Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 2614–2618, April 1992
Evolution

What is the optimum size for the genetic alphabet?

(ribozymes/genetic systems/origin of life/RNA world)

EÖRS SZATHMÁRY

Laboratory of Mathematical Biology, Medical Research Council, National Institute for Medical Research, The Ridgeway, Mill Hill, NW7 1AA, London, United Kingdom

Communicated by John Maynard Smith, December 11, 1991

ABSTRACT **An important question in biology is why the genetic alphabet is made of just two base pairs (G·C and A·T). This is particularly interesting because of the recent demonstration [Piccirilli, J. A., Krauch, T., Moroney, S. E. & Benner, S. A. (1990) *Nature (London)* 343, 33–37] that the alphabet can in principle be larger. It is possible to explain the size of the present genetic alphabet as a frozen character state that was an evolutionary optimum in an RNA world when nucleic acids functioned both for storing genetic information and for expressing information as enzymatically active RNA molecules—i.e., ribozymes. A previous model [Szathmáry, E. (1991) *Proc. R. Soc. London Ser. B* 245, 91–99] has described the principle of this approach. The present paper confirms and extends these results by showing explicitly the ways in which copying fidelity and metabolic efficiency change with the size of the genetic alphabet.**

The question why the genetic alphabet consists of exactly two base pairs instead of, say, one or four has become highly relevant through the design and partial realization of novel, replicatable base pairs (1). I have attempted to construct a model to account for the actual state of two base pairs in terms of evolutionary optimality (2). The model assumed the existence of an RNA world (3) with ribozymes (RNA enzymes) catalyzing all sorts of reactions in intermediate metabolism (4). In such a world nucleic acids have dual functionality: as replicatable information carriers (templates) and as enzymes. The fitness of such a ribo-organism was broken down into two components: overall copying fidelity (Q) and overall reproduction rate (A), following Eigen's (5) original formulation for the analogous case of replicating nucleic acids. The fitness W is simply defined as $W = AQ$, assuming an aspecific death rate. Q can be calculated from the per-base copying fidelity q, which decreases roughly exponentially with increasing size (N pairs) of the alphabet (keeping the length of the genome fixed). The reason for this is that as one adds more letters to the alphabet, they will resemble each other more and more, and hence the chance of mispairing and mutagenesis increases. A increases with N because A increases with metabolic efficiency, which increases with the number of monomer types used in building the enzymes. A increases slower than exponentially with N. It was found that W has a maximum that lies at $N = 2$ for most cases (2).

The study summarized above had two methodical elements that will be replaced here: (*i*) actual or estimated data for all base pairs were entered into numerical calculations, which precluded being more analytic in presenting the A and Q values, and (*ii*) the substrate set was replaced with the totality of letters in the maximal genetic alphabet (2). In this paper I re-develop my model in order to arrive at a more analytic description.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

COPYING FIDELITY

It is assumed, as previously (2), that the equation $Q = q^{\nu}$ is valid, where ν is the number of nucleotides in the genome (5). The task is to find out how q depends on N. To determine this relationship, assumptions must be made about the structure of and binding strength between the letters. It is assumed that there are $\bar{n}$ hydrogen bonding groups on each letter (for the alphabet designed in Benner's laboratory $\bar{n} = 3$), and each group can be either a donor or an acceptor. As a simplification it is assumed that the binding Gibbs free energy increases linearly with the number of complementary (donor–acceptor) groups between two letters. Thus a configuration such as

```
A—D—A
|   |   |
D—A—D
```

would have the highest negative energy, whereas pairs such as

```
D—D—D    A—A—A

D—D—D    A—A—A
```

