in blue, embedded into a messenger RNA) and the cognate amino acid (in magenta). The adaptor defining the genetic code and ensuring its arbitrariness and evolvability is an aminoacyl-tRNA, namely a transfer RNA reversibly charged at the acceptor end (in green) with an amino acid (ester linkage with an hydroxyl group of the terminal adenosine) and embedding an anticodon (in red). The recognition and pairing between the codon and the anticodon are fully determined by physical and stereochemical rules. They are possibly tuned by tRNA editing, e.g. modification of the first nucleoside in the anticodon or chemical modifications distorting the anticodon loop conformation. The bare tRNA is first charged with the amino acid, by means of the catalytic action of a dedicated aminoacyl-tRNA synthetase (see Fig. 3), then both the association of aminoacyl-tRNAs with the mRNA template and the assembly of successive amino acids into a polypeptide chain take place within the ribosome

Codon-anticodon pairing obeys physical necessity but can be modulated by tRNA editing

The recognition of a codon by a tRNA is based on the presence, in the tRNA sequence, of a corresponding nucleotidic triplet termed the 'anticodon' (Fig. 2). Codon/anticodon pairing is not a code: it follows from a physicochemical interaction that can in principle be predicted *ab initio* given the codon and tRNA anticodon loop molecular structure. The degeneracy with which a single aminoacyl-tRNA recognized several codons is often explained by the presence in the anticodon, at the first position, of a modified nucleoside, inosine, able to pair with either U, A or C in the third position in the codon, what has been termed the *wobble interaction* [3, 51]. This is an instance of *tRNA editing*, that is, post-transcriptional modification of the tRNA [52-54]. Anticodon loop and stem shape the anticodon conformation which in turn affects codon recognition and thus the reliability and efficiency of translation [55, 56].

Nature today achieves the bridge between amino acids and codons in two steps: a symbolic association between amino acids and anticodons mediated by *aminoacyl-tRNAs* [57] (Crick's adaptors) and a physical pairing between codons and anticodons (Fig. 2). The genetic code thus comprises both an arbitrary part and part entirely ruled by physical laws. Overall, there is no direct and obligatory (i.e. physical, chemical or stereochemical) relationship between codons and amino acids: the correspondence is indeed a code.

The genetic code relies on the nested action of two adaptors

The specific aminoacylation of tRNAs relies on assignment enzymes, the *aminoacyl-tRNA synthetases* (aaRSs) [40, 41, 58]. It constitutes a second code nested in the genetic code: the *acceptor code* [58-60], bridging bare tRNAs and their cognate amino acids into aminoacyl-tRNAs [57]. An aaRS is uniquely associated with an amino acid [61]. Like codon-anticodon pairing, the relationship between aaRS and the cognate amino acid is a binding *rule*. Arbitrariness of the acceptor code lies in the contingent presence within a unique entity, aaRS, of the amino-acid recognition site and the tRNA recognition site (Fig. 3).

Recognition of tRNAs by aaRSs involves the tRNA acceptor stem and possibly the anticodon loop (Fig. 3). The mechanism has been dissected using truncated parts of tRNA excluding the anticodon [60]. aaRSs are divided into two classes, according to the attachment site (hydroxyl group) of the amino acid to the terminal adenosine of tRNAs, and exhibiting very different features. These two classes may correspond to two different evolutionary origins [47].

Synthetases of class I, represented by Ala-RS (alanyl tRNA synthetase), are composed of one or two subunits and recognize only the acceptor stem [60, 62]. For the alanine system, the rate of charging tRNA is the rate of charging the acceptor stem, proving the absence of any contribution of the anticodon loop (Fig. 3A); cross-linking experiments demonstrate that there is no contact between the aaRS and the anticodon loop. Transfer of the Ala-tRNA determinant, namely the base pair G3-U70 in the acceptor stem, to other tRNAs confers alanine acceptance on them [58, 63, 64].

Synthetases of class II, represented by Gln-RS (glutaminyl tRNA synthetase), are composed of two or four subunits and recognize both anticodon loop and the acceptor stem [60, 65]. In the glutamine system, recognition of the bare tRNA by Gln-RS involves both anticodon loop and acceptor stem (Fig. 3B), and the rate of charging tRNA is five orders of magnitude larger than the rate of charging the truncated acceptor stem [60, 65].

