OPINION

Why are there four letters in the genetic alphabet?

Eörs Szathmáry

We list, without thinking, the four base types that make up DNA as adenine, guanine, cytosine and thymine. But why are there four? This question is now all the more relevant as organic chemists have synthesized new base pairs that can be incorporated into nucleic acids. Here, I argue that there are theoretical, experimental and computational reasons to believe that having four base types is a frozen relic from the RNA world, when RNA was genetic as well as enzymatic material.

In 1930, the eminent population geneticist Sir Ronald Fisher wrote: "No practical biologist interested in sexual reproduction would be led to work out the detailed consequences experienced by organisms having three or more sexes; yet what else should he do if he wishes to understand why the sexes are, in fact, always two?" [1]. By the same token, it could be asked why the size of the genetic alphabet is always four, both in DNA (A, T, C and G) and RNA (A, U, C and G). The elegance of the Watson–Crick model[2] of DNA indicates that no significant deviation from it is possible (FIG. 1). This seems especially true for the genetic alphabet: the constraints on base pairing make it difficult to imagine an alphabet consisting of, for example, eight letters. Yet this is exactly what was successfully proposed by Steven Benner and colleagues, who observed that the order and combination of hydrogen-bond acceptors and donors allows organic chemists to design new base pairs[3] (BOX 1). This 1990 paper has opened up serious experimental and theoretical research aimed at artificially creating alternative genetic alphabets and, at the same time, explaining why we have the four-letter alphabet.

Prompted by Benner's work, further investigation into our genetic alphabet has come mainly from synthesizing new base pairs: these can be used to probe the extent to which alternative genetic alphabets are, or were, feasible (as discussed below), as well as to test ideas about the mechanism of DNA/RNA polymerization by polymerase enzymes[4]. In science, it is generally rewarding to question the foundations of a discipline, provided that this can be done in a constructive manner; relativity theory, to mention a well-known case in physics, has questioned and put into context crucial ideas of Newtonian mechanics such as time, space, velocity and mass. If we consider alternative/extended genetic alphabets, our knowledge of the existing one can be expected to become deeper and better founded.

In this article, I summarize the practical constraints on creating extended or alternative genetic alphabets, and describe the most promising experimental attempts to do so. I argue that there are theoretical reasons to believe that the optimum size for a genetic alphabet in an RNA WORLD (in which the present alphabet is thought to have evolved) is four. Finally, I discuss prospects for the future: I believe that the field will develop by taking a more systematic experimental and theoretical approach to investigating the alternative alphabets that have been created. Such investigations seem to have speeded up in the past two years and there is hope that they will confirm old insights and yield some new ones. Alternatives to the phosphodiester backbone are not the subject of this article and have been addressed elsewhere (see REF. 5 for discussion).

Lessons from our genetic alphabet

Studies of the present genetic alphabet and of the contemporary DNA replication machinery highlight certain constraints that any alternative/extended genetic alphabet is expected to be subject to. These should guide us when attempting to synthesize new base pairs; conversely, the successful addition of new base pairs to the alphabet can sharpen our understanding of these constraints. There are four main constraints on the successful incorporation of a new base pair[6–8]: chemical stability (the base should not readily decompose); thermodynamic stability (new base pairs should not destabilize nucleic-acid structures); enzymatic processability (polymerases should accept the base pairs as substrates, catalyse addition to the primer and be able to carry on the process); and kinetic selectivity (ORTHOGONALITY to other base pairs). All four criteria are important but the combination of the last two, which we might call replicability, has received particular attention because it is the main obstacle to adding to the genetic alphabet.

Replicability. As Watson and Crick observed in their classic paper[2], complementarity of the hydrogen-bonding patterns of the bases is important not only for stability but also for replicability of the structure. Replication is carried out by polymerase enzymes, which form a tight pocket around the complex of the template and the primer strand (reviewed in REF. 4). This tightness is thought to contribute strongly to the kinetic selectivity of polymerization: the better the steric match between the opposite base in the template and the incoming base to be attached to the end of the growing (primer) strand, the more accurate this step of polymerization will be. This requirement for shape complementarity allows, as we shall see, the insertion of 'bases'

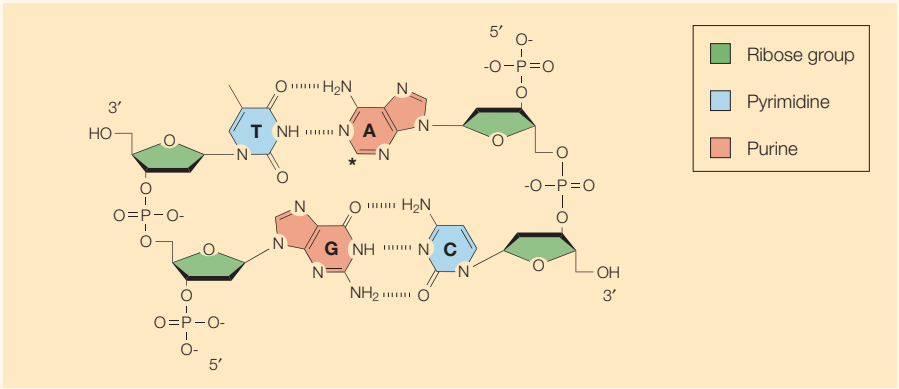


Figure 1 | **Base-pairing pattern of a DNA molecule.** A piece of DNA showing the Watson–Crick base pairs A•T and G•C. Note that A could have a third hydrogen bond at its position 2 (asterisk). The 'minor groove' in this representation (the narrower of the two helical grooves that are formed by the intertwined DNA strands) is to the bottom of the base pairs; in fact, the base pairs are stacked on top of one another and orientated perpendicular to the viewer.

without hydrogen bonds. Effectively, polymerases allow DNA to recognize complementary nucleotides more selectively. Therefore, Watson–Crick pairing does not seem to be important for insertion, but has a role in maintaining the accuracy of replication.

