DNA, dichotomic classes and frame synchronization: a quasi-crystal framework

BY SIMONE GIANNERINI1,*, DIEGO LUIS GONZALEZ2,1 AND RODOLFO ROSA1,2

1Dipartimento di Scienze Statistiche, Università di Bologna, Via delle Belle Arti 41, 40126 Bologna, Italy
2CNR-IMM, Sezione di Bologna, Via Gobetti 101, 40129 Bologna, Italy

In this article, we show how a new mathematical model of the genetic code can be exploited for investigating the almost periodic properties of DNA and mRNA protein-coding sequences. We present the main mathematical features of the model and highlight its connections with both number theory and group theory. The group theoretic framework presents interesting analogies with the theory of crystals. Moreover, we exploit the information provided by dichotomic classes, binary variables naturally derived from the mathematical model, in order to build statistical classifiers for retrieving and predicting the normal reading frame used by the ribosome in protein synthesis. The results show that coding sequences possess a local informational structure that can be related to frame synchronization processes. The information for retrieving the normal reading frame, which implies the existence of short-range correlations and almost periodic structures related to the organization of codons, offers an interesting analogy with the properties of quasi-crystals. From a theoretical point of view, our results might contribute to clarifying the relation between biological information and shape in nucleic acids and proteins. Also, from the point of view of applications, we present new promising tools for designing efficient algorithms for frame synchronization, which plays a crucial role in faithful synthesis of proteins.

Keywords: frame synchronization; dichotomic classes; ribosome; supervised learning; DNA coding sequences

1. Introduction

An important problem in modern biology is the study and prediction of stereochemical properties of molecules and their relation with normal and pathological biological functions. Prediction of shape in long polymeric molecules, such as nucleic acids and proteins, is a very difficult task. In addition to the specific physico-chemical features, biological molecules exhibit a new, probably emergent, property, i.e. they code for biological information. Thus, in some complex way, the biological information that necessarily codes for a function is also related to the shape because shape is a key feature for implementing such functions.

*Author for correspondence (simone.giannerini@unibo.it).

One contribution of 14 to a Theme Issue ‘Beyond crystals: the dialectic of materials and information’.
Biological forms are, in general, much more complex than simple crystals. However, in some cases, the frontier between them is not so well defined. Crystals are characterized by fixed periodical structures that repeat again and again. Nevertheless, more complex structures, indeed aperiodic ones, have gained the status of crystals. In fact, we know that quasi-crystals can attain dodecahedral and icosahedral symmetries not allowed by the canonical laws of crystallography. Many viruses can also attain analogous shapes [1,2] characterized by self-similar structures.

Some quasi-crystals are thermodynamically stable and this represents a challenge concerning the explanation of the mechanisms that govern their formation and their stability [3]. In this respect, two main approaches can be taken; the first one is related to energy minimization, while the second one is related to entropy maximization. Interestingly, entropy maximization offers a natural contact point with certain biological problems. In fact, biological information can be put in terms of entropy. However, at a biological level, information is a subtle entity. This is because besides the need of a physical carrier with associated entropy properties, it conveys also a meaning, usually associated to a specific biological function. Thus, a very important problem is to understand how biological information can influence physico-chemical aspects of the systems involved (including shape) and, inversely, how physical constraints can affect the biological function conveyed by such information.

In general, the biological information is not related directly to shape and form of biological structures, but acts indirectly by setting a series of parameters that represent the initial and contour conditions of some dynamical process of auto-organization. This is the case of embryology—that is, the attainment of complex shaped organisms from embryonal stages. There are cases, though, where genetic information can be related directly to the structural properties of nucleic acids [4]. Also, in the template-guided protein synthesis, the nucleotide sequence that determines the amino acid chain can also influence the secondary and tertiary structures of proteins through mechanisms still not completely understood [5,6]. In its simplest form, the genetic information contained in DNA or mRNA genes consists of a linear sequence of four different bases, i.e. T, C, A and G. In the crystal paradigm, this primary structure of the DNA had been considered from the beginning as some kind of aperiodic crystal, even before the discovery of the double helical structure of the DNA by Watson and Crick (e.g. [7]).

The DNA sequence, besides coding for the correct amino acid that will be incorporated to the nascent protein, carries information on reading frame synchronization. In turn, this implies the existence of lag 3 correlations in coding parts of DNA. In the crystal paradigm, such correlations can be interpreted in terms of almost periodic functions and of symmetric structures. As mentioned before, other almost periodicities can be related to structural features of the nucleic acid. In particular, the 10–11bp periodicity can be associated to wrapping around nucleosomes [8,9]. Moreover, the existence of a privileged frame is necessarily associated to some anisotropy along the DNA or mRNA molecules that implies a near-crystal structure composed of different elements. Such regularity is not so evident from the sequence of bases that form the helical structure.

