Ribosome evolution and structural capacitance

Ashley M Buckle[a] **and Malcolm Buckle**[b]

[a]Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia; [b]LBPA, ENS Paris-Saclay, CNRS, Université Paris-Saclay, F-94235, Cachan, France

Running Title: Ribosome evolution

Keywords: structural capacitance | ribosome evolution | protein disordered region | disorder-order transition | codon-anticodon | genetic code

Abstract

   In addition to the canonical loss-of-function mutations, mutations in proteins may additionally result in gain-of-function through the binary activation of cryptic 'structural capacitance elements'. Our previous bioinformatic analysis allowed us to propose a new mechanism of protein evolution - structural capacitance – that arises via the generation of new elements of microstructure upon mutations that cause a disorder-to-order (D➔O) transition in previously disordered regions of proteins. Here we propose that the D➔O transition is a necessary follow-on from expected early codon-anticodon and tRNA acceptor stem-amino acid usage, via the accumulation of structural capacitance elements - reservoirs of disorder in proteins. We develop this argument further to posit that structural capacitance is an inherent consequence of the evolution of the genetic code.

Like all cellular life, proteins evolved by Darwinian natural selection. Once a primitive enzyme acquired even very weak catalytic activity, genetic mutations followed by selection, did the rest. But here is the catch - how would classic Darwinian evolution proceed in the absence of pre-existing seed structures and functions? Nothing evolves unless it already exists. Darwinian selection needs something, some function, to select for (or against). When life started more than three billion years ago, what was the molecular event that created this "something" from randomness, and ignited the evolution of a protein fold?

We recently postulated a hypothesis that identifies a mechanism whereby microstructure is generated *de novo* in hitherto disordered regions of protein [1]. This process, termed structural capacitance (SC), provides a mechanism to accelerate Darwinian natural selection by taking advantage of a disorder-to-order (D→O) causing mutation that will generate a function conferring neostructure in a hitherto poorly structured region of a protein. This idea was supported by a thorough database analysis of over 68000 human disease-associated mutations which serve as reservoirs, or 'structural capacitance elements' in which a single gene mutation may create nucleating seeds that can act as a 'feedstock' for evolution to proceed. Structural capacitance is compatible with emerging evidence for significant structural changes induced by mutation [2-4], the 'dormant foldon' hypothesis [5], concepts of early protein evolution driven by peptide-world 'foldamers' and Dayhoff's hypothesis [6], as well as oligomerisation-duplication-fusion events of short peptide [7] [8].

Here we suggest that the D→O transition is a necessary follow-on from expected early codon-anticodon and tRNA acceptor stem-amino acid usage. Irrespective of how existing codon-anticodon and stem-amino acid usage evolved, our essential hypothesis is that it allowed the accumulation of reservoirs of disorder in proteins (structural capacitance elements) that have been maintained in the present ecosystem.

Many factors are evoked to explain codon-anticodon and stem-amino acid configurations as they exist today; these necessarily represent a compromise between mutational bias and protein translation requirements optimised through natural selection. However, for our purposes we need to wind the clock back to the very early stages in the development of the post RNA, early ribosomal world. We will first think about these two central aspects of ribosomal mechanisms; codon:anticodon interactions and acceptor stem:amino acid recognition.

From a purely chemical point of view, the idea that the stability of codon-anticodon interactions was predominant in the development of the genetic code was initially brokered by Eigen and Schuster [9] and given strong support by the application of the thermostability rule [10] to coding triplets determined by the stacking energies of adjacent nucleotides [11] [12]. Furthermore, it was recently postulated that stem-amino acid recognition is correlated with the size of the amino acid and that codon-anticodon recognition is based on polarity with a strong correlation between polar/non-polar and Purine/Pyrimidine frequency at the 2nd anticodon position [13].

 In the prebiotic world, as originally suggested by Jukes [14] and by Crick [15], the initial code would have been a 2-letter triplet (XYNXYN…) where the XY base step controlled the codon:anticodon interaction, thus XYN specified a particular amino acid with a usage of GC (Ala) GG/CC (Gly, Pro), GT/AC (Val, Thr), GA/TC (Asp or Glu).

