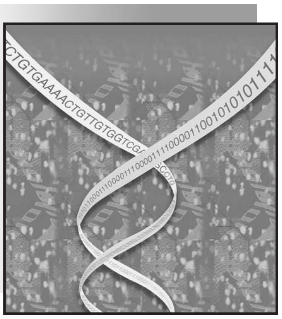
COMMUNICATION THEORY AND MOLECULAR BIOLOGY


© EYEWIRE

Digital Parity and the Composition of the Nucleotide Alphabet

Shaping the Alphabet with Error Coding

BY DÓNALL A. MAC DÓNAILL

> We not only want to know *how* nature is (and *how* her transactions are carried through), but we also want to reach, if possible, a goal which may seem utopian and presumptuous, namely, to know why nature *is such and not otherwise.*
>
> Albert Einstein [1]

When in 1953, James Watson and Francis Crick published their insight into the structure of deoxyribonucleic acid [2] and the implicit explanation of how A, C, G, and T (Figure 1) interact to form a replicating system, it signaled not only a new era in biology but, more generally, in society. The half century or so following Watson and Crick's seminal paper has produced an almost unimaginable wealth of data on the molecular mechanisms of biological systems, and as our grasp of the mysteries of molecular biology has deepened, DNA become one of the most powerful cultural icons of the 20th century [3]. Yet, for all the impact of Watson and Crick's discovery on modern society, the core concepts in DNA replication are surprisingly simple. Information is expressed as a one-dimensional string of letters written using the bases A, C, G, and T, where A and G are the larger bases, termed *purines*, and C and T are the smaller monocyclic bases, termed *pyrimidines*. The order of the letters is preserved by chemical attachment to a sugar-phosphate backbone. It can be observed, for example, that the size and shape of nucleotide T complements the corresponding features in A, so that A and T fit snugly together (Figure 1), with their association stabilized by hydrogen-bonds (weak stabilizing interactions between hydrogen atoms (H) and regions rich in accessible electron density. Similarly, nucleotides C and G complement each other, and this familiar presentation of A to T and C to G is commonly labeled the Watson-Crick arrangement. Replication of a DNA strand proceeds by using the original strand as a template against which a second, complementary strand is constructed by inserting the complementary nucleotide—say G for C or A for T—of whatever nucleotide is written in the primary strand. The resulting daughter strand may be regarded as a "negative" of the original strand. When this daughter strand is itself used as the template, the result, in the absence of error, is a perfect copy of the original strand.

Somewhat less heralded is the fact that the same 50 years has witnessed considerably less progress regarding the *why* of molecular biology. Is hydrogen-bonding necessary for nucleic acid replication, or might some other molecular interaction prove suitable? Why is the genetic code a triplet code and not a doublet code or even a quadruplet code? And, of course, the question we expand on here: Why are there four letters in the genetic alphabet and why A, C, G, and T in particular? It is one thing to explain how a system works, but it may prove quite another to explain why a particular strategy was favored over other conceivable solutions. Only when we can answer questions such as these with some certainty, can we begin to be satisfied with our grasp of the processes underlying life.

Reverse engineering offers a conceptually simple strategy, which may go some way to addressing questions such as these. By modifying a feature in the system of interest and observing the consequences, the benefit afforded by that feature, often far from obvious, may be identified. This final question, relating to the composition of the natural alphabet, is, as observed by Crick, one of the most fundamental issues in our understanding of the emergence of living matter [4]. Eschenmoser has adopted this in pursuing a series of studies exploring the chemical etiology of nucleic acid structure:

> The strategy is to conceive (through chemical reasoning) potentially natural alternatives to the nucleic acid structure, to synthesize such alternatives by chemical methods, and to compare them with the natural nucleic acids with respect to those chemical properties that are fundamental to the biological function of RNA and DNA. [5]

The sugar employed in DNA is deoxyribose, a sugar containing five carbon atoms and therefore a pentose, but the particular basis for nature's preference for this sugar over the many alternatives is not self-evident. In order to explore this, Eschenmoser synthesized a hexose analogue, a larger sugar with six carbon atoms, and observed that complementary strands of hexose-DNA do not fit together according to classical Watson-Crick rules, and that helices based on purine-purine associations, including A:A and G:G, are possible [6].

Thus, it would appear that the "choice" of ribose was not arbitrary, and that modifications of the polymer backbone, in this case the sugar, can in fact significantly affect nucleotide association and recognition. Further studies by Eschenmoser addressed the choice of the particular pentose, and it was observed that nucleic acids derived from alternative pentoses, that is, other than ribose, could result in even-stronger Watson-Crick pairing than in either RNA or DNA, with pyranosyl RNA in particular pairing exclusively in the Watson-Crick mode. Eschenmoser was able to conclude "that Nature did not choose her genetic system by the criterion of maximal base-pairing strength" [7]. Recently, and adopting a similar approach, a more complete explanation for the choice of ribose was offered by Springsteen and Joyce, who explored the reaction of cyanamide with ribose and a variety of other sugars. They reported a preferential formation of ribose-cyanamide, which can react with cyanoacetylene to form pyrimidine nucleosides), and a particular propensity of ribose-cyanamide to crystallize in aqueous solution [8].

The size of the nucleotide alphabet is a related problem, which might be amenable to a reverse-engineering approach. One expects, a priori, that in the absence of other constraints, a larger alphabet is to be preferred over a smaller one since the informational significance of a letter increases with alphabet size. It is in fact quite easy to conceive how the nucleotide alphabet might, in principle, be expanded through the inclusion of nucleotides similar in size and shape to those employed in replication but differing from the natural alphabet in the patterns of hydrogens and lone pairs (e.g., Figure 2). (In the chemical literature, for reasons relating to acid-base chemistry, such patterns are often referred to as donor-acceptor patterns, where the hydrogens attached to oxygen, O, or nitrogen, N, are potential hydrogen donors, and the electron-rich lone pairs, which may accept hydrogens from elsewhere, are potential hydrogen acceptors.) That the natural alphabet, the product of billions of years of evolution, consisted of just the four letters, A, C, G, and T, suggests that this particular set of nucleotides is somehow optimal, or close to optimal, although the reasons are not self-evident.