The way in which the information for protein synthesis is coded in nucleic acids is crucial also to the retrieval and maintenance of the correct reading frame. Clearly, the genetic code plays a central role in this aim. Before the
discovery of the standard genetic code, Crick et al. proposed a code based on the so-called comma-freeness property [10]. Such kinds of codes aroused interest from the point of view of coding theory because they are a particular type of error-correction codes [11]. In comma-free codes, a subset of the 64 possible codons (which correspond to all the combinations of the four bases in groups of three) are used for coding the 20 amino acids. The subset is chosen in such a way that a unique natural reading frame is allowed: the reading of a sequence out-of-frame produces invalid codons (codons that do not belong to the subset). This framework allows us to discriminate the correct reading frame and to reject the invalid codons, that is, the detection of errors in coded sequences. Unfortunately, the proposal of Crick itself turned out to be invalid [12], but recent works have shown that a particular kind of related codes, i.e. circular codes, are indeed used in protein-coding sequences. Circular codes are a less restrictive version of comma-free codes and can be used for normal reading frame retrieval [13,14]. They could be relics of some primeval comma-free codes, and some of them seem to be preferred over other possible ones, depending on the type of organism. One such instance is the so-called $X_0$ code empirically found both in eukaryotes and in prokaryotes [15]. In a recent study [16], we have revisited the problem from a statistical point of view and found that, on average, the code $X_0$ has the best covering capability, but there is a great variability among sequences.

In this article, we study the connections between the genetic information of protein-coding sequences and reading frame synchronization by using a recently developed mathematical theory of the genetic code [17,18]. The study is motivated by the assumption that, if circular codes are reminiscent of a primeval kind of coding, more information can be obtained by using the intrinsic mathematical properties (if any) of the genetic code(s). In particular, we investigate the possibility of retrieving the reading frame of a protein-coding sequence by using the information of dichotomic classes, quantities that are derived from the earlier-mentioned model of the genetic code rooted in the number theory, more specifically, on non-power number representations [19]. The model allows us to uncover many symmetry properties of the genetic code and to define binary variables, which we call dichotomic classes. Dichotomic classes can be defined as nonlinear functions of the information contained in a dinucleotide—that is, a group of two adjacent bases. Interestingly, such classes, which represent precise biochemical interactions, emerge naturally from the mathematical model. Moreover, dichotomic classes possess precise symmetry properties. In analogy with the crystallographic approach, such properties allow us to study the mathematical structure of the genetic information from a group theoretic perspective.

In order to assess whether the local information conveyed by dichotomic classes can be exploited by a mechanism that achieves synchronization in protein synthesis, we also develop statistical models for reading frame prediction. Such models are based on a supervised learning framework and can be used as a paradigm for interpreting the means used by the ribosome for retrieving and maintaining the correct reading frame when synthesizing a protein molecule. The results show that the information coming from dichotomic classes can be used to predict successfully the reading frame and detect possible errors due to frameshifts.

*Phil. Trans. R. Soc. A* (2012)
In the following section, we describe briefly the mathematical model of the genetic code and define the dichotomic classes and the associated group theoretic framework, in §3, we present the statistical models used, and in §4, we present the results obtained by analysing a set of 3248 protein-coding sequences. In the last section, we discuss the results and report some conclusions, together with some interesting open problems.

2. The mathematical model

What is a mathematical model of the genetic code? Since the middle of the twentieth century, when this strange translation table that defines the meaning of codons as amino acids was uncovered, different hypotheses on its origin have been proposed. One of the first proposals is that by Crick: the ‘frozen accident’ hypothesis. The idea is that the genetic code originated as a quasi-random correspondence between codons and amino acids and successively evolved until its present form. Once attained, such form remained frozen without further evolution because any change is potentially deleterious. In fact, a simple mutation in the meaning of a particular codon produces a change in any protein that has been coded using such codons and we know that this can produce serious or even lethal diseases.

Among the alternatives to the frozen accident, we mention the coevolution theory and the stereochemical theory. The first one proposes that the present form of the genetic code is related to its coevolution with the biochemical synthesis pathways of amino acids. The stereochemical theory assumes that the genetic code is determined by stereochemical affinities between anti-codons and amino acids. The frozen accident hypothesis is less compatible with any mathematical structure of the genetic code because of its alleged random origin. The latter two theories are somehow more prone to this possibility, but they do not offer any clear framework for building a precise mathematical model.

In any case, through the years, different approaches have shown that the genetic code is indeed a highly structured correspondence between codons and amino acids [20–25]. The most striking property of the genetic code is its degeneracy distribution—that is, the number of codons assigned to every amino acid. Because there are 64 codons and only 20 amino acids, such distribution is necessarily degenerated, namely, at least some amino acids needs to be coded by more than one codon. From a mathematical point of view, this amounts to saying that the genetic code is a non-injective mapping between a domain of 64 codons and a codomain of 20 amino acids.\(^1\) The first theoretical studies of this problem are due to the Russian theoretical physicist Yurii Borisovich Rumer [20]. Rumer showed that exactly one half of the quartets of the genetic code (a quartet is a group of four codons sharing the first two letters, for example, \([TTN] = \{TTT, TTC, TTA, TTG\}\)) specifies amino acids with degeneracy 4, while the other half specifies amino acids with non-4 degeneracy (i.e. 1, 2 or 3). Rumer noted that a global

---

\(^1\) Notice that, in addition to the 20 canonical amino acids, we have the stop codons that mark the end of the protein synthesis; moreover, two additional amino acids (selenocysteine and pyrrolysine) can be incorporated into the protein chain and are usually coded by ambiguous stop codons. If we take into account the degeneracy inside quartets, the codomain for the Euplotes nuclear code contains 24 different objects.