Polymerases seem to be idiosyncratic as to which other features of base pairs they are sensitive to. A recurring feature is the recognition of some of the groups that face the minor groove of DNA (FIG. 1) by hydrogen bonds between these groups and the polymerase. This is why different polymerases can be 'choosy' in different ways when challenged with alternative nucleotides.

These features must be carefully considered when devising new bases. Insertion might be efficient but the polymerase might then stall, or it might act processively but with low fidelity because the new bases are not sufficiently orthogonal to (different from) the pre-existing bases. The canonical genetic alphabet and the polymerase enzymes have co-evolved; it is therefore to be expected that existing polymerases might not be ideal for experiments on extended alphabets.

Experimental approaches

Following the pioneering and visionary 1990 paper from the Benner group[3] (BOX 1), several attempts have been made to create an extended genetic alphabet. These have met with encouraging, but by no means overwhelming, success. So far, something has always been missing: either chemical or thermodynamic stability, kinetic orthogonality or replicative processivity. Modest replicability seems to be the limitation that is hardest to overcome. These limitations notwithstanding, it is instructive to survey some new 'base' candidates that either obey the Watson–Crick mechanism of selectivity (mediated by complementary hydrogen-bonding patterns) or in which complementarity is extended to include a more general shape complementarity (without hydrogen bonds).

Base-pair complementarity. The new bases synthesized by Benner's group and shown in BOX 1 were designed to pair with one another according to their hydrogen-bonding patterns. However, as the examples below show, this approach has not been very successful.

Consider the case of the isocytosine: isoguanine (*iso*-C•*iso*-G) pair[9–11] (BOX 1). *Iso*-C decomposes by DEAMINATION and *iso*-G assumes various tautomeric forms (whereby its hydrogen-bonding pattern is rearranged). As both processes strongly undermine the maintenance of genetic information, it is unlikely that this base pair had a role in early genetic systems[11].

The xanthosine (X, puADA):2,4-diamopyrimidine (pyDAD) pair (BOX 1) has been studied at length. Although this base pair

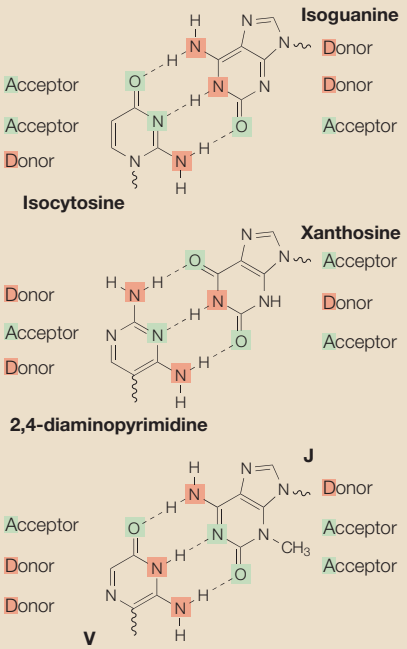
Box 1 | **Landmark experiments aimed at extending the genetic alphabet**

The figure shows 6 out of a total of 12 members of the extended DNA alphabet that was created by Benner[3] (who is an organic chemist with a deep interest in the chemical foundations of biology) and his group. The crucial idea behind the design of these bases is a permutation of the normal hydrogen-bond acceptor (A) and donor (D) groups, so that they appear in various serial orders on the ring facing the other half of a potential pair. In the two natural base pairs the pattern is DAA for C, ADA for U, ADD for G and DA- for A (the hyphen indicates a missing amino group that could be there; see asterisk in FIG. 1). For example, the hydrogen-bonding principle behind the new base pair isocytosine:isoguanine (*iso*-C•*iso*-G; see figure) is AAD and DDA, respectively. The breakthrough was not only in the experimental design, but also in the technical expertise that is required for the partial synthesis of new base pairs such as *iso*-C•*iso*-G[3].

The bases shown in the figure have been synthesized by organic chemical means and partly enzymatically incorporated into natural nucleic-acid molecules by organic chemical methods. However, the greatest challenge is to make such new base pairs replicable in the context of nucleic acids. The best result so far has been achieved for the xanthosine (X, puADA):2,4-diaminopyrimidine (pyDAD) pair (A. M. Sismour *et al.*, manuscript in preparation) in Benner's laboratory.

Besides opening up a new field, a crucial merit that is shared by Benner's original paper[3] and those that describe the incorporation of the new base pairs into nucleic acids (REFS 9,11 and A. M. Sismour *et al.*, manuscript in preparation) is the emphasis on the importance of the PROCESSIVITY by known or mutated forms of polymerase enzymes. Processivity is crucial if the enzyme is expected to extend the growing strand beyond an incorporated new base.

Future efforts should primarily aim to achieve the successful extension of the genetic alphabet, and to assess the template and enzymatic capabilities of macromolecules that are built from bases of the extended alphabet.



was chemically synthesized a long time ago[12], attempts at incorporating it into DNA have been made only relatively recently (A. M. Sismour *et al.*, manuscript in preparation): a doubly mutant version of the human immunodeficiency virus type I (HIV-1) reverse transcriptase has been shown to accept and copy this base pair in an oligonucleotide through several rounds of the polymerase chain reaction (PCR). The use of mutant polymerases is likely to be the most successful way to obtain replicases and polymerases that are acceptable to new base pairs.