Recognition or aminoacylation sites and template-reading anticodon loop are spatially segregated both along the tRNA sequence and in the 3-dimensional L-shaped tRNA conformation. The two tRNA domains interact with different parts of the ribosome [60]. These two facts support an evolutionary design of tRNA involving two domains of different origins. Their co-occurrence in a single entity has no necessity; it is presumably a product of molecular evolution.

AaRSs embed amino acid recognition and adenylation site, RNA recognition sites, and a non-specific active site recognizing tRNA acceptor end (a 4-nucleotide sequence NCCA) and catalyzing ester linkage of the amino acid on the terminal adenosine. All tRNAs embed the anticodon, identity determinants in the acceptor stem and possibly in the anticodon loop, and a non-specific acceptor site

in its 3' end. The arbitrariness of the genetic code comes from this coexistence of functionally different domains within each adaptor: the nature and characteristics of this code comes from the very existence of aaRSs and tRNAs, as they are.

Modifications of the adaptors allows the evolution or artificial design of code variants

Various alternative or expanded associations between codons and amino acids have been observed e.g. in mitochondria [66] or *Candida* species [48]. These *natural variants* follow from tRNA editing changing the codon-anticodon pairing rule [52, 67], from a modification of tRNA acceptor stem letting it be charged by another aaRS [64], or from the evolution of an aminoacyl-tRNA targeting a stop codon and standard or non-standard amino acids (selenosysteine and pyrrolysine) [4, 49, 55]. The presence of code variants supports the conventional nature and coevolutionary (*sensu* Darwin) origin of the genetic code adaptors [68]. The demonstration has been completed by the artificial design of code variants, allowed by a dedicated modification of the adaptors [69]. *Synthetic code variants* are achieved by modifying an aaRS so that it binds another or a modified amino acid [70], or by modifying tRNA anticodon or tRNA recognition sites for aaRS by site-directed mutagenesis. They are used to probe the structure and function of existing proteins or to devise novel proteins, e.g. for therapeutic purposes [69]. They testify that we have properly understood the code determinants and its flexible adaptors.

