COMMUNICATION

ChemComm www.rsc.org/chemcomm

A parity code interpretation of nucleotide alphabet composition†

Dónall A. Mac Dónaill*

Department of Chemistry, Trinity College, Dublin 2, Republic of Ireland. E-mail: dmcdonll@tcd.ie; Fax: +353-1-671 2826; Tel: +353 1 608 1456

Received (in Cambridge, UK) 12th June 2002, Accepted 1st August 2002
First published as an Advance Article on the web 15th August 2002

The purine–pyrimidine and hydrogen donor–acceptor patterns governing nucleotide recognition are shown to correspond formally to a digital error-detecting (parity) code, suggesting that factors other than physicochemical issues alone shaped the natural nucleotide alphabet.

The confluence of factors underlying the particular composition of the nucleotide alphabet, A, C, G and T/U, is one of the most fundamental issues in our understanding of the emergence of living matter. The alphabet is not self-evidently optimal, and alternative replicating nucleotides and nucleotide analogues have been demonstrated in a number of studies[1,2] while yet other reports have suggested that the natural alphabet was preceded by a different, possibly two-letter alphabet.[3]

Investigations into nucleotide alphabet composition have tended to focus on physicochemical and related issues. Yet nucleotide replication is at heart an information transmission phenomenon, and it seems reasonable to postulate that the evolutionary pressures shaping the nucleotide alphabet might not have been confined to physicochemical issues alone, and that considerations relating to informatics might have had a constraining evolutionary role, acting concurrently but independently of the physics and chemistry. Surprisingly therefore, with the exception of Szathmáry's pioneering work recognising the importance of hydrogen donor–acceptor (D/A) patterns,[4] informatics aspects of the problem have been largely neglected. In this study therefore we consider the fitness of nucleotides as molecular information carriers in the formal terms of error-coding theory.[5] It is hoped that this molecular informatics approach will clarify the debate by bringing the analytical power and formal descriptive facility of computer science to bear on this most fundamental of scientific issues.

Information in nucleotides is molecularly encoded in the hydrogen D/A patterns and purine/pyrimidine motifs. Elementary error-coding theory considerations inform us that not all combinations of patterns are equivalent with respect to error-resistance. In a molecular context this suggests that not all combinations of nucleotides would be equally error-resistant, allowing selection pressure to select an optimal alphabet with respect to informatics.

Error-coding theory is concerned with codes in which the codewords are conventionally binary numbers. In pursuing an error-coding analysis it is therefore convenient to construct a numerical description of nucleotides capturing molecular recognition features. Hydrogen D/A patterns are readily expressed in binary notation; an acceptor or lone-pair may be (arbitrarily) interpreted as 0, and a donor or hydrogen as 1. A further binary dimension may be associated with the puRine/pYrimidine (R/Y) size motif; R = 0, Y = 1, giving each informationally distinct nucleotide a unique 4-bit numerical representation. This is exemplified for the complementary pair G:C in Fig. 1. The potential nucleotide alphabet of 16 letters corresponds to the set of all 16 four-bit numbers (the binary space $\mathbf{B}^4$). Nucleotides, so interpreted, may be depicted as positions on a hypercube, represented by a cube within a cube (Fig. 2). The position of a nucleotide in a cube is determined by its D/A pattern, while the purine/pyrimidine nature determines whether it belongs to an inner cube for 'pyrimidines', final bit = 1), or an outer cube for 'purines' (final bit = 0); note that the terms pyrimidine and purine are employed here somewhat loosley as a convenient shorthand for monocyclic and bicyclic nucleotides respectively. Labels for additional nucleotides beyond the natural alphabet are taken from Szathmáry.[6]

One of the most fundamental concepts in error-coding theory is that of codeword (or nucleotide) parity, which is said to be odd or even according to whether the total number of 1's in its binary representation is odd or even. Thus C, numerically interpreted as (100,1), has even parity, whereas X, interpreted as (010,0), has odd parity (Fig. 2). A code in which all codewords have the same parity is termed a parity code, and possesses simple but effective error-resistant properties. In data transmission, where the data to be transmitted is of mixed parity, a parity code may be formed by addition of a single bit, termed a parity bit, and set to 0 or 1 as necessary to yield a set of codewords of desired parity.