A third example is that of the new base pair V•J (BOX 1). This pair has been successfully synthesized and incorporated, but is neither sufficiently selective in pairing nor stable enough. It is problematic that V efficiently mispairs with A, and J mispairs with U. Worse still, V EPIMERIZES into a form that is not suitable for a genetic role. Interestingly, this might not occur in the absence of the sugar-ring oxygen[13], which raises the possibility of a useful interplay with a modified nucleic-acid backbone and an extended alphabet. For the time being, the V•J pair is ruled out by chemical instability and insufficient thermodynamic orthogonality to the present-day alphabet.

Shape complementarity. Another line of investigation builds on the concept of shape complementarity alone. An impressive number of non-hydrogen-bonding base analogues have been synthesized at the Scripps Research Institute in California. Here, I focus on the most successful achievements of this laboratory. Note that a new base pair that satisfies all four conditions (chemical and thermodynamic stability, enzymatic processability and kinetic selectivity) has not yet been achieved.

A good example is the self-pair of 7-aza-indole-nucleoside (7AI; FIG. 2a). It might, at first, seem surprising to consider a self-pair, but in fact this solution relaxes the constraints that must be satisfied (with complementary base pairs, the requirements of stability, orthogonality and processivity must be satisfied twice, once for each base). Testing of 7AI has been done innovatively in the sense that two polymerases have been used: mammalian polymerase-β and the KLENOW FRAGMENT[14]. The Klenow fragment is responsible for efficient insertion and polymerase-β extends the primer efficiently, even following a 7AI•7AI pair. Extension proceeds with the same range of efficiency as when following canonical bases.

Derivatives of isocarbostyryl (ICS; FIG. 2b) have also been tested thoroughly. Among these, 10-thio-6-*N*-isocarbostyryl (SNICS; FIG. 2c) — which also forms a SNICS•SNICS self-pair— stands out. Its thermodynamic stability, insertion selectivity and extension efficiency by the Klenow fragment are in the promising range (the extension efficiency is $2 \times 10^4\,M^{-1}\,min^{-1}$)[15]. Although it is true that A is too frequently inserted opposite SNICS and that the extension past the SNICS•SNICS pair is two orders of magnitude less efficient than that past an A•T pair, the SNICS•SNICS pair represents an important advance towards expanding the alphabet.

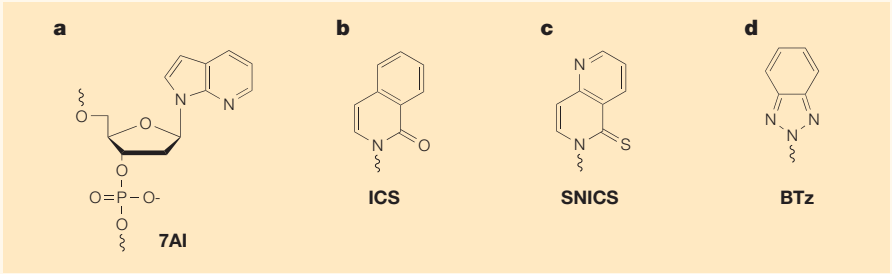


Figure 2 | **Base-pairing pattern dependent on shape complementarity.** Shown are four surprising new 'bases' that can be inserted into nucleic acids according to shape rather than hydrogen-bonding complementarity. These structures form self-pairs (such as 7AI•7AI) and have a strongly hydrophobic character (see main text for details). **a** | 7-aza-indole-nucleoside (7AI). **b** | Isocarbostyryl (ICS). **c** | 10-thio-6-*N*-isocarbostyryl (SNICS). **d** | Benzotriazole (BTz).

As mentioned above, difficulties with extension might be the result of the insufficient interaction of new bases with the polymerase in the minor groove. Systematic investigation of this effect has resulted in the synthesis of benzotriazole (BTz; FIG. 2d). Although ATP is inserted opposite BTz more efficiently than BTz is inserted opposite itself, extension past the BTz•BTz pair by the Klenow fragment is only 200 times slower than for natural pairs in the same context[16].

The efficient use of a new base pair — when one is ultimately found that meets all criteria — will require not only efficient polymerase action but also enzymatic phosphorylation of the nucleoside analogues (the unit formed by the 'bases' linked to sugar molecules). Promising experiments have been carried out towards this goal using *Drosophila melanogaster* nucleoside kinase[17].

Finally, some unnatural bases are used only in transcription rather than in replication[18,19], which is an application that will be important for the incorporation of new amino acids. These cases are not discussed further here.

Theoretical arguments

The feasibility of alternative base pairs raises the question: why are there four bases in the natural genetic alphabet? As Orgel pointed out, there are two types of answer: either evolution has never experimented with alternative base pairs or four bases 'were enough'[20]. The first option might hold for the hydrophobic base pairs discussed above (an adequate early synthesis might be lacking), but it is unlikely to be true for all of the hydrogen-bonding bases in a prebiotic 'chemical mayhem'. At any rate, it does not explain why we do not have only two bases[21–24]. It therefore seems worthwhile to pursue the second option: why might four bases be enough?

If 'enough' is understood in terms of evolutionary stability, it means optimality within the frame of the structural constraints that are afforded by natural selection. Here, I describe attempts to show that four bases are optimal under STABILIZING SELECTION, especially when we consider MUTATION–SELECTION EQUILIBRIUM. I then discuss evidence for the optimal size of the genetic code obtained from *in silico* DIRECTIONAL SELECTION and finally analyse a more abstract contribution from so-called ERROR-CODING THEORY.

