Downloaded from http://rsif.royalsocietypublishing.org/ on June 22, 2017

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2008) **5**, 953–956
doi:10.1098/rsif.2008.0085
Published online 15 April 2008

REPORT

How much of protein sequence space has been explored by life on Earth?

David T. F. Dryden*, Andrew R. Thomson and John H. White

School of Chemistry, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JJ, UK

We suggest that the vastness of protein sequence space is actually completely explorable during the populating of the Earth by life by considering upper and lower limits for the number of organisms, genome size, mutation rate and the number of functionally distinct classes of amino acids. We conclude that rather than life having explored only an infinitesimally small part of sequence space in the last 4 Gyr, it is instead quite plausible for all of functional protein sequence space to have been explored and that furthermore, at the molecular level, there is no role for contingency.

Keywords: protein sequence; evolution; contingency

1. INTRODUCTION

Two assumptions are generally made when considering the molecular evolution of functional proteins during the history of life on Earth. Firstly, the size of protein sequence space, i.e. the number of possible amino acid sequences, is astronomically large and, secondly, that only an infinitesimally small portion has been explored during the course of life on Earth (e.g. Salisbury 1969; Maynard Smith 1970; Mandecki 1998; Luisi 2003; Carrier 2004; de Duve 2005). Luisi and Chiarabelli have termed the unexplored part of sequence space as containing the 'never born proteins' (Luisi *et al.* 2006; Chiarabelli & de Lucrezia 2007). We wish to discuss these two assumptions by estimating how much of this space could have been explored since the origin of life some 4 Gyr ago. As will be described below, others have concluded that the first assumption is incorrect and we agree with this conclusion. However, we also conclude that the second assumption is incorrect and calculate that most of the sequence space may have been explored.

Before turning to a discussion of the second assumption, we wish to summarize information showing the first assumption, namely that the sequence space is vast, to be false. A typical estimate of the size of sequence space is 20^{100} (approx. 10^{130}) for a protein of 100 amino acids in which any of the normally occurring 20 amino acids can be found. This number is indeed gigantic but it is likely to be a significant overestimate of the size of protein sequence space. For example, Dill and colleagues used simple theoretical models to suggest (Lau & Dill 1990; Chan & Dill 1991; Dill 1999), and experimental or computational variation of protein sequence provides ample evidence (Cordes *et al.* 1996; Riddle *et al.* 1997; Plaxco *et al.* 1998; Larson *et al.* 2002; Guo *et al.* 2004; Doi *et al.* 2005), that the actual identity of most of the amino acids in a protein is irrelevant. An example in nature could be the prokaryotic DNA methyltransferases which each contain a target recognition domain (TRD) of approximately 150 amino acids that recognizes specific DNA sequences usually of 3–6 bp in length, and a conserved catalytic domain. The thousands of known TRD sequences show negligible amino acid sequence conservation despite the rather limited number of nucleotide sequences they are required to recognize (e.g. Sturrock & Dryden 1997; O'Neill *et al.* 1998; Bujnicki 2001; Roberts *et al.* 2007). As an extreme method to reduce the size of sequence space, Dill (1999) suggested that only two types of amino acid were needed to form a protein structure, hydrophilic and hydrophobic, and that furthermore it was critical to define only the surface of the protein. These two suggestions reduce the size of sequence space to 2^{100} and 2^{33}, respectively (i.e. approx. 10^{30} and approx. 10^{10}). It is noteworthy that recent coarse-grained 'tube' models go even further and remove all atomic information leaving only a potential energy function for interaction with other parts of the tube. Despite the extreme coarse graining of this model, recognizable 'protein' structures can still be found (Banavar *et al.* 2006). Although this may appear to go against Anfinsen's dogma that a protein structure is determined by its amino acid sequence (Anfinsen 1973), it is really only a case of an extreme reduction in the size of the amino acid 'alphabet'. The tube structures obtained are rather similar to the short folded segments adopted by sequences apparently conserved since the last universal ancestor (Sobolevsky & Trifonov 2006). The assumption that a protein chain needs to be at least 100 amino acids in length also rather inflates the size of sequence space when it is known that many proteins are modular and contain domains of as few as approximately 50 amino acids thereby reducing the space to 20^{50} or approximately 10^{65} (e.g. Sobolevsky & Trifonov 2006). The conclusion from all of these coarse-graining approaches is that a reduced alphabet of amino acids is quite capable of producing all protein folds (approx. a few thousand discrete folds; Denton 2008) and providing a scaffold capable of supporting all protein functions (we will ignore the space of natively unfolded proteins for this current discussion but since such proteins usually fold upon performing their function, the distinction is not important for our purposes; Dyson & Wright 2005). The phase space of function may be some orders of magnitude greater than the size of the folding space as metagenomics projects are revealing increasing numbers of unknown protein families as adjudged by the number of novel

*Author for correspondence (david.dryden@ed.ac.uk).