The most stable codon-anticodon pairs would have been selected for the restricted set of amino acids likely to have been present, namely Asp, Glu, Ala, Gly, Val and Thr in decreasing order of polarity [16]. Indeed, all of these amino acids are coded by single changes in the GCU triplet [10]. Trifonov has established a chronological order of amino acid appearance based on codon:anticodon thermostability (Gly/Ala, Val/Asp, Pro, Ser, Glu/Leu, Thr, Arg, Asn, Lys, Glu, Ile, Cys, His, Phe, Met, Tyr and Trp) [10]. If the most stable codon:anticodon couples were therefore selected by the simple prebiotic amino acids this suggests an inverse correlation

between codon-anticodon stability and amino acid complexity. Thus, at the beginning of the transition from the RNA world one would expect most early proteins to have been of low complexity and thus disordered or poorly structured. Eventually, moving down this list, a critical threshold of amino acids would have been reached, at which point moving to more complex protein structures would be limited by the lack of chemically diverse amino acids. Catalysis is required for efficient replication and, in the RNA world this would have arisen from RNA secondary structure. A current hypothesis is that acceptor stem recognition preceded anticodon recognition [13] and that this selection, based essentially on size discrimination led to a prevalence of beta structure with alternating small and large amino acids. However, this argument does not address the question of how these amino acetylated tRNA molecules then lined up on an mRNA template in the absence of anticodons.

Independent, however, of whichever mechanism arrives first - stem recognition or codon recognition - the limited subset of amino acids originally present (Asp, Glu, Ala, Val, Gly, and Thr of which two are polar/charged, two are non-polar, and two are indifferent, respectively) occupying necessarily the most stable codon:anticodon arrangements argues that the first primordial proteins were relatively poorly structured.

The elephant in the room here is that, as pointed out by Beloussof *et al* [17], the initial proto ribosome would require early peptides to interact with the RNA molecules and for that, the positively charged amino acids (His, Lys and Arg) would be highly selected, however, none of these amino acids are expected to be present in the prebiotic broth. Andrew Travers suggested two possible work arounds for this [12], either the early biotic environment somehow contained sufficient Arg so that it 'grabbed' a strong codon:anticodon slot (CGN), or that since one of the products in the Miller reaction was found to be norleucine and that CGN was originally taken by norleucine; subsequent cyanide and ammonia derivation of the unbranched norleucine generated Arg that then displaced the norleucine from the CGA codon:anticodon slot.

4

It follows therefore that the D→O transition is implicit in the genetic code simply because all the stable codon:anticodon interactions were initially taken by simple amino acids less capable of producing ordered structure. Consequently, any mutation would invariably lead to a less stable codon:anticodon interaction allowing other amino acids to come in and fill the void. Since other amino acids with few exceptions are chemically more complex than the limited prebiotic subset, then this would allow for more and more order to appear in protein structure. Nonetheless early proteins would have had tracts of disorder intrinsically present, with little evolutionary pressure to remove them and huge codon:anticodon pressure to remain!

Eventually, the initial set of proteins resulting from this process and making up the phenotype of a primitive organism would therefore already have a set of structure/function motifs in its proteome within tracts of disorder. Subsequent O→D transitions would thus be probably deleterious if they occurred within the structured folds; D→O transitions could still occur, but would require moving from stable to unstable codon:anticodon interactions. Since this interaction, however, is no longer important in terms of codon usage as the ribosome already exists, the effect would then be simply on protein function, and would therefore lead to structural capacitance. So in other words, the original evolutionary pressure to select codon use (i.e. to select stable codon:anticodon interactions) is no longer present and a consequence of the way that this initially evolved means that all D→O mutations which mostly involve stable to unstable codon:anticodon are refractive to evolutionary pressure, *except of course if that pressure is now exerted at the level of the ensuing new fold or structure*. The codon change involved in D→O transitions is neutral as far as selection at the level of codon usage is concerned. Once a primitive ribosome has started to attribute codon usage under the sole selective pressure of codon:anticodon stability, there is no going back. It is remarkable but logical that codon:anticodon stability favours simple amino acids. Structural capacitance therefore is an inherent consequence of ribosome evolution.

