ESSAY

OBC
www.rsc.org/obc

Genetic alphabetic order: what came before A?

Jay S. Siegel[a] and Yitzhak Tor[b]

[a] *Organisch-chemisches Institut, Universität Zürich, Winterthurerstr. 190, Zürich, CH-8057, Switzerland*

[b] *Department of Chemistry, University of California, San Diego, La Jolla, CA, 92093-0358, USA*

Received 19th January 2005, Accepted 14th March 2005
First published as an Advance Article on the web 14th April 2005

Simple and elegant, four "letters" (ATGC) universally encode the entire genome of every contemporary living organism. But what came before A? How did we arrive at this fortuitous mixture of purines and pyrimidines given the thin gruel that must have constituted the primordial soup? Could a single heterocycle have been the stock for a catalytic consommé that nourished the pre-RNA world? In the absence of a catastrophic reconstitution of life's biomolecular basis somewhere along its molecular evolution, the precursor nucleobase should have been compatible with the present system, but less fit for the environmental pressures that motivated mutational evolution. Such a structure would likely have been isosteric with the modern code but thermodynamically less stable. In addition it would be much easier to accept a fundamental progenitor system if a single heterocycle could have established a protocode. These are tough specs for a primitive system to meet. But by weathering these rocks of refutation well, the successful structural hypothesis could found a new approach in thinking about the molecular evolution of the genetic alphabet.

Despite this apparent paradox of structural and functional requirements, consideration of the existing letters shows that key clues may have been staring us in the face all along. Examination of the structure of cytosine, C, and uracil, U, reveals that U is the formal hydrolysis product of C. But then what is C if not the hydrolysis product of diaminopyrimidine, which we can call D (Fig. 1)? As such it is chemically feasible that a large mass of D plus water could lead directly to a pool of D, C, and U through "hydrolytic mutation", if you will. D therefore survives the single source criterion.

So, how does that solve anything? Let's consider the special character of purines that allows them to form glycosides through their endocyclic nitrogens. The tautomeric preference that Jerry Donahue pointed out to Watson and Crick made it clear that nitrogens adjacent to exocyclic oxygens adopted "amide" functionality, whereas an adjacent exocyclic nitrogen preferred aminoimine forms.[1] Thus, our new letter D could only participate in the code if it formed a glycoside through an exocyclic amino group.

Yes. . .are you still waiting for the nickel to drop? Well, a D nucleoside would therefore be isosteric to A, complementary to U and a progenitor of C and U! (Fig. 2) Two additional criteria, isosterism and compatibility, are thus addressed without excluding D. Is D then the molecular missing link to the primal genetic alphabet?

The most obvious difference between the four-letter code of today and the three-letter code we are now postulating is one letter. Could this suffice to create a catalytic pre-RNA world? Indeed, two letters are all one needs to establish RNA-like secondary structure with dangling bases for catalytic activity.[2] The third letter is a surplus for the catalytic mission and a motivation to evolve toward a more stable four-letter system. Polymers of D, C, and U-based protonucleotides could well adopt functionally rich and architecturally specific nucleic acid structures. The further criteria of form and function are met and D is still in the race.

Crick has raised the question of an all purine genetic code.[3,4] Joyce has discussed the possibility of a three-letter or even two-letter nucleic acids,[2,5] but again placing more importance on the role of the purines. Miller has analyzed the formation and hydrolysis of diaminopyrimidine but omitted any discussion of its potential role as a progenitor base.[6] The puzzle has been on the table for a long time, but the "aha" has been missing. Given the compelling relationship among D, C, and U, we formulate our central but very simple hypothesis around the idea that D was the first nucleobase, that C and U "evolved" from D by hydrolysis in an aqueous environment, and that although proximity favored a DCU genetic alphabet at the start, ultimately the thermodynamic benefit of complete pairing motivated a molecular mutation process that resulted in substitution of D by purines like A, and the incorporation of purines like G to complement C.