Received 27 February 2008
Accepted 25 March 2008

This journal is © 2008 The Royal Society

Downloaded from http://rsif.royalsocietypublishing.org/ on June 22, 2017

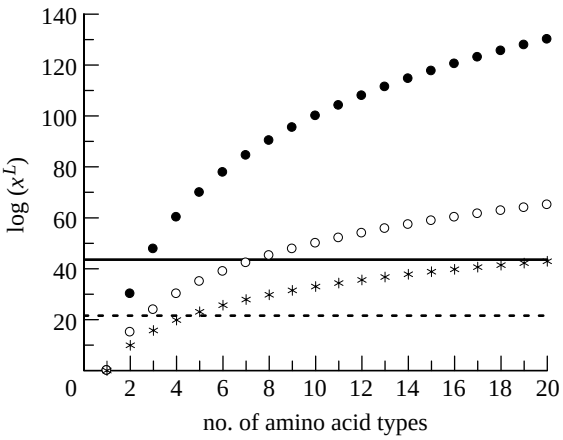

Figure 1. The size of protein sequence space $\log(x^L)$ as a function of the size of the amino acid alphabet (i.e. the number of different types of amino acids, x) for proteins containing 33 (asterisks), 50 (open circles) or 100 (filled circles) amino acids (length L). The horizontal lines represent estimates of the maximum (solid line) and minimum (dashed line) number of sequences explored during the 4 Gyr since the origin of life on Earth.

protein sequences (Raes *et al.* 2007). However, it is not clear that new folds are present as a conserved fold, such as the TIM barrel, is capable of displaying many functions (Nagano *et al.* 2002).

To further support this idea of a reduced alphabet of amino acids, there are also very plausible suggestions that the original amino acid repertoire consisted of only four or five amino acids like those found in the Miller–Urey experiments and the Murchison meteorite (Miller *et al.* 1976), and that the genetic code was initially limited to these few amino acids that still predominate in proteins to the current day (e.g. Trifonov 2000; Brooks *et al.* 2002; Ikehara 2002). Proteins with reduced amino acid repertoires can fold and function successfully (e.g. Cordes *et al.* 1996; Riddle *et al.* 1997; Plaxco *et al.* 1998; Guo *et al.* 2004; Doi *et al.* 2005; López de la Osa *et al.* 2007).

Figure 1 shows the number of possible sequences as a function of the number of different amino acids (or classes of amino acids, 1–20) and the length of the functionally important amino acid chain (33, 50 or 100). It highlights the drastic reduction in the size of sequence space if one limits the number of available amino acid types to less than the 20 usually found today, a limitation that appears to be justified experimentally.

2. RESULTS

We now wish to consider the second assumption commonly made about protein sequence space: that only an infinitesimal fraction has been explored by life on Earth. To examine how much of sequence space could have been explored, it is simplest to make upper and lower limit estimates for the number of unique amino acid sequences produced since the origin of life using some liberal assumptions. Considering the upper limit, it is clear that bacteria dominate the planet in terms of the product of the number of cells (10^{30}; Whitman *et al.* 1998) multiplied by the number of genes in each genome (10^4, a small overestimate). Let us assume that every single gene in this total of 10^{34} is unique and that evolution has been working on these genes for 4 Gyr completely changing each gene to some other unique, new gene every single year. This gives an extreme upper limit of 4×10^{43} different amino acid sequences explored since the origin of life. The contribution to this number of sequences by viral and eukaryotic genomes is difficult to estimate but it is very unlikely to be orders of magnitude greater than the 4×10^{43} sequences from bacteria. If their contribution is similar or smaller, then it can be ignored in our rough calculation. For comparison with our calculation, Mandecki (1998) gave a limit of 10^{50} protein sequences since the origin of life. A lower limit to the number of sequences explored is more difficult to estimate but it has been estimated that there are 10^9 different bacterial species on Earth (Whitman *et al.* 1998; Medini *et al.* 2005; Simonson *et al.* 2005). If we assume that each species has a unique complement of 10^3 sequences (an underestimate) and that only one sequence has changed per species per generation (a reasonable estimate based upon analysis of mutation rates in bacteria; Perfeito *et al.* 2007), and that the generation time is 1 year (a considerable underestimate for many modern bacteria (Ochman *et al.* 1999), but perhaps reasonable for an ancient organism or one growing slowly in a poor environment), then we arrive at a figure of 4×10^{21} different protein sequences tested since the origin of life.

