About the Protein Space Vastness

Jorge A. Vila[1]

Abstract

An accurate estimation of the Protein Space size, in light of the factors that govern it, is a long-standing problem and of paramount importance in evolutionary biology, since it determines the nature of protein evolvability. A simple analysis will enable us to, firstly, reduce an unrealistic Protein Space size of $\sim 10^{130}$ sequences, for a 100-residues polypeptide chain, to $\sim 10^9$ functional proteins and, secondly, estimate a robust average-mutation rate per amino acid ($\xi \sim 1.23$) and infer from it, in light of the protein marginal stability, that *only* a fraction of the sequence will be available at any one time for a functional protein to evolve. Although this result does not solve the Protein Space vastness problem frames it in a more rational one and illustrates the impact of the marginal stability on protein evolvability.

Keywords Protein Evolution · Protein Marginal Stability · Anfinsen · Protein Space · Maynard Smith

At first glance, the universe of possible sequences populating the Protein Space [1], for a 100 residues polypeptide chain, should contain 20^{100} or $\sim 10^{130}$ sequences [2–5]; where 20, the number of naturally occurring amino acids, represent the mutation rate per amino acid (ξ). In light of protein evolvability, a Protein Space (PS) of such size is absurdly huge. This problem is analog to the "Levinthal paradox" [6–8]. According to this paradox, exploring the whole conformational space, in search of the native-state of a 100 residues polypeptide chain would require $\sim 10^{27}$ years [7], *i.e.*, a time larger than the age of the universe, according to the Big Bang theory; yet proteins can fold in solution within microseconds to seconds. What do the Levinthal paradox and the Protein Space problem have in common? Certainly, it is not the size nor the time needed to explore it but our ignorance of the main forces that govern it. As far this work is concerned, the protein marginal stability is one of the forces limiting the size of the PS and, although it as being studied profusely in the last 50 years [4, 9–15] many questions remain to be answered. Among others an apparently simple one, which is a reasonable number of functional sequences in the PS? Hence, we are aimed to estimate, for an initially functional protein, all possible functional sequences in the

Protein Space in light of the proteins' marginal stability. It is worth noting that the navigation routes over the Protein Space [4, 13, 16] and, hereupon, the epistasis effects on it [17–19] or estimate which Protein Space fraction could have been explored since life began on Earth [3] will not be addressed here.

One of the central problems in the study of the evolvability of proteins is the influence of the 'protein stability' on the Protein Space (PS) size. As a first step in the analysis, a definition of 'protein stability' must be adopted because there are two ways to measure it. Specifically, by determining either the "denaturation" free-energy, ΔG^D [20] or the protein "marginal" stability [21–24]. The latter but not the former definition is a more accurate way to study the PS size, and the reason for this assertion follows. In the Protein Space model "…two sequences are neighbors if one can be converted into another by a single amino-acid substitution …" [1]. This simple requirement enables us to use the $\Delta\Delta G^D$ free-energy changes between the wild-type and the mutated protein as a tool to estimate the feasibility of a single amino-acid substitution. These $\Delta\Delta G^D$ values reflect, primarily, changes in the native-state rather than in the unfolded-states [12]. Because of this, we consider that the 'marginal stability' is, in light of the PS model, a more accurate way to characterize the 'protein stability'. In this regard, we have been able to demonstrate, based on simple concepts from statistical thermodynamics, the Gershgorin

✉ Jorge A. Vila
  jorgevila84@gmail.com

1 IMASL-CONICET, Universidad Nacional de San Luis,
   Ejército de Los Andes 950, 5700 San Luis, Argentina

theorem and a heuristic argument, the existence of an upper limit to the protein marginal stability [23].

In his seminal 1970 article, John Maynard Smith [1] wrote: "…if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates…" Implicit in this proposal is that any functional protein that pertains to the Protein Space must fulfill Anfinsen's dogma [25]; consequently, prion and IDP proteins will be excluded from our analysis. For the practical purpose, we depict the PS as a network of strings containing all functional proteins that nature devised. Each string contains (m) sequences, with a fixed number of residues (n), conforming a protein ensemble that fulfill Anfinsen's dogma and where each one of them can be converted into another by a single amino acid substitution. Nearest neighbor strings contain ($n+1$; $n-1$) residues. The string with the minima number of residues able to form a stable protein structure is assumed to contain $n = 16$ residues, *e.g.*, as the β-harping substructure of the immunoglobin binding domain of streptococcal protein *G*, because it forms a stable native state in the absence of the rest of the protein [26]. It is worth noting that metamorphic proteins, as protein *G* [27], may appear in any string. However, this is not an obstacle for the analysis because this kind of proteins fulfill Anfinsen's dogma [28].

