

Nils Aall Barricelli
 Department of Mathematics
 University of Oslo
 Blindern, Norway

Received January 18, 1980; April 28, 1980

ABSTRACT: The purine-pyrimidine (Pu-Py) pairing hypothesis (Barricelli 1977 and 1979) holds that, at an early stage of evolution, pairing rules for nucleotides may have not been as restrictive as they are in most cases today, and may have permitted some forms of unconventional purine-pyrimidine pairing (cf. wobbling). This hypothesis is used to interpret the frequent cases of Uracil-Guanine (U-G) pairing in the secondary structure of RNA molecules. The interpretation is based on the assumption that the avoidance of U-G pairing in RNA copying processes and in codon-anticodon pairing (first two bases) is due not to an intrinsic property of the RNA molecules or their nucleotides, but to the properties of the specialized devices (RNA polymerases, ribosomes, etc.) used by living organisms today. This interpretation is considered applicable to RNA, but not necessarily, or not in the same measure, to DNA molecules.

Several implications, with supporting evidence, are indicated.

*

*

*

In a preceeding paper (Barricelli 1977) some degenerancy properties of the genetic code were interpreted by assuming that wobbling may at some time in the past have been a much more common property not only in the third base but also in the first base of each codon. A reconstruction of several earlier stages in the evolution of the genetic code has been proposed on the assumption that wobbling may originally have been a common phenomenon for all of the three bases in the codon triplets. I also made the suggestion that the pairing of uracil with guanine, which is a main characteristic of the wobbling rules, may have been a common characteristic of pairing in primordial RNA molecules not only in condon-anticodon pairing but in general in all pairing processes (cf. the "Pu-Py hypothesis", Barricelli 1977 and 1979, section 6).

Perhaps the most important implication of this hypothesis is that the pairing rules which we are accustomed to consider normal today for RNA molecules, may not always have been normal in the past, and may be the result of a specialization obtained in a long evolutionary process, reflecting a property of modern polymerases and other emzymes, rther than a property of the four RNA bases.

However, chemical laws are not easily changed by evolution. If the Pu-Py hypothesis was valid in the past, it should be valid also today except in those cases in which we would expect to find some enzymatic activity with the explicit function of preventing uracil-guanine (U-G) pairing. This has a series of implications which I shall now discuss.

Emzymatic activity preventing or avoiding U-G pairing is expected to operate in those cases in which U-G pairing would be harmful for survival, notably:

- (1) In the process of copying RNA into RNA, in which case U-G pairing could produce mutations or change the hereditary characteristics of the RNA copy.
- (2) In other pairing phenomena, such as codon-anticodon pairing, where the avoidance of U-G pairing in the first two codon bases is important for the correct identification of amino acids.

We expect no activity preventing U-G pairing in those paired regions of single-stranded RNA viruses and RNA molecules in which two RNA segments are kept together by hydrogen bonding for purposes which would not be compromised by U-G pairing. An important implication of the Pu-Py hypothesis is therefore:

* * * * *

Evolutionary Theory 5:29-33 (August, 1980)

The editors thank H. Hartman and M. Ishigami for help in evaluating this paper.

© 1980, the author

Table 1: Total number of C-G, U-A, U-G and C-A pairs in the secondary structures of (1) tRNA molecules, (2) RNA-virus segments, (3) other RNA segments published by Barrell and Clark (1974).

	C-G	U-A	U-G	C-A
tRNA	708	261	51	3
VIRAL RNA	334	234	52	1
OTHER RNA	342	233	101	11

* * * * *

Implication 1: In the paired segments of many RNA viruses and RNA molecules of living organisms one should expect to find U-G pairs linked by hydrogen bonds, just as one finds U-A and C-G pairs, although not necessarily with the same frequency. Because of the triple hydrogen bonding in C-G pairs, they will be more efficient than U-G pairs as a means of keeping the two RNA strands together. U-G pairs may therefore have a tendency to be replaced by C-G pairs whenever that is possible without interfering with the other functions of the RNA segment. We may therefore expect fewer U-G and more C-G than U-A pairs. Nevertheless, the frequency of U-G pairs should be substantial and, for example, much greater than the frequency of C-A pairs, if the (Pu-Py) hypothesis is correct.

