

Experimentally testing the hypothesis of a limited amino acid repertoire in primitive proteins. S. Akanuma¹, Y. Nakajima¹, S. Yokobori¹ and A. Yamagishi¹, ¹Department of Applied Life Sciences, Tokyo University of Pharmacy and Life Sciences, 1432-1 Horinouchi, Hachioji, Tokyo 192-0392, Japan.

Introduction: The genetic code is an essential element of life because it links genetic information to proteins that express biological functions. The modern genetic code, which encodes the standard 20 amino acids (and the termination signal) using 64 triplet codons, is shared by all of the extant organisms on the earth with a few exceptions. Therefore, the genetic code is thought to have been established at the age of the last universal common ancestor, which we call the Commonote (Fig. 1).

As to the origin and evolution of the genetic code, a number of theories have been proposed. Crick proposed the *frozen accident theory* [1] in which he suggested two things; i) The code is universal because, at the present time, any change would be lethal, or at least very strongly selected against; and ii) The shape of the genetic code table was entirely a matter of *chance*. Theories that rationalize the evolution of the genetic code have been also proposed: e.g. *Stereochemical interaction* [2], *Co-evolution with amino acid biosynthesis* [3, 4], *Error minimization*, and *Expanding codons theories* [5]. These theories all suggest that only a few amino acids were used in primitive proteins and later the amino acid repertoire gradually increased up to 20 through the course of evolution.

If it was the case, how many number of and which types of amino acids were involved in the earliest protein synthesis system? By comparing homologous amino acid sequences from different organisms or referring to various criteria for amino acid chronology, chronological order of appearance of amino acids in the early evolution have been predicted [6–8]. However, only limited experimental studies have partially tested these predictions [9–12].

Given simpler protein synthesis system, the primitive proteins, which might have comprised a reduced set of amino acids, must have had a sufficiently adequate structure for functional interactions and catalysis. To address this issue experimentally, we created reduced amino acid set proteins to examine whether a protein composed of less than 20 types of amino acids can form a stable structure and express a biological function. To this end, we first resurrected several ancestral proteins and then restricted the amino acid usage of a resurrected protein to reduced amino acid sets (Fig. 1).

Resurrection of ancient proteins: We targeted ancestral nucleoside diphosphate kinase (NDK) se-

quences for resurrection. The kinase catalyzes the transfer of a phosphate from a nucleoside triphosphate to a nucleoside diphosphate. The ancestral NDK may have arisen early because at least one gene that encodes NDK is present in most extant organisms. The first step in the reconstruction of ancestral NDK sequences is to prepare multiple amino acid sequence alignments using homologous sequences of NDK from extant species, which is then used to build phylogenetic trees. Then, ancestral sequences of NDK that seem to represent the last common ancestors of Archaea and of Bacteria were designed using the information contained in predictive phylogenetic trees. These ancestral kinases display extreme thermal stabilities, suggesting thermophilic ancestries for Archaea and Bacteria [13].

Experimentally testing the limited amino acid repertoire hypothesis: It is generally impossible to infer amino acid sequences that existed before the Commonote by currently well-used phylogenetic analysis. Instead, using the most thermally stable reconstructed NDK, Arc1 [13], as the starting molecule, we restricted its amino acid usage to several reduced sets. First, we followed the chronological order of appearance of amino acids proposed by Trifonov [7, 8]; Met, Gln, Lys, Tyr and Asn, which have been proposed to appear later, were replaced by other amino acids, thus

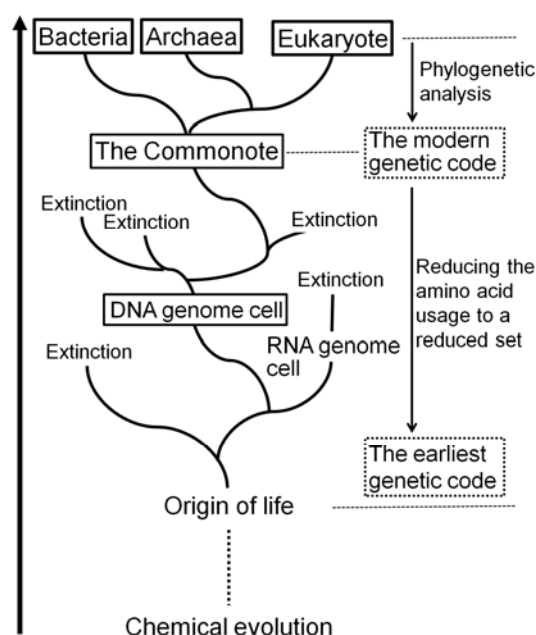


Fig. 1. A model for the history of life on the earth and our experimental strategy to address the limited amino acid repertoire hypothesis.

creating Arc1ΔKMNQY. Because Cys is absent from Arc1, Arc1ΔKMNQY consists of only 14 amino acid species. Arc1ΔKMNQY retains high thermal stability; whereas, no detectable level of catalytic activity was observed. Therefore, the fourteen amino acid types are sufficient to encode a thermally stable protein but more amino acid types would be required for its function.

In order to choose a reduced amino acid set more objectively and systematically, we evaluated the individual contributions of the 19 amino acid types to the stability and activity of Arc1. We reconstructed 19 Arc1 variants in which one of the 19 amino acid types was all replaced by other amino acids, thus creating proteins that consisted of 18 amino acid species. As the result, we found that some amino acid species would be easily lacked but others would be important for the NDK's structure and function. Based on the observation, we reconstructed two Arc1 variants, both of which consisted of respective 13 amino acid types. One of the variant is stable at up to 70°C and exhibits detectable level of catalytic activity. The other variant must be catalytically inactive because it lacks some catalytically important amino acids. However, the conformational stability of the variant is similar to that of a hyperthermophilic NDK.

Perspectives: In future, we will further restrict the amino acid usage to identify the minimum amino acid set that are required to encode a stable and catalytically active NDK. The result will provide an important insight into the amino acid usage in primitive proteins as well as the origin and evolution of the genetic code.

References: [1] Crick. F. H. C. (1968) *J. Mol. Biol.* 38, 367–379. [2] Woese C. R. (1967) *The genetic code*, Haper and Row, New York. [3] Wong J. T. (1975) *PNAS* 72, 1909–1912. [4] Wong J. T. (2005) *BioEssays* 27, 416–425. [5] Eigen M. and Schuster P. (1977) *Naturwissenschaften* 64, 541–565. [6] Jordan I. K. *et al.* (2005) *Nature* 433, 633–638. [7] Trifonov E. N. (2000) *Gene* 261, 139–151. [8] Trifonov E. N. (2004) *J. Biomol. Struct. Dyn.* 22, 1–11. [9] Akanuma S. *et al.* (2002) *PNAS* 99, 13549–13553. [10] Walter K. U. *et al.* (2005) *J. Biol. Chem.* 280, 37742–37746. [11] Doi N. *et al.* (2005) *Protein Eng. Des. Sel.* 18, 279–284. [12] Tanaka J. *et al.* (2011) *PLoS ONE* 6, e18034. [13] Akanuma S. *et al.* (2013) *PNAS* 110, 11067–11072.