The depiction of the full space of nucleotides, partitioned into even-parity (Fig. 2(a)) and odd-parity (Fig. 2(b)) subsets, exhibits a similar pattern. When the parity is even, the D/A pattern 100 must be associated with a final parity bit set to 1 (a pyrimidine), yielding C, whereas the D/A pattern 011 can only be expressed on a purine, *i.e.* the parity bit is set to 0, yielding G. One of the more striking features of Fig. 2 therefore is that the natural alphabet, U, C, G and aA (amino-adenine, an idealized form of A) belong to the even-parity subset. Thus, it would appear that in nature the purine/pyrimidine nature of a nucleotide is strictly and intriguingly related to the D/A pattern as a parity bit. The critical question is whether the parity-code structure is accidental, or shaped by selection through evolutionary advantage.

In conventional error-coding theory the advantage afforded by a parity code structure lies in the number of features which must be changed to convert one codeword into another; a transmission error in any one bit changes the parity of the transmitted element whereby the error may be detected. The difference between codewords may be expressed in terms of the Hamming distance, ∂, defined as the number of bits in which two codewords differ. It is equivalent to the number of bits set to 1 in the Boolean exclusive OR product XOR. In nucleotide terms this corresponds to the difference between the binary interpretations of molecularly encoded patterns. Thus C and U,

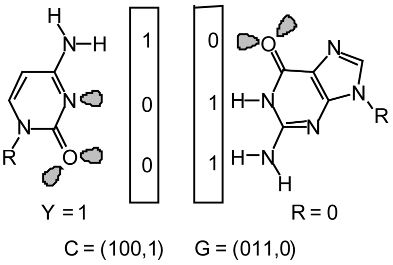

Fig. 1 Numerical interpretation of C and G.

† Electronic supplementary information (ESI) available: expanded background to information and error-coding theory; computational details. See http://www.rsc.org/suppdata/cc/b2/b205631c/

DOI: 10.1039/b205631c

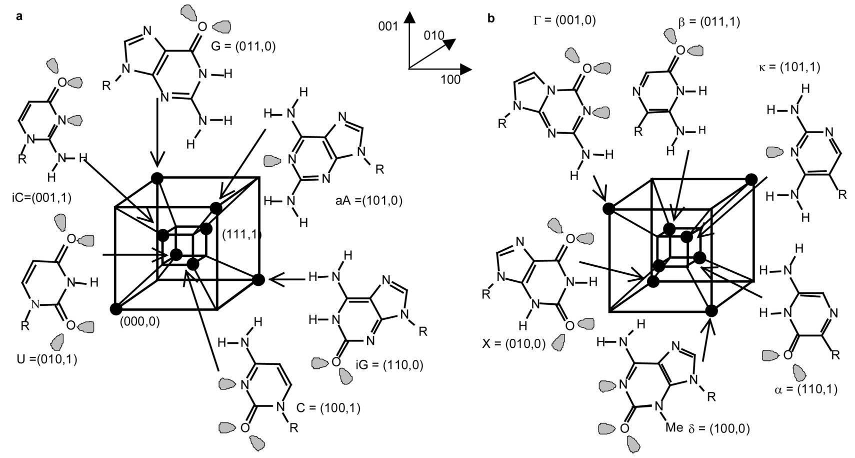

Fig. 2 Numerical interpretation of nucleotides depicted as positions on the **B**^{4} hypercube: (a) even-parity code; (b) odd-parity code.

C = 1001	G = 0110
U = 0101	X = 0100
XOR(C, U) = 1100	XOR(G,X) = 0010
∂(C,U) = 2	∂(G,X) = 1

Fig. 3 Calculation of distances between the pyrimidines C and U, ∂(C,U), and the purines G and X, ∂(G,X).

numerically interpreted as (100,1) and (010,1) respectively, have a distance of two, ∂(C,U) = 2, whereas the Hamming distance between X = (010,0) and G = (011,0), is one, ∂(X,G) = 1, Fig. 3. In mixed parity systems the interpyrimidine or interpurine distances may be as little as one, and non-complementary purine–pyrimidine associations may be opposed in just a single D/A position. For example, the distance between the even-parity C (100,1) and the odd-parity κ (101,1) is just one, ∂(C,κ) = 1, so that the attempted association between C and X, the complement of κ, is opposed at a single position, and by the relatively weak repulsion between opposed lone-pairs. Indeed, calculations at the *ab initio* 6-31G* level of approximation give the mismatched complex C : X a net binding energy of −38.3 kJ mol^{-1}. The energetics of nucleotide association therefore is such that a single mismatch is insufficient to ensure fidelity.[7] By contrast, in a parity code the minimum distance between codewords is two. In a nucleotide context a minimum distance of two means that pyrimidines (and purines) differ from each other in the setting of two of the three D/A features, and attempted non-complementary pyrimidine–purine associations are opposed in two D/A positions. Thus, as the distance between C and U is two, ∂(C,U) = 2, the attempted association between G and U is opposed in two positions. With two opposed positions the repulsion is considerably greater; *ab initio* 6-31G* level calculations estimate a repulsive energy in the absence of wobble of 111.9 kJ mol^{-1}, for U and G.

