INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus
Biologies

Jean Weissenbach

Comments on the origin of the genetic code: a 27-codon hypothetical precursor of an intricate 64-codon intermediate shaped the modern code

Volume 343, issue 4 (2020), p. 7-9.
Published: 21st April 2021

<https://doi.org/10.5802/crbiol.46>

Les Comptes Rendus. Biologies *sont membres du*
Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org

Comptes Rendus
Biologies
2020, 343, nº 4, p. 7-9
https://doi.org/10.5802/crbiol.46

News and Views / *C'est apparu dans la presse*

Comments on the origin of the genetic code: a 27-codon hypothetical precursor of an intricate 64-codon intermediate shaped the modern code

Commentaires sur l'origine du code génétique : un hypothétique précurseur de 27 codons d'un intermédiaire complexe de 64 codons a façonné le code moderne

Jean Weissenbach [a]

[a] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France
E-mail: Jean.Weissenbach@genoscope.cns.fr

Manuscript received and accepted 25th February 2021.

The origin of the genetic code and its evolution remains one of the great enigmas of contemporary biology. Since the mid-1960s, following the deciphering of the genetic code, hypotheses and models have followed one another. But the debate remains totally open today. A great deal of data, which sheds new light on the issue, has also accumulated in recent decades. It is in this context that this article by Bernard Dujon presents a new possible evolutionary path based on a set of facts, sometimes antagonistic, known to date [1].

The central hypothesis of this model postulates that the current genetic code derives from an ancestral system in which nucleic acids contain only three nucleobases making 27 possible triplets. The combination chosen by Dujon would include bases G, U and C and would allow the formation of G–C and G–U basepairs in the translation process. This codon/anticodon recognition system would rely on the formation of a strict Watson–Crick basepair in central position 2, flanked by often less stable pairs in 1 and 3 of the codon-anticodon associations. This would make it possible to code for a limited number of amino acids among which we find 5 current amino acids. The main difficulty of such a code is its replication, since G could pair with C or U. It is likely that in the synthesis of the complementary strand, the C's, whose pairing with G is more stable, will be preferentially incorporated into the neo-transcribed strand. To restore the lost U's, an oxidative deamination of some C's could act as a more or less faithful editing process. One could also consider other mechanisms

for restoring the initial sequence or assume that these poly (U, C, G) RNAs are randomly assembled without the help of a template.

The operation of such a recognition system has a whole set of consequences on the reading process, in particular the existence of (1) ambiguities in the recognition of codons by anticodons of tRNAs carrying distinct amino acids and (2) intricacies in which the same anticodon can recognize codons encoding different amino acids. In support of his hypothesis, Dujon mentions two main types of observations: (1) that the enzymes that activate the tRNAs , carrying the 5 amino acid of this reduced code, belong all to the same class of tRNA synthetases and (2) the analysis of variant codes found in certain protists and in cellular organelles.

Dujon then envisages the progressive transition from the 27 triplet code to the modern 64-codon code. His scenario postulates the integration of a second purine (A) in the coding sequences. This integration of A will be accompanied by the progressive extension of the repertoire of amino acids composing the proteins up to the current 20 as well as important changes in the decoding/translation process. An increase in the matching possibilities will progressively allow a gain in robustness of the codon/anticodon pairing which will include 2 strict Watson–Crick pairs at least, resulting in a more accurate decoding accompanied by a complexification of the code. e.g.an increase in the stringency of the pairing between position 1 of the codons and 3 of the anticodons enabled by the establishment of strict Watson–Crick pairs will reduce and progressively eliminate coding ambiguities and intricacies. Dujon goes so far as to propose a chronology and dating of certain stages in the transition to the modern code in connection with environmental changes such as the beginning of oxygen accumulation in the atmosphere.

This article integrates many facts established in recent decades and proposes a coherent vision of the evolution of the genetic code from an early stage of its existence, certainly hypothetical, but possible. As is often the case, this novel hypothesis raises many new questions. As a result, it constitutes a new basis for discussion for anyone interested in the establishment of the very fundamental biological processes. Bernard Dujon's model provides a new view that further enriches this major debate that has been ongoing in the community for nearly 60 years.