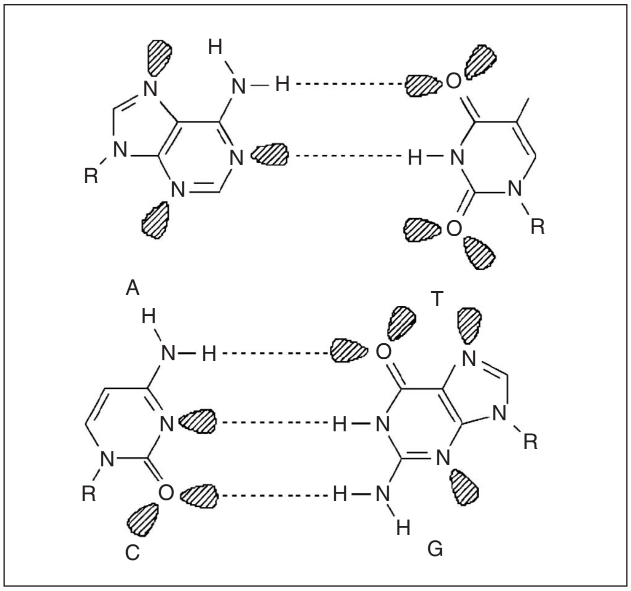


Fig. 1. The nucleotide alphabet, A (adenine), T (thymine), C (cytosine), and G (guanine), in conventional chemical notation. Interpreting the molecular representation as a graph, the vertices correspond to atoms, and the edges, or solid lines, to chemical bonds. By convention, carbon atoms (C) are not explicitly represented, and any vertex where the occupying atom is unspecified is taken to be a carbon. The shaded lobes are termed lone pairs, referring to a pair of electrons not involved in chemical bonds. Lone pairs are rich in electron density and participate in weak bonding with hydrogens (H), which are attached to nitrogen (N) or oxygen (O) atoms. Such interactions are termed hydrogen bonds, and are indicated in the figure by broken lines. The symbol R represents the sugar-phosphate backbone, to which the nucleotides are attached.

Adopting a reverse-engineering philosophy, the laboratories of both Benner [9] and Switzer [10] explored expanded alphabets, considering the noncanonical pairs κ:X and iC:iG (Figure 2). A comparison of Figures 1 and 2, will reveal that the κ:X pair is quite similar to the naturally occurring pair A:T, the most significant difference being that the hydrogen/lone-pair or donor-acceptor (D/A) pattern has been exchanged between the larger and smaller nucleotides. As their symbols suggest, iC and iG, or iso-C and iso-G, are similar to the naturally occurring C and G, differing from them in having inverted hydrogen/lone-pair patterns. The results, which showed that the additional nucleotides were accepted and copied appropriately by the natural replication apparatus, were very significant, as they implied that larger alphabets were in fact possible. One possibility was that the natural genetic alphabet might not after all be optimal and that nature had failed to discover the additional nucleotides, possibly because the biochemical apparatus which had evolved around the canonical alphabet was such that the extension of the alphabet was practically impossible. One can easily imagine that such a situation might in fact arise, but it is nonetheless

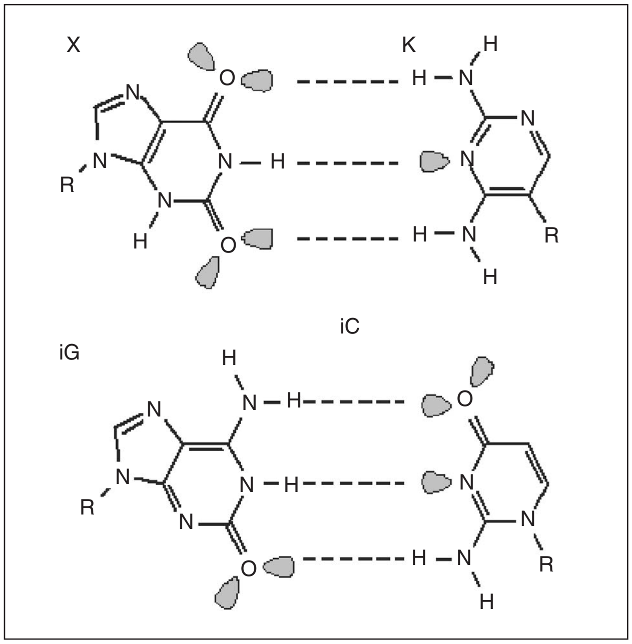


Fig. 2. Additional nucleotides κ and X (9) and iC and iG (10). These letters differ form the canonical set (Figure 1) in the hydrogen/lone-pair patterns. The symbol R represents the sugar-phosphate backbone, to which the nucleotides are attached.

Information transmission and information processing, central features of the living state, are subject to rules relating to error-coding theory.

appropriate to seek positive causes that might constrain the composition of the nucleotide alphabet.

Most studies addressing the composition of the alphabet have explored what we may loosely consider as "hardware" or physicochemical issues, such as the prebiotic availability of nucleotides [11], the role of hydrogen-bonding [12], tautomerism [13], or the effect of thermodynamic binding and mismatch energies [14]. But could there also be a "software" aspect to this and other problems in molecular biology? Stahl and Goheen suggested as much, demonstrating as far back as 1963 that some molecular biological processes could, in a strict and formal sense, be interpreted as computational processes [15], but their insight failed to have the impact that, in retrospect, one might have expected. Forsdyke invoked Hamming's error-coding theory [16] as a potential explanation for the role of introns [17], an idea subsequently investigated by Liebovitch et al. [18]. Others considered the structure information within the genetic code [19]. One may detect in these and the work of Battail [20], [21] and, more recently, in that of Mac Dónaill [22] and May et al. [23] the emergence of a "software" aspect to molecular biological systems, succinctly summarized by Dawkins [24] as follows:

> If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology.

The key concept is that information transmission and information processing, central features of the living state, are subject to rules relating to error-coding theory, quite independent of any constraint relating to the particular medium in which life is encoded. In other words, life exists at the intersection of chemistry and informatics; see Figure 3.