Stabilizing selection. All present models to explain the fact that we have four base types in our genetic alphabet hinge, in covert or overt form, on the assumption that the genetic alphabet evolved in an RNA world[25,26]. As with every enzyme, the enzymatic capacity of RNA rests on the three-dimensional positioning of functional groups. The primary sequence cannot be used to predict the three-dimensional shape of RNA, but sophisticated algorithms can turn a primary sequence into its two-dimensional RNA structure (such as the cloverleaf structure of transfer RNA). Gardner *et al.*[27] (on the basis of previous attempts by other researchers[28–30]) aimed to statistically characterize the effect of genetic-alphabet size on the features of sequence-to-shape mappings using many explicitly calculated secondary structures. They considered three measures: the fraction of paired bases required to obtain the optimally folded structure, whether there are many or a few nearly optimal structures for the same sequence and the difference between the optimally folded structure and a completely unfolded one.

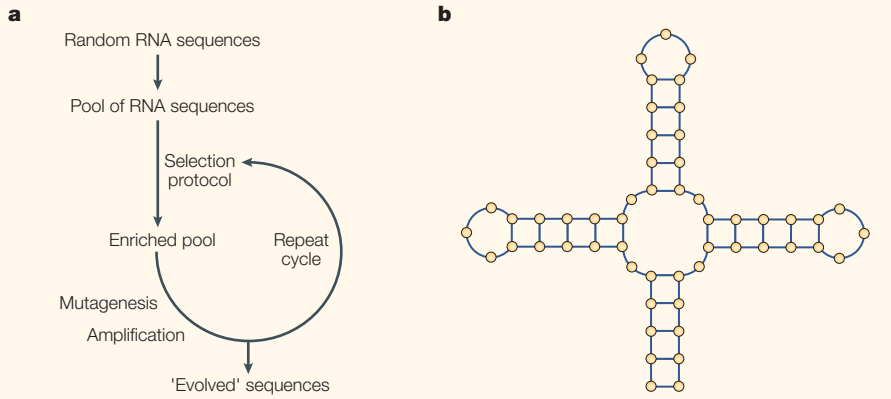


Figure 3 | ***In silico* evolution of RNA.** This *in silico* evolution experiment, developed by Gardner *et al.*[28], consisted of the following steps. An initial population of random RNA sequences evolved a population of molecules (**a**) towards a target cloverleaf structure (**b**) through mutation and selection. The fitness of each molecule was measured according to its distance in 'shape space' from the target: the farther away the shape of an individual was from the target, the lower its fitness. Runs were made with a population of 100 individuals and a fixed mutation rate of 0.009 per base per replication. The fitness of an alphabet is the average fitness of the population with that alphabet at generation 1,000. In this test, the fittest alphabet was that with four bases. Reproduced with permission from REF. 27 © (2003) The Royal Society.

All three of these measures decrease with increasing alphabet size[27], which is obvious with hindsight. Take the example of two bases, A and U. If a random sequence is assembled composed of equal amounts of A and U, then a randomly chosen base can potentially pair with 50% of all bases in the sequence. For the natural alphabet the same exercise gives 25%, and so on. For this reason, the optimal structures will become more and more clearly defined (and therefore fold more accurately) as the alphabet size increases. This model would therefore predict the optimum size of an alphabet to be large, possibly larger than the natural one. However, this expansion in size must be contained as it is probably offset by the increased risk of incorporating an incorrect base during replication. The above reasoning gives us an approximation, but the following model pins the optimal alphabet size down to $N = 4$.

A complementary approach, which aims to determine the optimum size of the genetic alphabet by assessing metabolic and replicative features in an RNA world, is now more than 10 years old[31,32]. The basic assumption is that the RNA world was metabolically complex[26,33].

Imagine a RIBO-ORGANISM with complex metabolic and replicative features. The fitness (W) of such an organism can be expressed[34] as the product of the fidelity of replication (Q) and the MALTHUSIAN GROWTH RATE (A); that is, $W(N) = A(N) \times Q(N)$. As the alphabet size (N) increases, the accuracy of insertion of cognate bases decreases (which is affected by the difference in base-pairing energy between correct and incorrect pairs); importantly, the decrease of Q with N must be faster than exponential. This by itself would select for a small alphabet size[31,32].

However, on average, an increase in N leads to a faster than linear, but slower than exponential, gain in overall catalytic efficiency, which is proportional to the second component of the equation: the growth rate A[31,32]. The reason for this type of increase is the fact that new introduced bases are, on average, decreasingly dissimilar to existing ones, and therefore are less likely to make a significant contribution to the modulation of the active sites of the ribozymes. The product of the two functions $Q(N)$ and $A(N)$ therefore results in a fitness curve with a hump: there must be an optimal N (N^*) that confers the highest fitness on ribo-organisms. Under a wide range of parameter values (such as temperature) $N^* = 4$, as in the natural alphabet.

Are there ways to test this hypothesis? Yes, but progress in experimental tractability is slow. The effect of N on the accuracy of base insertion cannot be assessed yet because of the choosiness of available replicases (see above), which unfortunately extends to alphabets that are composed of only a single base pair.

The situation is slightly better, although by no means conclusive, with testing the effect of N on enzymatic efficiency. Recently, several RNA ligases with the same function, composed of four, three[35] (A, U and C) and two[36] (AMINO-A and U) bases, have been produced by *in vitro* evolution. Although this example is not by itself sufficient for estimating the shape of $A(N)$, a robust conclusion is that catalytic efficiency increases with alphabet size.

Glossary

AMINO-A
An adenine molecule with a second amino (-NH_2) group attached to its carbon in position 2, which acts as an extra hydrogen-bond donor.

DEAMINATION
The reaction of a water molecule with the amino-group on position 4 of the pyrimidine ring of cytosine, which results in the conversion of cytosine to uracil.

DIRECTIONAL SELECTION
Natural selection that acts to promote the fixation (an increase in frequency in the population to 100%) of a particular allele.

