Exact correspondence between walks in nucleotide and protein sequence spaces

Dmitry N. Ivankov

Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, 08003 Barcelona, Spain; Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya str., Pushchino, Moscow region, 142290, Russia; ivankov13@gmail.com

In the course of evolution, genes traverse the nucleotide sequence space, which translates to a trajectory of changes in the protein sequence in protein sequence space. The correspondence between regions of the nucleotide and protein sequence spaces is understood in general but not in detail. One of the unexplored questions is how many sequences a protein can reach with a certain number of nucleotide substitutions in its gene sequence. Here I present an algorithm to calculate the volume of protein sequence space accessible to a given protein sequence as a function of the number of nucleotide substitutions made in the protein-coding sequence (1). The algorithm utilizes the power of the dynamic programming approach (2, 3), and makes all calculations within a couple of seconds on a desktop computer.

The algorithm is described as follows. The number of amino acid sequences $N_{i,j}$, comprising the increment of protein sequence space for i nucleotide substitutions introduced in the sequence of j codons can be calculated recursively:

$$N_{i,j} = \sum_{k=0}^{\min\{3,i\}} N_{i-k,j-1} \cdot M_{k,codon_j}. \qquad (1)$$

where values $M_{k,codon_j}$ are the numbers of the newly accessible amino acid variants for $codon_j$ by k (k = 0, 1, 2, 3) nucleotide substitutions. Values $M_{k,codon_j}$ are calculated directly from the genetic code table. Clearly, $N_{0,0} = 1$ because only one sequence exists of zero length with zero substitutions, the empty sequence. Now, it is straightforward to go from the incremental characteristics to the integral one. That is, the volume of the protein sequence space $L_{p,j}$ accessible for a protein sequence after introducing a given number of nucleotide substitutions p, equals to:

$$L_{p,j} = \sum_{i=0}^{p} N_{i,j}. \qquad (2)$$

It sums up all increment volumes obtained by up to *p* nucleotide substitutions.

An approximate method can be used as an alternative to the exact approach presented here. To check the accuracy of the approximate method I applied both methods to nucleotide sequences coding different proteins. Taking into account the astronomically huge size of the protein sequence space, the approximate solution gives estimate which can be considered as acceptable as an order of magnitude estimation.

Overall, the presented algorithm is fast and calculates exactly the volume of protein sequence space accessible to a given protein sequence as a function of the number of nucleotide substitutions made in the protein-coding sequence. It can have practical applications in the study of evolutionary trajectories in sequence space.

I thank Mikhail Roytberg, Alexei Finkelstein, Guillaume Filion, Natalya Bogatyreva, Fyodor Kondrashov, Dinara Usmanova, and Karen Sarkisyan for discussions.

The study was in part funded by HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2015-68723-P), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME).

1. D.N.Ivankov (2017) Exact correspondence between walk in nucleotide and protein sequence spaces. *PLoS ONE*, under review.

2. T.H.Cormen, C.E.Leiserson, R.L.Rivest, C.Stein (2009) Introduction to algorithms.

3. A.V.Finkelstein, M.A.Roytberg (1993) Computation of biopolymers: a general approach to different problems. *Biosystems* **30:**1–19.