would have zero Gibbs free energy in association [it is understood that large groups (= "purines") and small groups (= "pyrimidines") occupy the upper and lower positions, respectively]. The energetic effect of the steric clash between two donor groups is not considered here; such effects were amply treated in the earlier model (2). The closer are the donor–acceptor configurations of two base pairs in the same alphabet, the lower is their copying fidelity because of the increased mutation rate. For this reason, given fixed $\bar{n}$ and N, one should choose those pairs that would maximize the distance, calculated as the number of differing hydrogen-bonding groups, among the letters. Thus, when $\bar{n} = 2$ and $N = 2$, one should choose pairs such as

```
A—A    D—D
|  |   |  |
D—D    A—A
```

and not, say

```
A—A    A—D
|  |   |  |
D—D    D—A
```

since the distance between the two purines is 2 in the first and is only 1 in the second case (within-alphabet distances are the same for the complementary pyrimidines). It is an interesting question how one should position base pairs for arbitrary combinations of $\bar{n}$ and N. Note, however, that $N = <2^{\bar{n}}$. In the following I shall not deal with wobble pairing (see ref. 2 for its incorporation).

First we consider the number of possible pairs d distance from a given pair. It is easy to check that the answer is

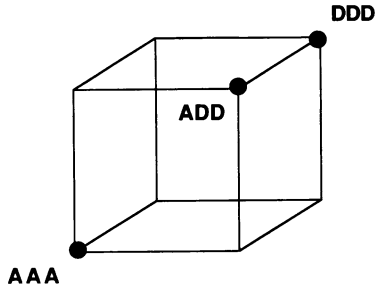

FIG. 1. Configuration of the best possible genetic alphabet with $\tilde{n} = 3$ and $N = 3$. Each corner of the cube corresponds to a unique hydrogen-bonding pattern of "purines". The structure of "pyrimidines" is complementary.

$$m_d = \binom{\tilde{n}}{d}. \quad [1]$$

Let $\tilde{n} = 3$ for the time being. For $N = 1$ and 2 the choice of the base pairs is trivial. This is not so for $N = 3$. To obtain the most favorable combination we must consider the functions that determine q. For a given base the so-called insertion fidelity q'' (cf. ref. 6) can be calculated as

$$q'' = e^{G/(RT)} \Big/ \left\{ \sum_{d=0}^{d_{max}} n_d \exp[G(d_{max} - d)/(RTd_{max})] \right\}, \quad [2]$$

where n_d is the number of the base pairs in the alphabet d distance from the given base, G is the Gibbs free energy of base pairing (shown as positive since a negative value must be multiplied by -1), R is the gas constant, and T is the absolute temperature. In the present example we know that $n_0 = 1$ and $n_3 = 1$. q'' is maximized if for the third base pair $d = d_{max}/2$; i.e., it should be as far as possible from the previous two. Clearly, 3/2 cannot be realized in the present situation, hence the third pair must be at a distance of 1 and 2 units, respectively, from one and the other pair (Fig. 1). Table 1 shows how the alphabets with maximum q'' occupy the different positions with increasing N. Asymmetries have artificially been corrected for by taking $n_d = 1/2$ values when necessary. One can check that the distributions shown are isotropic in the sense that they are the same viewed from any occupied point on the respective (hyper) cube.

Having thus determined the fidelity-maximizing alphabets, we may list the formulae necessary for the calculation of Q.

Table 1. Genetic alphabets maximizing replication

	d					d				
N	0	1	2	3	N	0	1	2	3	4
$\tilde{n} = 2$					$\tilde{n} = 4$					
1	1	0	0		1	1	0	0	0	0
2	1	0	1		2	1	0	0	0	1
3	1	1	1		3	1	0	1	0	1
4	1	2	1		4	1	0	2	0	1
$\tilde{n} = 3$					5	1	0	3	0	1
1	1	0	0	0	6	1	0	4	0	1
2	1	0	0	1	7	1	0	5	0	1
3	1	0.5	0.5	1	8	1	0	6	0	1
4	1	1	1	1	9	1	0	6	1	1
5	1	1.5	1.5	1	10	1	1	6	1	1
6	1	2	2	1	11	1	1.5	6	1.5	1
7	1	2.5	2.5	1	12	1	2	6	2	1
8	1	3	3	1	13	1	2.5	6	2.5	1
					14	1	3	6	3	1
					15	1	3.5	6	3.5	1
					16	1	4	6	4	1