Phil. Trans. R. Soc. A (2012)
Dichotomic classes and synchronization

Table 1. Degeneracy distribution of the Euplotes version of the genetic code.

<table>
<thead>
<tr>
<th>degeneracy</th>
<th>number of amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

transformation (called Rumer’s transformation) of the bases, i.e. T,C,A,G ↔ G,A,C,T, transforms a codon of class 4 into a codon of class 1, 2 or 3, and vice versa. In this respect, Rumer’s transformation uncovers the existence of an intrinsic antisymmetric property of the genetic code. In figure 1, we show the Euplotes nuclear genetic code and in table 1, its distribution of degeneracy inside quartets. For instance, there are two amino acids that have degeneracy 3; namely, each of these two amino acids is represented/coded by three different codons.

At a first glance, both the genetic code and its degeneracy distribution might look arbitrary; however, we will show that such distribution can be described exactly by using a model based on the theory of integer number representation systems. Positional representation systems usually use the powers of a given base \(b\) as positional weights. These systems are called ‘power positional number representation systems’. Our usual digital system (base 10) and the binary numeration system used in computers (base 2) belong to this class. A main
property of such systems is that they are univocal, that is, any integer number has only one representation and any representation is associated to one and only one integer number. Mathematically speaking, such systems define a bijective mapping between the numbers and their representations. For example, taking \( b = 2 \), the number 17 is represented univocally by the string 010001,

\[
\begin{array}{ccccccc}
  0 & 1 & 0 & 0 & 0 & 1 & \text{string} \\
  2^5 & 2^4 & 2^3 & 2^2 & 2^1 & 2^0 & \text{weights} \\
  0 & 16 & 0 & 0 & 0 & 1 & = 17
\end{array}
\]

Now, we have mentioned that the genetic code is not univocal so that power positional systems cannot be used for its description. However, there exist other types of representation systems that are not univocal. These rely on at least one of the following two possibilities.

— The values of the positional digits go beyond the admissible range of power positional systems \((0, 1, \ldots, b - 1)\). A known example used in electronic digital systems is the so-called signed-binary representation where, in addition to the allowed digits 0 and 1, the use of the digit \(-1\) is allowed.

— The positional weights grow more slowly than the powers of a given base. A known example is the binary Fibonacci non-power positional system in which the positional weights grow as the Fibonacci numbers, i.e. \(1, 2, 3, 5, 8, \ldots\) (note that the weights grow more slowly than the powers of two).

The first possibility does not allow the degeneracy of the genetic code to be described. On the contrary, there exists a unique solution based on non-power integer number representation systems. The model associates a length-6 binary string to every codon of the genetic code and a whole number (from 0 to 23) to the corresponding amino acid (including the \textit{stop} signals). The non-power weights used in the representation are a slight modification of the 6 first numbers of the Fibonacci series, i.e. 1, 1, 2, 4, 7, 8. For instance, number 17 has two representations, 110011 and 110100,

\[
\begin{array}{ccccccc}
  1 & 1 & 0 & 0 & 1 & 1 & \text{string} \\
  8 & 7 & 4 & 2 & 1 & 1 & \text{weights} \\
  8 & 7 & 0 & 0 & 1 & 1 & = 17 = \\
\end{array}
\]

The degeneracy distribution associated to the representation system is exactly that of the Euplotes nuclear genetic code (\textit{figure 1}). Notice that the Euplotes code is more symmetric than the standard code; the latter can be derived from the first with a small symmetry break.\(^2\) It is possible to show that no other non-trivial solution exists. A scheme of the model that describes its salient features

\(^2\)The Euplotes and the standard code differ only by the Opal codon TGA; in the standard code, such codon codes for a stop instead of coding for Cysteine.
**Figure 2.** A representation of the first 24 whole numbers (outer columns) in the non-power representation defined by the positional weights \([112478]\) (length-6 binary strings, horizontal rows). The degeneracy number (D), the number of binary strings that represent the same whole number and the corresponding amino acids are shown at the centre of the table. The colours indicate the parity of each string (white = odd, grey = even).

is presented in figure 2. It is important to stress, by using different symmetry arguments, that this representation system is a model of the genetic code because it describes many of its biological properties. We refer the interested reader to Gonzalez [17,18] for a detailed treatment.

(a) **Dichotomic classes**

We have seen that the mathematical model described earlier assigns a length-6 binary string to each of the 64 codons and an integer number from 0 to 23 to the corresponding amino acid. Remarkably, in the studies of Gonzalez et al. [26,27], we have shown that the mathematical properties of these binary strings are deeply linked to the chemical properties of the bases of a codon. These findings led us to the definition of the *dichotomic classes*. The first dichotomic class is the *parity* of a codon, defined as the parity of the associated binary string. It is possible to show that this mathematical operation on a binary string is related to the chemical features of a codon. In fact, each base—T, C, A and G—can be classified according to chemical classes as follows:

- \{purine; pyrimidine\} \(\{R = A, G; \quad Y = C, T\}\),
- \{keto; amino\} \(\{K = T, G; \quad Am = A, C\}\),
- \{strong; weak\} \(\{S = C, G; \quad W = A, T\}\).