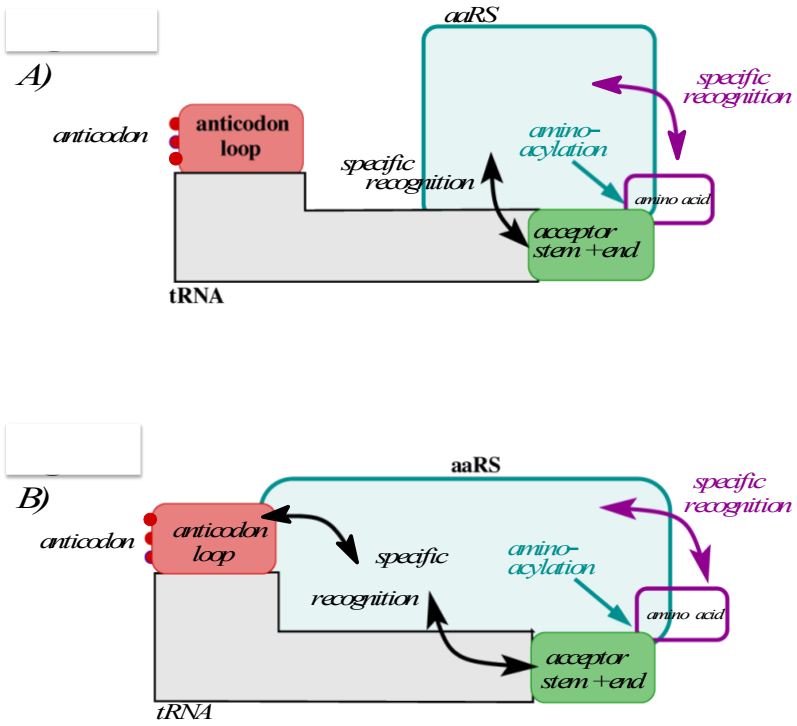

Figure 3 --- tRNA aminoacylation: the acceptor code
The figure sketches the process by which bare tRNAs are charged with their cognate amino acid to form aminoacyl-tRNAs establishing the genetic code by their very existence (see Fig. 2). The anticodon loop (in red) and acceptor stem (in green) are functionally and spatially segregated in the 3-dimensional L-shaped tRNA structure. The correspondence between anticodons and amino acids is actually a code, the acceptor code, whose adaptors are enzymes, aminoacyl-tRNA synthetases (aaRSs). Each enzyme recognizes and binds specifically a single amino acid (magenta arrow). As such, there are 20 different aaRSs, belonging to two classes. Two cases are encountered. ***(A)*** *tRNA recognition by the aaRS relies entirely on determinants in the acceptor stem (black arrow). It does not involve the anticodon loop (actually the aaRS does not overlap the anticodon loop). This is the case of e.g. the alanine system (class I aaRS).* ***(B)*** *the aaRS recognizes specifically tRNA determinants in both the acceptor stem and the anticodon loop (black arrows). This is the case of e.g. the glutamine system (class II aaRS).Then* ***(A, B)*** *aaRS recognizes the universal acceptor end NCCA where N is any nucleotide (see Fig. 1), and catalyzes the ester linkage of the amino acid with an hydroxyl group of the tRNA terminal adenosine (cyan arrow).*

All recognition and binding processes follow from physical, chemical or stereochemical necessity. The definition and arbitrariness of the acceptor code lies in the very existence of aaRS and its dual recognition, at distant sites, of the amino acid and the tRNA determinants. The genetic code thus relies on two nested adaptors: aaRS then aminoacyl-tRNA. (For visual clarity, the dimensions do not match reality: the aaRS, composed of hundreds of residues, is far larger than the amino acid, which is itself smaller than the anticodon).

Code origin and optimality are distinct from its symbolic nature

Exploring the origin of the genetic code and the evolutionary path, if any, towards the present situation is another issue [71, 72]. Neither the hypothesis of a primeval stereochemical bias in the relation between codons and amino acids [47, 68] nor the possibility of an evolved optimization [41, 49, 73, 74] question the current arbitrariness of the code [40]. Today, the association between codon and amino acids is solely ruled by the existence of aminoacyl-tRNAs. In the linguistic metaphor, the fact that a word etymologically originates in an onomatopoeia does not negate that it is today used conventionally within a symbolic language. Once the adaptors have been devised, the correspondence between codewords and objects is mediated in a gratuitous way, and does not depend on any direct interaction or affinity between them. Modifications of the adaptors yield variants of the code.

The notion of adaptor relates with those of allostery and signal transduction

A code relates to the notion of allostery insofar as adaptors embed two spatially and functionally segregated domains. They result from the evolved association of these domains within a single entity [39]. Allosteric causality thus goes far beyond what could be generically obtained from physical, chemical and stereochemical necessity. It embeds evolution within molecular processes. Allosteric enzymes with segregated effector and catalytic domains, hormonal responses or transduction pathways are codes, though trivial ones, with one object and one codeword. However, they present the required coevolved arbitrariness to speak of a code.

Sequence codes typesetting the genome are only physical rules and statistical associations

We can now discuss the relevance of calling codes various mappings between genomic sequences and biological entities or events, beyond the genetic code. Discussion is summarized in Table I, in annex.

Protein-DNA recognition should not be termed a code [6, 75] since it is entirely prescribed by DNA binding energy landscapes (one landscape along DNA for each protein). It can be (at least in principle) computed *ab initio* [76]. The mapping between proteins and their genomic binding sites is a nicely summarized set of physical, chemical and stereochemical rules, but it is not a code for the cell.

The *nucleosome code* describes how the bendability of the DNA molecule and ensuing nucleosome positioning are determined by the genomic sequence (basically due to the strength of A-T versus G-C pairings) [7, 8]. This correspondence is fully determined by the laws of physics and can be computed *ab initio* [5, 6, 76]. Recent genomic sequence analyses suggest that nucleosome positioning is also ensured by nucleosome exclusion motifs [77, 78]. In any case, there is no way to modify the correspondence as long as the laws of physics do not change. The positioning/excluding genomic sequences cannot be called a genomic code, but rather a rule, for nucleosome positioning. Neither do they "code for" the chromatin structure [6, 76, 78-80] since there is no arbitrariness in this relationship: chromatin structure imprinting in the genomic sequence is not a biological code. It is only a practical shortcut allowing us to predict chromatin features from analyses at the level of genomic sequences.