French version

L'origine du code génétique et son évolution restent une des grandes énigmes de la biologie contemporaine. Depuis le milieu des années 60, à la suite du décryptage du code génétique, les hypothèses et les modèles se sont succédé. Mais le débat reste aujourd'hui totalement ouvert. De nombreuses données, qui permettent d'éclairer la question sous un jour nouveau se sont aussi accumulées au cours des décennies récentes. C'est dans ce contexte que cet article de Bernard Dujon présente une nouvelle voie d'évolution possible qui s'appuie sur un ensemble des faits, parfois antagonistes, connus à ce jour [1].

L'hypothèse centrale de ce modèle postule que le code génétique actuel dérive d'un système ancestral dans lequel les acides nucléiques ne contiendraient que trois nucléobases, soit 27 triplets possibles. La combinaison retenue par Dujon inclurait les bases G, U et C et autoriserait la formation des appariements G–C et G–U dans le processus de traduction. Ce système de reconnaissance codon/anticodon reposerait sur la formation d'un appariement de type Watson–Crick strict en position centrale 2, flanqué de paires souvent moins stables en 1 et 3 des associations codon-anticodon. Ceci permettrait de coder pour un nombre limité d'aminoacides parmi lesquels on retrouve 5 aminoacides actuels. La principale difficulté d'un tel code est sa réplication, puisque G pourrait s'apparier avec C ou U. Il est vraisemblable que dans la synthèse du brin complémentaire, les C dont l'appariement avec G est plus stable, seront préférentiellement incorporés dans le brin néotranscrit. Pour rétablir les U perdus, une désamination oxydative de certains C pourrait agir comme un processus d'édition plus ou moins fidèle. On pourrait aussi envisager d'autres mécanismes de rétablissement de la séquence initiale ou supposer que ces ARN poly (U, C, G) sont assemblés au hasard sans l'aide d'une matrice.

Le fonctionnement d'un tel système de reconnaissance entraîne tout un ensemble de conséquences sur le processus de lecture, en particulier l'existence (1) d'ambiguïtés de reconnaissance de codons par

des anticodons de tRNA portant des aminoacides distincts et (2) des intrications dans lesquelles un même anticodon peut reconnaître des codons de sens différents. A l'appui de son hypothèse, Dujon mentionne principalement deux types d'observations : (1) l'appartenance des enzymes d'activation des tRNA, utiles à la traduction du code réduit à 5 aminoacides, à une seule des deux classes de tRNA synthétases et (2) l'analyse des codes variants trouvés chez certains protistes et dans les organites cellulaires.

Dujon envisage ensuite le passage progressif de ce code à 27 triplets au code moderne à 64 codons. Son scénario postule l'intégration d'une deuxième purine (A) dans les séquences codantes. Cette intégration de A va s'accompagner de l'extension progressive du répertoire d'aminoacides constituant les protéines jusqu'aux 20 actuels ainsi que de changements importants au niveau du processus de décodage/traduction. Une augmentation des possibilités d'appariement va progressivement permettre un gain de robustesse des liaisons codon/anticodon qui comprendra au moins 2 paires Watson–Crick strictes, d'où un décodage plus précis s'accompagnant d'une complexification du code. Parmi ces gains de robustesse on peut prendre l'exemple de l'augmentation de la rigidité de l'appariement en position 1 des codons et 3 des anticodons grâce à l'établissement de vraies paires Watson–Crick qui réduiront puis supprimeront les ambiguïtés de codage et les intrications. Dujon va jusqu'à proposer une chronologie et une datation de certaines étapes dans la transition vers le code moderne en liaison avec les changements environnementaux tels que le début d'accumulation d'oxygène dans l'atmosphère.

Cet article intègre de nombreux faits établis au cours des décennies récentes et propose une vision cohérente de l'évolution du code génétique à partir d'une phase précoce de son existence, certes hypothétique, mais possible. Cette hypothèse inédite, comme souvent, suscite de nombreuses questions nouvelles. De ce fait elle constitue une nouvelle base de discussion pour toute personne intéressée par la mise en place des processus biologiques fondamentaux. Le modèle de Bernard Dujon apporte un nouvel angle d'attaque à ce débat majeur qui anime la communauté depuis bientôt 60 ans.

References

[1] B. Dujon, "On the origin of the genetic code: a 27-codon hypothetical precursor of an intricate 64-codon intermediate shaped the modern code", *C. R. Biol. Acad. Sci.* **343** (2021), no. 4, p. 15-52.