These two limits are shown in figure 1. Although the oft-quoted 20^{100} (approx. 10^{130}) size of sequence space is far above these limits, the other more plausible estimates for the size of sequence space, particularly with limited amino acid diversity or reduced length, are near to or within these two limits. Considering the upper limit, all sequences containing 20, 8 and 3 types of amino acids have been explored if the chains are 33, 50 and 100 amino acids in length, respectively. Considering the lower limit, then virtually all chains of length 33 and 50 amino acids containing five or three types of amino acid, respectively, could have been explored. (The exploration of longer chains of 100 amino acids with only two types of residue is obviously much less complete but it is not a negligible fraction of the total.) Therefore it is entirely feasible that for all practical (i.e. functional and structural) purposes, protein sequence space has been fully explored during the course of evolution of life on Earth (perhaps even before the appearance of eukaryotes).

3. DISCUSSION

Protein sequence space is often viewed as a limitless desert of maladjusted sequences with only a few oases of working sequences linked by narrow pathways (Axe 2000, 2004). The navigation over this space by natural selection is difficult and could take many different routes thus resulting in organisms with largely different protein compositions. This idea of contingency, if taken at the level of species, led Gould to suggest that if one was to rerun the 'tape of life' then evolution would take a totally different path and we,

Downloaded from http://rsif.royalsocietypublishing.org/ on June 22, 2017

as a species, would only appear as a highly improbable accident (Gould 1991; Luisi 2003; de Duve 2007*a*,*b*). However, if there is any merit to our simple calculation then protein sequence analysis provides no support for the idea of contingency at a molecular level and it provides strong support for the ideas of convergence (Conway Morris 2000, 2004; Dawkins 2005; Vermeij 2006; de Duve 2007*a*,*b*). If one was to rerun the tape, then the protein composition of organisms would be similar. Our calculation removes the almost impossibly unrealistic pressure on natural selection to navigate through protein sequence space avoiding the vast number of functionless sequences by simply indicating that most sequences have been tried are useful in some way, and that there are many possible routes to obtain proteins with desirable functions (Nagano *et al.* 2002; Anantharaman *et al.* 2003; Holliday *et al.* 2007).

Finally, we conclude that the number 20^{100} and similar large numbers (e.g. Salisbury 1969; Maynard Smith 1970; Mandecki 1998; Luisi 2003; Carrier 2004; de Duve 2005) are simply 'straw men' advanced to initiate discussion in the same spirit as the 'Levinthal paradox' of protein folding rates (Levinthal 1969; Zwanzig *et al.* 1992). 20^{100} is now no more useful than the approximate $2\times10^{1\,834\,097}$ books present in Borges' (1999) fantastical 'Library of Babel' and has no connection with the real world of amino acids and proteins. Hence, we hope that our calculation will also rule out any possible use of this big numbers 'game' to provide justification for postulating divine intervention (Bradley 2004; Dembski 2004).

We thank the Engineering and Physical Sciences Research Council for the award of a grant to D. Dryden, M. Greaney, M. Bradley, D. A. Leigh and R. L. Baxter which partially funded this work. We gratefully acknowledge discussions with Simon Conway Morris (Cambridge), Tom McLeish (Leeds) and Wilson Poon (Edinburgh). We also thank the referees, including Geerat J. Vermeij, for their detailed comments and opinions. This work was initiated at the Isaac Newton Institute for Mathematical Sciences workshop on 'Statistical Mechanics of Molecular and Cellular Biological Systems', January–July 2004.

REFERENCES

Anantharaman, V., Aravind, L. & Koonin, E. V. 2003 Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20. (doi:10.1016/S1367-5931(02)00018-2)

Anfinsen, C. B. 1973 *Studies on the principles that govern the folding of protein chains*. Nobel Foundation Lectures, pp. 103–119.

