A possible molecular metric for biological evolvability

Aditya Mittal[1],* and B Jayaram[1,2]

[1]Kusuma School of Biological Sciences, [2]Department of Chemistry and Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology Delhi, New Delhi 110 016, India

*Corresponding author (Fax, +91-11-26591052; Email, amittal@bioschool.iitd.ac.in)

Proteins manifest themselves as phenotypic traits, retained or lost in living systems via evolutionary pressures. Simply put, survival is essentially the ability of a living system to synthesize a functional protein that allows for a response to environmental perturbations (adaptation). Loss of functional proteins leads to extinction. Currently there are no universally applicable quantitative metrics at the molecular level for either measuring 'evolvability' of life or for assessing the conditions under which a living system would go extinct and why. In this work, we show emergence of the first such metric by utilizing the recently discovered stoichiometric margin of life for all known naturally occurring (and functional) proteins. The constraint of having well-defined stoichiometries of the 20 amino acids in naturally occurring protein sequences requires utilization of the full scope of degeneracy in the genetic code, i.e. usage of all codons coding for an amino acid, by only 11 of the 20 amino acids. This shows that the non-availability of individual codons for these 11 amino acids would disturb the fine stoichiometric balance resulting in non-functional proteins and hence extinction. Remarkably, these amino acids are found in close proximity of any given amino acid in the backbones of thousands of known crystal structures of folded proteins. On the other hand, stoichiometry of the remaining 9 amino acids, found to be farther/distal from any given amino acid in backbones of folded proteins, is maintained independent of the number of codons available to synthesize them, thereby providing some robustness and hence survivability.

[Mittal A and Jayaram B 2012 A possible molecular metric for biological evolvability. *J. Biosci.* **37** 573–577] **DOI** 10.1007/s12038-012-9210-x

1. Introduction

Exploration of relationships between the genetic code and occurrence of amino acids in primary sequences of proteins is fundamental to biological evolution. While there have been numerous thought-provoking investigations towards understanding the codon usage and stoichiometric occurrences of amino acids in primary sequences of proteins (Kimura 1968; King and Jukes 1969; Knight *et al.* 2001; Ermolaeva 2001; Yampolsky and Stoltzfus 2005; Hershberg and Petrov 2008; Yang *et al.* 2009), none (without exception) of the findings and/or discussions arising from them are universally applicable. Several investigations have been severely limited as they considered sequences of (specific) proteins in only specific species of living organisms. Others have been limited by either the number, and/or length, and/or type (i.e. classified structurally and/or functionally) of the protein sequences analysed. Thus, in the absence of any universally applicable results, the underlying implications for understanding evolvability of species or populations (or genomes or protein structures), at the molecular level, are still debated under the context of 'complexity' arising out of every investigation providing a 'new' species-/population-specific insight (for example, see Singer and Hickey 2003; Casanueva *et al.* 2010). However, in view of our recent observations on some remarkably universal aspects of all known protein sequences and structures (independent of origin, species, sequence length and structural/functional classifications), we decided to explore evolvability in terms of a possible universally applicable understanding at the molecular level.

Recently, we showed that naturally occurring functional proteins are characterized by well-defined stoichiometries (percentage of the times that each of the amino acids occurs in primary sequences) for the 20 amino acids (table 1). While the initial observation was obtained from 3718 protein

Keywords. Codon; evolution; genetic code; proteins; stoichiometry

Published online: 25 June 2012

sequences of natural proteins with crystal structures having 2.5 Å resolution or better (Mittal *et al.* 2010), the results have now been confirmed for more than 130,000 functional proteins (Mittal and Jayaram 2011a). Since the standard deviation of average occurrence of each amino acid in primary sequences of functional proteins in nature is much smaller compared to a random distribution (Mezei 2011), we called the tight deviations as the margin of life. It is interesting to note that the 20-letter alphabet for naturally occurring functional proteins is not degenerate, and required. If it were degenerate, then the margin of life would not exist (i.e. if one amino acid could be replaced by another, standard deviations of percentage occurrences of amino acids would be much larger).