The role of D/A patterns is therefore twofold, serving to bind associating complementary pairs, while simultaneously opposing non-complementary associations. Any set of complementary nucleotides is approximately equivalent with respect to the former, however, a parity code alphabet is optimal with respect to the latter, ensuring that the association of non-complementary pyrimidine–purine pairs will be opposed in two of three D/A positions. As information integrity is likely to have been a powerful evolutionary factor in the emergence of replicating nucleotide alphabets, selection pressure should favour parity code structured alphabets. The parity model resolves a problem raised by the study of Benner and co-workers[2] which found that κ and Π (an analogue of X) were apparently reliably replicated by polymerase, leading Orgel[8] to suggest that nature had simply failed to discover them. Error-coding analysis however suggests that mixed parity alphabets with interpurine or interpyrimidine distances of one have an inherently low fidelity.

Informatics considerations would in principle permit an alphabet of up to eight letters. Nature however uses a subset of the potential even-parity alphabet. iC and iG are not employed because of tautomeric instability.[9] Moreover, a nucleotide analogue corresponding to the D/A pattern 000 (three lone-pairs) can only be expressed using an oxygen in the central position, giving an acid anhydride readily subject to hydrolysis, thus reducing the viable alphabet to aA, U/T, C and G. Nature's choice of A instead of aA may relate to the deselection of iC. Inspection shows that the 2-amino group would be critical in opposing iC:aA associations. However, in an alphabet from which iC : iG is excluded, the 2-amino group, while not quite redundant, may not offer sufficient advantage to be particularly favoured by selection pressure.

Finally, we note that in error-coding terminology[5] an alphabet spanning the entire even-parity space is a linear systematic (4,3) code with a total of four bits, namely three information bits and a fourth parity bit. However, as nature uses just half of these the information content is just 2 bits/letter, and the code is technically a (4,2) code, a subset of the larger (4,3) code. Details may be found in the electronic supplementary information.

To summarise, error-coding considerations show how a parity code structure might offer a replication fidelity advantage. The natural alphabet appears to be structured like a parity code, and it would appear that the error-coding theory proposed by Hamming in 1950[5] was actually anticipated by nature.

I am grateful to Drs T. Gunnlaugsson and T. McKevitt for useful discussions.

Notes and references

[1] C. Y. Switzer, S. E. Moroney and S. A. Benner, *J. Am. Chem. Soc.*, 1989, **111**, 8322; S. Moran, R. Ren, S. Rumney and E. T. Kool, *J. Am. Chem. Soc.*, 1997, **119**, 2056.

[2] J. A. Picirilli, T. Krauch, S. E. Moroney and S. A. Benner, *Nature*, 1990, **343**, 33.

[3] F. H. C. Crick, *J. Mol. Biol.*, 1968, **38**, 367; M. Levy and S. L. Miller, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 7933; G. L. Wächtershäuser, *Proc. Natl. Acad. Sci. USA*, 1988, **85**, 1134.

[4] E. Szathmáry, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 2614.

[5] R. W. Hamming, *Bell Syst. Tech. J.*, 1950, **26**, 147; J. F. Humphreys and M. Y. Prest, *Numbers Groups and Codes*, Cambridge University Press, Cambridge, 1989; J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 1995.

[6] E. Szathmáry, *Proc. R. Soc. London, Ser. B*, 1991, **245**, 91.

[7] Full details of binding and mismatch energies between nucleotides will be reported elsewhere, D. A. Mac Dónaill and D. Brocklebank, manuscript in preparation.

[8] L. E. Orgel, *Nature*, 1990, **343**, 18.

[9] C. Roberts, R. Bandaru and C. Switzer, *J. Am. Chem. Soc.*, 1997, **119**, 4640.