Borrowing the formulae from ref. 2, the copying fidelity with proofreading is

$$q' = 2q'' - q''^2, \quad [3]$$

and with mismatch repair it becomes

$$q = 2q' - q'^2. \quad [4]$$

The relevance of these component processes will be discussed below. Example curves for Q are shown in Fig. 2. It is apparent from the logarithmic plots that Q decreases more than exponentially with N. The lack of mismatch repair considerably decreases Q (case C versus case A). For the same N (e.g., 3), increase in $\tilde{n}$ increases Q, since more hydrogen bonds result in stronger binding and increased fidelity.

OVERALL METABOLIC EFFICIENCY

The task is to determine the efficiency of enzymes made of $2N$ monomers in general. We lack direct experimental evidence to address this problem. Rather, the following calibration procedure is chosen. We know that at least some protein

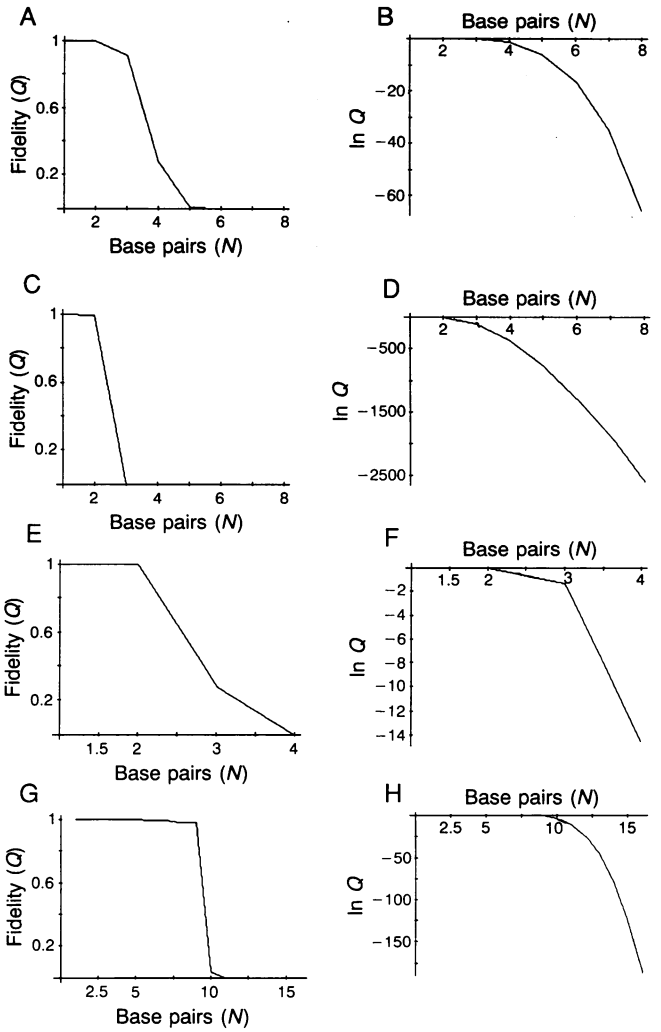

FIG. 2. Copying fidelity of genetic alphabets. (A) $\tilde{n} = 3$, with mismatch repair, $G = 5$ kcal/mol (1 cal = 4.18 J). (C) $\tilde{n} = 3$, without mismatch repair, $G = 5$ kcal/mol. (E) $\tilde{n} = 2$, with mismatch repair, $G = 3.33$ kcal/mol. (G) $\tilde{n} = 4$, with mismatch repair, $G = 6.66$ kcal/mol. (B, D, F, and H) logarithmic plots corresponding to A, C, E, and G, respectively. $T = 298$ K and $\nu = 10^5$ in all cases. G values are based on ref. 2.

enzymes are perfect, in the sense that the rate of the catalyzed step equals that of diffusion of the substrate to, and products from, the enzyme (see ref. 7 for a discussion of this point). However, in certain other cases, post-translationally modified amino acids seem to play an important role in the increase of catalytic efficiency (8). We make the assumption that catalysts made of 32 monomer types are practically perfect for any intermediate metabolism.

