Adaptive Properties of the Amino Acid Alphabet and its Subsets

Rudrarup Bose[1], Markus Meringer[2], Melissa Ilardo[3,4] and H. James Cleaves[5,6,7,8]

[1]National Institute of Science Education and Research, Bhubaneshwar, India
[2]German Aerospace Center, Oberpfaffenhofen, Germany
[3]University of Copenhagen, Copenhagen, Denmark, [4]University of Utah, Salt Lake City, UT 84112
[5]Tokyo Institute of Technology, Japan, [6]Institute for Advances Study, Princeton, NJ 08540
[7]Blue Marble Space Center of Science, Seattle, WA 98154, [8]Georgia Institute of Technology, Atlanta, GA 30332
rudrarup.bose@niser.ac.in

Abstract

The standard alphabet of the 20 genetically encoded amino acids is considered to have been selected during early evolution from a larger pool of α–amino acids based on its coverage of the chemical space. Chemical space is here defined by charge, size and hydrophobicity, leading to 6–tuples representing coverage, which is composed of range and evenness in these three physico–chemical properties. We summarize findings of previous studies on the adaptive properties of the 20 encoded amino acids and show how we extend these computational experiments to subsets of the standard alphabet.

Introduction

The modern genetically encoded alphabet is believed to be highly optimized, by means of a stepwise growth of earlier simple alphabets, for a number of features including codon mapping and coverage of chemical space (Philip and Freeland, 2011; Ilardo et al., 2015; Freeland and Hurst, 1998), as opposed to a random expansion (Wong, 2005). The origin of the genetic code has been of interest to both origin of life and artificial life community (Froese et al., 2018). We explored here the optimality of smaller alphabets based on maximum coverage of chemical space.

Chemical structure space is defined as the set of compounds, hypothetical or actual, which fulfil a given set of property criteria, such as molecular formula, chemical property or chemical substructure (Eberhardt et al., 2011).

Non–randomness of the Standard Alphabet

Philip and Freeland (2011) hypothesised that natural selection, instead of random incorporation, would have favoured a set of amino acids that is better covering chemical space in terms of charge, size and hydrophobicity. They proved their hypothesis with a computer experiment based on 76 known amino acids of abiotic and biosynthetic origin. The coordinates in chemical space were computed as pK_a, van der Waals volume V_{vdw} and partition coefficient $logP$. They randomly sampled sets of $n = 20$ amino acids from the background set of 76, and calculated their coverage of chemical space in terms of range and evenness as follows:

Let $p_1 \leq p_2 \leq \ldots \leq p_n$ be the sorted values of a property $P \in \{pK_a,\ V_{vdw},\ logP\}$ for a set A of n amino acids. Then the range of A w.r.t. P is defined as difference of maximum and minimum value

$$\varrho(A, P) = p_n - p_1, \tag{1}$$

and the evenness of A w.r.t. P is computed as variance of differences of successive property values

$$\varepsilon(A, P) = \text{Var}\,\{p_i - p_{i-1} : 1 < i \leq n\}. \tag{2}$$

Better coverage means higher range and lower evenness. One significant result of Philip and Freeland (2011) was that none of the sampled sets outperformed the coded set in terms of coverage in all three properties. We refer to the method described above as *adaptive analysis*.

Extraordinarily Adaptive Properties

In order to conduct even more rigorous testing of the hypothesis of Philip and Freeland (2011) virtual libraries of α–amino acids were prepared using *in silico* molecular structure generation (Meringer et al., 2013; Meringer and Cleaves, 2017). From these libraries 1913 xeno amino acids were chosen as extended background set for adaptive analysis (Ilardo et al., 2015). 10^8 random sets of size 20 were sampled and compared to the coded set in terms of coverage of chemical space as described above. It turned out that better sets do exist, but they are extremely rare.

Subsets of the Standard Alphabet

A frequently asked question on the studies summarized above was whether such adaptive properties can also be found among subsets of the standard alphabet. In the following we describe how we extended adaptive analysis to handle not only one reference set, but the $\binom{20}{n}$ subsets of size n, short called n–subsets.

Using equations (1) and (2) we define the coverage of a set A of amino acids as

$$\begin{aligned}\kappa(A) \quad = \quad & (\varrho(A, pK_a), \varrho(A, V_{vdw}), \varrho(A, logP), \\ & -\varepsilon(A, pK_a), -\varepsilon(A, V_{vdw}), -\varepsilon(A, logP)),\end{aligned}$$

the 6–tupel composed of range and negative evenness values in the three considered properties, charge, size and hydrophobicity. Using this formalism, we can say that a set A is better than a set B iff $\kappa(A) > \kappa(B)$, where the 'greater' relationship for two 6–tuples $a = (a_1, \ldots, a_6)$ and $b = (b_1, \ldots, b_6)$ is defined in the following natural way:

$$a \geq b \quad :\Longleftrightarrow \quad a_1 \geq b_1 \wedge \ldots \wedge a_6 \geq b_6, \tag{3}$$

$$a > b \quad :\Longleftrightarrow \quad a \geq b \wedge a \neq b. \tag{4}$$

Let C denote the standard alphabet of the 20 coded amino acids, and X the set of 1913 xeno amino acids. Our basic approach to find subsets of X that have better coverage than subsets of C is straightforward.