Unfortunately, even in those RNA molecules in which the nucleotide sequence has been identified there is seldom enough information to give an unambiguous reconstruction of the secondary structure. Nevertheless several hundred proposed reconstructions of RNA secondary structures have been published in the literature. Common to them is a substantial representation of U-G pairs as shown in Table 1, based on secondary structures published by Barrell and Clark (1974), including the secondary structures of about 50 tRNA molecules and comparable numbers of other RNAs.

Among the most reliable secondary structures are probably those of tRNA molecules, whose familiar cloverleaf configuration imposes an extra constraint on the selection of secondary structures and does not leave as many possibilities of choice as other RNAs.

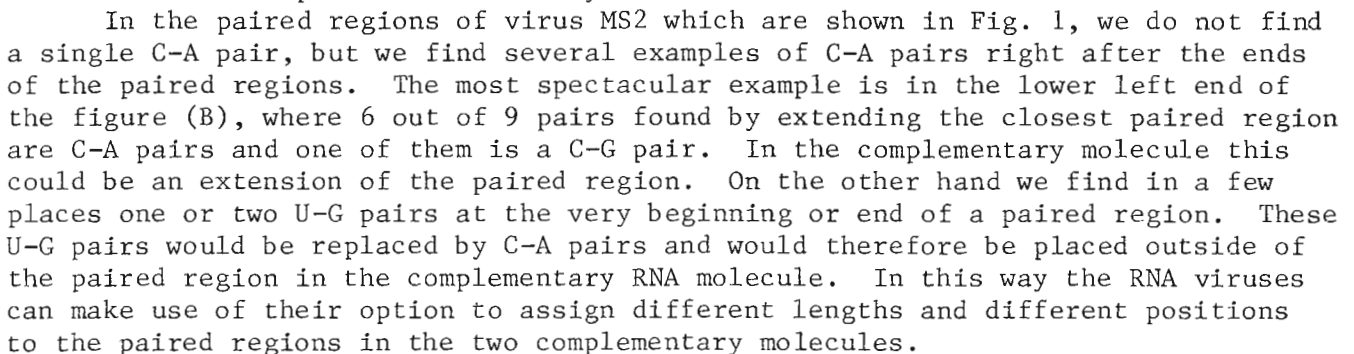
All the three classes of RNA listed in Table 1 show a substantial representation of U-G pairs, whose frequency is at least one order of magnitude greater than the frequency of C-A pairs. One example is presented in Fig.1, showing the secondary-structure models of two segments of the RNA bacteriophage MS2, namely: (A) the "end of protein intercistronic region to coat-initiation region" and (B) the "coat-protein region".

Quite a few U-G pairs, marked with arrows, can be seen in the figure. The ratio of U-G to U-A pairs is $\frac{UG}{UA} = \frac{10}{20} = 0.50$ in the upper figure (A) and $\frac{UG}{UA} = \frac{12}{60} = 0.20$ in the lower figure (B), numerators and denominators giving the number of U-G pairs and U-A pairs respectively.

In order to weaken the evidence presented in Table 1 and in other published RNA secondary structures one would have to show that meaningful secondary structures can be constructed without including a substantial number of U-G pairs, as a substitute for the published ones.

This has not been done yet. As long as it is not done we may be entitled to consider the published RNA secondary structures as a strong argument in support of the (Pu-Py) hypothesis. They also show that a large majority of the authors who have reconstructed secondary RNA structures consider the use of U-G pairs legitimate and/or unavoidable.

Implication 2: By making use of U-G pairs, RNA viruses have also the option of giving different extensions to each paired region in the two complementary versions of their molecule. In fact if pairing is wanted between two segments of an RNA molecule, but not in its complement (conditional pairing), all one has to do is to fill the would-be conditionally paired regions with U-G pairs and avoid U-A as well as



No overrepresentation of C-A pairs is found near the ends of paired regions in those RNA molecules which are copied from DNA, in which case it would not serve any obvious purpose. A slight over-representation of U-G pairs might, however, be found also in this kind of RNA molecules at the very ends of paired regions, if in many of these places there should happen to be a lowered selection for replacing U-G by -G pairs, compared with the rest of the paired regions (see implication 1). Implication 3: Annealing should be possible between UGUGUG strands but not so easily between poly-U and poly-G, because self structures of poly-G are too stable.*

Other annealing experiments involving also A's or C's can be designed in order to investigate U-G pairing properties in the presence of these nucleotides. Annealing experiments of this kind are well within reach of present technology, and we expect that the question will soon be answered. If the experiments should give a positive result, present views about the origin and the chemical basis for the 4-letter hereditary alphabet would have to be drastically changed.