For the sake of simplicity, and following some previous work on the evolution of enzymatic function (9, 10), we may consider a metabolic pathway of monomolecular reactions. First let us calculate the catalytic efficiency of an average enzyme. To this end we must calculate the probability that a randomly chosen substrate will be converted by an appropriate enzyme. Here the relationship between substrate and enzyme space becomes relevant.

Enzyme space can be quantitatively characterized as follows. Imagine that, somewhat similar to the assumption in ref. 10, the substrates are represented by boxes. I assume that exactly four faces of this box are to be recognized by an active site. In accordance with the foregoing, this means that we find a perfect active site for every substrate among the $32^4 = 10^6$ possible active sites. This also means that we may assume that the substrates are embedded in a 20-dimensional binary chemical space (cf. ref. 11). Although not strictly true, it is assumed that substrates are distributed in this space randomly. In fact, of course, metabolites have to be close to their transformed products in chemical space.

For calculating the catalytic activity, we determine the probability that a randomly chosen substrate will have a corresponding catalyst within distance δ in chemical space. I borrow the method of calculation from the work of Perelson and Oster (11) on the clonal selection theory of the immune system. Let the current alphabet size be $2N$, because letters come in pairs. The density of active sites/enzymes in chemical space is thus $\rho = (2N)^4/32^4$. It is assumed that the number of enzymes within distance from the substrate follows a Poisson distribution (11). Under the given assumptions chemical space has 2^{20} discrete points, so the maximum distance on a representing hypercube between two points is 20. There are

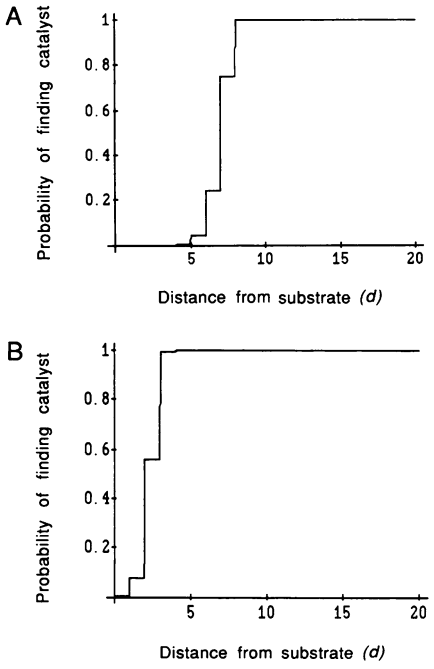

FIG. 3. Plot of Eq. 6: probability of finding at least one enzyme in chemical space distance from the substrate. (A) $N = 2$. (B) $N = 4$.

$$\binom{20}{r}$$

points that are r distance units from chosen point. Within a volume having radius δ, the number of points is thus:

$$s(\delta) = \sum_{r=0}^{\delta} \binom{20}{r}. \quad [5]$$

The probability of finding no enzyme within this volume is (cf. ref. 11)

$$\exp[-(2N)^4 s(\delta)/32^4]$$

and the probability that there is at least one such enzyme is

$$u(\delta) = 1 - \exp[-(2N)^4 s(\delta)/32^4]. \quad [6]$$

The way in which this probability increases with N is shown in Fig. 3. I shall use this expression in the further calculations, being aware that it is only an approximation, which is rather crude at low and high δ but which nevertheless reflects the correct tendency: the probability of finding a catalyst should increase with the radius of the ball drawn around the substrate.