Nucleic acid replication, the central reaction responsible for the transmission of hereditary information, is the quintessential molecular biological information transmission phenomenon, and it is here that one might expect "software" or informatics considerations to play a constraining evolutionary role in an emergent nucleotide alphabet, in addition to any physicochemical or "hardware" constraints. A couple of studies approached this aspect of nucleotide alphabet composition. Szathmáry recognized the importance of hydrogen donor-acceptor (D/A) patterns [25], while a potential role for error-coding theory was implicitly suggested by Yockey [26] in the context of a discussion on error-coding theory in molecular biology. Yockey even assigned 5-b numerical representations to A, C, G, and U/T, but the assignments were arbitrary, and based on equating each nucleotide with elements of a four-letter code from a simple problem in error detection. For an error-coding analysis to usefully address the problem of alphabet composition, the representation of nucleotides should reflect the expression of information or patterns inherent in the nucleotides, and it is in this respect Yockey's assignments were lacking.

A potential role for informatics in constraining the composition of the alphabet seems quite plausible. In the model discussed below, we summarize an approach which acknowledges Yockey's implicit suggestion of a role for error-coding theory, but one in which the association between codewords and nucleotides is not arbitrary but based on patterns in the nucleotides themselves, echoing Szathmáry's insight into the significance of hydrogen/lone-pair patterns [22], [27]. A digital representation of nucleotides is constructed, and the problem of alphabet composition is approached from the perspective of error-coding theory.

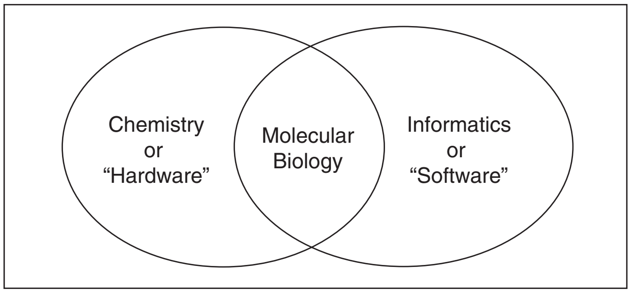


Fig. 3. A schematic representation of the relationship of molecular biology to chemistry and informatics.

Error-Coding Theory

Error-coding theory is concerned with error detection and correction in data transmission and storage systems and was proposed by Hamming more than half a century ago [16]. In its most elementary form, it involves the judicious selection of a set of binary numbers with a view to minimizing the possibility that a transmission error could go undetected, and that, subject to appropriate conditions, an error might be corrected. In the example depicted in Figure 4(a), the set of two-digit codewords $C_1 = \{00,01,10,11\}$, encompassing the set of all-two-digit binary numbers B^2, is employed to represent the four cardinal directions. Although economical in the sense that two-digits are employed to transmit 2 b of information, the code is error prone; if noise flips a bit in a codeword $c_i \in C_1$, it necessarily changes it into another codeword, $c_j \in C_1$. In the example when 01 = down is transmitted, but 00 = up is received, the received word may go undetected since it belongs to the set of possibly transmitted words.

A simple but effective remedy is to add an additional bit such that all codewords have the same parity. Parity is determined by counting the number of bits set to 1. Figure 4(b) depicts the creation of the even-parity code C_2, a subset of the available 3-b numbers in B^3 in which no two members are adjacent. An error in any single bit necessarily

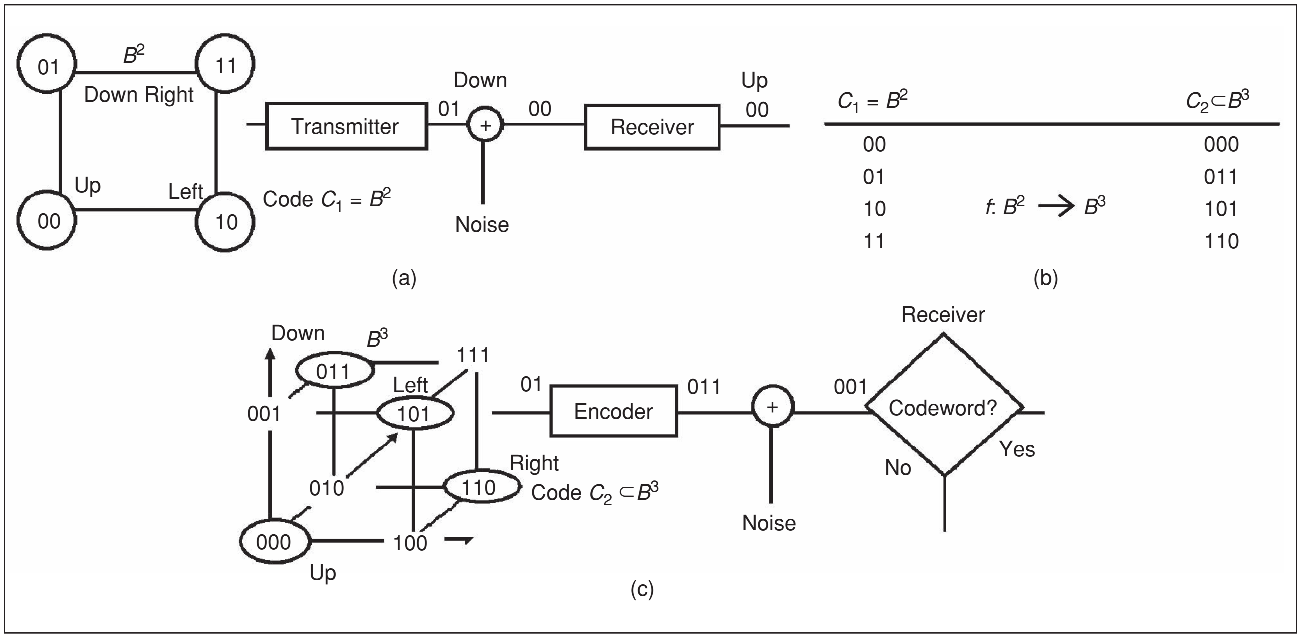


Fig. 4. (a) The set of four cardinal directions are depicted together with binary interpretations, forming the code C_1. All four possible codewords in B^2 are employed, and a change in any one bit will change one codeword into another recognized codeword. In the schematic representation of transmission, the transmitter converts the input Down to the corresponding codeword 01. However, a transmission error converts 01 into 00, and the error remains undetected since the received word 00 also belongs to code C_1. (b) The addition of a parity bit to the elements of C_1 to yield the even-parity code C_2 is represented. In this case, a change in any one bit will change one codeword into a noncodeword, and such an error would be recognized. Such an eventuality is depicted in (c), where an equivalent transmission error to that in (a) converts the codeword 011 (even) to the noncodeword 001 (odd), so that the error is detected.