The combinatorial rules determining mRNA from the features of the raw transcript (pre-mRNA) composed of coding exons and non-coding introns have been called the *splicing code* [9, 10]. It has

been partly unraveled by an algorithm that combines more than 200 features of DNA (that may be codewords) with predictions of mRNA sequences (the encoded objects) [11, 12]. At this point, the result is a contingency table that provides us, using a computer, the mRNA associated with a segment of DNA and given conditions or cell type. To assess that it is actually a biological code, with the required level of gratuity, one would have to evidence the presence of an adaptor within the splicing machinery and provide a mechanistic understanding of its action. The ultimate proof, as or the genetic code, would be to artificially modify the correspondence.

Proteomic codes result from physical necessity and are only a summary of our computations

Protein 3-dimensional structure results from the peptidic sequence in a direct and obligatory way. This statement of principle is now partly confirmed by protein folding simulations [81]. Stating that the amino acid sequence "codes for" the native fold means only "directly determines" (both the pathway and the folding kinetics [82]). Neither is the *proteomic code* [83] associating structural motifs to protein subsequences a true code. It is only a shortcut, insofar as it enables to avoid long *ab initio* computations and predict (e.g. from statistical association studies) structures based on the sequences.

The relationship between the protein sequence and its post-translational modifications might be a code but its actual nature has yet to be investigated: for the *posttranslational code* [84] to be a true code, one has to evidence an adaptor, presumably an enzyme, recognizing some features of the sequence and, at another site, catalyzing the posttranslational modifications of the protein. Monitoring a change in this enzyme would then provide a code variant and demonstrate the nature of the code, if any.

There is potentially a transcriptional regulatory code but it is not yet demonstrated

The proclaimed *cis-regulatory code*, encoding gene expression levels in TF binding patterns of cis-regulatory modules (a genomic sequence upward the gene) is currently supported only by statistical association [21]. The actual mechanisms underlying the association are unknown. It would be a code if an intermediary and coevolved entity is involved that recognizes the cis-regulatory module and its occupancy, and accordingly controls RNA-polymerase recruitment.

The eukaryote genome is organized by histone proteins into chromatin. A major step has been the discovery of histone post-translational modifications (PTMs), that is, covalent modifications occurring on specific residues and catalyzed by dedicated enzymes [13, 20]. Histones and chromatin structure provide additional levels of transcription regulation. Parallel additional levels of coding have been recently proposed, with partly overlapping and ambiguous names: histone codes [13-15], epigenetic codes [22-24] or transcriptional regulatory codes [18, 19], all relying on histone PTMs. The physicochemical rules by which histone PTMs are read cannot form a code on their own, hence the name of *histone code* is in any case misleading [17].

The name *transcriptional regulatory code* presently designates the association between enhancer sequences and TFs binding patterns [18, 19], or between TF PTMs and their binding sites [20], or between histone PTMs in enhancers and the propensity of these genome regions to bind TFs [13, 14]. These associations have direct physical or stereochemical determinants hence are definitely not codes. The wording ‘transcriptional regulatory code’ also refers to the mapping between histone modifications and transcriptional activity of neighboring genes [19, 22, 26], or between signals coming from transduction pathways and changes in gene expression levels [20]. These mappings may be codes, but have not yet been investigated in this regard. Candidates for adaptor(s) are cofactors recognizing histone PTMs and preparing the initiation site for RNA-polymerase [84], homeodomain finger proteins [26], protein bromodomains or chromodomains [14] mediating histone PTMs and chromatin remodeling, or the RNA polymerase itself, and/or the chromatin architecture and conformational dynamics coordinating distant events regulating transcription initiation [25].

The chromatin itself, through its coevolved architecture and conformational dynamics, could possibly

act as a decoding device associating regulatory events to sets of histone PTMs [25]. Histone PTM patterns generally occur in broad domains, at a scale larger than the nucleosome [85], hence rather act at the chromatin level. Experimental data suggests that some PTMs can operate as a binary switch controlling chromatin folding, favoring or preventing binding of other proteins [27]. Such observations support the idea of an allosteric behavior of the chromatin, proposed in [25] on theoretical grounds. According to this idea, a chromatin conformational change mediates a relationship (today established but *a priori* arbitrary) between sets of histone PTMs and DNA binding of some proteins. The influence of histone PTMs on chromatin conformation alone is not a code, but the control exerted on distant binding events has enough arbitrariness to qualify as a code. It has been proposed that such a *chromatin cod*e (or *epigenetic code* [22-24]) may inform long-term modifications of gene expression involved in memory formation and storage [86, 87].

