Why are there 20 amino acids and 4 nucleotides?

Ezequiel Galpern, Diego Ferreiro, Ignacio Sánchez

https://doi.org/10.1590/SciELOPreprints.11013

Submitted on: 2024-12-30
Posted on: 2025-01-10 (version 1)
(YYYY-MM-DD)

Why are there 20 amino acids and 4 nucleotides?

Ezequiel A. Galpern, Diego U. Ferreiro*, Ignacio E. Sánchez*

Laboratorio de Fisiología de Proteínas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN-CONICET), Buenos Aires CP1428, Argentina.

* To whom correspondence may be addressed, email**:** isanchez@qb.fcen.uba.ar, ferreiro@qb.fcen.uba.ar

Abstract

Evolution of folded biopolymers requires effective and fast search of both the conformational space for folding and the sequence space for evolution. Molecular information theory and energy landscape theory show that the alphabet size of extant proteins and RNA is just enough to fulfill these requirements, given the constraints posed by the chemical physics of these polymers. Empirical estimations of the size of the effective sequence and conformational spaces of natural biopolymers support the theoretical predictions.

Keywords

Biopolymer, random energy model, genetic code

Author contributions

EAG DUF and IES designed research, performed research, analyzed data, and wrote the manuscript.

Conflict of interests

none declared

Ezequiel A. Galpern https://orcid.org/0000-0001-9516-3985

Diego U. Ferreiro https://orcid.org/0000-0002-7869-4247

Ignacio E. Sánchez https://orcid.org/0000-0003-4284-9013

Introduction

Terrestrial biochemistry is based on the existence of linear polymers of just a handful of monomers. A fundamental aspect of Biology is that the functional structures of these molecules can be *coded* in linear strings of units which, in the appropriate conditions, spontaneously fold into stable structures (1). In turn, the exploration of the vast sequence spaces that encodes structures has to be efficient enough such that evolution can proceed (2). It has long been debated why are current polypeptides and polynucleotides encoded with just 20 and 4 monomers respectively, out of the many chemically plausible alternatives (3, 4). Here we show that the conditions for fast folding and effective evolution strongly constraint the coding alphabet sizes to just the currently used by proteins and RNA.

A simple theory for biopolymer alphabet size

We have previously integrated molecular information theory with energy landscape theory to analyze the relationship between configurational entropy, sequence entropy and alphabet size in biopolymer folding (5, 6). From the viewpoint of molecular information theory, folding is a self-recognition molecular process where the information gained by finding the native configuration ($R_{Levinthal}$) matches the information gained upon going from a random sequence to an evolved sequence ($R_{sequence}$). Empirical estimations show that this is indeed the general case for protein folding, at 2.2 ± 0.3 bits/(site·operation) (5). $R_{Levinthal}$ can be related to the effective number of configurations per site in the unfolded and folded configurations ($N_{unfolded}$ and N_{native}) and $R_{sequence}$ can be related to the size of the monomer alphabet and the effective number of monomers per site in the evolved sequences (*Alphabet size* and $N_{evolved}$).

$$R_{Levinthal} = log_2\left(\frac{N_{unfolded}}{N_{native}}\right) = R_{sequence} = log_2\left(\frac{Alphabet\ size}{N_{evolved}}\right); \frac{Alphabet\ size}{N_{evolved}} = \frac{N_{unfolded}}{N_{native}} \quad [1]$$

The application of energy landscape theory to a random energy model indicates that it is feasible to find sequences that fold into a given structure as long as the number of monomers per site of an evolved sequence ($N_{evolved}$) is larger than the effective number of configurations per site in the unfolded biopolymer ($N_{unfolded}$) (5, 7):

$$N_{evolved} > N_{unfolded} \quad [2]$$

Combining [1] and [2] at the limit of $N_{evolved}$=$N_{unfolded}$, it follows that

$$\left(N_{unfolded}\right)^2 = Alphabet\ size \cdot N_{native} = \left(N_{evolved}\right)^2 \quad [3]$$

Extant folded biopolymers can be usually well described using a single reference structure and small fluctuations around it. In this case, we can approximate N_{native} to 1 and

$$\left(N_{unfolded}\right)^2 = Alphabet\ size = \left(N_{evolved}\right)^2 \quad [4]$$

Equation [1] constraints the ratio of *Alphabet size* and $N_{evolved}$, but not their absolute values. In contrast, equation [4] fully constraints $N_{evolved}$ and *Alphabet size*, given a value of $N_{unfolded}$. This result should be valid for any evolved and folded biopolymer (8).