Now, it is possible to prove that the parity of the binary strings can be obtained from the chemical classes of the last two bases of the codon. The algorithmic representation of the parity is shown in figure 3a. In words, the rule can be described as follows. If the last letter of the codon is a purine \((R = A, G)\), the parity of the binary string is obtained immediately: an A corresponds to an odd string and a G to an even string. If the last letter is a pyrimidine \((Y = T, C)\), in order to determine the parity, we need to observe the chemical character of the previous base in the codon—that is, the second or middle base. If the second base belongs to the amino class \((Am = C, A)\), the corresponding string is even; if, instead, it belongs to the keto class \((K = T, G)\), the corresponding string is odd.

In the study of Gonzalez et al. [27], we have shown that a similar rule can be used for deriving the other two dichotomic classes: Rumer’s degeneracy class (figure 1) and a new dichotomy that we call the *hidden class*. In order to derive

Phil. Trans. R. Soc. A (2012)
the Rumer class: (i) shift the analysis window to the first two bases of the codon, (ii) consider the amino–keto dichotomy for the middle base, and (iii) use the chemical dichotomy strong ($S = C, G$) or weak ($W = T, A$) for the first base (figure 3b). Again, it can be shown that the Rumer class can be obtained from a partial parity of the associated binary string. A third dichotomic class, the hidden class, can be obtained by a further shift to the left of the window. Such a class emerges naturally from the extension of the other two classes. Its existence implies a connection between adjacent codons because it is defined outside the domain of a single codon. The algorithmic representation of the hidden class is shown in figure 3c.

For example, consider the sequence

$$\text{TCA GGT AAG GGC}$$

we have three possible reading frames,

- frame 0: TCA GGT AAG GGC,
- frame 1: CA GGT AAG GGC,
- frame 2: A GGT AAG GGC.

We compute the parity class on the frame-1 sequence and the Rumer class on the frame-0 sequence as follows:

The same analysis can be applied to the complementary reversed sequence

$$\text{GCC CTT ACC TGA}.$$
In the work of Gonzalez et al. [27], we have shown that the dichotomic classes can be derived by means of nonlinear functions of the extended codon (a set of four bases formed by a given codon plus the last base of the previous one). Also, we have uncovered the existence of universal, strong, short-range correlations between specific combinations of dichotomic classes (see also [28]). Such correlations are universal because they have been observed in almost all the sequences considered, no matter their base composition or GC content. They are short range since the lags involved imply an interaction inside a codon or at most between adjacent codons. The fact that these short-range correlations have not been found previously is not surprising because the dichotomic classes derived from the mathematical model imply a new (and nonlinear) encoding of genetic information.

An important result that highlights the connections between the mathematical model and the quasi-crystal framework is related to a group theoretic interpretation of the set of global transformations of a codon. This set is presented in table 2. For instance, if we apply the K/A (keto/amino) transformation to the codon TCA, we obtain the codon GAC. Now, it is possible to show that the Rumer class is antisymmetric with respect to such transformations. In our example, the codon TCA has Rumer class 0 (degeneracy non-4), whereas the codon GAC has class 1 (degeneracy 4). The same argument holds for the other two dichotomic classes, that is, parity and hidden classes are antisymmetric with respect to the purine/pyrimidine (Y/R) and strong/weak (S/W) transformations, respectively. As shown in Gonzalez et al. [27], the set of global transformations \( \Gamma \) together with the usual matrix product asterisk (*) form an Abelian (commutative) group isomorphic to the Klein V group \((Z_2 \otimes Z_2)\). In order to see this, denote the bases in vector notation as

\[
T' = (1000), \quad C' = (0100), \quad A' = (0010) \quad \text{and} \quad G' = (0001).
\]

The transformations of the bases can be implemented through the following permutation matrices:

\[
L = \begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{pmatrix}, \quad M = \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0
\end{pmatrix}, \quad N = \begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{pmatrix} \quad \text{and}
\]

\[
I = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}.
\]

Now, \( \{\Gamma, \ast\} \), where \( \Gamma = \{L, M, N, I\} \) is an Abelian group. In fact, for each \( x, y, z \in \Gamma \), we have

- \( I \) is the neutral element;
- \( x \ast x = I \) (indeed, \( L, M, N, I \) are orthogonal);
- \( x \ast (y \ast z) = (x \ast y) \ast z \) (associativity); and
- \( x \ast y = y \ast x = z \) (commutativity and closure).
Table 2. A set of possible global transformations of a codon. The last column reports the associated antisymmetric dichotomic class.

