

Tautomerism as a Constraint on the Composition of Alternative Nucleotide Alphabets

D. A. Mac Dónaill

Department of Chemistry, Trinity College, Dublin 2, Republic of Ireland. E-mail: dmcdonll@tcd.ie

Abstract

The factors determining the composition of the nucleotide alphabet are not self-evident. While, straightforward extensions to the alphabet such as iC:iG have received some theoretical attention, more exotic nucleotide pairs, such as α : Γ , β : δ , and κ :X, have been ignored. At the current state of knowledge the possibility of viable alternative nucleotide alphabets remains an intriguing possibility. In this paper the broader space of potential letters is considered in terms of hydrogen donor-acceptor patterns. Quantum chemical simulations at the PM3 semi-empirical level of approximation are employed to examine the stability of different tautomeric forms of candidate letters. The results suggest that tautomeric instability favoring hydroxyl elements over keto groups is one of the critical constraints on viable alternative alphabets. Rotation of a hydroxyl group changes the expressed hydrogen donor/acceptor pattern, destroying the integrity of the principal recognition feature. Two distinct alphabets appear viable, one containing the familiar aA:T/U and C:G, and a second containing just κ :X. Thus tautomerism offers an explanation for the composition of the natural alphabet, while posing challenges for the engineering of alternative alphabets.

Introduction

The replication of nucleotide texts is based on molecular recognition through the association of complementary hydrogen donor/acceptor (D/A) patterns and pyrimidine/purine motifs. Three positions with the capacity for H-bonding give a potential alphabet of 16 informationally distinct nucleotides, of which the natural nucleotide alphabet of A, C, G and T/U is a subset. This subset however is not self-evidently optimal; the successful incorporation by polymerase of the additional base pairs, iC:iG (Switzer, Moroney, & Benner 1989) and κ : π (Picirilli *et al.* 1990) has been demonstrated, and thus viable alternative nucleotide alphabets seem possible.

Nature's failure to avail of a larger alphabet has prompted some discussion of the features shaping the natural alphabet (Szathmáry 1991; 1992). An alphabet of n letters, all equally employed, contains $\log_2 n$ bits

of information per letter, and thus an alphabet of eight nucleotides would contain three bits of information per letter, compared to just two bits in the natural alphabet. A larger alphabet would appear to confer an advantage on a replicating system, allowing complex functionality to be specified in a shorter polymer than would the natural nucleotide alphabet.

The composition of an emergent alphabet is likely to be shaped by two parallel considerations, the prebiotic availability of potential letters, and the suitability of candidate letters as information carriers. Orgel (1990) suggested that perhaps nature had simply failed to discover additional letters. However, this cannot be assumed, and it seems reasonable to explore any factors which might afford advantage, with regard to both the evolutionary emergence of the natural alphabet, and the engineering of alternative alphabets.

Information integrity is the most fundamental requirement of a molecular alphabet. As information in nucleotides is partly expressed in hydrogen D/A patterns, the phenomenon of tautomerism, whereby the position of hydrogen may be exchanged between different positions, is of critical importance. Tautomerism may affect information integrity in two ways; firstly, the simple existence of two or more thermally accessible forms destroys the integrity of information expressed in D/A patterns. Secondly, in the case of oxygen, whenever the enol form is favoured over the keto form, the expressed D/A pattern can be changed simply by the rotation of the hydroxyl group. The role of tautomerism in the iC:iG base pair was considered by Switzer group (Roberts, Bandaru, & Switzer 1997). Their results concurred with experimental observation of the existence of stable tautomeric forms for iG, one of which had a pattern analogous to aA (amino-adenine) and therefore complementary to U/T. Thus an instance of iG written, by polymerase or its equivalent, as the complement of iC, might in turn express a U, giving an iC \rightarrow U transition (Fig. 1). The D/A pattern of iG may therefore be regarded as volatile, lowering the replication fidelity of any alphabet to which it belongs, so that the iC:iG base pair is likely to be excluded by selection pressure.