EPIMERIZATION
The spontaneous change of configuration of chemical groups that are attached to a so-called asymmetric carbon atom. Such isomers are not mirror images of each other.

ERROR-CODING THEORY
A theory that was developed by Hamming to analyse the detection and correction of errors in messages consisting of 'zeros' and 'ones'.

KLENOW FRAGMENT
The *Escherichia coli* DNA polymerase, without the exonuclease subunit.

MALTHUSIAN GROWTH RATE
The *per capita* rate of growth of a population modelled in continuous time.

MUTATION–SELECTION EQUILIBRIUM
The equilibrium at which selection that decreases the frequency of an unfavourable allele exactly balances mutations that increase its frequency.

ORTHOGONALITY
Features of natural and/or artificial bases that in a given set (alphabet) decrease the degree of incorporating non-cognate base pairs.

PROCESSIVITY
The ability of polymerases to repeatedly add bases to the primer, extending even a new type of base.

RIBO-ORGANISM
A cell in the RNA world.

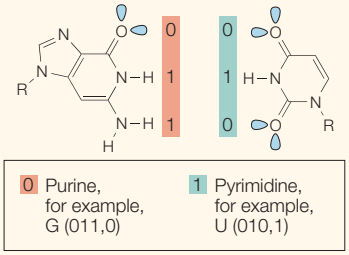
RNA WORLD
A hypothetical, but widely believed, era in early evolution when RNA-like molecules were not only genetic but also enzymatic material.

SIMULATED PROTOCELL MODEL
An *in silico* implementation of a ribo-organism.

STABILIZING SELECTION
Selection for the mean or intermediate phenotype; consequently, peripheral variants are eliminated, which maintains an existing state of adaptation in a stable environment.

Box 2 | **An error-coding approach to the genetic alphabet**

Mac Dónaill developed an alternative way of assessing the optimal size of the genetic alphabet[38]. His approach relies on abstractly defining alternative alphabets, which can readily be interpreted according to error-coding theory[46].



Representing the genetic alphabet

Mac Dónaill realized that nucleotides can be encoded by a binary vector, in which zeros and ones stand for hydrogen-bond acceptors and donors, respectively (see figure). A fourth digit indicates whether the nucleotide is a purine (0) or a pyrimidine (1). In information transmission, the parity of a codeword or number is defined as the number of 'ones' in the abstract 'codewords'. The number of bits in which two codewords differ is the so-called Hamming distance. For example, in the natural alphabet the parity of both G and U is even (it is 2; see figure), and the Hamming distance between, for example, C and U is 2. A pure-parity alphabet has only bases with similar parity. By contrast, mixed-parity alphabets consist of bases with even and odd parity. Bases that are separated by a Hamming distance of 1 are more readily mistaken for one another than are bases that are farther apart.

Assessing alternative alphabets

A pure-parity alphabet, such as the natural genetic alphabet, has the advantage of efficient error detection because the letters are less likely to be confused. In the natural code, a transition (pyrimidine→ pyrimidine or purine→purine) alters two hydrogen-bonding groups (distance 2) and a transversion (pyrimidine↔purine) changes the size of the base and at least one hydrogen-bonding group (minimal distance 2).

A good summary of alphabet categories is given in the table, which lists the information density and error vulnerability of various genetic-alphabet categories. It is notable that the natural alphabet belongs to category II, which is error resistant but has a relatively high information density (with even parity). The reason that the natural alphabet does not contain all of the bases with even parity in category II is that the pattern (0,0,0) is not feasible as it is vulnerable to hydrolysis. This potential even-parity alphabet would be reduced to amino-A, U, C and G, which is almost identical to the actual alphabet, with amino-A replaced by A in the latter.

So, general considerations of the abstract properties of various subsets of the full Benner alphabet are in agreement with the findings discussed previously: information density and error resistance are in conflict. When realistic chemical constraints are taken into account, once again an alphabet of size four seems optimal.

Category	Minimum*	Parity	Maximum‡	Vulnerabilty to mutations
I	1	Mixed	4	Transitions
II	2	Same	3	Resistant
III	3	Mixed	2	Transversions
IV	4	Same	1	Resistant

*Minimum Hamming distance between bases. ‡Maximum information bits per nucleotide.

In summary, two-dimensional RNA-like structures (and, presumably also the three-dimensional structures) become better defined as alphabet size increases, whereas the accuracy of replication decreases. A counteracting force is enzymatic efficiency: theoretical and experimental investigations indicate that this should increase with alphabet size. The above arguments indicate that the optimal number of bases is four (pending experimental verification), at least in an RNA world with bases that are chosen from the Benner alphabet. Qualitatively, a similar conclusion should hold for any alphabet, provided that genes also act as enzymes.

Directional selection and evolvability. As discussed in the previous section, the statistics of RNA secondary structures indicate that nucleic acids that are built using a larger alphabet are more stable and, therefore, are less likely to change during evolution. Lineages that more readily respond to directional selection can gain increased representation in the global population because their offspring are more likely to survive in the long run. Might this force favour an increased alphabet size? Probably not — as we shall see, stabilizing selection and selection for evolvability might well favour the same alphabet size of four.

This expectation is borne out by recent simulations[27]. The results of an *in silico* evolution study (FIG. 3) showed that populations with an alphabet size of four (two base pairs) were the most evolvable (that is, reached the highest fitness).

A possible limitation of the above study is that it assumes a fixed mutation rate, which, as discussed in the previous section, increases with alphabet size. The authors circumvented this problem by re-running simulations for a number of different mutation rates. $N = 2$ wins only at extremely low mutation rates, $N = 6$ is the winner between mutation rates of 0.001 and 0.003 per base per replication, and $N = 4$ is the winner above a mutation rate of 0.003. As primordial mutation rates could not have been low[34], the authors conclude that alphabets of four letters, including the natural one, have the highest evolvability.