<table>
<thead>
<tr>
<th>from</th>
<th>to</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>K/A</td>
<td>T,C,A,G</td>
<td>G,A,C,T</td>
</tr>
<tr>
<td>Y/R</td>
<td>T,C,A,G</td>
<td>C,T,G,A</td>
</tr>
<tr>
<td>I</td>
<td>T,C,A,G</td>
<td>T,C,A,G</td>
</tr>
</tbody>
</table>

3. Frame synchronization as a classification problem

The understanding of the mechanisms that govern the synchronization and that detect and/or prevent frameshift errors is a crucial task that can produce a major advance in the comprehension of the management of genetic information. A frameshift is a change of reading frame in protein-coding genes; we can classify a frameshift as either natural or artificial. A natural frameshift occurs in programmed locations in pseudogenes or as a consequence of errors, whereas an artificial frameshift is mainly due to errors in sequencing and assembling genome information. Of course, the study of frame synchronization techniques has both a theoretical and a practical interest, especially for the accuracy and efficiency of genome sequencing. From the bioinformatics point of view, the design of efficient algorithms for frameshift detection can improve the performance of the assembly process, decrease requirements on the sequencing coverage and reduce the cost of sequencing. See also Kislyuk et al. [29] and references therein, for approaches from a bioinformatics point of view, and Weindl & Hagenauer [30] for a communication theory approach. Another important applicative aspect is related to the detection of programmed frameshifts. Programmed frameshifts depend on particular informational words situated in the neighbourhood of the transition point and produce a slipping of the ribosome of one or two bases. From the frameshift point onwards, the sequence completely changes its meaning (in terms of amino acid coding). For this reason, detection of programmed frameshifts is important for the understanding of alternative coding of proteins. Moreover, in order to ensure faithful protein synthesis, accidental frameshifts must be prevented by the synthesis machinery so that the processes of frame detection and retrieval are crucial.

From a statistical point of view, the frame synchronization problem can be seen as a classification (or supervised learning) problem. When attached to a binding site, the ribosome synchronizes with the sequence. Our assumption is that the local chemical interaction at the binding site produces enough information to achieve the synchronization and possibly detect and correct frameshift errors. Hence, coding sequences should carry sufficient information to achieve this aim. In the following, we show that the dichotomic classes can play a role in the frame synchronization process. In our previous work [27], we have shown the existence of universal, strong-range correlations between dichotomic classes; here, we will show that the local information provided by the dichotomic classes can be used to predict the correct frame. We start with a general description of the problem and present some notation and tools.
In a supervised learning problem, we have a response variable $Y_t$, $t \in \mathbb{N}$, and a set of $p$ predictors $(X_1, \ldots, X_p)$ that can be used to predict $Y_t$. This procedure usually involves the building of a statistical model of the kind

$$Y_t = f(X_1, \ldots, X_p) + \varepsilon_t,$$ (3.1)

where $\varepsilon_t$ can be a white noise process. Hence, $Y_t$ is a periodic sequence that contains the information on the frame. In the simplest case, assuming that the sequence starts in the frame, we have $Y_t = 100100100 \ldots$. At each site of the sequence $t$ and based on the information contained in the regressors, we would like to predict whether we are in the frame ($Y_t = 1$) or not ($Y_t = 0$). In the classification literature, the performance of the model is measured by resorting to the confusion matrix that compares observed values $Y_t$ and predicted values $\hat{Y}_t$ computed through the model on a test set, namely, a different sample from that used to fit the model (training set). The confusion matrix is

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted</td>
<td>$n_{00}$</td>
<td>$n_{01}$</td>
</tr>
<tr>
<td></td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
</tr>
<tr>
<td>observed</td>
<td>$n_{0}$</td>
<td>$n_{1}$</td>
</tr>
</tbody>
</table>

where $N = n_{00} + n_{01} + n_{10} + n_{11}$ is the sample size of the test set. A measure of the overall performance of the classifier is the misclassification rate

$$M = \frac{n_{10} + n_{01}}{N} \quad \text{and} \quad C = 1 - M = \frac{n_{11} + n_{00}}{N}.$$ (3.2)

The rate $M$ is the proportions of errors over the test set. Alternatively, one can use the proportion of correctly predicted observations $C = 1 - M$. In the following, we use the classification rate $C$.

Now, the crucial issue is the choice of the set of $p$ predictors of the frame to be used in the model. We start with the following set of 72 variables presented in table 3. Here, $r$, $p$, $h$ stand for the Rumer, parity, hidden classes, respectively; their barred version indicates the classes computed on the reversed complemented sequence, whereas the primed version indicates the classes computed on the reversed sequence. From a biological point of view, the issue is to assess (i) how the information given by the dichotomic classes can help the synchronization process and (ii) the characteristic length of the interaction between the ribosome and the sequence. From a statistical point of view, this is tantamount to a model selection procedure: in fact, the selected model carries the information on the number and on the kind of variables involved so that it would be possible to understand whether there are specific classes that play a preeminent role in the synchronization process. The performance of the selected models as measured by the classification rate $C$ completes the portrait.

As concerns the choice of the classifier, we opted for two different approaches: (i) logistic regression and (ii) neural networks (NNs). The logistic regression model makes more assumptions and produces stable but possibly inaccurate predictions. The method based on NNs makes mild structural assumptions: its predictions are often accurate but can be unstable.
where \( n \) is the length of the training set and RSS is the residual sum of squares.