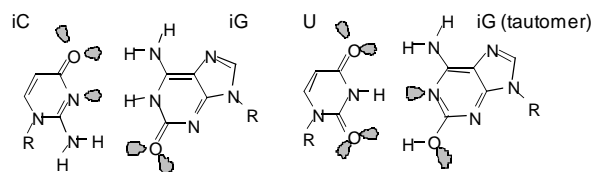


Figure 1: Tautomerism in iG gives it the capacity to bond with different pyrimidines, lowering fidelity.

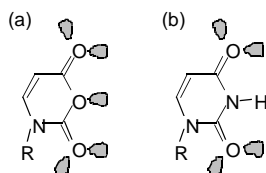


Figure 2: (a) The triple acceptor motif expressed on a ‘pyrimidine’, (b) a nitrogen in the central position must express a hydrogen.

Curiously, other potential nucleotides have not figured significantly in the debate, and it is the purpose of this study to consider the broader family of nucleotides, beyond the natural four, which together with iC:iG, have been the most widely studied.

We preface our consideration of other nucleotides by a few observations which it is hoped will facilitate later discussion. With three H-bonding positions there are eight possible hydrogen D/A patterns, which in turn may be expressed as either ‘purines’ or ‘pyrimidines’, giving a total of 16 possible letters. Of these, four correspond to hypothetical nucleotides expressing patterns of three donors (hydrogens), and three acceptors (lone pairs). Simple chemical considerations show that the chemical expression of three acceptors requires an O in the central H-bonding position giving an acid anhydride, readily subject to hydrolysis (Fig. 2), since a N atom in the central position, as in U, must express a hydrogen. Because of its ease of hydrolysis nucleotides expressing the triple-acceptor or complementary triple donor motifs are not considered; for all other potential nucleotide letters however, the issue of prebiotic availability is set aside, the study focusing instead on the issue of tautomeric stability.

The reduced potential alphabet, now consisting of 12 letters, is usefully divided into two sets. Set I (Fig. 3a) consists of pyrimidines expressing two hydrogen acceptors and a single hydrogen donor, together with their complementary purines. This set consists of the natural alphabet, where A is represented by the ‘ideal’ form amino-adenine (aA), as well as iC and iG. Set II (Fig. 3b) consists of potential alphabet nucleotides in which the ‘pyrimidines’ now express two hydrogens, and their ‘purine’ complements. Set II is a digital mirror image of

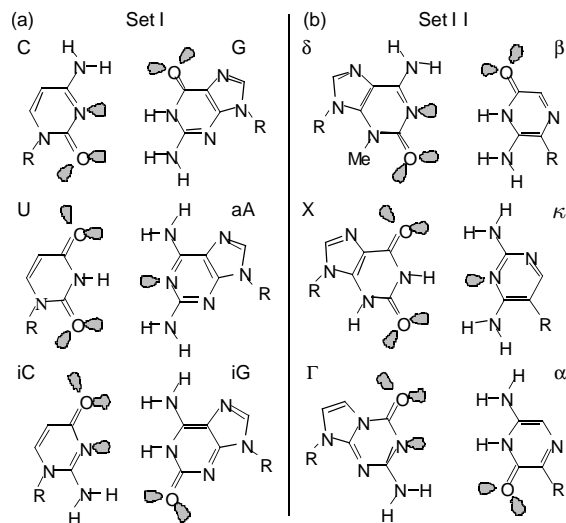


Figure 3: (a) Nucleotide set I, consisting of the natural alphabet and iC and iG. Purines express two hydrogens and a single lone pair, while pyrimidines express a single hydrogen; (b) Set II. Purines express a single hydrogen and two lone pairs. ‘Pyrimidines’ express two hydrogens.

set I, the essential difference between the two sets being that the D/A patterns expressed on pyrimidines in set I are expressed on purines in set II, and vice versa. Labels for set II nucleotides are taken from Szathmáry (1991).