(i) run through all set sizes $n = 3, \ldots, 19$

(ii) for each n sample 10^8 random sets A of size n from X

(iii) for each A run through all n–subsets B of C and check if $\kappa(A) \geq \kappa(B)$. If any such B is found then output A.

However, we want to improve step (iii) to avoid relationships $\kappa(B) \leq \kappa(A) \leq \kappa(B')$ for any other n–subsets B' of C. For this purpose we take advantage of the partial order introduced in definition (3). Instead of running through all n–subsets $B \subset C$ we can simplify this step by checking against n–subsets of maximum coverage in C. A n–subset $M \subset C$ has maximum coverage in C iff there is no other n–subset $B \subset C$ with $\kappa(B) > \kappa(M)$. Note that partially ordered sets may have more than one maximum. Taking this into account, step (iii) can be replaced by the more efficient

(iii') for each A run through all n–subsets M of maximum coverage and check if $\kappa(A) \geq \kappa(M)$. If any such M is found then output A.

Our first computations have shown that the number of n–subsets of C with maximum coverage is much smaller than $\binom{20}{n}$, see Figure 1. This way computation time can be reduced to a few days. Independent implementations in Matlab and Python show consistent results pointing to strong adaptive properties of subsets of the amino acid alphabet. Details are currently being prepared for publication (Ilardo et al.).

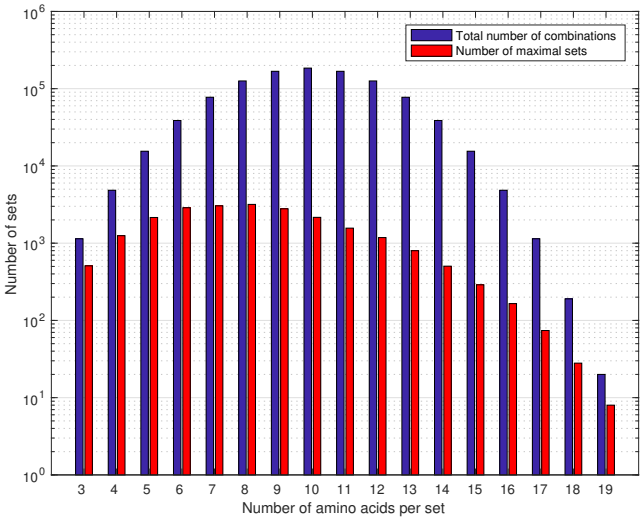

Figure 1: The number of subsets with maximum coverage (red) is for every considered set size n smaller than the number of n–subsets (blue). In total this results in a factor of about 50 and renders computations feasible in realistic time.

Acknowledgements

The authors acknowledge the contributions of Professors Stephen Freeland and Bakhtiyor Rasulev and Dr Natalie Grefenstette to the work described here. The authors would like to thank the Earth–Life Science Institute (ELSI) for support during the preparation of this work, and EON for generous travel support for MI, MM and RB. This work was also supported by JSPS KAKENHI Grant–in–Aid for Scientific Research on Innovative Areas "Hadean Bioscience", Grant Number JP26106003, and by the ELSI Origins Network (EON), which is supported by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation.

References

Eberhardt, L., Kumar, K., and Waldmann, H. (2011). Exploring and exploiting biologically relevant chemical space. *Current drug targets*, 12(11):1531–1546.

Freeland, S. J. and Hurst, L. D. (1998). The genetic code is one in a million. *Journal of molecular evolution*, 47(3):238–248.

Froese, T., Campos, J. I., Fujishima, K., Kiga, D., and Virgo, N. (2018). Horizontal transfer of code fragments between protocells can explain the origins of the genetic code without vertical descent. *Scientific reports*, 8(1):3532.

Ilardo, M., Meringer, M., Freeland, S., Rasulev, B., and Cleaves II, H. J. (2015). Extraordinarily adaptive properties of the genetically encoded amino acids. *Scientific reports*, 5:9414.

Ilardo, M., Rudrarup, B., Meringer, M., Grefenstette, N., Freeland, S., Rasulev, B., and Cleaves, H. J. (in preparation). Chemoinformatic exploration of the expansion of the genetically encoded set of amino acids.

Meringer, M. and Cleaves, H. J. (2017). Exploring astrobiology using in silico molecular structure generation. *Phil. Trans. R. Soc. A*, 375(2109):20160344.

Meringer, M., Cleaves, H. J., and Freeland, S. J. (2013). Beyond terrestrial biology: Charting the chemical universe of α-amino acid structures. *Journal of chemical information and modeling*, 53(11):2851–2862.

Philip, G. K. and Freeland, S. J. (2011). Did evolution select a nonrandom alphabet of amino acids? *Astrobiology*, 11(3):235–240.

Wong, J. (2005). Coevolution theory of the genetic code at age thirty. *BioEssays*, 27(4):416–425.