Axe, D. D. 2000 Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. *J. Mol. Biol.* **301**, 585–595. (doi:10.1006/jmbi.2000.3997)

Axe, D. D. 2004 Estimating the prevalence of protein sequences adopting functional enzyme folds. *J. Mol. Biol.* **341**, 1295–1315. (doi:10.1016/j.jmb.2004.06.058)

Banavar, J. R. *et al.* 2006 Geometry of proteins: hydrogen bonding, sterics, and marginally compact tubes. *Phys. Rev. E* **73**, 031921. (doi:10.1103/PhysRevE.73.031921)

Borges, J. L. 1999 The Library of Babel. In *Collected fictions* (transl. A. Hurley). New York, NY: Penguin.

Bradley, W. L. 2004 Information, entropy and the origin of life. In *Debating design: from Darwin to DNA* (eds W. A. Dembski & M. Ruse), pp. 331–351. New York, NY: Cambridge University Press.

Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. 2002 Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655.

Bujnicki, J. M. 2001 Understanding the evolution of restriction–modification systems: clues from sequence and structure comparisons. *Acta Biochim. Pol.* **48**, 935–967.

Carrier, R. C. 2004 The argument from biogenesis: probabilities against a natural origin of life. *Biol. Philos.* **19**, 739–764. (doi:10.1007/s10539-005-6860-1)

Chan, H. S. & Dill, K. A. 1991 "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.* **95**, 3775–3787. (doi:10.1063/1.460828)

Chiarabelli, C. & de Lucrezia, D. 2007 Question 3: the worlds of the prebiotic and never born proteins. *Orig. Life Evol. Biosph.* **37**, 357–361. (doi:10.1007/s11084-007-9075-4)

Conway Morris, S. 2000 Evolution: bringing molecules into the fold. *Cell* **100**, 1–11. (doi:10.1016/S0092-8674(00)81679-7)

Conway Morris, S. 2004 *Life's solution: inevitable humans in a lonely universe*. Cambridge, UK: Cambridge University Press.

Cordes, M. H., Davidson, A. R. & Sauer, R. T. 1996 Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10. (doi:10.1016/S0959-440X(96)80088-1)

Dawkins, R. 2005 *The ancestor's tale: a pilgrimage to the dawn of life*. London, UK: Phoenix, Orion Books Ltd.

de Duve, C. 2005 *Singularities: landmarks on the pathways of life*. Cambridge, UK: Cambridge University Press.

de Duve, C. 2007*a* Chemistry and selection. *Chem. Biodivers.* **4**, 574–583. (doi:10.1002/cbdv.200790051)

de Duve, C. 2007*b* Chance and necessity revisited. *Cell. Mol. Life Sci.* **64**, 3149–3158. (doi:10.1007/s00018-007-7442-y)

Dembski, W. A. 2004 The logical underpinnings of intelligent design. In *Debating design: from Darwin to DNA* (eds W. A. Dembski & M. Ruse), pp. 311–330. New York, NY: Cambridge University Press.

Denton, M. J. 2008 Protein-based life as an emergent property of matter: the nature and biological fitness of the protein folds. In *Fitness of the cosmos for life; biochemistry and fine-tuning* (eds J. D. Barrow, S. Conway Morris, S. J. Freeland & C. L. Harper), pp. 256–279. Cambridge, UK: Cambridge University Press.

Dill, K. A. 1999 Polymer principles and protein folding. *Protein Sci.* **8**, 1166–1180.

Doi, N., Kakukawa, K., Oishi, Y. & Yanagawa, H. 2005 High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng. Des. Sel.* **18**, 279–284. (doi:10.1093/protein/gzi034)

Dyson, H. J. & Wright, P. E. 2005 Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208. (doi:10.1038/nrm1589)

Gould, S. J. 1991 *Wonderful life: the Burgess shale and the nature of history*. London, UK: Penguin Books Ltd.

Guo, H. H., Choe, J. & Loeb, L. A. 2004 Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210. (doi:10.1073/pnas.0403255101)

Holliday, G. L., Almonacid, D. E., Mitchell, J. B. & Thornton, J. M. 2007 The chemistry of protein catalysis. *J. Mol. Biol.* **372**, 1261–1277. (doi:10.1016/j.jmb.2007.07.034)

Ikehara, K. 2002 Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC–SNS primitive genetic code hypothesis. *J. Biosci.* **27**, 165–186. (doi:10.1007/BF02703773)

Downloaded from http://rsif.royalsocietypublishing.org/ on June 22, 2017

Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. 2002 Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci.* **11**, 2804–2813. (doi:10.1110/ps.0203902)

Lau, K. F. & Dill, K. A. 1990 Theory for protein mutability and biogenesis. *Proc. Natl Acad. Sci. USA* **87**, 638–642. (doi:10.1073/pnas.87.2.638)

Levinthal, C. 1969 How to fold graciously. Mossbauer spectroscopy in biological systems. In *Proc. Meeting held at Allerton House, Monticello, IL* (eds P. Debrunner, J. C. M. Tsibris & E. Munck), pp. 22–24. Urbana, IL: University of Illinois Press.