Viable alphabets may in principle arise within one or other of the sets, but are unlikely to contain pairs from both sets. Inspection will show that attempted purine-pyrimidine associations between non-complementary nucleotides within the same set may be opposed by mismatches in two positions, for example U:G or $\kappa:\delta$. However, potential alphabets drawing on pairs from both sets I and II are opposed by mismatches in just a single position, facilitating replication errors. For example, an alphabet containing the pairs C:G and $\kappa:X$, admits the possibility of a mismatch between C and X, notwithstanding the opposed lone pairs (Fig. 4). Using PM3 calculations we estimate a binding energy of 4.9 kJ mol^{-1} . Net binding energies between non-complementary nucleotides is clearly quite untenable. Hypothetical alphabets drawing from both sets I and II would have little fidelity, and sets I and II may therefore be regarded as orthogonal. See Mac Dónaill (2002) for a discussion in terms of error-coding theory.

The alphabet of our own biological world is a subset of set I. Arguments based on tautomeric stability alone account for the deselection of iC and iG. Set II offers an alternative menu of potential nucleotides from which an alphabet might be constructed, and offers the possibility of an alternative replicating molecular alphabet. This paper examines the potential of set II nucleotides as viable information carriers, focusing on the integrity

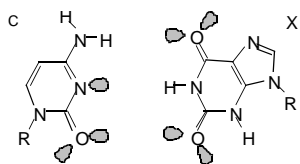


Figure 4: Possible mismatch in a hypothetical alphabet containing C:G from set I and κ :X from set II.

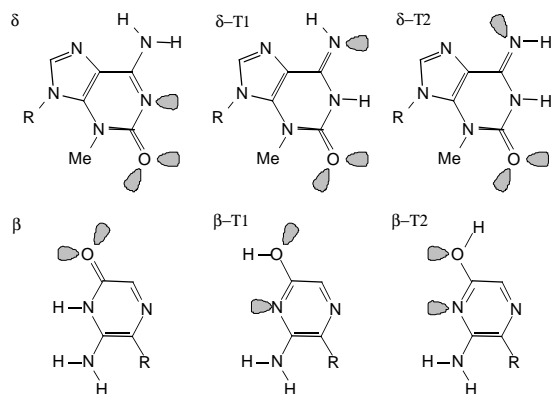


Figure 5: Tautomeric forms of Set II nucleotides δ and β .

of the D/A patterns, as reflected in the relative stability of tautomeric forms.

Calculations

Many different hypothetical nucleotides could in principle express the D/A patterns required in set II. Those chosen for consideration in this preliminary study, and depicted in Fig. 3, are those already considered in the literature (Picirilli *et al.* 1990; Szathmary 1991). Quantum chemical calculations are performed at the PM3 level of approximation, one of the more reliable semi-empirical quantum computational methods. The backbone, represented in figures by R, is represented in calculations by H.

The δ : β Nucleotide Pair

We begin by considering the δ : β complementary pair (Fig. 5). The energies from PM3 calculations for nucleotides δ and β are given in table 1. The conventional form of δ is the most stable, tautomeric forms being less stable, although δ -T1 is thermally accessible. However, in the case of nucleotide β , hydroxyl forms exhibit greater stability than the conventional keto form.

The D/A pattern encoded by a hydroxyl group is not stable, and a simple, thermally accessible, rotation about the C–O bond of the hydroxyl group, exchanges tautomers β -T1 and β -T2 (Fig. 6). Tautomer β -T1 can bind with δ , with a binding energy of 33.4 kJ mol^{-1} , whereas tautomer β -T2 can bind to X with a binding

Nucleotide	Heat of Formation	Relative Energy
δ	44.92	0.00
δ -T1	50.60	5.69
δ -T2	73.01	28.09
β	-22.31	0.00
β -T1	-42.84	-20.53
β -T2	-35.41	-13.10

Table 1: Heats of Formation and Relative Energies of the tautomeric forms of δ and β in kJ/mol.