At this point it is worth discussing the fascinating results of a study by the Root-Bernsteins [18]. The authors discuss the origin of ribosomes in terms of three hypotheses: firstly, that ribosomes evolved prior to cellular life and was to a large extent genetically self-sufficient, secondly that rRNA is purely structural (which is the classical text book definition and that they call the null hypothesis) and finally that any genetic information present in ribosomes is purely random. Their study strongly argues for the first hypothesis and we would like to discuss the ramifications of this in terms of SC and its implicit origin in early codon:anticodon usage. They showed that rRNA regions possessed the capacity to code for all 20 tRNAs in an overlapping fashion. They also showed that rRNA coded for proteins associated with ribosomes and furthermore they identified between 30 and 55% of identified rRNA that encoded proteins with similarities to functional sites of identified active regions of proteins that they mimic, such as polymerases, helicases, phosphodiesterases and so forth. Thus, in existing ribosomes we find that codon:anticodon use is common for both tRNA synthesis from rRNA and protein synthesis from rRNA. Consequently, would this not then argue that codon recognition predates stem recognition since tRNA synthesis from rRNA would not necessarily require proteins? If this were the case then as we argue above, the most stable codon:anticodon pairs *that already exist* would be grabbed by the most abundant amino acids available (D, E, A, G, V, T) with little potential for structure.

This is in fact borne out by an extensive energetic appraisal of codon interactions carried out by Grosjean and Westhof [19] that supports the concept that complementary RNA:RNA duplexes constituted the original proto ribosome in which stable complimentary G:C rich triplets coded for small polypeptides consisting of alternating Ala and Gly amino acid. As Grosjean and Westhof succinctly explain, weaker codon:anticodon pairs would associate with more diverse emerging amino acids and interestingly enough would therefore constitute a pool

of more and more deviations from the genetic code and therefore leave place for the insertion of additional amino acids and thus structure!

An interesting test of our hypothesis in the light of the Root-Bernstein results would be to examine the structures of ribosomal proteins (and the active sites of proteins sharing homology with rRNA encoded proteins) to determine the potential for SC in disordered regions.

As mentioned earlier, one of the factors dictating ribosome function is protein translation requirements, in other words how fast can the protein fold as it peels away from the ribosome. Both sequential and non-sequential folding mechanisms dictate that disordered structures will impose less of a kinetic barrier to this than ordered regions. Thus, the presence of disordered regions will act as 'fast lanes' during protein synthesis and there will be a certain selection pressure to maintain them. Another exciting possibility, and here there may be metaphoric parallels with nucleosome organisation and the control of gene expression, is that the distribution of these disordered regions in proteins may have been selected to ensure a specific overall rate of synthesis, in other words if synthesis of a particularly highly structured protein becomes economically difficult because of the time limitations on folding then maintaining some optimally spaced disordered regions in the protein to allow the ribosome to put its foot down a bit, will alleviate the road block! Again, of course these are not designed; according to our hypothesis, early proteins would have had huge areas of disorder, these would have been reorganised with selection but many will have been retained precisely for the reason outlined above. Indeed, there is no way of knowing in advance when conditions will arise such that selective capacitance will, through a random mutation in the disordered region, provide a useful fold/function; using disordered regions as "reverse governors" during protein synthesis will ensure their retention in the gene pool.

Concluding remarks

Whilst the early codon usage may not be the "best" for the whole range of amino acids it will have dictated that structural capacitance is embedded in the system. In fact, had prebiotic amino acids been more polar and complex, then structural capacitance would not be present and it is highly unlikely that evolution would have the time to have produced the huge range of diversity in life that exists today. A prediction of this is, therefore, if or when we find extra-terrestrial life, it will be complex only if it has evolved from simple structures!!

Author Contributions

Conceptualization, Ashley M Buckle and Malcolm Buckle; Writing – original draft, Ashley M Buckle and Malcolm Buckle; Writing – review & editing, Ashley M Buckle and Malcolm Buckle

Conflict of interest

The authors declare no conflict of interest.