Now let us calculate the effective energy of binding between the enzyme and the nonreacting parts of the substrate (cf. ref. 2), defined as follows:

$$G_{ES} = \sum_{\delta=0}^{\delta_{max}} \{G_c(\delta_{max} - \delta)\,[u(\delta) - \operatorname{sign}(\delta)u(\delta - 1)]\}/\delta_{max}. \quad [7]$$

This rather unconventional formula for a binding free energy is meant to express the following. As we move away from the substrate in chemical space, the binding energy decreases from G_c (perfect binding), but the chance that we find at least one catalyst increases. Therefore, energy is weighted by this incremental probability as we move toward the boundary of the ball centered around the substrate. These weighted energies are finally summed up to yield the effective binding energy. An indication of how G_{ES} changes with N—it increases with diminishing returns—is shown in Fig. 4.

Finally, we calculate the catalytic efficiency. First note that for an unsaturated linear enzymatic pathway the flux

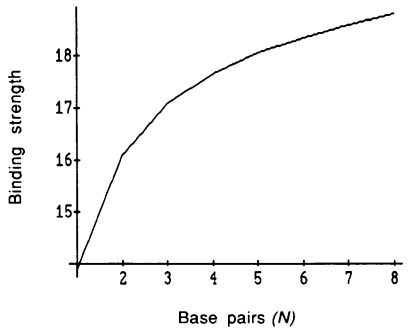

FIG. 4. Plot of Eq. 7: effective binding strength between a randomly chosen substrate and its candidate enzymes. $T = 298$ K, $\tilde{n} = 3$.

depends on the catalytic efficiencies of the enzymes as follows (12):

$$F = C/\sum_i 1/e_i, \quad [8]$$

where C is an environmental constant and

$$e_i = c_i E_i / K_i \quad [9]$$

where c_i and E_i are the catalytic efficiency and concentration of enzyme i in the chain, respectively, and K_i is an equilibrium constant. For calculational simplicity, it is assumed that $c_i = c$, $E_i = E$, and $K_i = K$, and hence the flux becomes

$$F = CcE/(n_E K), \quad [10]$$

where n_E is the number of enzyme species in the pathway. If we assume that the exponential growth rate constant of the cell is determined by this flux, we can then apply the formula in (cf. refs. 2 and 9)

$$A = F/(n_E E) \quad [11]$$

and after rearrangement we obtain

$$A = Cc/(n^2 K). \quad [12]$$

c depends on G_{ES} exponentially (9). Substituting this from Eq. 7 and absorbing all parameters treated as constants we obtain

$$A = me^{G_{ES}/(RT)}. \quad [13]$$

Metabolic efficiency increases less than exponentially with N (Fig. 5).

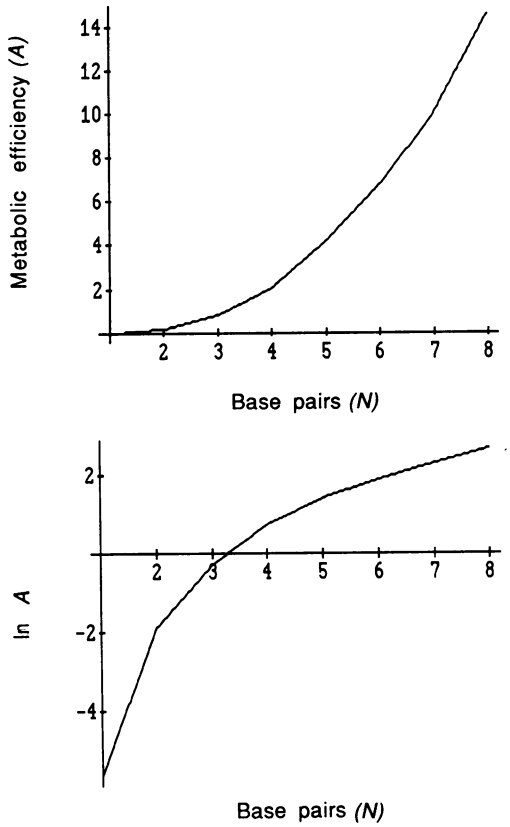

FIG. 5. Plot of Eq. 13: overall metabolic efficiency. $\tilde{n} = 3$, $T = 298$ K, $m = 2.4 \times 10^{-13}$ (cf. ref. 2), $G_c = 20$ kcal/mol.