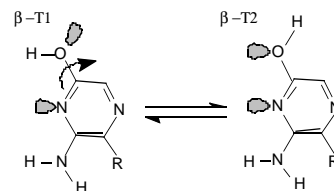


Figure 6: Rotation about the OH group changes the D/A pattern.

energy of 4.0 kJ mol^{-1} , subject only to a relatively weak repulsion between two lone pairs (Fig. 7). The capacity of β in its hydroxyl manifestations to bind with two different purines renders β unsuitable as an information carrier. We conclude that δ : β cannot participate in a nucleotide alphabet

The Γ : α Nucleotide Pair

Tautomeric forms of Γ and α are depicted in Fig. 8. PM3 calculations on Γ indicate that the tautomeric forms of Γ are inaccessible, and that Γ is therefore a potentially viable letter. α however is essentially similar to β , differing only in its point of attachment to the backbone. Calculations, using both -H and -Me to represent the backbone confirm similarly stable tautomeric forms for α and β ; α -T1 can bind to X, while α -T2 can bind to Γ (Fig. 9). Thus, as with δ : β , the nucleotide pair α : Γ may not participate in a molecular alphabet.

The κ :X Nucleotide Pair

Some simple tautomeric forms of nucleotides κ and X are depicted in Fig. 10. Calculations suggest that all tautomeric forms other than the conventional representation

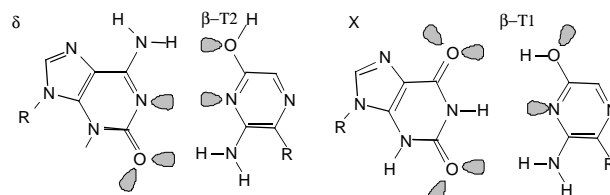


Figure 7: Tautomeric forms of β binding with both δ and X.

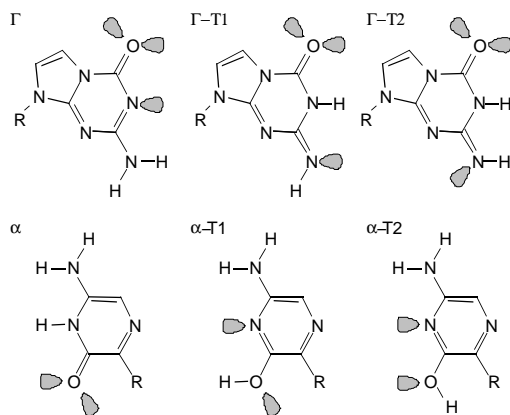


Figure 8: Tautomeric forms of Set II nucleotides Γ and α .

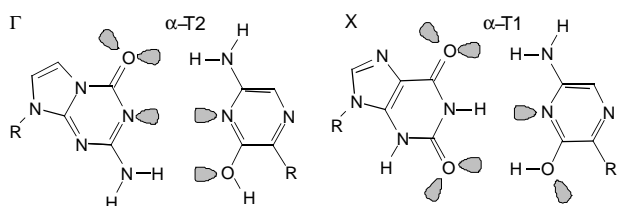


Figure 9: α -T2 can bind to Γ , while α -T1 can bind to X.

are thermally inaccessible. The D/A patterns encoded in κ and X are therefore stable, and κ :X are therefore a viable nucleotide pair.

Discussion

The stability of D/A patterns depends largely on the equilibrium between the keto and enol manifestations of oxygens. Carbonyl oxygens unambiguously express an acceptor, whereas hydroxyl groups ambiguously express an acceptor or donor, allowing binding with multiple pseudo-complements. Of the two sets of potential nucleotides, set I has four viable members, essentially the set we find in terrestrial biology today. Set II by contrast contains just two viable letters. In information terms set I alphabets have 2 bits/letter, whereas set II has just 1 bit/letter. The information necessary to express some

Nucleotide	Heat of Formation	Relative Stability
Γ	49.40	0.00
Γ -T1	98.33	48.93
Γ -T2	73.25	23.86
α	-22.31	0.00
α -T1	-42.84	-20.53
α -T2	-35.41	-13.10

Table 2: Heats of Formation and relative energies of nucleotides Γ and α in kJ mol^{-1} .