A contribution from error-coding theory. Recently, Mac Dónaill[37,38] took a fresh look at the problem of the genetic alphabet[37,38] (BOX 2). He created an abstract way of categorizing a genetic alphabet by using a series of ones and zeros (as well as some other concepts from error-coding theory), and in doing so showed that the natural alphabet has relatively high information density coupled with good error resistance (resistance to mismatches).

The question might be justifiably raised as to whether such abstract considerations adequately reflect the nature of alternative alphabets that are built of concrete molecules. An early analysis that was based on available data on binding energies and extrapolations thereof, confirmed the idea of the natural alphabet being optimal in terms of fitness[31]. Recently, Mac Dónaill and Brocklebank carried out an analysis based on *ab initio* quantum-mechanical calculations[39] (in which the structure and properties of molecules are calculated directly from atomic theory). They found an excellent correlation between the number of mismatched hydrogen-bonding positions and the calculated association energies between cognate and non-cognate base pairs, despite some variation owing to the idiosyncratic contributions of the various atoms in different

positions. An important conclusion is that a single mismatch between hydrogen-bonding positions, whether repulsive (two hydrogen atoms collide) or non-binding (two hydrogen-acceptor groups face each other) is not sufficient to guarantee accurate discrimination against mismatches[31,39]: at least two mismatches in a non-cognate partner are required to maintain acceptable replication fidelity.

It might therefore be concluded that all theoretical investigations point in the same direction: a certain alphabet size (probably four) seems to be optimal as a compromise between stability and evolvability, between fidelity and catalytic efficiency, and between information density and error resistance.

Conclusions and prospects

Experiments show that the project of extending the genetic alphabet is feasible. The main obstacle to overcome is the choosiness of contemporary (protein) replicases. Considering the fact that this character state is presumably the result of alphabet–enzyme co-evolution, we might expect *in vitro* genetics to be successful in producing variant replicases and/or polymerases that can cope with an extended genetic alphabet (such as those synthesized by Benner's group; BOX 1).

The feasibility of artificially extended alphabets sheds light on the question of why the natural genetic alphabet looks as it does. Theoretical investigations based on structural, energetic and information-theoretic studies confirm the view that increased alphabet size decreases copying fidelity while increasing information density. This indicates that there must be an optimum alphabet size in terms of fitness, whether we assume that the genetic alphabet was fixed in an RNA world or not. If we do, then a systematic investigation of the effect of alphabet size on the fitness of RNA-based cells (so-called 'ribocytes') is mandatory. In this respect, several obstacles need to be overcome. First, as mentioned above, we need generalized replicases. Second, we have to consider replicase–alphabet co-evolution: ideally, the replicase in question must be composed of the same letters as all other genes in the unicellular ribo-organism (if we adopt the RNA-world hypothesis). Third, we have to test the effect on ribozyme activity of various alphabets for a range of reactions.

Although there is mounting evidence for the general enzymatic capabilities of RNA[26], we still do not have an efficient RNA polymerase ribozyme, let alone a replicase ribozyme[40]. An efficient replicase would have to show a generic mechanism of substrate recognition, highly accurate monomer insertion and processivity. Added to this, a replicase must have the ability to separate the template and the copy strands. These stringent requirements have not yet been fulfilled by any known ribozyme, but it is encouraging that protein polymerases catalyse primer extension only indirectly, through the accurate positioning of divalent metal ions[41].

According to the RNA-world-based view, the genetic alphabet became fixed more than 3 billion years ago[31], and the origin of the genetic code and translation happened subsequently[42]. This line of reasoning indicates that the informational/operational division of labour between nucleic acids and proteins[43] has uncoupled the genetic alphabet from enzymatic functionality constraints. As the genetic code evolved in the context of a certain genetic alphabet, any further change of the alphabet would have been unnecessary and/or extremely unlikely.

If, however, the genetic code originated by the simultaneous co-evolution of nucleic acids and proteins (a much more complicated model), then the fixation of the genetic alphabet must be considered in this complex context. Here, the general insight of Mac Dónaill[38] helps: the information density of the alphabet is a useful concept, whether the exercised function is ribozymic or a messenger function in protein synthesis. In this case, the problem of the size of the 'catalytic alphabet' (the number of encoded amino acids) readily arises: why do we have 20 rather than, for example, 16 or 25 different amino acids? It has been pointed out that some of the considerations discussed in this article (effects on catalytic efficiency and translation fidelity) apply to this related problem[32]. However, another crucial factor is likely to be involved: the metabolic cost of producing amino acids. An amino acid that belongs to the same biosynthetic family[43] is expected to increase catalytic efficiency only modestly and its metabolic cost is likely to be small. By contrast, an amino acid from a new biosynthetic family is likely to confer a high enzymatic advantage, but is expected to incur high metabolic costs (for instance, many new ATP-requiring steps).

Considering the theoretical studies on evolvability, an analysis should be carried out with many shapes and alphabets, presumably in the context of some SIMULATED PROTOCELL MODEL[44]. Synthetic biology[45] might therefore offer insights into natural biology as well.