FITNESS

As previously defined, the formula for fitness is $W = AQ$, and various combinations are shown in Fig. 6. Thus it seems that with the present efficiency of replication (with proofreading and mismatch repair), $N = 3$ would be optimal for a genome made of 10^5 nucleotides (case A). Excluding mismatch repair, however, gives $N = 2$—i.e., two base pairs, as in our current genetic alphabet. Decreasing ν to 10^4 makes the optimal value of N shift to 5 (case C), but without mismatch repair, $N = 2$ stays optimal. An increase in temperature to $T = 348$ K (75°C) shifts the value in case C back to $N = 4$ (case E). With $T = 348$ K and $\nu = 10^5$, $N = 3$ is still optimal with mismatch repair (case G). Without it, $N = 2$ stays optimal.

The effect of temperature is understandable. In formulae 2 and 13 T decreases the efficiency of binding exponentially.

I also show fitness values when the number of hydrogen bonds per base pair ($\tilde{n}$) is altered (Fig. 7). With $\tilde{n} = 2$, $N = 2$ is still optimal, whereas $\tilde{n} = 4$ shifts N to 9 base pairs.

DISCUSSION

There are two crucial aspects of my previous study (2) that have been confirmed here: (*i*) Q decreases and A increases with N, and there is an optimum of W and (*ii*) without mismatch repair, $N = 2$ is optimal. The important differences are as follows: (*i*) Q is clearly shown to decrease with N faster than exponentially; (*ii*) catalytic efficiency increases not with diminishing returns, as claimed in ref. 2, but it does increase