changes a codeword $c_i \in C_2$ into a noncodeword $c_j \notin C_2$. The error can therefore be detected, although in this code, no correction is possible. It is important to note that not all subsets of B^3 would be equally effective; for example, the transmission error depicted in Figure 4(c) would not have been detected had the nonparity code, say $C_3 = \{000, 001, 011, 110\}$, containing both 001 and 011, been employed. The difference in codes C_2 and C_3 lies in the distance between the constituent codewords, which may be usefully 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. Inspection will show that the Hamming distance between any two codewords in code C_2 [Figure 4(b)], e.g., 011 and 110, is equal to 2. By contrast, the distance between codewords 001 and 011 in C_3 above is equal to 1, and the noise-induced conversion of 001 to 011 would go undetected.

Numerical Interpretation of Nucleotides

In nucleotide recognition hydrogens (hydrogen donors) are always opposed by lone pairs (hydrogen acceptors), and the monocyclic pyrimidines are always opposed by the two-cycle purines. This "lock-and-key" nature of molecular recognition readily admits a binary representation, and hydrogen/lone-pair (or donor/acceptor) patterns can be expressed in terms of zeros and ones, and purines and pyrimidines may be interpreted as 0 and 1, respectively. Thus, 4 b are sufficient to capture nucleotide recognition patterns (Figure 5). The particular choice of 0 or 1 for donors or acceptors, purines or pyrimidines, is of course arbitrary.

The noncanonical nucleotides considered by Benner [9] and Switzer [10] (Figure 2) are not the only ones possible. As nucleotides may present either a hydrogen donor or acceptor in each of three positions, eight different D/A patterns are available. As each pattern may be separately expressed on purines and pyrimidines, a total of 16 distinct patterns is possible, each corresponding to a potential nucleotide or nucleotide analogue (Figure 6). The binary interpretation of these nucleotides spans the binary space B^4. Error-coding theory informs us that not all sets of codewords are equally robust with respect to error resistance. If the recognition features of nucleotides may be reasonably interpreted as codewords, then perhaps not all sets of

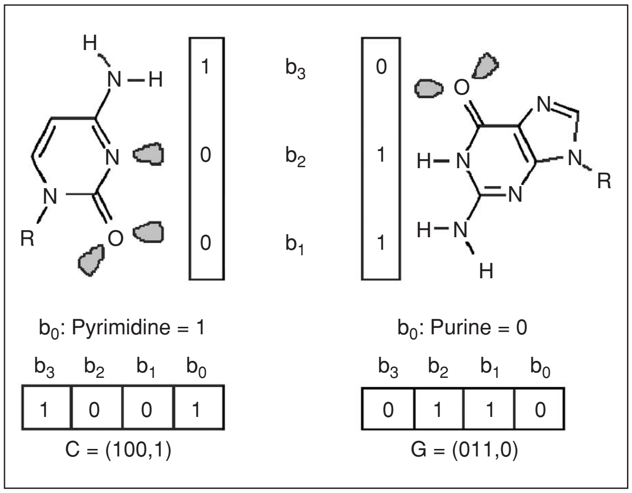


Fig. 5. Binary interpretations of nucleotides C and G.

Nucleic acid replication is the quintessential molecular biological information transmission phenomenon.

nucleotides are equally error resistant, implying a mechanism by which one alphabet might be preferred by evolution over others. Arranging nucleotides according to their numerical parity (Figure 6) reveals that members of the natural alphabet, U/T, C, G, aA (an idealized form of A), belong to the set of even-parity nucleotides. (Nature employs adenine, A, Figure 1, and not amino-adenine aA as depicted in Figure 6. However, by using aA we have a 4-b codeword in common with all the other potential codewords, simplifying exposition of the model without affecting the outcome. The difference between A and aA is discucssed in "Chemical Limitations.") For example, G and C correspond to the even-parity codewords 0110 and 1001, respectively, whereas κ and X, the noncanonical bases considered by Benner et al. [9], correspond to 0100 and 1011, both odd parity.