After clarifying some key definitions, the following question arises: given a string containing an arbitrary number of residues (n), which is the size of the corresponding Protein Space? Let us analyze it. As a simple and factual assumption, we envision that each amino acid in the sequence may be substituted by the one possessing the highest substitution rate in the Point Accepted Mutation (PAM1) matrix [29, 30]. The adoption of this simple approach ($\xi = 2.0$) will enable us to predict quickly an order of magnitude for the PS size. Indeed, for a sequence of $n = 100$ residue, there will be $m = 2^{100}$ or ~10^{30} different sequences. However, this size for the Protein Space can be predicted by an even simplest approach. In fact, the modeling of a protein as a collection of hydrophobic or polar (HP) beads [31, 32] also enables to predict a PS size of ~10^{30}. But, there are two observations to note about the accuracy of this prediction. Firstly, the mutation rate under the HP modeling is unsound in light of molecular evolution. Secondly, the protein representation under the HP model foresees a challenge to Anfinsen's dogma, *e.g.*, has been shown that a two-letter code is insufficient to give an energy landscape like that of a wild type protein [33]. Let us briefly discuss the relevance of each of these objections. Under the HP modeling a single-point mutation is simply a replacement of a hydrophobic (H) residue by a polar (P) one, and vice versa ($\xi = 2.0$). This approach ignores key mutations seen in nature, *e.g.*, mutation between polar residues in hemoglobin [34]. Moreover, whether or not the

protein HP model challenges the Anfinsen dogma it is not a minor problem. On the contrary, this is a serious problem because the marginal stability of the proteins is one of the most important factors restricting their evolvability and its existence is a consequence of Anfinsen's dogma [23]. In this regard, should be noted that the protein marginal stability is a universal property, that includes biomolecules and macromolecules complex [24], and their origin can be thoroughly understood from a purely physical point of view [23] or, alternatively, from specific evolutionary arguments [35].

Whatever the above-adopted model to estimate the Protein Space (PS) size is the impact of the protein marginal stability implies that its actual size shall be significantly smaller than foretold (~10^{30}) because most of the single-site mutations in real proteins are destabilizing [12, 13]. Moreover, it is well documented, from evolutionary changes observed in cytochromes *c* of various species, that many sites are invariant to mutations [36]. In other words, substitutions at specific sites in the protein sequence may not ever occur and the reason is the following. Proteins for which single-site substitutions lead to a free-energy change ($\Delta\Delta G^D$) larger than the marginal stability upper bound limit (~7.4 kcal/mol) will never fold or function [24]. This destabilization threshold value agrees with observations made on the green fluorescent protein from Aequorea victoria (avGFP), viz., "…The average fluorescence of single mutants of avGFP as a function of the predicted protein destabilization, $\Delta\Delta G$, reveals a threshold around 7–9 kcal mol^{-1} …" [37]. In particular, is worth noting that a sigmoid-like fitness function obtained by an independent neural-network analysis predicts a ~60% drop on the log-fluorescence if $\Delta\Delta G > $ ~7.5 kcal/mol (see Fig. 4 of Sarkisyan et al. [37]). Putting all together, the prediction of a reasonable PS size requires a model reappraisal in light of the proteins' marginal stability.

Evidences have been presented showing that "…more stable proteins are more evolvable because they are better able to tolerate functionally beneficial but destabilizing mutations…" [11]. Our concern about this proposal comes from the fact that the upper bound free-energy gap between the native-folds and the misfolded ones is very small; to be specific, equivalent to break, at most, ~5 hydrogen-bonds! [24]. In other words, protein evolvability need to overcome this drawback by using a very efficient mechanism. A possible solution to this problem has already been anticipated by Kimura [38] and highlight by Wagner [9] and Bloom et al. [11] by suggesting that neutral mutations may play a critical role in the transition from one structure to another in the Protein Space, *e.g.*, by counterbalance the effects of destabilizing mutations although beneficial from the functional point of view. From this perspective and considering that each amino acid is coded by a nucleotide triplet, an *upper bound limit* for the PS size can be estimated. Let do it by assuming, firstly, $n = 100$ as a trial length, although the analysis would