It is clear that naturally occurring proteins are different from each other due to fluctuation of the amino acid stoichiometries in primary sequences within the margin of life. Stated differently, sequence variations that characterize different proteins are bounded by the margins found for each amino acid. Further, we have shown that the stoichiometric margin of life (in primary sequences) manifests itself into a naturally occurring functional protein by adopting structural configurations that are constrained by an invariant spatial ordering of amino acids (Mittal and Jayaram 2011b). This invariant order has been captured as a 'wheel' of neighbourhoods (Mittal and Jayaram 2011b). Thus, in terms of naturally occurring proteins, it is now established that stoichiometry of amino acids leads to a structure (constrained by invariant wheel of neighbourhoods) that leads to function. But what are the origins of the stoichiometries and the fluctuations within the margin of life that give rise to the observed wonderful functional diversity in natural proteins leading to survival of living systems? Further, how do these origins get reflected in nucleic acid sequences (Sarma 2011) and transferred from generation to generation across species?

Table 1. Stoichiometric and spatial characteristics of amino acids in naturally occurring functional proteins

Amino acid	Number of codons	Percentage occurrence		Structural proximity – 'wheel position' (Mittal and Jayaram 2011b)
		Average	Std	
Cys (C)	2	1.8	1.5	1
Ile (I)	3	5.8	2.4	2
Val (V)	4	7.1	2.4	3
Leu (L)	6	9	2.9	4
Phe (F)	2	3.9	1.8	5
Met (M)	1	2.2	1.3	6
Ala (A)	4	7.8	3.4	7
Tyr (Y)	2	3.4	1.7	8
Trp (W)	1	1.3	1	9
His (H)	2	2.3	1.4	10
Gly (G)	4	7.2	2.8	11
Ser (S)	6	6	2.5	12
Thr (T)	4	5.5	2.4	13
Arg (R)	6	5	2.3	14
Gln (Q)	2	3.8	2	15
Pro (P)	4	4.4	2	16
Asn (N)	2	4.3	2.2	17
Lys (K)	2	6.3	2.8	18
Glu (E)	2	7	2.7	19
Asp (D)	2	5.8	2	20

2. Hypothesis

To answer the above, on the basis of fundamental principles of physical chemistry, we formulated a straightforward testable hypothesis: number of codons available in the genetic code (Nirenberg *et al.* 1965; Khorana *et al.* 1966) for each amino acid is responsible directly for dictating the number of times that an amino acid can occur in primary sequences of naturally functional proteins. The testing of this hypothesis was expected to yield two distinct (and extreme) possibilities:

1. The number of codons for every amino acid is correlated to its relative occurrence in primary sequences: This would mean that all the codons for any given amino acid are required and utilized to synthesize the amino acid during assembly of natural polypeptide chains. This would also mean that the apparent degeneracy in the genetic code only provides more codon choices for an amino acid that is required more. Hence, this would directly point towards the extremely 'delicate' nature of genomic assemblies, since changing the number of codon choices (in the extant genetic code) for synthesis of any amino acid would disturb the stoichiometric margin of life, leading to a non-functional protein. This would imply that base mutations affecting even a single codon would lead to non-functional proteins and hence extinction.
2. The number of codons for every amino acid is not correlated to its relative occurrence in primary sequences: In this case, the only option to obtain functional proteins, within the stoichiometric constraints for amino acids, would be to have a codon bias and repeat the codons, i.e. repeat certain bases in the genome, to synthesize an amino acid that is required more. This would control the overall ATGC composition. (The individual base compositions are not anchored at 25% for all genomes/genes, nor is there an equivalence of base composition in codons

coding for the same amino acid and hence the expected change in the base composition with changes in the codon bias. Codons are the same for different species but codon biases could be different for different species.) However, this would imply that the overall genomic (ATGC) composition must vary among species to account for variations in survival conditions by the virtue of variations in protein sequences (and hence functions).

Clearly, the former possibility pointed out to a lower control on survival under environmental perturbations leading to mutations, and the latter pointed towards more stable survival under different environmental conditions supporting specific types of genomic (ATGC) compositions. Therefore, to test our hypothesis we listed the number of codons in the genetic code for each amino acid along with its percentage occurrence in naturally occurring primary sequences in folded proteins (table 1). Although the order of amino acids was not important, we listed the amino acids in the order of their appearance in the wheel of invariant neighbourhoods since that represents a key structural constraint in naturally occurring functional proteins. Taking inspiration from the central dogma (Crick 1970), we were now ready to test the following hypothesis for naturally occurring functional proteins:

Number of codons → Stoichiometry of amino acids
→ Structure constrained by a specific order of
neighbourhoods for every amino acid → Function

3. Results and discussion

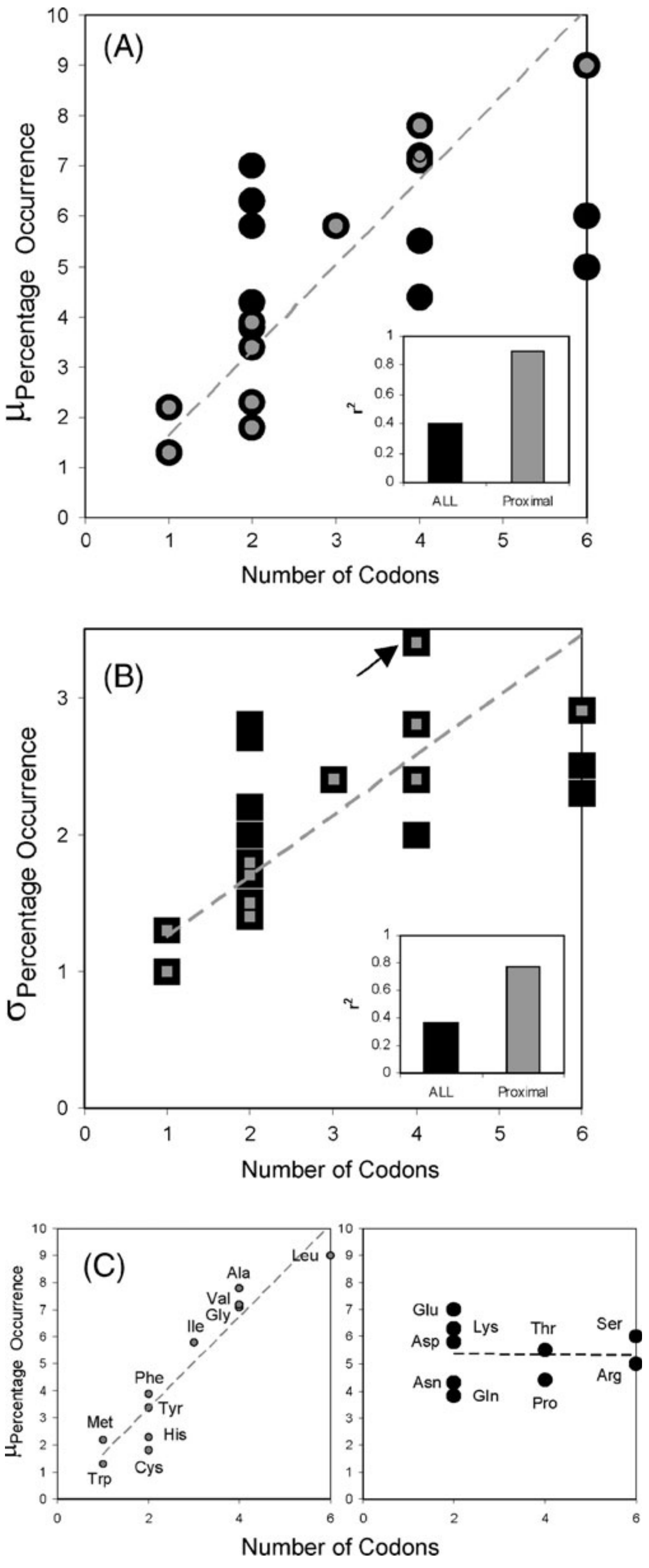
Figure 1A shows the average percentage occurrence of all 20 amino acids (●) as a function of number of codons corresponding to each of the amino acids. The inset (black bar) shows that number of codons for amino acids do not correlate well (r^2=0.40) with their respective percentage occurrences in primary sequences. Figure 1B shows the same lack of correlation (r^2=0.37) for the stoichiometric margin of life, i.e. standard deviations (■). This was a promising finding in that it was supported by the known codon biases leading to variations in the overall genomic compositions among different species. However, why were the correlations not negligible? The search for this answer yielded a remarkable result. The grey symbols in figure 1A and B show that stoichiometries of the first 11 amino acids as per the structural constraints in functional proteins (i.e. position in the invariant wheel of neighbourhoods; table 1) were highly correlated with the number of codons (in the genetic code) for each of the amino acids (r^2=0.90 in figure 1A, grey dashed line and grey bar in the inset; r^2=0.77 in figure 1B). Therefore, not only are these amino acids structurally important by being 'proximal' to every other amino acid in structures of naturally occurring functional proteins, but their stoichiometries in primary sequences are a direct consequence of the degeneracy in the genetic code. These proximal amino acids require every single codon in the genetic code to be able to maintain their required numbers in primary sequences.