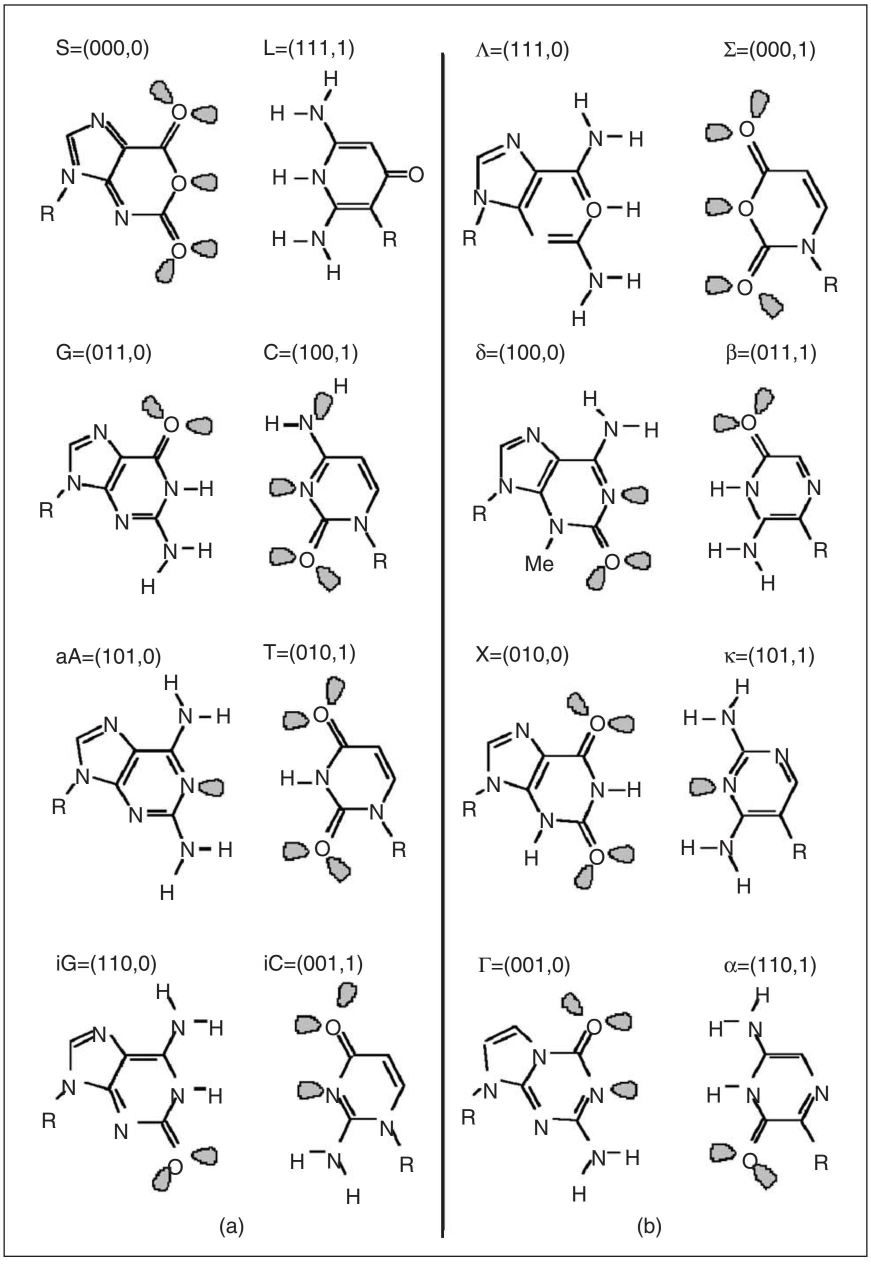


Fig. 6. The set of all 16 possible nucleotides or nucleotide analogues, together with corresponding 4-b interpretations: (a) even-parity nucleotides and (b) odd-parity nucleotides. For labels see (27).

Viewed in their binary interpretation, the recognition features of the natural alphabet appear to be structured as a parity code, in which the size of a nucleotide, its purine/pyrimidine nature, is related to the hydrogen D/A pattern as a parity bit. As one of the most elementary structures affording error resistance, a parity structure is also arguably the form most likely to be first discovered by nature. The question, therefore, is if and how such a structure might afford advantage or whether the observed structure is a mere coincidence, i.e., a frozen evolutionary accident.

Parity and Error Resistance

In a conventional transmission context the advantage of a parity code relates to the number of bits which must be changed to convert one codeword into another. No two like-parity members are adjacent (Figure 7, reproduced from [28]), and since an error in the transmission of any single bit changes the parity of the transmitted element, the corrupted codeword can be identified as not belonging to the alphabet. In terms of the Hamming distance, the minimum distance between any like-parity codewords is 2. A close molecular analogy may be observed in the phenomenon of tautomeric instability, in which the arrangement of hydrogen donors and acceptors—and, hence, the expressed information—is altered. Nucleotide G for example, equivalent to 0110 (even parity), has among its tautomers [Figure 8(a)] a pattern equivalent to 0010 (odd parity), expressing the recognition pattern of $\Gamma =$ (001,0). Here, however, the analogy ends, since although a parity-checking mechanism that might detect such tautomeric forms is conceivable, no such mechanism has been as yet been detected; and the advantage afforded by the parity code structure must be sought elsewhere.

When the role of hydrogen bonding in nucleic acid replication is considered, it is usually in terms of complementary hydrogen/lone-pair patterns which serve to stabilize the associations in C:G and A:U. Somewhat less recognized is the role of patterns in encoding the extent of repulsion between noncomplementary pairs. Inspection will reveal that in an alphabet composed of like-parity letters only, a purine presented with a noncomplementary pyrimidine finds the association resisted in two of the three hydrogen/lone-pair positions, e.g., G presented with T [Figure 9(a)]. However, in an alphabet composed of mixed parity nucleotides, where of course the purine/pyrimidine feature no longer mimics a parity bit, noncomplementary associations may be opposed in just a single position, as between C and X [Figure 9(b)]. Here, the single set of opposed lone pairs is insufficient to preclude binding [29], so that whereas G and U experience a repulsive interaction, C can bond not only with its complement G but also with X.

An alphabet in which X and C coexisted would be expected to experience a high error rate, errors which would be avoided in an alphabet composed of nucleotides of like parity. The parity-code argument suggests, therefore, that error-resistant alphabets may draw from the eight even-parity letters [Figure 6(a)] or the eight odd-parity letters [Figure 6(b)] but not both. The model sets a constraint to which the processes culminating in the development of a replicating alphabet, and ultimately the origin of life, were subject. An emergent replicating alphabet, based perhaps on just a single complementary pair, in principle could be of either parity; however once booted, the evolutionary advantage of increasing alphabet size could only be pursued by the inclusion of like-parity elements. Thus, constraints relating to information, and not to the physicochemical nature of nucleotides, preclude the natural alphabet of A, C, G, and T (even parity) from expansion through the incorporation of odd-parity elements such as κ and X.

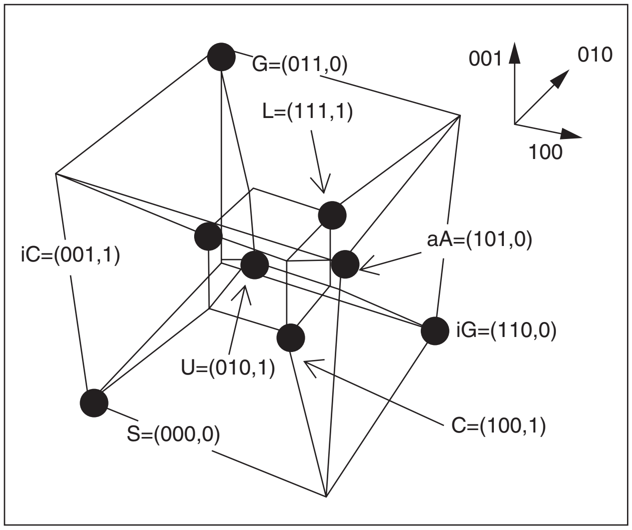


Fig. 7. The figure shows the subset of even-parity nucleotides interpreted as 4-b digits (in B^4) depicted as positions on a hypercube. It is convenient to partition the hypercube into two 3-D cubes; the outer cube represents purines (final bit = 0) and the inner cube pyrimidines (final bit = 1). The particular location of a nucleotide or codeword on a cube is determined by the three leftmost bits, expressing the binary representation of hydrogen/lone-pair patterns, which are used as coordinates. It may be observed that the distance between any two codewords (or nucleotides) is at least two bits. Figure adapted from (29).