Empirical test of the theory

The theoretical results pose three clear relationships between *Alphabet size*, $N_{unfolded}$, and $N_{evolved}$, namely that both $N_{unfolded}$ and $N_{evolved}$ should be equal to each other and to the square root of the *Alphabet size*. As the current *Alphabet size* is 20 amino acids in the case of proteins and 4 nucleotides in the case of RNA, equation [4] also predicts that $N_{unfolded}$=$N_{evolved}$=2 for RNA and $N_{unfolded}$=$N_{evolved}$≈4.47 for proteins (Figure 1).

We previously approximated $N_{unfolded}$ of proteins from six estimations of the polypeptide backbone configurational entropy, the effective secondary structure alphabet, structure prediction from residue burial layers and structure prediction from residue-residue contacts (5), obtaining an average value of 5.38±0.46 conformations per amino acid. To evaluate $N_{unfolded}$ of RNA we used estimations of the loss in backbone configurational entropy upon RNA folding. Analysis of the 23S ribosomal subunit yields a loss in backbone configurational entropy of 1.52 cal/(mol·K) per nucleotide (9). A similar analysis of several ribozymes and riboswitches yields an average loss in backbone configurational entropy of 2.20 cal/(mol·K) per nucleotide (10). Since $N_{unfolded} = e^{\frac{\Delta S}{R}}$, the estimated values for $N_{unfolded}$ are 2.13 and 3.00 conformations per nucleotide, respectively. We also approached RNA folding as the finding of one base pair among all possible pairs for each nucleotide of the chain. As a first approximation, we consider that the searches in the different sites in an RNA chain are independent. We can calculate the effective size of the base pair structural alphabet for RNA using the base pair abundances observed in natural RNA structures. Performing this calculation for the data in Table 3 of (11) yields a value for $N_{unfolded}$ of 2.48 conformations per nucleotide. From our three estimates of $N_{unfolded}$, we calculated an average $N_{unfolded}$ for RNA of 2.54±0.25 conformations per nucleotide. The values of $N_{unfolded}$ for both proteins and RNA are close to the square root of the respective *Alphabet size*, fulfilling the first prediction of the model (Figure 1, blue circles).

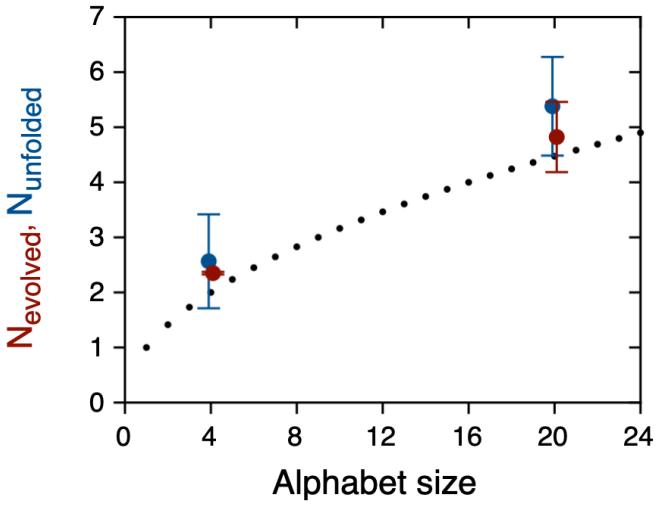

Figure 1. *Relationship between Alphabet size, $N_{evolved}$ and $N_{unfolded}$ for biopolymer folding and evolution. The black circles indicated the expected value for each Alphabet size according to equation [4]. The red circles indicate empirical estimates of $N_{evolved}$ for folded RNA and protein molecules. The blue circles indicate empirical estimates of $N_{unfolded}$ for folded RNA and protein molecules. The error bars indicate the mean plus/minus 1.96 times the standard error of the mean (95% confidence interval for estimation of the mean value). Red and blue circles are slightly displaced along the x-axis for clarity.*

We previously estimated the $N_{evolved}$ of proteins from an *Alphabet size* of 20 and seven estimations of the sequence entropy of natural proteins obtaining an average value of 4.82±0.32 amino acids per site (5). Here, we estimate the $N_{evolved}$ of naturally occurring folded RNA molecules which can be calculated from sequence variations within natural RNA families. We used the methodology in (5) to calculate the average $N_{evolved}$ over all sites with less than 50% gaps in over 1200 RNA alignments from the Rfam database (12), considering that the sequence restrictions at the different sites in an RNA chain are independent. The average value of $N_{evolved}$ over all alignments is 2.35±0.024 nucleotides per site. The values of $N_{evolved}$ for both proteins and RNA are close to the square root of the respective *Alphabet size*, fulfilling the second prediction from the model (Figure 1, red circles). Evolved protein and RNA sequences contain enough information to code for an efficient search of the conformational space during folding (5, 13). The blue and red circles in Figure 1 also show that the values of $N_{evolved}$ coincide with those for $N_{unfolded}$ for both proteins and RNA, fulfilling the third and last prediction of equation [4]. This ensures an efficient search of the sequence space in the evolution of foldable biopolymers (5).