Implication 4: Cytosine may not have been among the earliest bases synthesized by primordial RNA molecules. The basic implication of the Pu-Py hypothesis is that in the earliest ancestors of present RNA molecules each pyrimidine may have been replaceable by other pyrimidines and each purine may have been replaceable by other purines (see Barricelli 1977, last section, Table 12). The transition from this early stage to the present condition in which such substitutions are avoided, can be understood as a consequence of the gradual development of the biologic control of nucleotide synthesis. While the pre-biologic formation of some kind of nucleotides is required by all theories about the origin of polynucleotides, the question how biologic control of nucleotide production started is among the most controversial ones. Barring the unlikely possibility that all or most of the mechanisms for the production of the 5 nucleotides A, U, G, C, and T were created simultaneously, we are forced to the conclusion that the biologic control of nucleotide production may have been developed gradually, for one nucleotide at a time as the need arose. According to Crick (1968) "Adenine is likely to be the commonest base available in the primitive soup". If so, both Adenine and another nucleotide capable of pairing with Adenine may have been the first bases whose production was brought under biologic control. That other nucleotide cannot have been Cytosine, but Uracil is a very likely candidate during an early (A-U) stage, in which A and U may have been the two letters of a binary genetic alphabet. This applies even if there might have been a still earlier (A-I) stage in which Adenine and its deaminated form Inosine were the only two letters of the genetic alphabet, as suggested by Crick (1968), following a lead by L. Orgel. Before Cytosine could be introduced there must have been a (U-A,G) stage in which both Adenine and Guanine may have been used, possibly sometimes even substituting for one another in the same organism, since both would have been able to pair with U.

According to this evolutionary sequence Cytosine would have been the last one among the 4 main RNA nucleotides to be produced. This also fits well with results (Barricelli 1977, sections 6 and 7, Tables 10 and 11) presenting reconstructions of the earliest genetic codes strongly suggesting that the primordial RNA molecules may have had an early stage in their evolution in which Cytosine was absent. It is during the introduction of Cytosine that mechanisms designed to prevent U-G pairing (both by specific enzymes and possibly by the introduction of DNA genomes using Thymine, which is less prone to pair with G than Uracil is) would have been badly needed and were probably developed in order to avoid the substitution of U and A

* * * * *

*Stability constants for base pairings presented by Eigen (1971), Table 12, in two different non-polar media C_6H_6 and CCl_4 , $K_{Ass} (M^{-1})$ at $25^\circ C$, give respectively the values 30,000 for CG, 150 for UA, 1,200 for GG and $<1,200$ for UG in the first medium (C_6H_6), and the values 10,000 for CG, 550 for UA, 1,000 for GG and $<1,000$ for UG in the second medium (CCl_4). In either case the stability constant $<1,200$ or $<1,000$ for UG pairing does not have to be smaller than the constant 150 or 550 for UA pairing. The high value of 1,200 or 1,000 for the stability constant for GG pairing has prevented a better determination of the UG pairing constant.

bases by C and G respectively in those places in which the need for a U or an A had already been irreversibly established by preceding evolution. One example of a U base whose need may have been irreversibly established before abundant C production, has been presented in the preceding paper (Barricelli 1977, sections 6 and 7, Tables 9,10 and 11).

The present abundance of C-G pairs in the paired regions of RNA molecules can be explained by the tendency of U-G pairs to be replaced, whenever possible, by C-G pairs (see above implication 1) and is not a valid argument for drawing conclusions about the time in which C bases were originally introduced.

Other important implications of the Pu-Py hypothesis deal with the origin and evolution of primordial polynucleotides, and the methods one should use if attempts were made to reconstruct such processes in vitro.

REFERENCES

- Barrell, B.G. and B.F.C. Clark. 1974. Handbook of nucleic acid sequences. Joynson-Bruvvers Ltd. Oxford, England.
- Barricelli, N. Aall. 1977. J. theor. Biol. 67:85-109.
- Barricelli, N. Aall. 1979. Biosystems 11:19-28.
- Crick, F.H.C. 1968. J. Molec. Biol. 38:167.
- Eigen, M. 1971. Naturwissenschaften 33a:465-523.
- Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berge, G. Volckaert, and M. Ysebært. 1976. Nature 260:500-507.