Chemical Limitations

From an informatics perspective, therefore, it seems that the optimum alphabet would consist of eight letters and a corresponding information density of 3 b/letter ($\log_2 8 = 3$). However, the natural alphabet is not an abstract informational construct, and, when expressed in a molecular medium, it will be bounded by the physicochemical limitations of a molecular medium.

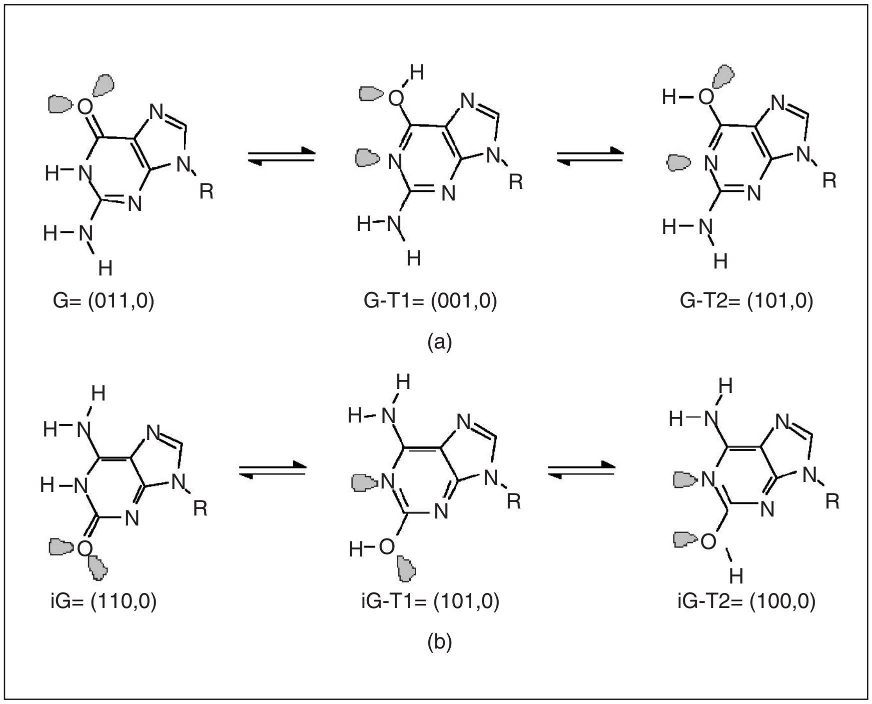


Fig. 8. (a) Some tautomeric forms of (a) G and (b) iG, together with corresponding binary interpretations. The tautomer G-T2 expresses a pattern equivalent to that on aA and would therefore match with T, while the tautomer G-T1 corresponds to the odd-parity nucleotide Γ. The pattern expressed by iG-T1 also mimics A (or aA), and again would be capable of matching with T. It readily interconverts, through internal rotation with tautomer iG-T2, yielding a pattern corresponding to δ.

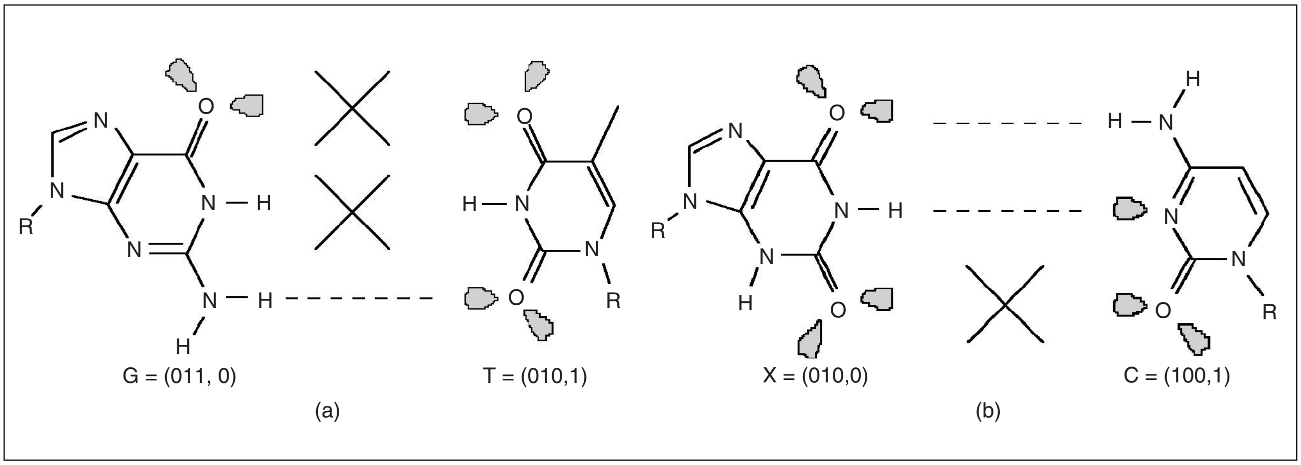


Fig. 9. (a) Association of even-parity nucleotides, U and G, (b) mixed-parity association of C (even) and X (odd). The dashed lines indicate an attractive hydrogen-bonding interaction; the crosses represent repulsive interactions between opposed lone pairs or opposed hydrogen atoms.

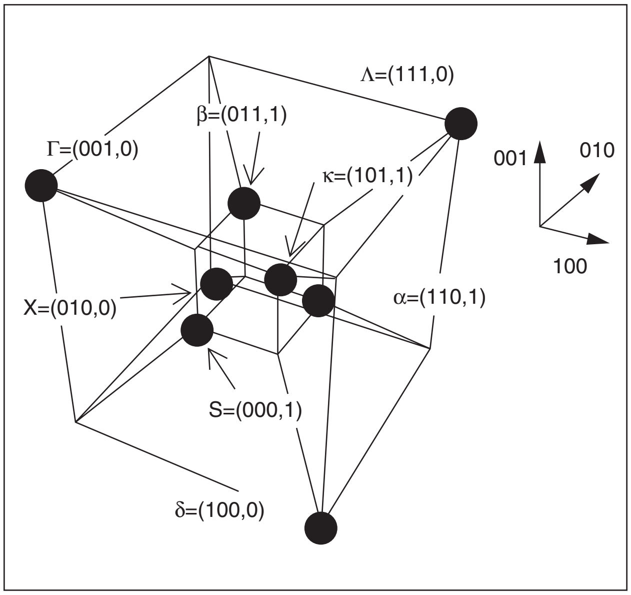


Fig. 10. Odd-parity nucleotides depicted on a hypercube. The subset of even-parity elements in the space of 4-b digits, B^4, depicted as positions on a hypercube; adapted from (29).