Following the ideas of protein evolution discussed by Doolittle, optimization of structure and function must co-evolve.[7] Thus, a prevailing hypothesis should help us understand the progression from random structure and catalytic function to regular secondary structure with specific catalytic (replicative) function to a significantly more stable form suitable for storage of genetic information and the associated increasingly more faithful replicative function. In principle, a two to three to four letter code progression and the proto-RNA to RNA to DNA world set exactly the same criteria.

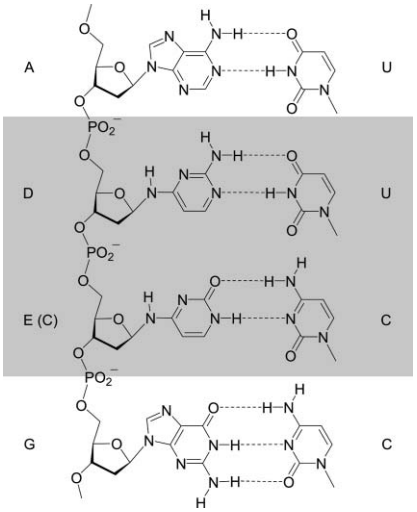

Fig. 2 Base pairing between the proto-letter D and U in a typical double-helix sequence.

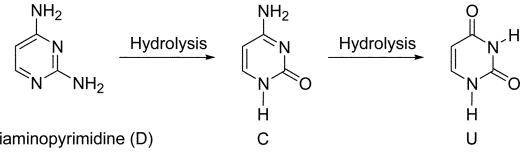

Fig. 1 Hydrolysis cascade from diaminopyrimidine (D) to C to U.

DOI: 10.1039/b500921a

An initial DCU code can get things started, albeit in a clunky and inefficient manner that would have to evolve. Already in the DCU alphabet there is a coding "error" possible, wherein C would form a glycosidic bond through its exocyclic amine; let's call this E (for error!). E would serve as a harbinger of modern G; in the protocode it could pair with C, as one might expect today, but could also pair with D, analogous to an A–G pair. Under the less restrictive structural energetic requirements of that time, anything that promotes structure could be a competitive advantage. Indeed, the fossils of such D–E pairing might be the A–G pairing regularly found in modern folded RNA structures.[8] Thus, the DCU code jump-starts the process at the 3-letter stage and has within its "dislexicon" the futuristic jargon necessary to give rise to a modern RNA with the advent of purines. As already noted, the evolution to T from U in the DNA world adds a specificity benefit, but would not exclude the use of U during the transition period.

Szathmary has postulated that the four-letter code is theoretically defensible as "optimal."[9] He cites Gardner's work on how the size of the genetic alphabet could "stabilize" the genetic code.[10] The three criteria used were the chances of folding, diversity of folded isomers, and difference between folded and unfolded structures. "More is less" was the conclusion. But if you want to encode information, two letters does not lead to good "replicability." In addressing Orgel's comments about whether there are four-letters because nature never experimented with more,[11] Szathmary suggests that there must have been a myriad of possible prebiotic base-pair possibilities and 'at any rate it does not explain why we do not have only two bases.' If D and hydrolysis were the progenitor pool, then three would be a natural local number of possibilities, and the use of the full array randomly would exceed two but still remain in a range acceptable to Gardner's criteria.

Independent of their possible role as precursor to the modern code, nucleosides formed through the exocyclic amines also offer an extension of the modern repertoire of nucleotide mimetics (Fig. 3).[12] Greenberg has shown that formamidopyrimidines are important pseudo-bases formed during the damage of DNA.[13] Although they are later edited out from the strand, they can function as isosteres of A in a nucleic acid. Various di- and tri-amino substituted azabenzenes could also serve as isosteres of A, and as such be A-mimetics in a variety of medical therapies and diagnostics, such as antisense agents, nucleoside phosphate mimetics, ribozyme inhibitors, or anywhere that a nucleobase cognate might assume the place of a "natural" partner.[14] New anti-virals or antibiotics would be obvious targets, but the role of nucleotidephosphates as ubiquitous co-factors in biological processes further expands the impact of these protobases beyond their potential historical significance.