For comparison purposes, let us mention the calcium code [88]. In some excitable cells, a constant stimulus sustains periodic oscillations of intracellular calcium concentration, whose frequency is proportional to the stimulus amplitude. In contrast to a periodic forcing where the output frequency is simply driven by the frequency of the input, it is here legitimate to speak of frequency encoding of the stimulus amplitude: the adaptor is the calcium dynamics itself.

Conclusion

Biological codes result from an evolutionary harnessing of physical laws achieving arbitrary correspondences. Arbitrariness results from the involvement of a general 'allosteric' mechanism, settled in the course of evolution via the design of chimeric objects, the adaptors. The genetic code fully illustrates this point. It in fact involves two nested codes, the acceptor code mediated by aaRSs charging tRNAs with their cognate amino acid, and the genetic code established by aminoacyl-tRNAs, associating to each codon a single amino acid through codon-anticodon pairing rules.

Mapping rules providing an efficient summary of intricate *ab initio* computations are not biological codes since their very derivation shows their physical, chemical or stereochemical necessity. Neither it is sufficient to produce contingency tables established by statistical association studies to speak of a biological code without evidencing a co-evolved adaptor.

The existence of adaptors endows a code with both robustness and evolvability. It offers a target for enhancement or therapeutic control through the design of code variants, opening a promising path in synthetic biology. Several potential molecular codes have been suggested, e.g. splicing codes or transcriptional regulatory codes or chromatin codes, but the demonstration that they are indeed codes remains to be done. The focus should be on identifying the adaptors, like the correspondence between codons and amino acids evolved from an association table to a code established by aminoacyl-tRNAs. Today, the genetic code is still the sole demonstrated molecular code.

Acknowledgements:
I acknowledge Finn Kjellberg and Julien Mozziconacci for their critical reading of the first version of the manuscript.

References

[1] Crick FHC, Barnett L, Brenner S, *et al.* 1961. General nature of the genetic code for proteins. *Nature* 192: 1227–32.
[2] Nirenberg M, Leder P, Bernfield M, *et al*. 1965. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci* 53: 1161–8.
[3] Söll D, Jones DS, Ohtsuka E, *et al.* 1966. Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique. *J Mol Biol* 19: 556–73.
[4] Söll D, RajBhandary UL. 2006. The genetic code – Thawing the `frozen accident'. *J Biosci* 31: 45963.
[5] Trifonov EN. 1989. The multiple codes of nucleotide sequences. *Bull Math Biol* 51: 41732.
[6] Trifonov EN. 1999. Sequence codes. In Creighton TE, ed; *Encyclopedia of Molecular Biology*. New York: Wiley. p 23246.

[7] Segal E, Fondufe-Mittendorf Y, Chen L, *et al*. 2006. A genomic code for nucleosome positioning. *Nature* 442: 772–8.
[8] Clark DJ. 2010. Nucleosome positioning, nucleosome spacing and the nucleosome code. *J Biomol Struct Dyn* 27: 781–93.
[9] Fu XD. 2004. Towards a splicing code. *Cell* 119: 736–738.
[10] Wang GS, Cooper TA. 2007. Splicing in disease: Disruption of the splicing code and the decoding machinery. *Nat Rev Gen* 8: 749–61.
[11] Tejedor JR, Valcárcel J. 2010. Breaking the second genetic code. *Nature* 465: 456.
[12] Barash Y, Calarco JA, Gao W, *et al.* 2010. Deciphering the splicing code. *Nature* 465: 53–9.
[13] Jenuwein T, Allis CD. 2001. Translating the histone code. *Science* 293: 1074.
[14] de la Cruz X, Lois S, Sanchez-Molina S, *et al.* 2005. Do protein motifs read the histone code? *Bioessays* 27: 164–75.
[15] van Attikum H and Gasser SM. 2005. The histone code at DNA breaks: a guide to repair? *Nat Rev Mol Cell Biol* 6: 757–65.
[16] Dion MF, Altschuler SJ, Wu LF, *et al.* 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci* 102: 5501–6.
[17] Henikoff S. 2006. Is it a code: The debate. No: Histone modifications, although diverse, do not constitute a complex code of chromatin states. *The Scientist* 20: 38–9.
[18] Harbison CT, Gordon DB, Lee TI, *et al.* 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431: 99–104.
[19] Benecke A. 2006. Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs. *Eur Phys J E* 19: 353–66.
[20] Benayoun BA, Veitia RA. 2009. A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends Cell Biol* 19:189–97.
[21] Rister J, Desplan C. 2010. Deciphering the genome's regulatory code: The many languages of DNA. *BioEssays* 32: 381–4.