Thus, figure 1C (a parsed representation of data in figure 1A) clearly shows a balance between the two distinct and extreme possibilities arising out of our hypothesis. Stoichiometry of only some (and not all) amino acids, interestingly in close proximity to every other amino acid in folded proteins, is dictated by the degeneracy in the genetic code. This balance provides a strong metric to gauge adaptability and survivability of primary sequences arising out of genomes. Tilting the balance towards the 11 proximal amino acids, whose percentage occurrence is a direct consequence of the number of ways they are synthesized, will result in a 'delicate' living system, prone to extinction with environmental perturbations. On the contrary, tilting the balance towards the remaining 9 amino acids would imply adaptability and survivability by selection (and reuse) of only a particular set of codons to maintain stoichiometric margin of life required for functional proteins. It would also imply a distinct genomic identity to a species surviving under different environments, and the ability to evolve in response to environmental perturbations by simply switching the codon choices of amino acids.

On a single protein level, our findings explain some very interesting and important experimental observations:

1. Cysteines are found with relatively lower frequencies in naturally occurring functional proteins. Structurally, they are found to be closest to everyone including themselves (Mittal and Jayaram 2011b). Thus, Anfinsen's historic experiments disturbed more than just disulfide bonds. They disturbed the overall packing of proteins, and hence 100% activity was never recovered on removal of the reducing agent. In the context of current findings, cysteines are one of the 'delicate' amino acids. Mutating cysteines is already known to be detrimental for proteins.
2. Alanine scanning is a routine molecular biology protocol for studying protein function. Figure 1B (arrow) shows that among the 'delicate' amino acids, alanine is farthest from the correlation line; i.e. from the perspective of margin of life, alanine can be utilized for mutational tolerances.

Our findings not only provide the first of its kind metric for evolvability of living systems at the molecular level, but they also have a major implication in the modern genomic and proteomic era, especially in synthetic biology. Once the genomes and proteomes of a species are known, then one simply needs to see whether the average percentage



◀ **Figure 1.** Stoichiometric occurrences of individual amino acids (in primary sequences of naturally occurring functional proteins) are plotted as a function of the number of codons in the genetic code for the respective amino acid (see table 1 for the data). (**A**) shows that average percentage occurrences of all the 20 amino acids (●) do not correlate well with the number of codons (r^2=0.40, black bar in inset). However, the same for first 11 amino acids (◍) in table 1 correlate extremely well with the number of codons (r^2=0.90, grey bar in inset). (**B**) shows that standard deviations of average percentage occurrences (margin of life, see text for details) of all the amino acids (■) do not correlate well with the number of codons (r^2=0.37, black bar in inset). However, the margin of life for the first 11 amino acids (▨) in table 1 correlate well with the number of codons (r^2=0.77, grey bar in inset). Arrow shows 'alanine' that is the farthest from the correlation line (in fact, removal of alanine from ▨ enhances the r^2 to 0.86). This distance from the correlation line indicates ability for mutational tolerances. Overall, the figure indicates a metric for adaptability and evolvability (see text for details). (**C**) shows the data from (**A**), parsed separately into two panels, the left corresponding to the first 11 amino acids and the right corresponding to the remaining 9. Overall, the figure indicates a metric for adaptability and evolvability (see text for details).

occurrences of individual amino acids in their proteomes are correlated to the number of codons utilized for synthesizing each of those amino acids. If the answer is yes, then clearly the species survival is extremely delicate and one would predict low adaptability as well as low survivability. From the perspective of synthetic biology, for designer genomes that result in increased 'survivability' of species, it would be suitable to create proteomes that are not synthesized by utilizing the degeneracy of the genetic code completely. On the other hand, to create designer genomes for which survivability needs to be controlled (e.g. mosquitoes for out-competing malarial parasite careers), it would be suitable to create proteomes that are synthesized by utilizing full degeneracy of the genetic code, thereby limiting biological recovery and the ability to evolve as a result of mutations. It may be mentioned that an explanation for the observed/evolved degeneracy at the level of codons in RNA can be found in mRNA–tRNA interactions in the ribosomal machinery (Jayaram 1997), but the correspondence between codons and amino acids eludes a simple interpretation.