References

1. Li, C.; Clark, L.V.T.; Zhang, R.; Porebski, B.T.; McCoey, J.M.; Borg, N.A.; Webb, G.I.; Kass, I.; Buckle, M.; Song, J., et al. Structural Capacitance in Protein Evolution and Human Diseases. *J Mol Biol* **2018**, *430*, 3200-3217, doi:10.1016/j.jmb.2018.06.051.

2. Arodz, T.; Plonka, P.M. Effects of point mutations on protein structure are nonexponentially distributed. *Proteins* **2012**, *80*, 1780-1790, doi:10.1002/prot.24073.

3. He, Y.; Chen, Y.; Alexander, P.A.; Bryan, P.N.; Orban, J. Mutational tipping points for switching protein folds and functions. *Structure* **2012**, *20*, 283-291, doi:10.1016/j.str.2011.11.018.

4. Toth-Petroczy, A.; Tawfik, D.S. Hopeful (protein InDel) monsters? *Structure* **2014**, *22*, 803-804, doi:10.1016/j.str.2014.05.013.

5. Uversky, V.N. Unusual biophysics of intrinsically disordered proteins. *Biochim Biophys Acta* **2013**, *1834*, 932-951, doi:10.1016/j.bbapap.2012.12.008.

6. Romero Romero, M.L.; Rabin, A.; Tawfik, D.S. Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew Chem Int Ed Engl* **2016**, *55*, 15966-15971, doi:10.1002/anie.201609977.

7.   Chessari, S.; Thomas, R.; Polticelli, F.; Luisi, P.L. The production of de novo folded proteins by a stepwise chain elongation: a model for prebiotic chemical evolution of macromolecular sequences. *Chem Biodivers* **2006**, *3*, 1202-1210, doi:10.1002/cbdv.200690121.

8.   Soding, J.; Lupas, A.N. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **2003**, *25*, 837-846, doi:10.1002/bies.10321.

9.   Eigen, M., Schuster, P. The Hypercycle A principle of Natural Self-Organisation. Part C. The realistic Hypercycle. *Naturewissenschafen* **1978**, *65*, 341-369.

10.   Trifonov, E.N. Consensus temporal order of amino acis and evolution of the triplet code. *Genes Chromosomes Cancer* **2000**, *261*, 139-151.

11.   Krueger, A.; Protozanova, E.; Frank-Kamenetskii, M.D. Sequence-dependent base pair opening in DNA double helix. *Biophys J* **2006**, *90*, 3091-3099, doi:10.1529/biophysj.105.078774.

12.   Travers, A. The evolution of the genetic code revisited. *Orig Life Evol Biosph* **2006**, *36*, 549-555, doi:10.1007/s11084-006-9041-6.

13.   Carter, C.W., Jr.; Wolfenden, R. tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc Natl Acad Sci U S A* **2015**, *112*, 7489-7494, doi:10.1073/pnas.1507569112.

14.   Jukes, T.H. Indications of an Evolutionary Pathway in the Amino Acid Code. *Bioch Biophys Res Commun* **1967**, *27*, 573-578.

15.   Crick, F. The Origin of the Genetic Code. *J Mol Biol* **1968**, *38*, 367-369.

16.   Miller S L; Urey H, C. Organic Compound Synthesis on the Primitive Earth. *SCIENCE* **1959**, *130*, 245-251.

17.   Belousoff, M.J.; Davidovitch.C.; Bashan, A.; Yonath, A. On the Development Towards the Modern World: A Plausible Role of Uncoded Peptides in the RNA World. *Origins of Life and Evolution of Biospheres* **2010**, *40*, 5.

18.   Root-Bernstein, M.; Root-Bernstein, R. The ribosome as a missing link in the evolution of life. *J Theor Biol* **2015**, *367*, 130-158, doi:10.1016/j.jtbi.2014.11.025.

19.   Grosjean, H.; Westhof, E. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res* **2016**, *44*, 8020-8040, doi:10.1093/nar/gkw608.