[22] Turner BM. 2000. Histone acetylation and an epigenetic code. *Bioessays* 22: 836–45.
[23] Turner BM. 2006. Is it a code: The debate. Yes? An epigenetic histone code may allow for unprecedented predictive power. *The Scientist* 20: 38–9.
[24] Turner BM. 2007. Defining an epigenetic code. *Nat Cell Biol* 9: 26.
[25] Lesne A. 2006. The chromatin regulatory code: beyond a histone code. *Eur Phys J E* 19: 375–7.
[26] Mellor J. 2006. It takes a PHD to read the histone cde. *Cell,* 126: 22–4.
[27] Georgatos SD, Markaki Y, Christogianni A, *et al*. 2009. Chromatin remodeling during mitosis: a structure-based code? *Front Biosci* 14: 2017–27.
[28] Weigmann K. 2004. The code, the text and the language of God. *EMBO Rep* 5: 116–8.
[29] Bensmaia SJ. 2008. Tactile intensity and population codes. *Behav Brain Res* 190: 165–73.
[30] Wang X, Lu T, Bendor D, *et al.* 2008. Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience* 157: 484–94.

[31] Baccus SA. 2007. Timing and computation in inner retinal circuitry. *Annu Rev Physiol* 69: 271–90.
[32] Johnson BA, Leon M. 2007. Chemotopic odorant coding in a mammalian olfactory system. *J Comp Neurol* 503: 1–34.
[33] Peirce CS. 1931–1935. *Collected Papers.* Cambridge MA: Harvard University Press.
[**34**] **Barbieri M**. 1998. The organic codes. The basic mechanism of macroevolution. *Rivista di Biologia*, 91:481.
[35] Barbieri M. 2003. *The Organic Codes – an introduction to semantic biology*. Cambridge: Cambridge University Press, Cambridge.
[**36**] **Barbieri M**. 2008. Biosemiotics: a new understanding of life. *Naturwissenschaften*, 95 : 577–599.
[**37**] **Barbieri M.** (2018). What is code biology? *Biosystems*, 164 : 1-10.
[38] Monod F. 1971. *Chance and necessity: An essay on the natural philosophy of modern biology*. New York: Knopf.
[39] Monod J, Changeux JP, Jacob F. 1963. Allosteric proteins and cellular control systems. *J Mol Biol* 6: 306–29.
[40] Stegmann U. 2004. The arbitrariness of the genetic code. *Biol Phil* 19: 20522.

[41] Kjosavik F. 2007. From symbolism to information? Decoding the genetic code. *Biol Phil* 22: 333–49.
[42] Crick FHC. 1958. On protein synthesis. *The Symposia of the Society for Experimental Biology* 12: 138–63.
[43] Levinthal C. 1959. Coding aspects of protein synthesis. *Rev Mod Phys* 3: 249–55.
[44] Gamow G, Rich A, Ycas M. 1956. The problem of information transfer from nucleic acids to proteins. *Adv Biol Med Phys* 4: 23–68.
[45] Crick FHC, Griffith JS, Orgel LE. 1957. Codes without commas. *Proc Natl Acad Sci USA* 43: 416.
[46] Söll D, Rajbhandary UL. 1967. Studies on polynucleotides. LXXVI. Specificity of transfer RNA for codon recognition as studied by amino acid incorporation. *J Mol Biol* 29: 113–24.