Strictly speaking, this is a model for the probability of occurrence of the frame \( E[Y_t|X] = \pi_t \). Hence, the predicted values lie in \([0, 1]\). In order to use the model in a classification context, we discretize the probability according to a threshold \( \tau = 0.5 \): \( \hat{Y}_t = 1 \) if \( \hat{\pi}_t \geq 0.5 \) and \( \hat{Y}_t = 0 \) otherwise. Then, the confusion matrix and the misclassification rate can be derived.

(b) Neural networks

NN models were introduced almost independently in the fields of statistics and artificial intelligence. Eventually, the terminology introduced in the artificial intelligence context has been retained and this contributed to give a magical character to them. As a matter of fact, they are just two-stage nonlinear regression models that consist of sums of nonlinearly transformed linear models. NN models are effective in problems where prediction without interpretation is the goal, which is why we chose to compare them with a parametric model that is less flexible and accurate but whose results can be interpreted.
Table 4. Classes of proteins whose sequences have been analysed. The third and fourth columns report the number of sequences and the number of kilobases (kb) of each class, respectively.

<table>
<thead>
<tr>
<th>Protein Type</th>
<th>Number of Sequences</th>
<th>kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>albumin</td>
<td>142</td>
<td>101.5</td>
</tr>
<tr>
<td>alpha-globin</td>
<td>57</td>
<td>15.2</td>
</tr>
<tr>
<td>beta-globin</td>
<td>141</td>
<td>50.3</td>
</tr>
<tr>
<td>carboxypeptidase A</td>
<td>31</td>
<td>35.2</td>
</tr>
<tr>
<td>globulin</td>
<td>6</td>
<td>2.9</td>
</tr>
<tr>
<td>glycogen synthase</td>
<td>346</td>
<td>458.3</td>
</tr>
<tr>
<td>heat shock protein 70</td>
<td>1022</td>
<td>1148.1</td>
</tr>
<tr>
<td>insulin</td>
<td>45</td>
<td>9.9</td>
</tr>
<tr>
<td>lactate dehydrogenase</td>
<td>164</td>
<td>113.7</td>
</tr>
<tr>
<td>lysozyme</td>
<td>327</td>
<td>161.5</td>
</tr>
<tr>
<td>phosphoglycerate kinase</td>
<td>1077</td>
<td>1221.7</td>
</tr>
<tr>
<td>phosphorylase kinase</td>
<td>10</td>
<td>18.2</td>
</tr>
<tr>
<td>troponin C</td>
<td>40</td>
<td>17.6</td>
</tr>
</tbody>
</table>

Table 5. Percentages over the 2966 sequences that the set of variables are inserted in the logistic regression models selected through the stepwise AIC.

<table>
<thead>
<tr>
<th></th>
<th>t − 4</th>
<th>t − 3</th>
<th>t − 2</th>
<th>t − 1</th>
<th>t</th>
<th>t + 1</th>
<th>t + 2</th>
<th>t + 3</th>
<th>t + 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>59.3</td>
<td>59.6</td>
<td>49.7</td>
<td>60.1</td>
<td>53.1</td>
<td>44.2</td>
<td>50.5</td>
<td>50.0</td>
<td>47.8</td>
</tr>
<tr>
<td>p</td>
<td>77.2</td>
<td>49.0</td>
<td>56.4</td>
<td>57.9</td>
<td>43.0</td>
<td>51.5</td>
<td>49.8</td>
<td>39.2</td>
<td>41.0</td>
</tr>
<tr>
<td>h</td>
<td>56.3</td>
<td>50.1</td>
<td>63.4</td>
<td>62.0</td>
<td>42.3</td>
<td>56.1</td>
<td>54.0</td>
<td>46.7</td>
<td>43.0</td>
</tr>
<tr>
<td>r′</td>
<td>66.5</td>
<td>57.4</td>
<td>52.2</td>
<td>67.2</td>
<td>59.4</td>
<td>55.2</td>
<td>68.2</td>
<td>59.9</td>
<td>52.2</td>
</tr>
<tr>
<td>p′</td>
<td>49.5</td>
<td>35.1</td>
<td>43.0</td>
<td>43.5</td>
<td>34.7</td>
<td>41.4</td>
<td>38.2</td>
<td>39.4</td>
<td>41.8</td>
</tr>
</tbody>
</table>

For our problem, we have fit a single layer feed-forward model of the kind

\[ f(X_1, \ldots, X_p) = b_0 + \sum_{j=1}^{q} \beta_j \psi \left( \sum_{i=1}^{p} \omega_{ij} X_i + \omega_{oj} \right), \]  (3.5)

where \( \beta_j \) represents the hidden-to-output weight, \( \omega_{ij} \) is the connection strength, \( u_j = \sum_{i=1}^{p} \omega_{ij} X_i \) \((j = 1, \ldots, q)\) is the q hidden unit and \( \psi \) is the activation function taken as the logistic function \( \psi(x) = \exp(x)/(1 + \exp(x)) \). Parameter estimation has been performed through least-square fitting.

4. Data analysis

In order to assess the presence of a frame synchronization mechanism, we used 3248 nucleotide sequences obtained from GenBank (http://www.ncbi.nlm.nih.gov).
Table 6. Statistics of the classification rates over the 2966 sequences for the logistic regression model (GLM) and for the NN model.