Discussion

We present a simple theoretical result for the restriction of the *Alphabet size* of biopolymers. The potential for folding and evolution is fundamentally constrained by the physico-chemical degrees of freedom of the soluble polymer that determine $N_{unfolded}$. For spontaneous folding, the information contained in evolved sequences must be enough to *code* for the structures of the native folded states ($R_{sequence}$ = $R_{Levinthal}$). This is mathematically equivalent to the quantification of their functional information (14). The minimum mapping between the configurational and the evolutionary landscapes is thus $N_{evolved}$ = $N_{unfolded}$ such that one effective digital character is matched with one effective analog conformation. This directly bounds the minimum digital *Alphabet size*. The simplest codable and foldable polymer would be that one for which $N_{unfolded}$ ≈2, restricting the *Alphabet size* to 4 digits. Foldable RNA molecules appear to fall in this category. Detailed computational studies on the design of foldable RNA sequences coincide to show that an *Alphabet size* of 4 was the minimum required (13). For proteins, the polypeptide backbone geometry dictates that $N_{unfolded}$ = 5.38 ± 0.46, and as such *Alphabet size* must be larger than for RNA, in the order of 20~39. The fact that the *Alphabet size* of modern genetically coded proteins is 20 may reflect constraints on the size of the molecular systems involved in protein synthesis and that further information besides folding needs to be coded for function (1). For proteins $N_{evolved}$ = 4.82 ± 0.32, which is in line with the theoretical and experimental findings for the minimal number of amino acids needed to specify a fold (5, 15). The chemical nature of polynucleotide and polypeptide chains imply that a minimal non-overlapping genetic code relating them must be one of degenerate triplets.

Acknowledgements

This work was supported by the Consejo de Investigaciones Científicas y Técnicas (CONICET) (IES and DUF are CONICET researchers and EAG is a postdoctoral fellow); CONICET Grant PIP2022-2024—11220210100704CO and Universidad de Buenos Aires grant UBACyT 20020220200106BA. We call the attention of the international scientific community about the catastrophic erosion of Argentina's strong scientific tradition due to current funding constraints and the sudden termination of long term policies.

References

1. D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration in biomolecules. *Quart. Rev. Biophys.* **47**, 285–363 (2014).
2. P. G. Wolynes, Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218–230 (2015).
3. F. H. Crick, The origin of the genetic code. *J Mol Biol* **38**, 367–379 (1968).
4. E. V. Koonin, A. S. Novozhilov, Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet* **51**, 45–62 (2017).
5. I. E. Sánchez, E. A. Galpern, M. M. Garibaldi, D. U. Ferreiro, Molecular Information Theory Meets Protein Folding. *J. Phys. Chem. B* **126**, 8655–8668 (2022).
6. T. D. Schneider, A brief review of molecular information theory. *Nano Communication Networks* **1**, 173–180 (2010).
7. G. Magi Meconi, I. Sasselli, V. Bianco, J. Onuchic, I. Coluzza, Key aspects of the past 30 Years of protein design. *Rep. Prog. Phys.* (2022).
8. I. E. Sánchez, E. A. Galpern, D. U. Ferreiro, Solvent constraints for biopolymer folding and evolution in extraterrestrial environments. *Proc Natl Acad Sci U S A* **121**, e2318905121 (2024).
9. C. H. Mak, L. L. Sani, A. N. Villa, Residual Conformational Entropies on the Sugar–Phosphate Backbone of Nucleic Acids: An Analysis of the Nucleosome Core DNA and the Ribosome. *J. Phys. Chem. B* **119**, 10434–10447 (2015).
10. C. H. Mak, T. Matossian, W.-Y. Chung, Conformational Entropy of the RNA Phosphate Backbone and Its Contribution to the Folding Free Energy. *Biophysical Journal* **106**, 1497–1507 (2014).
11. J. Stombaugh, C. L. Zirbel, E. Westhof, N. B. Leontis, Frequency and isostericity of RNA base pairs. *Nucleic Acids Research* **37**, 2294–2312 (2009).
12. N. Ontiveros-Palacios, *et al.*, Rfam 15: RNA families database in 2025. *Nucleic Acids Research* gkae1023 (2024). https://doi.org/10.1093/nar/gkae1023.
13. B. Burghardt, A. K. Hartmann, RNA secondary structure design. *Phys. Rev. E* **75**, 021920 (2007).
14. R. M. Hazen, P. L. Griffin, J. M. Carothers, J. W. Szostak, Functional information and the emergence of biocomplexity. *Proc Natl Acad Sci U S A* **104 Suppl 1**, 8574–8581 (2007).
15. P. G. Wolynes, As simple as can be? *Nat Struct Mol Biol* **4**, 871–874 (1997).

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.