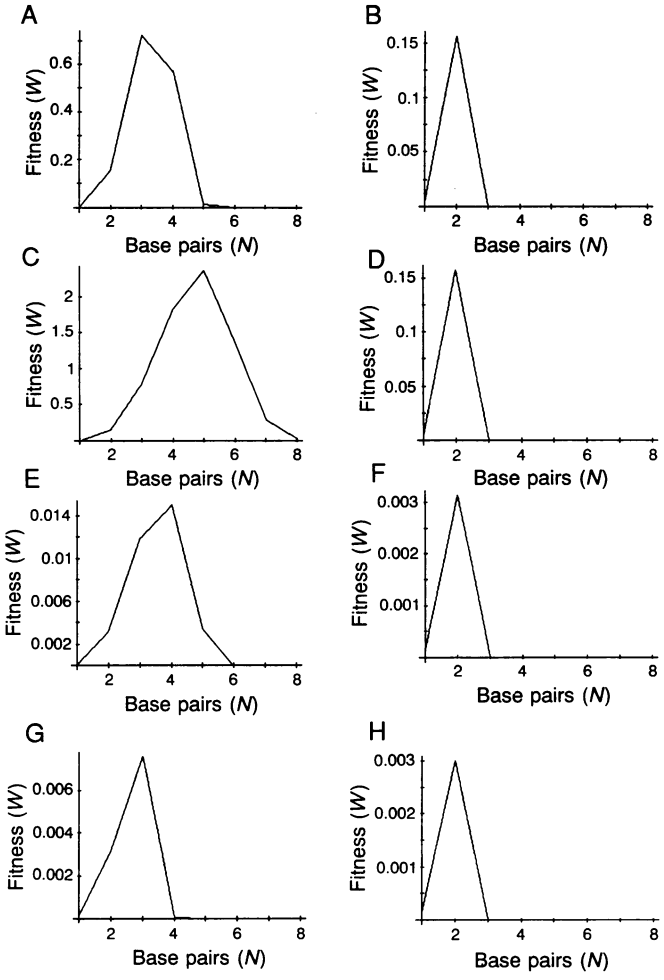

FIG. 6. Fitness values for various systems with $\tilde{n} = 3$. (A, C, E, and G) With mismatch repair. (B, D, F, and H) Without mismatch repair. (A and B) $T = 298$ K, $\nu = 10^5$. (C and D) $T = 298$ K, $\nu = 10^4$. (E and F) $T = 348$ K, $\nu = 10^4$. (G and H) $T = 348$ K, $\nu = 10^5$.

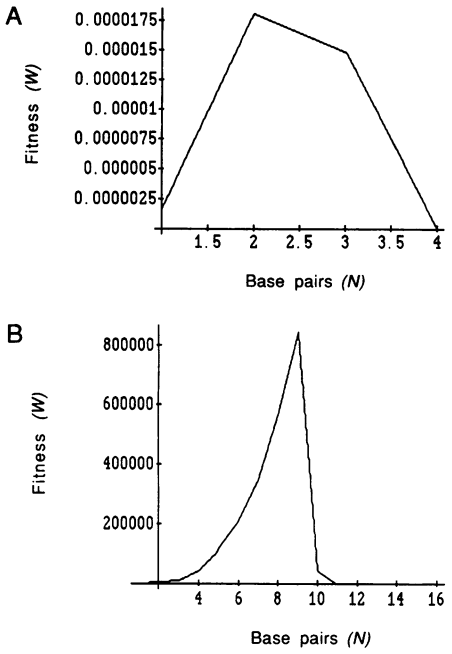

FIG. 7. Fitness values for systems with $\bar{n} = 2$, $G_c = 13.33$ kcal/mol (A) and with $\bar{n} = 4$, $G_c = 26.66$ kcal/mol (B).

slower than exponentially with N; and (*iii*) with mismatch repair, $N = 3$ seems to be optimal for room and high temperature as well.

Before evaluating these differences, I must call attention to some important limitations of the present model. (*a*) It does not calculate the efficiency of the copying machinery from the intrinsic enzymatic efficiencies. (*b*) In contrast to the earlier model (2), wobble pairing is not included in the calculations, which erroneously increases Q. (*c*) Similar to the earlier model, transversions are excluded from mutations. This again increases Q. (*d*) Unlike the earlier model, the present work does not incorporate competitive inhibition among the different reactions by related substrates. This overestimates catalytic power. (*e*) Similar to the earlier model, there is no energetic or time cost assigned to the synthesis of more letters.

From the above-listed limitations it must be clear that the optimal value $N = 3$ (Fig. 6, case A) for the genetic alphabet designed by Piccirilli *et al.* (1) is likely to be artefactual. Since both A and Q are overestimated, W is overestimated as well. The assumption that is probably the least realistic is to suppose that ribozymes can copy nucleic acids as efficiently as present-day protein enzymes, which are capable of proofreading as well as mismatch repair. It makes more sense to suppose that the best ribozymes performed only at about the efficiency of protein-catalyzed replication without mismatch repair. A future goal is to modify this model so that it can account for the dependence of replicase activity on N.

As to the performance of the alphabets with $\bar{n} = 2$ and 4, a few observations can be made. The former are less fit and the latter, more fit than alphabets with $\bar{n} = 3$. The $\bar{n} = 4$ alphabet may appear unrealistic from a chemical point of view, but this is not so. A hypothetical base pair of this type is presented in Fig. 8. It seems that others could be designed. Why such an alphabet is unlikely to be fitter than one with $\bar{n} = 3$ can be explained by two considerations: (*i*) the increased metabolic load associated with synthesizing such large letters and (*ii*) the strong overestimation of the catalytic power of enzymes made of these letters. By the latter I mean that monomers in enzymes cannot be very useful if they become too big; if one has a small bathroom, then it is unwise to buy too-large tiles. This is a reason why proteins are more efficient enzymes than ribozymes: the constraint of pairing during replication does not apply to amino acids, hence they can be rather different from each other and reasonably small at the same time.