[47] Woese CR, Olsen GJ, Ibba M, *et al.* 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* 64: 202–36.
[48] Santos MA, Tuite MF. 1995. The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. *Nucl Acids Res* 23: 1481–6.
[49] Berleant D, White M, Pierce E, *et al.* 2009. The genetic code – More than just a table. *Cell Biochem Biophys* 55: 107–16.
[50] Alvager T, Graham G, Hilleke R, *et al.* 1989. On the information content of the genetic code. *Biosystems* 22: 189–96.
[51] Crick FHC. 1966. Codon-anticodon pairing: The wobble hypothesis. *J Mol Biol* 19: 548–55.
[52] Janke A, Pääbo S. 1993. Editing of a tRNA anticodon in marsupial mitochondria changes its codon recognition. *Nucl Acids Res* 21: 1523–5.
[53] Yokobori SI, Pääbo S. 1995. tRNA editing. *Nature* 377: 490.
[54] Brennicke A, Machfelder A, Binder S. 1999. RNA editing. *FEMS Microbiol Rev* 23: 297–316.
[55] Bossi L, Roth JR. 1980. The influence of codon context on genetic code translation. *Nature* 286: 123–7.
[56] Yarus M. 1982. Translational efficiency of transfer RNA's: Uses of an extended anticodon. *Science* 218: 646–52.
[57] Ibba M, Söll D. 2004. Aminoacyl-tRNAs: Setting the limits of the genetic code. *Genes Dev* 18: 731–8.
[58] Schimmel P. 1996. Origin of the genetic code: A needle in the haystack of tRNA sequences. Proc Natl Acad Sci USA 93: 4521–2.
[59] de Duve C. 1988. The second genetic code. *Nature* 333: 117–8.
[60] Musier-Forsyth K, Schimmel P. 1999. Atomic determinants for aminoacylation of RNA minihelices and relationship to genetic code. *Acc Chem Res* 32: 368–75.
[61] Cavarelli J and Moras D. 1993. Recognition of tRNAs by aminoacyl-tRNA synthetases. *FASEB J.* 7: 79–86.
[62] An S, Barani G, Musier-Forsyth K. 2008. Evolution of acceptor stem tRNA recognition by class II prolyl-tRNA synthetase. *Nucl Acids Res* 36: 2514–2521.
[63] Chapeville F, Lipmann F, von Ehrenstein G, *et al.* 1962. On the role of soluble ribonucleic acid in coding for amino acids. *Proc Natl Acad Sci USA* 48: 1086–92.
[64] Hou YM, Schimmel P. 1988. A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* 333: 140–5.
[65] Sherman JM, Söll D. 1996. Aminoacyl-tRNA synthetases optimize both cognate tRNA recognition and discrimination against noncognate tRNAs. *Biochemistry* 35: 601–7.
[66] Jukes TH, Osawa S. 1990. The genetic code in mitochondria and chloroplasts. *Experientia* 46: 1117–26.
[67] Laforest MJ, Roewer I, Lang BF. 1997. Mitochondrial tRNAs in the lower fungus *Spizellomyces punctatus*: tRNA editing and UAG 'stop' codons recognized as leucine. *Nucl Acids Res* 25: 626–32.
[68] Knight RD, Freeland SJ, Landweber LF. 1999. Selection, history and chemistry: three faces of the genetic code. *Trends Biochem Sci* 24: 241–7.
[69] Wang L, Schultz PG. 2005. Expanding the genetic code. *Angew Chem Int Ed* 44: 34–66.
[70] Cropp TA, Anderson JC, Chin JW. 2007. Reprogramming the amino-acid substrate specificity of orthogonal aminoacyl-tRNA synthetases to expand the genetic code of eukaryotic cells. *Nature protocols* 2: 2590–600.
[71] Crick FHC. 1968. The origin of genetic code. *J Mol Biol* 38: 367–79.
[72] Di Giulio M. 2005. The origin of the genetic code: Theories and their relationships, a review. *BioSystems* 80: 175–84.
[73] Chechetkin VR. 2006 Genetic code from tRNA point of view. *J Theor Biol* 242: 922 –34.
[74] Tlusty T. 2010. A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. *Phys Life Rev* 7: 362–76.
[75] Benos PV, Lapedes AS, Stormo, GD. 2002. Is there a code for protein-DNA recognition? Probab(ilistical)ly... *BioEssays* 24: 466–75.
[76] Lavery R, Lafontaine A. 2000. Optimisation of nucleic acid sequences. *Biophys J* 79: 680–5.
[77] Vaillant C, Audit B, Arneodo A. 2007. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys Rev Lett* 99: 218103.
[78] Arneodo A, Vaillant C, Audit B, *et al.* (2011) Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys Rep* 498: 45–188.
[79] Ben Haïm E, Lesne A, Victor J.M. 2001. Chromatin : a tunable spring at work inside chromosomes, *Phys Rev E* 64: 051921.
[80] Lesne A, Victor JM. 2006 Chromatin fiber functional organization: some plausible models. *Eur. Phys. J. E* 19: 279–90.
[81] Lazaridis T, Karplus M. 2003. Thermodynamics of protein folding: a microscopic view.*Biophys Chem* 100: 367–95.
[82] [RHM] Rumbley J, Hoang L, Mayne L, *et al.* 2001. An amino acid code for protein folding.*Proc Natl Acad Sci* 98: 105–11.
[83] Trifonov EN, Berezovsky IN. 2002. Proteomic code. *Mol Biol* 36: 239–43.