Eörs Szathmáry is at the Institute for Advanced Study, Berlin (Wissenschaftskolleg zu Berlin), on leave from the Institute for Advanced Study, Budapest (Collegium Budapest), 2 Szentháromság, H-1014 Budapest, Hungary.
e-mail: szathmary@colbud.hu

doi:10.1038/nrg1231

1. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, London, 1930).
2. Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737 (1953).
3. Piccirilli, J. A., Krauch T., Moroney, S. E. & Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33–37 (1990).
4. Kool, E. T. Hydrogen bonding, base stacking, and steric effects in DNA replication. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 1–22 (2001).
5. Benner, S. A. *et al.* Redesigning nucleic acids. *Pure Appl. Chem.* **70**, 263–266 (1998).
6. Mathis, G. & Hunziker, J. Towards a DNA-like duplex without hydrogen-bonded base pairs. *Angew. Chem. Int. Ed.* **41**, 3203–3205 (2002).
7. Ogawa, A. K., Wu, Y., Berger, M., Schultz, P. G. & Romesberg, F. E. Rational design of an unnatural base pair with increased kinetic selectivity. *J. Am. Chem. Soc.* **122**, 8803–8804 (2000).
8. Kool, E. T. Synthetically modified DNAs as substrates for polymerases. *Curr. Opin. Chem. Biol.* **4**, 602–608 (2000).
9. Switzer, C. Y., Moroney, S. E. & Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.* **111**, 8322–8323 (1989).
10. Roberts, C., Bandaru, R. & Switzer, C. Theoretical and experimental study of isoguanine and isocytosine: base pairing in an expanded genetic system. *J. Am. Chem. Soc.* **119**, 4640–4649 (1997).
11. Switzer, C. Y., Moroney, S. E. & Benner, S. A. Enzymatic recognition of the base pair between isocytidine and isoguanine. *Biochemistry* **32**, 10489–10496 (1993).
12. Chu, C. K., Reichmann, U., Watanabe K. A. & Fox, J. J. Nucleosides 104. Synthesis of 4-amino-5-(D-ribofuranosyl) pyrimidine *C*-nucleosides from 2-(2,3-*O*-isopropylidene-5-*O*-trityl-D-ribofuranosyl)acetonitrile. *J. Org. Chem.* **42**, 711–714 (1977).
13. Voegel, J. J. & Benner, S. A. Nonstandard hydrogen bonding in duplex oligonucleotides. The base pair between an acceptor–donor–donor pyrimidine analog and a donor–acceptor–acceptor purine analog. *J. Am. Chem. Soc.* **116**, 6929–6930 (1994).
14. Tae, E. L., Wu, Y., Xia, G., Schultz, P. G. & Romesberg, F. E. Efforts toward expansion of the genetic alphabet: replication of DNA with three base pairs. *J. Am. Chem. Soc.* **123**, 7439–7440 (2001).
15. Yu, C., Henry, A. A., Romesberg, F. E. & Schultz, P. G. Polymerase recognition of unnatural base pairs. *Angew. Chem. Int. Ed.* **41**, 3841–3844 (2002).
16. Matsuda, S. *et al.* The effect of minor-groove hydrogen-bond acceptors and donors on the stability and replication of four unnatural base pairs. *J. Am. Chem. Soc.* **125**, 6134–6139 (2003).
17. Wu, Y. *et al.* Enzymatic phosphorylation of unnatural nucleosides. *J. Am. Chem. Soc.* **124**, 14626–14630 (2002).
18. Ohtsuki, T. *et al.* Unnatural base pairs for specific transcription. *Proc. Natl Acad. Sci. USA* **98**, 4922–4925 (2001).
19. Hirao, I. *et al.* An unnatural base pair for incorporating amino acid analogs into proteins. *Nature Biotech.* **20**, 177–182 (2002).
20. Orgel, L. E. Nucleic acids — adding to the genetic alphabet. *Nature* **343**, 18–20 (1990).
21. Orgel, L. E. Evolution of the genetic apparatus. *J. Mol. Biol.* **38**, 381–393 (1968).
22. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
23. Wächtershäuser, G. An all-purine precursor of nucleic acids. *Proc. Natl Acad. Sci. USA* **85**, 1134–1135 (1988).
24. Zubay, G. An all-purine precursor of nucleic acids. *Chemtracts* **2**, 439–442 (1991).
25. Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
26. Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418**, 214–221 (2002).
27. Gardner, P. P., Holland, B. R., Moulton, V., Hendy, M. & Penny, D. Optimal alphabets for an RNA world. *Proc. R. Soc. Lond. B* **270**, 1177–1182 (2003).
28. Fontana, W., Konings, D., Stadler, P. & Schuster, P. Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404 (1993).
29. Schuster, P. RNA-based evolutionary optimization. *Orig. Life Evol. Biosphere* **23**, 373–391 (1993).
30. Grüner, W. *et al.* Analysis of RNA sequence and structure maps by exhaustive enumeration. *Monatshefte Chem.* **127**, 355–374 (1996).
31. Szathmáry, E. Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proc. R. Soc. Lond. B* **245**, 91–99 (1991).
32. Szathmáry, E. What is the optimum size for the genetic alphabet? *Proc. Natl Acad. Sci. USA* **89**, 2614–2618 (1992).