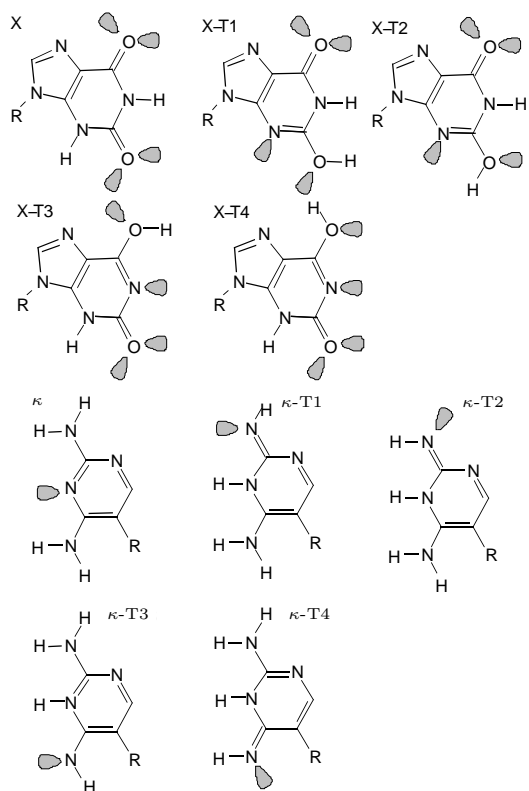


Figure 10: Some tautomeric forms of nucleotides κ and X.

biological functionality can be more succinctly expressed in set I, yielding superior fidelity, perhaps a partial explanation of why the natural alphabet derives from set I nucleotides.

If artificial alphabets are to be developed, by extension of set I or by exploitation of set II, problems associated with tautomerism must be overcome. This is quite a subtle problem; it is not self-evident for example, why κ is tautomericly stable whereas β is not. Moreover, the members of set II chosen for study are not definitive and it is quite possible that analogues with stable and desirable tautomeric properties might be forthcoming. Additionally, the difficult problem of the role of solvation must be considered. These and related problems are the focus of ongoing high level *ab initio* calculations.

References

- Mac Dónaill, D. A. 2002. A parity code interpretation of nucleotide alphabet composition. *Chem. Comm.* in press.
- Orgel, L. 1990. Adding to the genetic alphabet. *Nature* 343:18–20.
- Picirilli, J. A.; Krauch, T.; Moroney, S. E.; and Benner, S. A. 1990. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* 343:33–37.

Nucleotide	Heat of Formation	Relative Stability
X	94.43	0.00
κ -T1	154.52	60.09
κ -T2	177.48	83.05
κ -T3	166.88	72.45
κ -T4	196.99	102.56
X	-187.25	0.00
X-T1	-146.04	41.21
X-T2	-169.22	18.03
X-T3	-145.48	41.77
X-T4	-142.06	45.19

Table 3: Heats of Formation for tautomers of X and κ in kJ mol⁻¹.

Roberts, C.; Bandaru, R.; and Switzer, C. 1997. Theoretical and experimental study of iso-C and iso-G: Base-pairing in an expanded genetic system. *J. Am. Chem. Soc.* 119:4640–4649.

Switzer, C. Y.; Moroney, S. E.; and Benner, S. E. 1989. Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.* 111:8322–8323.

Szathmáry, E. 1991. *Proc. Roy. Soc. London Ser. B* 245:91–99.

Szathmáry, E. 1992. What is the optimum size for the genetic alphabet? *Proc. Natl. Acad. Sci. USA* 89:2614–2618.