[84] York B, Yu C, Sagen JV, *et al.* 2010. Reprogramming the posttranslational code of SRC-3 confers a switch in mammalian systems biology. *Proc Nat Acad Sci USA* 107: 11122–7.
[85] Liu CL, Kaplan T, Kim M, *et al.* 2005. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biology* 3: e328.
[86] Chwang WB, O'Riordan KJ, Levenson JM, *et al.* 2006. ERK/MAPK regulates hippocampal histone phosphorylation following contextual fear conditioning. *Learn Mem* 13: 322–8.
[87] Wood MA, Hawk JD, Abel T. 2006. Combinatorial chromatin modifications and memory storage: a code for memory? *Learn Mem* 13: 241–4.
[88] Tang Y, Othmer HG. 1995. Frequency encoding in excitable systems with applications to calcium oscillations, Proc Natl Acad Sci USA 92: 7869-7873.
[89] Verhey KJ, Gaertig J. 2007. The tubulin code. *Cell Cycle* 6: 2152–60.

Table 1 --- Potential molecular codes.
Numerous situations are named codes in the literature. In this table, we summarize the nature of some of them, identifying the encoded objects, the codewords and the two potential levels of combinatorics --- in composing codewords with letters, or in concatenating codewords into a message to be parsed during reading. In some cases, the code is simply a mapping fully determined by chemical, stereochemical or physical laws. In other cases, it is only a shortcut for us scientists, providing an elegant summary of heavy ab initio computations or statistical association studies. Most often, the term of code has been used without demonstrating the existence of an adaptor ensuring the arbitrariness of the relationship: whether these are true biological codes remains to be demonstrated. Although it is not a molecular code but rather akin to neural codes, the calcium code is mentioned for comparison.

Potential code	Encoded objects	Codewords	Combinatorics in the codewords	Concatenation into a message	Adaptor	Is it a code ?	References
genetic code	amino acids	codons	yes (letters = nucleotides)	concatenation into isomorphic sequences	charged tRNA (aminoacyl-tRNA)	yes	[1, 41, 42, 49, 57]
acceptor code	amino acids	tRNA features	maybe (several determinants)	no	aminoacyl-tRNA synthetase	yes (e.g. alanine)	[59-60, 62, 64, 65]
DNA-protein recognition	protein binding	DNA sequence	yes (letters = nucleotides)	maybe (network)	no	no	[75]
nucleosome code	nucleosome positioning	DNA sequence	yes (letters =nucleotides)	yes (sequential)	no	no	[5, 7, 8, 77, 78]
splicing code	mRNA sequence	more than 200 DNA features	yes	no	not yet addressed	maybe	[9-12]
amino acid code for protein folding	protein 3D structure (native fold)	protein sequence	yes (letters = residues)	no	no	no	[82]
proteomic code	protein structural motifs	protein subsequences	yes (letters = residues)	maybe (array of motifs)	no	no	[83]
protein post-translational processing	posttranslational protein modification	protein sequence	maybe	no	not yet addressed undescribed enzyme ?	maybe	[84]
histone code	transcription factor binding	histone covalent modifications	yes	no	no	no	[13, 16, 17, 22]
transcriptional regulatory code	gene expression level	histone modifications in the enhancer	yes	no	not yet addressed TFs ? finger proteins ?	maybe	[14, 19, 26, 84, 85]
cis-regulatory code	gene expression level	TF occupancy of cis-regulatory modules	yes	maybe	not yet addressed	maybe	[18, 19, 21]
chromatin code or epigenetic code	events regulating transcription initiation	histone covalent modifications	yes	no	chromatin fiber structure ?	maybe	[19, 22-24, 25-27]
tubulin code	microtubule function	tubulin posttransl. modifications	yes	no	microtubule associated proteins ?	maybe	[88]
calcium code	amplitude of the incoming stimulus	frequency of calcium oscillations (output)	no	no	excitable cell calcium dynamics ?	probably	[89]