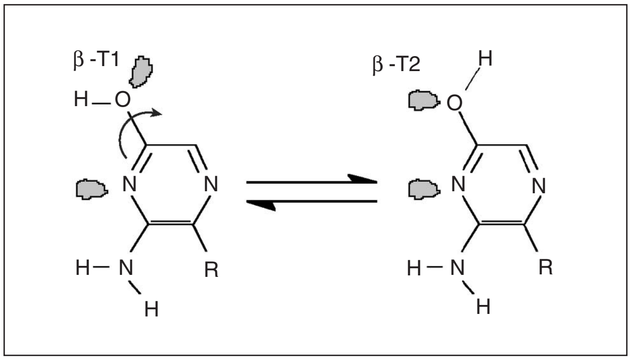


Fig. 11. Rotation about the -OH group changes the hydrogen/lone-pair pattern. As the bit pattern is unstable, tautomeric forms possessing such -OH groups are undesirable.

A viable molecular medium for genetic information demands a measure of chemical robustness with respect to both chemical degradation, in which nucleotides are destroyed and, perhaps more importantly, against pattern instability, which might cause one letter to be read for another.

Elementary chemical considerations inform us that the D/A motif of three lone pairs, encoded in S = (000,0) can only be expressed by oxygen situated between two keto groups (Figure 6). This particular motif is in fact a carboxylic acid anhydride, and is readily subject to hydrolysis in an aqueous environment. Being hydrolytically unstable, S is excluded from viable alphabets, accompanied by its complement, L = (111,1). Unlike G, the nucleotide iG proves tautomerically unstable [13], having accessible forms mimicking other nucleotides [Figure 8(b)]. Thus, whereas κ and X are excluded for error-coding reasons, iG and iC are excluded for reasons of tautomeric instability. The potential alphabet of eight even-parity letters is reduced to aA, C, G, and T by physicochemical constraints and differs from the natural alphabet, which employs adenine A (Figure 1) in preference to aA. The two-amino group in aA (the lower – NH_2 unit, Figure 6) would be required to oppose association with iC. However, in an alphabet from which iC has been deselected, this proves superfluous, and the natural alphabet of A, C, G, and T is essentially degenerate with the optimal even-parity alphabet (a more detailed discussion of the argument maybe found in [27].)

The Choice of Parity: Odd Versus Even

The parity-code model simply requires that a nucleotide alphabet be composed of nucleotides of like parity. While the arguments above explain the particular composition of the even-parity alphabet, the possibility of an odd-parity alphabet is also admitted (Figures 6 and 10). It is difficult to state with any certainty why nature selected the even-parity solution, and it must be recognized that it may simply be an accident of evolution. Nevertheless, tautomeric instability offers a possible explanation; quantum chemical simulations at the PM3 semiempirical level of approximation suggest that of the odd-parity letters depicted in Figure 6, α and β are tautomerically unstable (Figure 11) [30]. In fact, the tautomeric forms of β labeled β-T1 and β-T2 (Figure 11) are thermodynamically more stable than the reference form β. Moreover, just as in the even-parity alphabet, the 000 motif, corresponding with the three lone pairs, nucleotide Σ, is vulnerable to hydrolysis. Thus, α, β, and Σ, together with their complements, are eliminated, and the viable odd-parity alphabet is limited to κ and X.

Of the two competing parities, the even-parity set has four physicochemically viable members, each expressing 2 b/letter, compared to the odd-parity set containing just two viable letters and expressing 1 b/letter. The information necessary to express some biological functionality would be more succinctly expressed by the even-parity alphabet, with concomitant advantage in fidelity and efficiency, offering perhaps a partial explanation of why the natural alphabet is even parity. It should also be noted that the elements of the hypothetical odd-parity alphabet depicted in Figure 6 are not definitive, and as the basis of tautomeric instability is not always self-evident, it is possible that analogues with equivalent recognition patterns, yet possessing desirable tautomeric properties, might be forthcoming. However, if we assume for the moment that the preliminary conclusions are sound, then there may exist elsewhere a primitive biology based on κ and X or their close analogues. We might expect that, possessing an information density of just 1 b/letter, development of a genetic code is less likely, and the system would remain trapped in an RNA world.

Conclusions

In retrospect, a potential role for error coding in shaping the nucleotide alphabet seems obvious, and yet, with two notable exceptions, it appears to have been largely ignored; Szathmáry recognized the relationship between D/A patterns and replication errors [25], while Yockey implicitly implied a role for error coding in nucleotide transmission, assigning 5-b representations to nucleotides [26]. Unfortunately, these assignments had no physicochemical basis, being based on mapping the natural alphabet to a code employed in an error-coding text [27]. The model outlined in this paper melds these approaches, embracing a role for error coding, but one based on hydrogen/lone-pair patterns. The attraction of the error-coding description is that it offers a strikingly simple explanation of nature's choice of alphabet from among the set of potential nucleotides; optimal alphabets correspond to those in which the purine/pyrimidine feature relates to the D/A pattern as a parity bit. When this error-coding approach is coupled with chemical constraints, the natural alphabet of A, C, G, and T emerges as the optimal solution for nucleotides.