be valid for sequences of any length and, secondly, that: (*i*) each nucleotide pair replacement entails an amino acid substitution; (*ii*) each nucleotide pair replacement occurs, after removing synonymous mutations, every ~ 2 years [38]; and (*iii*) almost all mutations will be neutral [38] or nearly neutral [39]. If the starting point is a functional protein, adoption of these rules will assure that the Protein Space will be a "…continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates…" [1]. Consequently, if life began on earth around a billion years ago [40] then the Protein Space should contain $\sim 10^9$ functional sequences. Therefore, the average mutation rate per amino acid will be $\xi \sim 1.23$. Before we continue, let us analyze some nucleotide replacement alternatives to the one proposed above, in section (*ii*). For example, we could have considered one nucleotide pair replacement per day or $\sim 10^{14}$ per second rather than one every ~ 2 years [38]. These alternatives will lead to Protein Space sizes containing $\sim 10^{11}$ ($\xi \sim 1.30$) or $\sim 10^{30}$ ($\xi \sim 2.0$) functional sequences, respectively. These Protein Space sizes seem to be reasonable although the nucleotide pair replacement rates are not. This simple example illustrates that a reliable estimation of the PS size, in light of molecular evolution, is not just a combinatorial problem. Indeed, time is an essential variable to find out a reasonable answer, *i.e.*, as in the search of solutions for the Levinthal paradox [7]. So, an average mutation rate per amino acid of $\xi \sim 1.23$ is equivalent to think that *only* a fraction of the protein sequence is variable at any one time, *e.g.*, 30% with $\xi = 2.0$; 13% with $\xi = 4.0$; and 5% with $\xi = 8.0$, while mutations on the remaining of the sequence would lead to conformations that do not fold; in other words, will be nonfunctional. That each functional protein in the Protein Space, independent on the fold-class or sequence, could tolerate only a fraction of their sequence to be variable at any one time, is consistent with the conjecture that protein marginal stability's is the main force that confines the Protein Space (PS) size; basically, because the marginal stability sets up a physical limit to the amount and type of mutations that a protein can admit while still folding into the native structure [24]. Hence, it is not surprising to read that $\sim 25\%$ of the single-site mutations on the green fluorescent protein (from avGFP) have not deleterious effects on fluorescence [37], or that many sites are invariant to mutations in cytochrome *c* [36].

So far, we have illustrated how to estimate, in light of the protein marginal stability, the PS size of a 100-residues functional protein belonging to a given specie. This PS size ($\sim 10^9$) actually is a small sub-space of the evolutionary Proteins Space, that include *all* functional proteins, and whose size could be estimated based on the existence of $\sim 10^{10}$ unique protein sequences [41] belonging to $\sim 10^7$ species on Earth [42]. The resulting evolutionary Protein Space size would be several orders of magnitude larger than $\sim 10^9$, although a tiny fraction of the colossal protein sequence space.

There are at least two hidden assumptions in this proposal, just to mention a few. Firstly, the estimated PS size ($\sim 10^9$) represents *all* functional forms compatible with one functional protein that existed since life started on Earth. This does not necessarily mean that all the functional proteins come from a single starting point. Indeed, as noted by John Maynard Smith 50 years ago [1], is possible that "…there are two or more distinct networks, or that there is one network with multiple starting points…" Secondly, the use of the beginning of the life on Earth as a reference point does not mean, nor imply either, that evolution was able to explore a fraction, or the whole, evolutionary Protein Space or how evolution occurred during such a period of time, *i.e.*, either by the natural (Darwinian) selection of favorable mutations or by random fixation of neutral or nearly neutral mutants.

Overall, the existence of intermedia steps during the molecular evolution, like those brought by neutral or nearly neutral mutations, enable us, firstly, to conjecture a conceivable Protein Space size of $\sim 10^9$ for a starting 100-residues functional protein and, secondly to estimate a robust average-mutation rate per amino acid ($\xi \sim 1.23$), *i.e.*, independent of the protein fold-class, or sequence, and infer, in light of the protein marginal stability, that *only* a fraction of the sequence will be available at any one time for a functional protein evolve.

Note: I am honored to dedicate this Letter to the memory of Harold A. Scheraga, Todd Professor of Chemistry, Emeritus at Cornell University. He achieved leadership in the world of science, and high respect among colleagues [43], as a result of his colossal experience in Experimental and Theoretical Chemistry, Physics and Mathematics and, in particular, due to his tireless effort in the search of a possible solution to the, yet unsolved, Protein Folding problem. Harold Scheraga passed away at the age of 98 in Ithaca, NY, on August 1, 2020.