López de la Osa, J., Bateman, D. A., Ho, S., Gonzalez, C., Chakrabartty, A. & Laurents, D. V. 2007 Getting specificity from simplicity in putative proteins from the prebiotic Earth. *Proc. Natl Acad. Sci. USA* **104**, 14 941–14 946. (doi:10.1073/pnas.0706876104)

Luisi, P. L. 2003 Contingency and determinism. *Phil. Trans. R. Soc. A* **361**, 1141–1147. (doi:10.1098/rsta.2003.1189)

Luisi, P. L., Chiarabelli, C. & Stano, P. 2006 From never born proteins to minimal living cells: two projects in synthetic biology. *Orig. Life Evol. Biosph.* **36**, 605–616. (doi:10.1007/s11084-006-9033-6)

Mandecki, W. 1998 The game of chess and searches in protein sequence space. *Trends Biotechnol.* **16**, 200–202. (doi:10.1016/S0167-7799(98)01188-3)

Maynard Smith, J. 1970 Natural selection and the concept of a protein space. *Nature* **225**, 563–564. (doi:10.1038/225563a0)

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. 2005 The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594. (doi:10.1016/j.gde.2005.09.006)

Miller, S. L., Urey, H. C. & Oro, J. 1976 Origin of organic compounds on the primitive earth and in meteorites. *J. Mol. Evol.* **9**, 59–72. (doi:10.1007/BF01796123)

Nagano, N., Orengo, C. A. & Thornton, J. M. 2002 One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765. (doi:10.1016/S0022-2836(02)00649-6)

Ochman, H., Elwyn, S. & Moran, N. A. 1999 Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA* **96**, 12 638–12 643. (doi:10.1073/pnas.96.22.12638)

O'Neill, M., Dryden, D. T. F. & Murray, N. E. 1998 Localization of a protein–DNA interface by random mutagenesis. *EMBO J.* **17**, 7118–7127. (doi:10.1093/emboj/17.23.7118)

Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. 2007 Adaptive mutations in bacteria: high rate and small effects. *Science* **317**, 813–815. (doi:10.1126/science.1142284)

Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. 1998 Simplified proteins: minimalist solutions to the 'protein folding problem'. *Curr. Opin. Struct. Biol.* **8**, 80–85. (doi:10.1016/S0959-440X(98)80013-4)

Raes, J., Harrington, E. D., Singh, A. H. & Bork, P. 2007 Protein function space: viewing the limits or limited by our view? *Curr. Opin. Struct. Biol.* **17**, 362–369. (doi:10.1016/j.sbi.2007.05.010)

Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. 1997 Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809. (doi:10.1038/nsb1097-805)

Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. 2007 REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res.* **35**, D269–D270. (doi:10.1093/nar/gkl891)

Salisbury, F. B. 1969 Natural selection and the complexity of the gene. *Nature* **224**, 342–343. (doi:10.1038/224342a0)

Simonson, A. B., Servin, J. A., Skophammer, R. G., Herbold, C. W., Rivera, M. C. & Lake, J. A. 2005 Decoding the genomic tree of life. *Proc. Natl Acad. Sci. USA* **102**, 6608–6013. (doi:10.1073/pnas.0501996102)

Sobolevsky, Y. & Trifonov, E. N. 2006 Protein modules conserved since LUCA. *J. Mol. Evol.* **63**, 622–634. (doi:10.1007/s00239-005-0190-4)

Sturrock, S. S. & Dryden, D. T. F. 1997 A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes. *Nucleic Acids Res.* **25**, 3408–3414. (doi:10.1093/nar/25.17.3408)

Trifonov, E. N. 2000 Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151. (doi:10.1016/S0378-1119(00)00476-5)

Vermeij, G. J. 2006 Historical contingency and the purported uniqueness of evolutionary innovations. *Proc. Natl Acad. Sci. USA* **103**, 1804–1809. (doi:10.1073/pnas.0508724103)

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. 1998 Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583. (doi:10.1073/pnas.95.12.6578)

Zwanzig, R., Szabo, A. & Bagchi, B. 1992 Levinthal's paradox. *Proc. Natl Acad. Sci. USA* **89**, 20–22. (doi:10.1073/pnas.89.1.20)