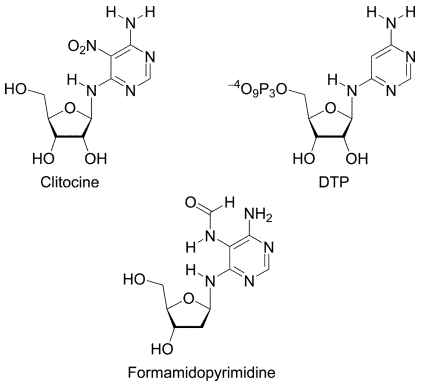

Fig. 3 Structures of clitocine (fungal metabolite), DTP (isostere of ATP), and the formamidopyridines (recently studied by Greenberg).

The beauty and the frustration of prebiotic chemistry is that we can never really know its true history. Our postulate of diaminopyrimidine D as the first genetic letter is intended to provoke criticism and motivate new studies. Experiments are now underway to prepare nucleotidic D and to incorporate it into DNA and RNA sequences appropriate for studying its relevance to genetic molecule evolution. We welcome additional vigorous and critical experimental tests.

References

1 J. D. Watson and F. H. C. Crick, *Nature*, 1953, **171**, 737–738.
2 J. S. Reader and G. F. Joyce, *Nature*, 2002, **420**, 841–844.
3 F. H. C. Crick, *J. Mol. Biol.*, 1968, **38**, 367–379.
4 G. Waechtershaeuser, *Proc. Natl. Acad. Sci. USA*, 1988, **85**, 1134–5.
5 J. Rogers and G. F. Joyce, *Nature*, 1999, **402**, 323–325.
6 M. P. Robertson, M. Levy and S. L. Miller, *J. Mol. Evol.*, 1996, **43**, 543–550.
7 M. Simon, J. Zieg, M. Silverman, G. Mandel and R. Doolittle, *Science*, 1980, **209**, 1370–1374.
8 P. B. Moore, *Annu. Rev. Biochem.*, 1999, **68**, 287–300.
9 E. Szathmary, *Nat. Rev. Genet.*, 2003, **4**, 995–1001.
10 P. P. Gardner, B. R. Holland, V. Moulton, M. Hendy and D. Penny, *Proc. R. Soc. London., Ser. B*, 2003, **270**, 1177–1182.
11 L. Orgel, *Nature*, 1990, **343**, 18–20.
12 C. R. Geyer, T. R. Battersby and S. A. Benner, *Structure (London)*, 2003, **11**, 1485–1498.
13 (*a*) M. M. Greenberg, Z. Hantosi, C. J. Wiederholt and C. D. Rither, *Biochemistry*, 2001, **40**, 15856; (*b*) K. Haraguchi and M. M. Greenberg, *J. Am. Chem. Soc.*, 2001, **121**, 8636; (*c*) K. Haraguchi, M. O. Delaney, C. J. Wiederholt, A. Sambandam, Z. Hantosi and M. M. Greenberg, *J. Am. Chem. Soc.*, 2002, **124**, 3263.
14 A related nucleoside based on a nitrodiaminopyrimidine, clitocine, is a bioactive fungal metabolite, and supports this hypothesis, see: (*a*) R. J. Moss, C. R. Petrie, R. B. Meyer, L. D. Nord, R. C. Willis, R. A. Smith, S. B. Larson, G. D. Kini and R. K. Robins, *J. Med. Chem.*, 1988, **31**, 786–790; (*b*) T. Kamikawa, S. Fujie, Y. Yamagiwa, M. Kim and H. Kawaguchi, *J. Chem. Soc., Chem. Commun.*, 1988, 195–196.