References

1. Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225:563–564
2. Mandecki W (1998) The game of chess and searches in protein sequence space. Trends Biotechnol 16:200–202

3. Dryden DTF, Thomson AR, White JH (2008) How much of protein sequence space has been explored by life on Earth? J R Soc Interface 5:953–956

4. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10:866–876

5. Ivankov DN (2017) Exact correspondence between walk in nucleotide and protein sequence spaces. PLoS ONE 12(8):e0182525

6. Levinthal C (1968) Are there pathways for protein folding? J de Chim Phys 65(1):44–45

7. Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. Proc Natl Acad Sci USA 89:20–22

8. Finkelstein AV, Garbuzynskiy SO (2013) Levinthal's question answered…again? J Biomol Struct Dyn 31(9):1013–1015

9. Wagner A (2005) Robustness, evolvability, and neutrality. FEBS Lett 579:1772–1778

10. DePristo M, Weinreich D, Hartl D (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6:678–687

11. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci USA 103:5869–5874

12. Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc Natl Acad Sci USA 104:16152–16157

13. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 19:596–604

14. Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoS Comput Biol 4(2):e1000002

15. Kurahashi R, Sano S, Takano K (2018) Protein evolution is potentially governed by protein stability: directed evolution of an esterase from the hyperthermophilic archaeon sulfolobus tokodaii. J Mol Evol 86(5):283–292. https://doi.org/10.1007/s00239-018-9843-y

16. Otwinowski J (2018) Biophysical Inference of epistasis and the effects of mutations on protein stability and function. Mol Biol Evol 35(10):2345–2354

17. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. Nature 490:535–538

18. Starr TN, Thornton JW (2016) Epistasis in protein evolution. Protein Sci 25:1204–1218

19. Miton CM, Chen JZ, Ost K, Anderson DW, Tokuriki N (2020) Statistical analysis of mutational epistasis to reveal intramolecular interaction networks in proteins. Methods Enzymol 643:243–280. https://doi.org/10.1016/bs.mie.2020.07.012

20. Koehl P, Levitt M (2002) Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 99(3):1280–1285

21. Privalov PL, Tsalkova TN (1979) Micro- and macro-stabilities of globular proteins. Nature 280:694–696

22. Hormoz S (2013) Amino acid composition of proteins reduces deleterious impact of mutations. Sci Rep 3:1–10

23. Vila JA (2019) Forecasting the upper bound free energy difference between protein native-like structures. Phys A 533:122053

24. Martin OA, Vila JA (2020) The marginal stability of proteins: how the jiggling and wiggling of atoms is connected to neutral evolution. J Mol Evol 88:424–426

25. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–323

26. Cecchini M, Krivov SV, Spichty M, Karplus M (2009) Calculation of free-energy differences by confinement simulations. Application to peptide conformers. J Phys Chem B 113:9728–9740

27. Spichty M, Cecchini M, Karplus M (2010) Conformational free-energy difference of a miniprotein from nonequilibrium simulations. J Phys Chem Lett 1(13):1922–1926

28. Vila JA (2020a) Metamorphic proteins in light of Anfinsen's Dogma. J Phys Chem Lett 11(13):4998–4999

29. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

30. Gillespie JH (1994) The causes of molecular evolution. Oxford University Press, Incorporated, Oxford

31. Lipman DJ, Wilbur WJ (1991) Modelling neutral and selective evolution of protein folding. Proc Royal Soc Lond B 245:7–11

32. Bornberg-Bauer E (1997) How are model protein structures distributed in sequence space? Biophys J 73(5):2393–2403

33. Wolynes P (1997) As simple as can be? Nat Struct Mol Biol 4:871–874

34. Perutz MF (1983) Species adaptation in a protein molecule. Mol Biol Evol 1(1):1–28

35. Wilson AE, Kosater WM, Liberles DA (2020) Evolutionary processes and biophysical mechanisms: revisiting why evolved proteins are marginally stable. J Mol Evol 88:415–417

36. Margoliash E, Smith EL (1965) Structural and functional aspects of cytochrome *c* in relation to evolution. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins: a symposium. Academic Press, New York, London, pp 221–242

37. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV et al (2016) Local fitness landscape of the green fluorescent protein. Nature 533:397–401

38. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

39. Otha T (2006) Slightly deleterious mutant substitutions in evolution. Nature 246:96–97

40. Schopf JW (2006) The first billion years: When did life emerge? Elements 2:229–233

41. Koonin E, Wolf Y, Karev G (2002) The structure of the protein universe and genome evolution. Nature 420:218–223. https://doi.org/10.1038/nature01256

42. Sweetlove L (2011) Number of species on Earth tagged at 8.7 million. Nature. https://doi.org/10.1038/news.2011.498

43. Vila JA (2020b) Harold A. *Scheraga Legatum*. Protein J. https://doi.org/10.1007/s10930-020-09917-w