Dónall A. Mac Dónaill is a senior lecturer in advanced materials in the School of Chemistry, in the University of Dublin, Trinity College, Ireland. He received his bachelor's degree in chemistry in 1980 and completed his Ph.D. on the quantum chemical simulation of molecular solvation environments in 1984 in Trinity College. From 1984–1986 he was a postdoctoral fellow in chemistry in the University of Western Ontario, in London, Ontario, where he worked on the simulation of fast-ion conductors, with a particular interest in superionic oxides. In 1986, he took up a faculty position in the new degree program in materials science at Trinity College. In 1989–1990 he took a Hitachi Fellowship (HIVIPS) in Advanced Computation, at Hitachi Central Research Laboratory, Kokubunji, Tokyo, where he worked on the development of fast algorithms for application in materials simulation. From 1996–2002 he was the first director of the computational chemistry degree program. During this same period, he became increasingly interest in the interpretation of molecular biological phenomena as computational processes.

Address for Correspondence: Dónall A. Mac Dónaill, School of Chemistry, Trinity College, Dublin 2, Ireland. Phone: +353 1 608 1456. Fax: +353 1 671-2826, E-mail: dmcdonll@tcd.ie.

References

[1] A. Einstein, *in Festschrift für Aurel Stodola*. E. Honegger, Ed. Zürich: Orell Füssli Verlag, 1929, p. 126.

[2] J.D. Watson and F.H.C Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, pp. 737–738, 1953.

[3] D. Nelkin and M.S. Lindee, *The DNA Mystique: The Gene as Cultural Icon*. New York: W.H. Freeman, 1995.

[4] F.H.C. Crick, "Origin of the genetic code," *J. Mol. Biol.* vol. 38, no. 3, pp. 367–379, 1968.

[5] A. Eschenmoser, "Chemical etiology of nucleic acid structure," *Sci.*, vol. 284, no. 5423, pp. 2118–2124, 1999.

[6] A. Eschenmoser, "Hexose nucleic acids," *Pure Applied Chemistry*, vol. 65, no. 6, pp. 1179–1193, 1993.

[7] N. Hall, "The quest for the chemical roots of life," *Chem. Comm. (11)*, pp. 1247–1252, 2004.

[8] G. Springsteen and G.F. Joyce, "Selective derivatization and sequestration of ribose from a prebiotic mix," *J. Am. Chem. Soc.*, vol. 126, no. 31, pp. 9578–9583, 2004.

[9] J.A. Piccirilli, T. Krauch, S.E. Moroney, and S.A. Benner, "Enzymatic incorporation of a new base pair into DNA and RNA extends the generic alphabet," *Nature*, vol. 343, no. 6253, pp. 33–37, 1990.

[10] C.Y. Switzer, S.E. Moroney, and S.A. Benner, "Enzymatic incorporation of a new base pair into DNA and RNA," *J. Am. Chem. Soc.*, vol. 111, no. 21, pp. 8322–8323 1989.

[11] G. Zubay and T. Mui, "Prebiotic synthesis of nucleotides," *Origins Life Evol. Bio.*, vol. 31, no. 1–2, pp. 87–102, 2001.

[12] K.M. Guckian, T.R. Krugh, and E.T. Kool, "Solution structure of a nonpolar, non-hydrogen-bonded base pair surrogate in DNA," *J. Amer. Chem. Soc.*, vol. 122, no. 29, pp. 6841–6847, 2000.

[13] C. Roberts, R. Bandaru, and C. Switzer, "Theoretical and experimental study of isoguanine and isocytosine: Base pairing in an expanded genetic system," *J. Am. Chem. Soc.,* vol. 119, no. 20, pp. 4640–4649, 1997.

[14] E. Szathmáry, "4 Letters in the genetic alphabet—a frozen evolutionary optimum," in *Proc. Roy. Soc. Lon Ser.*, 1991, vol. B 245, no. 1313, pp. 91–99.

[15] W.R. Stahl and H.E. Goheen, "Molecular algorithms," *J. Theor, Biol.,* vol. 5, no. 2, pp. 266–287, 1963.

[16] R.W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Technical J.*, vol. 29, no. 2, pp. 147–160, 1950.

[17] D. Forsdyke, "Are introns in-series error-detecting sequences?," *J. Theor. Biol.*, vol. 93, no. 4, pp. 861–866, 1981.

[18] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine, "Is there an error-correcting code in the base sequence of DNA?," *Biophys. J.,* vol. 71, no. 3, pp. 1539–1544, 1996.

[19] J. Rzeszowska-Wolny, "Is genetic code error-correcting?," *J. Theor. Biol.*, vol. 104, no. 4, pp 701–702, 1983.

[20] G. Cullmann and J.-M. Labouygues, "Noise immunity of the genetic code," *Biosyst.*, vol. 16, no. 1, pp. 9–29, 1983.

[21] G. Battail, "Does information theory explain biological evolution?," *Europhys. Lett.*, vol. 40, no. 3, pp. 343–348, 1997.

[22] D.A. Mac Dónaill, "A parity code interpretation of nucleotide alphabet composition," *Chem. Comm.,* no. 18, p. 2062–2063, 2002.

[23] E.E. May, M.A. Vouk, D.L. Bitzer, and D.I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *Biosyst.*, vol. 76, no. 1–3, pp. 249–260, 2004.

[24] R. Dawkins, *The Blind Watchmaker.* Longman Scientific & Technical, 1986.

[25] E. Szathmáry, "What is the optimum size for the genetic alphabet?," in *Proc. Natl. Acad. Sci., USA*, vol. 89, no. 7, pp. 2614–2618, 1992.

[26] H.P. Yockey, *Information Theory and Molecular Biology*. Cambridge, UK: Cambridge Univ. Press, p. 102, 1992.

[27] D.A. Mac Dónaill, "Why nature chose A, C, G and U/T: An error-coding perspective of nucleotide alphabet composition," *Origins Life Evol. Bio.*, vol. 33 no. 4-5, pp. 433–455, Oct. 2003.

[28] D.A. Mac Dónaill, "The concept of parity in nucleotides: Implications for the possible existence of alternative alphabets," in *Proc. 2nd European Workshop on Exo-Astrobiology*, Graz, 2002, pp. 99–102.

[29] D.A. Mac Dónaill and D. Brocklebank, "An ab initio quantum chemical investigation of the error-coding model of nucleotide alphabet composition," *Mol. Phys.*, vol. 101, no. 13, pp. 2755–2762, 2003.

[30] D.A. Mac Dónaill, "Tautomerism as a constraint on the composition of alternative nucleotide alphabets," *Artificial Life VIII—Proc. 8th International Conference on Artificial Life*, Sydney, 2002, pp. 106–110.