33. Benner, S. A., Ellington, A. D. & Tauer, S. A. Modern metabolism as a palimpsest of an RNA world. *Proc. Natl Acid. Sci. USA* **86**, 7054–7058 (1989).
34. Eigen, M. Self-organization of matter and the evolution of biological macromolecules. *Naturwiissenschaften* **58**, 465–523 (1971).
35. Rogers, J. & Joyce, G. F. The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA* **7**, 395–404 (2001).
36. Reader, J. S. & Joyce, G. F. A ribozyme composed of only two different nucleotides. *Nature* **420**, 841–844 (2002).
37. Mac Dónaill, D. A. A parity code interpretation of nucleotide alphabet composition. *Chem. Commun.* **18**, 2062–2063 (2002).
38. Mac Dónaill, D. A. Why nature chose A, C, G and U/T: an error-coding perspective of nucleotide alphabet composition. *Orig. Life Evol. Biosphere* **33**, 433–455 (2003).
39. Mac Dónaill, D. A. & Brocklebank, D. An *ab initio* quantum chemical investigation of the error-coding model of nucleotide alphabet composition. *Mol. Phys.* **101**, 2755–2763 (2003).
40. McGinness, K. E. & Joyce, G. F. In search of an RNA replicase ribozyme. *Chem. Biol.* **10**, 5–14 (2003).
41. Brautigam, C. A. & Steitz, T. A. Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.* **8**, 54–63 (1998).
42. Szathmáry, E. The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet.* **15**, 223–229 (1999).
43. Wong, J. T. A coevolution theory of the genetic code. *Proc. Natl Acad. Sci. USA* **72**, 1909–1912 (1975).
44. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
45. Benner, S. A. Synthetic biology: act natural. *Nature* **421**, 118 (2003).
46. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Techn. J.* **29**, 147–160 (1950).

Acknowledgements

I thank the biologists at the Wissenschaftskolleg zu Berlin for vivid discussions. Also, B. Papp and V. Müller who kindly read the manuscript before submission.

Competing interests statement

The author declares that he has no competing financial interests.

Online links

FURTHER INFORMATION

Eörs Szathmáry's web page: http://www.colbud.hu/fellows/szathmary.shtml
Scripps Research Institute: http://www.scripps.edu
Steven Benner's web page: http://www.chem.ufl.edu/benner.html
Access to this interactive links box is free online.

INNOVATION

Digital genotyping using molecular affinity and mass spectrometry

Sobin Kim, Hameer D. Ruparel, T. Conrad Gilliam and Jingyue Ju

The goal of DNA sequencing and genotyping is to efficiently generate accurate high-throughput digital genetic information that unambiguously identifies sources of genetic variation and clearly distinguishes heterozygous from homozygous variants. Recent advances in mass-spectrometry-based DNA sequencing and genotyping bode well for meeting these criteria. Pilot studies show that these recently developed approaches allow unambiguous multiplex detection of heterozygous variants and the identification of deletion and insertion variants.

The completion of the Human Genome Project has set the stage for screening genetic mutations to identify disease genes on a genome-wide scale[1]. Accurate high-throughput methods for resequencing the intron/exon regions of candidate genes are needed to explore the complete human genome sequence for disease-gene discovery. State-of-the-art technology for high-throughput DNA sequencing, such as that used in the Human Genome Project, uses capillary-array DNA sequencing with LASER-INDUCED FLUORESCENCE DETECTION[2–5]. Although this technology meets the throughput and read-length requirements of large-scale DNA sequencing projects, the accuracy that is required for mutation detection needs to be improved for a wide range of applications, ranging from disease-gene discovery to personalized medicine. For example, the unambiguous detection of heterozygotes is difficult with electrophoresis-based DNA sequencing methods. Problems also arise in GC rich regions owing to COMPRESSION[6,7], which leads to poor resolution in sequencing DNA fragments. Also, the first few bases downstream of the priming site are often masked by high levels of fluorescence from the excess dye-labelled primers or dye-labelled terminators, and are therefore difficult to identify.

Many recent advances in DNA sequencing technology address these limitations. In particular, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has emerged as a rapid and efficient analytical tool in DNA sequencing and genotyping. Here, we review these advances and discuss new approaches that use molecular affinity for the accurate and simultaneous detection of genetic variations, which have wide applications in PHARMACOGENOMICS and clinical diagnostics.

DNA sequencing by MALDI-TOF MS

MALDI-TOF MS has been widely used in DNA sequencing[8–12] (FIG. 1). SANGER DNA SEQUENCING is generally used to produce DNA sequencing fragments in MALDI-TOF MS[13]. Compared with gel electrophoresis-based sequencing systems, MALDI-TOF MS produces high-resolution short DNA sequencing fragments of <100 base pairs (bp), rapid fragment separation on microsecond timescales and the complete elimination of the compressions that are associated with gel electrophoresis.

An important challenge for DNA sequencing using mass spectrometry is the stringent purity requirement for the sequencing fragments that are introduced into the mass detector. Because DNA sequences are determined by accurately measuring the mass of the DNA fragments, DNA must be free from alkaline earth salts and other contaminants.

Approaches for purifying DNA samples that rely on the strong interaction of a small molecule (biotin) and a protein (streptavidin[14]) on solid surfaces (such as magnetic beads) are widely used[15,16]. In DNA sequencing using MALDI-TOF MS, Monforte and Becker obtained read lengths of 100 bp by purifying DNA sequencing samples using a cleavable biotinylated primer[11]. In this method, the primer-extension fragments are captured at their 5′ end on streptavidin-coated magnetic beads, whereas the other components in the sequencing reaction are washed away. Fu *et al.* reported the sequencing of exons 5–8 of the human tumour suppressor *p53* gene (also known as *TP53*) by MALDI-TOF MS using a DNA template that was immobilized on a solid phase for one cycle of extension. In this study, extended DNA fragments were hybridized on the immobilized templates, whereas the other components in the sequencing reaction were removed. Neither method eliminates falsely terminated DNA sequencing fragments. Falsely terminated DNA fragments (false stops) are generated in Sanger DNA sequencing reactions when a DNA fragment is terminated by the incorporation of a deoxynucleotide (dNTP) rather than a dideoxynucleotide (ddNTP). False stops and dimerized primers can produce extra peaks in the mass spectra that prevent accurate base identification[9]. Also, four separate reactions are carried out in both methods, one for each dideoxynucleotide terminator, which is analogous to dye-labelled primer sequencing.