In terms of understanding evolution at the molecular level, our findings point towards a new and thought-provoking direction. We show that constraint on evolution is actually the stoichiometry of amino acids to obtain functional proteins. The stoichiometric margin of life imposes restrictions on genomic assemblies (DNA compositions). Hence, proteins must have come first in the primordial soup and what we are observing now is simply a feedback loop with functional proteins dictating the selection of genomic assemblies. If a protein, in spite of having catalytic activity, did not fold properly due to stoichiometric and structural (wheel of neighbourhoods) constraints, it was simply degraded and there was no feedback loop. Strong experimental evidence for this feedback loop has been recently reported where the function of a protein is different due to a difference in the genetic code despite the same primary sequence (Komar 2007; Kimchi-Sarfaty *et al.* 2007; Sharma *et al.*

2008; Weygand-Durasevic and Ibba 2010; Zhang *et al.* 2010). Nucleic acids, on the other hand, could not have initiated a feedback loop since there is no stoichiometric dependence (ATGC composition) on structure and function. Finally, it is important and interesting to note that stoichiometric dependence is the basic rule for any reaction to occur in chemistry.

References

Casanueva A, Tuffin M, Cary C and Cowan DA 2010 Molecular adaptations to psychrophily: the impact of 'omic' technologies. *Trends Microbiol.* **18** 374–381

Crick, F 1970 Central dogma of molecular biology. *Nature* **227** 561–563

Ermolaeva MD 2001 Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **3** 91–97

Hershberg R and Petrov DA 2008 Selection on codon bias. *Annu. Rev. Genet.* **42** 287–299

Jayaram B 1997 Beyond the wobble: the rule of conjugates. *J. Mol. Evol.* **45** 704–705

Khorana HG, Büchi H, Ghosh H, Gupta N, Jacob TM, Kössel H, Morgan R, Narang SA, Ohtsuka E and Wells RD 1966 Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **31** 39–49

Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV and Gottesman MM 2007 A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315** 525–528

Kimura M 1968 Evolutionary rate at the molecular level. *Nature* **217** 624–626.

King JL and Jukes TH 1969 Non-Darwinian evolution. *Science* **164** 788–798

Knight RD, Freeland SJ and Landweber LF 2001 A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2** research0010.1–research0010.13

Komar AA 2007 SNPs, Silent but not invisible. *Science* **315** 466–467

Mezei M 2011 Discriminatory power of stoichiometry-driven protein folding? *J. Biomol. Struct. Dyn.* **28** 625–626

Mittal A and Jayaram B 2011a The newest view on protein folding: Stoichiometric and spatial unity in structural and functional diversity. *J. Biomol. Struct. Dyn.* **28** 669–674

Mittal A and Jayaram B 2011b Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *J. Biomol. Struct. Dyn.* **28** 443–454

Mittal A, Jayaram B, Shenoy SR and Bawa TS 2010 A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding? *J. Biomol. Struct. Dyn.* **28** 133–142

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F and O'Neal C 1965 RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci. USA* **53** 1161–1168

Sarma RH 2011 A conversation on protein folding. *J. Biomol. Struct. Dyn.* **28** 587–588

Sharma M, Hasija V, Naresh M and Mittal A 2008 Functional control by codon bias in magnetic bacteria. *J. Biomed. Nanotechnol.* **4** 44–51

Singer GA and Hickey DA 2003 Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317** 39–47

Weygand-Durasevic I and Ibba M 2010 New roles for codon usage. *Science* **329** 1473–1474

Yampolsky LY and Stoltzfus A 2005 The exchangeability of amino acids in proteins. *Genetics* **170** 1459–1472

Yang D, Jiang Y and He F 2009 An integrated view of the correlations between genomic and phenomic variables. *J. Genet. Genomics* **36** 645–651

Zhang F, Saha S, Shabalina SA and Kashina A 2010 Differential arginylation of actin isoforms is regulated by coding sequence–dependent degradation. *Science* **329** 1534–1537