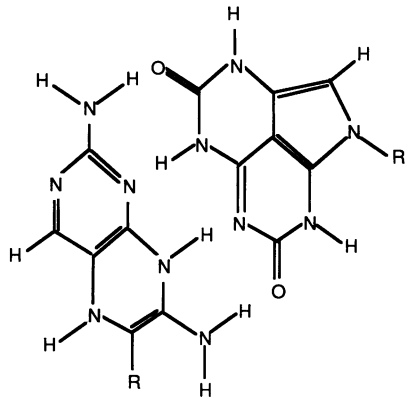

FIG. 8. Hypothetical "base pair" from an alphabet with $\bar{n} = 4$.

I concluded from my previous analysis (2) that the current genetic alphabet size $N = 2$ is likely to be a frozen character state, since protein enzymes would enable cells to maintain a genome with more base pairs, but this trait must have become fixed in the RNA world where it was an evolutionary optimum. Although it may be possible to imagine takeover scenarios for a transition to a higher N after the origin of translation, such a transition would not have been selectively advantageous since the main catalytic power shifted to proteins.

This theory of optimal genetic alphabets is testable (2): Piccirilli *et al.* (1) have shown that one of the novel base pairs is replicated in RNA as well as DNA, and elsewhere (13, 14) I have outlined an experimental system to make novel ribozymes at will through artificial selection. Ellington and Szostak (15) have recently reported some interesting progress in this direction. Copying fidelities and catalytic efficiencies could thus be both measured.

Finally, I am aware of the fact that in our current genetic alphabet $\bar{n} = 2$ for A. As discussed in ref. 2, this may have a fitness-increasing effect.

I thank Pál Juhász-Nagy and Elemér Lábas for stimulating discussion and Tom Kirkwood and two anonymous referees for helpful comments on the manuscript.

1. Piccirilli, J. A., Krauch, T., Moroney, S. E. & Benner, S. A. (1990) *Nature (London)* **343**, 33–37.
2. Szathmáry, E. (1991) *Proc. R. Soc. London Ser. B* **245**, 91–99.
3. Gilbert, W. (1986) *Nature (London)* **319**, 618.
4. Benner, S. A., Ellington, A. D. & Tauer, A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7054–7058.
5. Eigen, M. (1971) *Naturwissenschaften* **58**, 465–523.
6. Goodman, M. F. & Branscomb, E. W. (1986) in *Accuracy in Molecular Processes*, eds. Kirkwood, T. B. L., Rosenberger, R. F. & Galas, D. J. (Chapman & Hall, London), pp. 191–232.
7. Koshland, D. E., Jr. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 1–7.
8. Wold, F. (1981) *Annu. Rev. Biochem.* **50**, 783–814.
9. Kacser, H. & Beeby, R. (1984) *J. Mol. Evol.* **20**, 38–51.
10. Beeby, R. & Kacser, H. (1990) in *Organisational Constraints on the Dynamics of Evolution*, eds. Maynard Smith, J. & Vida, G. (Manchester Univ. Press, U.K.), pp. 57–75.
11. Perelson, A. S. & Oster, G. F. (1979) *J. Theor. Biol.* **381**, 645–670.
12. Kacser, H. & Burns, J. A. (1973) *Symp. Soc. Exp. Biol.* **27**, 65–104.
13. Szathmáry, E. (1989) in *Oxford Surveys in Evolutionary Biology, Vol. 6*, eds. Harvey, P. H. & Partridge, L. (Oxford Univ., Oxford, U.K.), pp. 169–205.
14. Szathmáry, E. (1990) *Nature (London)* **344**, 115.
15. Ellington, A. D. & Szostak, J. W. (1990) *Nature (London)* **346**, 818–822.