<table>
<thead>
<tr>
<th></th>
<th>minimum</th>
<th>1st quartile</th>
<th>median</th>
<th>mean</th>
<th>3rd quartile</th>
<th>maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM, training</td>
<td>68.0</td>
<td>80.2</td>
<td>84.6</td>
<td>85.0</td>
<td>89.8</td>
<td>100.0</td>
</tr>
<tr>
<td>GLM, test</td>
<td>45.7</td>
<td>72.3</td>
<td>77.6</td>
<td>77.3</td>
<td>83.4</td>
<td>97.0</td>
</tr>
<tr>
<td>NN, training</td>
<td>66.7</td>
<td>98.4</td>
<td>99.5</td>
<td>98.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>NN, test</td>
<td>45.7</td>
<td>77.0</td>
<td>82.0</td>
<td>81.0</td>
<td>86.2</td>
<td>96.3</td>
</tr>
</tbody>
</table>

We have extracted all the coding sequences from the 13 classes of proteins reported in table 4, where for each class, we list the number of sequences and the kilobase weight. Such an ensemble has been reduced by eliminating duplicate and short (less than 300 bp) sequences. The final set consists of 2966 sequences. Each sequence has been split in two parts, the first 75 per cent of the sequence is used as the training set (model fit), whereas the last 25 per cent of it is used as the test set (classification rates).

In table 5, we present the percentages of inclusion of the variables of table 3 in the logistic regression model selected through the stepwise AIC of equation (3.4). Notably, the Rumer class seems to play an important role for frame prediction. In fact, the variable $r'_{t+4}$ occurs in 83.8 per cent of the 2966 models. For each lag, except for $t-4$ where the parity class comes first, the Rumer class (either reversed or reversed complemented) has the highest percentages of inclusion. Also, it is clear that lags up to 4 are needed in order to achieve good results. The performance of the models measured through the classification rate $C$ (see equation (3.2)) is presented in table 6. The columns of the table present the minimum, the first quartile, the median, the mean, the third quartile and the maximum of the distribution of the rates over the 2966 sequences. The table shows that, on average, the logistic regression model predicts correctly the frame in 77.3 per cent of the length of the test set. Such a percentage increases to 81.0 per cent for the NN model. This means that a combination of predictions over a time window would allow us to predict perfectly the frame. The length of such a window might be protein or organism dependent. Also, these results might indicate that it is possible to design a new bioinformatic algorithm for frameshift prediction motivated by our mathematical model of the genetic code. The algorithm might complement or compete with existing proposals such as those of Frey & Michel [13] and Lassez et al. [32]. In particular, our preliminary results compare favourably with those of Lassez et al. [32] because they achieve a classification rate of 75 per cent by using the local information of 10 codons. In our case, we achieve a correct prediction rate of 81.0 per cent by exploiting only nine codons.

In order to test further whether there is a local structure that allows our models to synchronize the frame, we have repeated the whole analysis on a permuted version of the 2966 sequences. In this way, the correlation structure is destroyed, while the global proportion of bases and the length of the sequences are preserved. In figure 4, we show the boxplots of the differences between classification rates

\(^3\)The query script is available upon request.
Dichotomic classes and synchronization

Figure 4. Boxplots of the differences of classification rates \( C_{\text{real}} - C_{\text{perm}} \) computed between real and permuted sequences. The label GLM indicates logistic regression, while NN indicates the NN model. (Online version in colour.)

computed on the original sequences \( C_{\text{real}} \) and those computed on the permuted sequences \( C_{\text{perm}} \). On average, there is a 10 per cent gain in predictive ability. To assess whether such a difference is significant, we have performed a paired Student’s \( t \)-test for the difference of the means between the original and the permuted sequences. We can apply such a test, even if we cannot assume that the variances of the two populations are equal, because the sample size is large and the central limit theorem approximation holds. For both the logistic regression and the NN model, the \( p \)-value of the test results is smaller than \( 2.2 \times 10^{-16} \). Thus, the hypothesis of a local correlation structure that can play a role in the frame synchronization process is plausible.

In order to investigate the existence of a common mechanism for frame retrieval, we have merged 2273 of the 3248 sequences in a single chain of more than 2 million bases and have performed a model fit upon it by using both the logistic regression (GLM) and the NNs. Then, we have computed the classification rates over the 975 remaining sequences (ca 1 million bases overall). The boxplots of the rates are presented in figure 5. The results are remarkable: first, the mean predictive power of the NN settles to 96.7 per cent and does not fall below the minimum value of 93.3 per cent. We take this result as strong evidence of the existence of a universal frame synchronization mechanism. Second, the GLM-based classifier fails completely; the mean classification rate of the 66.7 per cent is smaller than that based on individual sequences. This result indicates that such a hypothetical mechanism is complex and most probably nonlinear.

Phil. Trans. R. Soc. A (2012)
We can extend the approach presented here by differentiating between the two frameshift situations. In such instances, the variable to be predicted would be a three-state periodic sequence, i.e. \( Y_t = 1\ 2\ 3\ 1\ 2\ 3\ 1\ 2\ 3 \ldots \). From a statistical point of view, this would not present additional complications as it would require multinomial models that are natural extensions of the logistic regression models to the case of a number of categories greater than two. Also, the NN model extends seamlessly to the multi-nomial case. Preliminary investigations indicate that a three-state classification would lead to finer and more accurate frame prediction. In conclusion, the results open the door to the possibility of developing new high-performance algorithms for determining the normal reading frame of coding sequences of DNA.

5. Conclusions

In this article, we have shown how a recent mathematical model of the genetic code can be used as a tool for studying the almost periodic properties of DNA and mRNA protein-coding sequences from the point of view of their informational content. We have described the main mathematical features of the model and have stressed its connections with both number theory and group theory. The group theoretic framework presents interesting analogies with the theory of crystals.
Moreover, we have used the information contained in dichotomic classes, binary variables that are naturally derived from the mathematical model, in order to build statistical classifiers for retrieving and predicting the normal reading frame used by the ribosome in protein synthesis.

Different results have highlighted the role of periodicity 3 in frame retrieval and maintenance. The correct frame is equivalent to a zero-phase of the periodicity 3 found in coding sequences. In particular, Trifonov [33] found abundance of letter G in the third position of the codon and lag 3 periodicity of C in binding sites between rRNA and mRNA (evidence of energy-based mechanisms of frame detection). On the same line of thought, some authors have searched for amplitude and phase features related to almost-periodicity in the interaction energy at such binding sites [34]. Arquès & Michel [15], by studying the preferential use of codons in the three different reading frames, discovered the existence of circular codes in Prokaryotes and Eukaryotes organisms. Such codes can be used to implement algorithms for synchronizing the reading frame [13]. Note that theoretical hypotheses about the relation of circular codes and dichotomic classes can be advanced: in fact, in analogy with dichotomic classes, also circular codes are related to the group of global transformations.

Our approach for the problem of frame detection and maintenance is based on the use of a new coding scheme that derives naturally from the mathematical model of the genetic code described in §2. The scheme includes some of the symmetry features that are naturally present in the structure of the genetic code. Moreover, because the coding is based upon dichotomic classes that are nonlinear functions of chemical classes, the information produced cannot be trivially reduced to that of the mere sequence of single nucleotides. In this respect, analogously to what happens in nonlinear dynamical systems, a complex behaviour can be obtained from simple rules at a symbolic level. Our results are promising and compare favourably with the existing proposals, including those based upon the theory of circular codes. In fact, on average, the NN model fitted on a single long chain (obtained by merging different sequences) produces a correct classification in 96.7 per cent of the cases. Our findings show that coding sequences possess a local informational structure that can be related to frame synchronization mechanisms. No doubt, it would be interesting to explore whether our approach can convey additional information when compared with other methods. If the answer is even only partially affirmative then, in principle, it will be possible to derive a combined method with a better performance. Another interesting point is the persistence of properties found in coding sequences, such as the almost-3 periodicity, outside coding regions (in the so-called intra-genic and inter-genic regions of the genome).

The genetic code is one of the most universal features of extant life and as such, needs to be also a feature of the common predecessor of life on the Earth. However, the genetic code has some variants, which can be a product of successive evolution. In any case, it is surprising that the mathematical model based on non-univocal representations of numbers is able to encompass almost all these variants in a unified framework [35]. This fact points to an origin of genetic coding of proteins based on strong organizational principles. Thus, properties that can be observed by studying dichotomic classes should be also a relic of such primeval organization. In particular, the algorithmic representation of the Rumer class implied by the mathematical model gives new information about the origin.
of such mechanisms. Several authors favour the idea that degeneracy is linked to the codon–anti-codon interaction energy of the first two letters of the codon (see [36] and references therein). In this view, degeneracy-4 codons are associated to the most stable interactions between codons and anti-codons. Of course, because the algorithm for describing Rumer’s class is a nonlinear function of the chemical classes, it can be analysed also in terms of such interaction energies. However, the algorithmic representation proves that when the second base is of the amino type, such letters alone are sufficient for determining the degeneracy of a specific codon. Moreover, the energy interaction argument presupposes an origin of genetic coding in the form of a triplet in which only the first two bases are read. Our recent results point to a radically new hypothesis (not necessarily in contrast with former theories) about the origin of protein coding: the degeneracy of the genetic code arises from key symmetry properties of codons in a non-triplet assignment [35]. In such schemes, the degeneracy assumes a clear physical meaning—that is, the existence of different states of a system with the same energy. In fact, if the interaction energy of two different adaptors that carry the same amino acid is the same, both of them can be used indistinctly for adding such amino acids to the nascent chain—that is, the codons are synonymous. In turn, in the transition to a triplet code, the correspondence between symmetries and degeneracy is expressed by the Rumer dichotomic class, which implies also energy considerations. In this respect, we suspect that the ability related to the retrieval of the correct reading frame in coding sequences of DNA (mRNA) points to the origin of the genetic code.

The authors thank the Guest Editor Julyan Cartwright for encouragement and support.

References

Dichotomic classes and synchronization


Phil. Trans. R. Soc. A (2012)

