From genetic coding to protein folding

Jean-Luc Jestin

FROM GENETIC CODING TO PROTEIN FOLDING

Jean-Luc Jestin

ABSTRACT

A discrete classical mechanics (DCM) is established here. It consists of classical mechanics principles and equations, which are solved over the field of rational numbers *Q*, and not over the field of real numbers *R*. The need for a DCM results from observations of the genetic code. A DCM model for protein folding allows a set of folding nuclei to be derived for each protein. Applications in protein structure prediction are envisioned.

INTRODUCTION

The genetic code is quasi-universal among living organisms [1], [2], [3]. This property is consistent with the fact that the genetic code can be considered as an optimum, as a product of evolution. Optimization models tend to identify selection criteria which might have been essential in shaping the genetic code [4], [5], [6].

For example, codons with the second base being thymine (uracil) are known to encode amino acids whose residues are hydrophobic [7]. Base substitutions at the first codon base tend therefore to minimize the deleterious effects of amino acid mutations by conservation of the hydrophobic nature of the residue.

Deciphering of the genetic code yielded the observation that transitions at the third base of codons are generally neutral at the amino acid level. The effects of the most frequent base substitutions are thereby minimized [8].

In another model, it was argued that amino acid pairs whose codons differ only by the third base are biosynthetically related in five cases out of seven [9], [10]. The biosynthetic relationship is either a precursor-product relationship or a relationship between two products with a common precursor within the metabolic pathways.

In a fourth model, the stop codons were found to minimize the deleterious effects of single-base deletions during nucleic acid polymerization [5]. The hotspots for single-base deletions are typically opposite a purine R within YTRV template sequences during DNA-dependent DNA-polymerization by family 1 or 2 DNA polymerases. Assigning these hotspot sequences to the end of genes (i.e. stop codons) minimizes the mutations impact on protein by the synthesis of full-length proteins with a peptide added at the C-terminus likely to be functional, instead of truncated proteins unlikely to be functional if the hotspot sequence had been assigned to codons encoding amino acids.

THE GENETIC CODE'S OPTIMIZATION FOR PROTEIN FOLDABILITY MODEL

It was noted that the amino acids arginine, glycine and histidine have amino acid residue masses which are squares (100, 1 and 81 respectively). The corresponding codons are highlighted in Fig.1. As shown in the table, the assignment of these amino acids to codons is

far from a random assignment: the codons tend to differ from one another by unique base substitutions.

The following model is suggested to account for this optimization of the genetic code. If proteins were not able to fold into functional three-dimensional compact structures, then there would be no selective advantage to genetic coding for living organisms. The genetic code may accordingly be optimized to facilitate the folding of proteins.

The kinetic energy of an amino acid residue of mass m can be noted $mX^2/2$. Let us consider a protein with the mutation of an amino acid with residue mass m into another amino acid with residue mass $m' = m / a^2$. The kinetic energy $m'X^2/2$ can then be rewritten as $m(X/a)^2/2$: both terms represent the same numbers over the field of rational numbers [12]. The deleterious effects of mutation on protein foldability are minimized if the folding energetics is conserved, that is, if the same numbers are represented by the kinetic energy term.

A PROTEIN FOLDING MODEL

Let us consider the following protein folding model. A chemical group of mass m, the folding unit, such as an amino acid or an amino acid residue folds onto a folding nucleus of mass (M-m) to yield a larger folding nucleus or the folded protein of mass M.

The reference frame is fixed with respect to the rotating folding nucleus, so that its kinetic energy equals 0 in this frame. The kinetic energy of the folding unit in this frame equals $mX^2/2$. The kinetic energy of the larger folding nucleus after folding is $MY^2/2$. The internal energy relased during folding is noted Ui. The difference in energy due to bond breaking and bond formation between the folding unit, the solvent and the folding nucleus is noted Ep.

Energy conservation during the folding step can then be written :

$$mX^2/2 = MY^2/2 + E$$

where $E = Ui + Ep$.

While most analyses in bioinformatics have been carried out so far over the field of real numbers because of the ease of calculations, new and interesting conclusions can be derived from an arithmetical analysis over the field of rational numbers as shown below.

For a given E, if this equation has no solution in X and Y, folding cannot occur.
If this equation has an infinite number of solutions in X and Y, the foldability of the folding unit onto the folding nucleus is optimized. For any given E, this equation has an infinite number of solutions in X and Y if (m/M) is a square.

Let us consider a folding unit whose mass m is a square. Folding nuclei are then polypeptide sequences whose mass M is a square. A set of folding nuclei can then be derived easily for any protein. In Figure 2, this set was established for *Escherichia coli* cold shock protein A (CspA). Tertiary structure information can be extracted from such graphs which shall find important applications in protein structure prediction.

FIGURE LEGENDS

Legends to Figure 1.

Standard representation of the genetic code.
The pink cylinder corresponds to codons and their amino acids, whose residues are strictly hydrophobic. The blue squares correspond to codons and their amino acids, which are closely related within biosynthetic pathways. The green blobs correspond to codons and their amino acids whose residues' masses are squares. The broken green line corresponds to codons encoding cysteine, whose mass is a square.

Legends to Figure 2.

Representation of the set of 22 folding nuclei for the cold shock protein A (CspA) from *Escherichia coli.*
On the x-axis: numbering of the 69 amino acids from CspA as referenced in the PDB (1mjc).
On the y-axis: each red segment does correspond to a folding nuclei (cf. text).

ACKNOWLEDGEMENTS

A. Guilloux and A. Kempf contributed respectively to the mathematical and physical analyses of the folding model.

REFERENCES

1. Ninio, J., *Divergence In The Genetic-Code.* Biochemical Systematics And Ecology, 1986. **14**(5): p. 455-457.
2. Osawa, S. and T.H. Jukes, *Evolution of the genetic code as affected by anticodon content.* Trends Genet., 1988. **4**(7): p. 191-198.
3. Osawa, S., et al., *Recent evidence for evolution of the genetic code.* Microbiol. Rev., 1992. **56**(1): p. 229-264.
4. Sonneborn, T.M., *Degeneracy of the genetic code: extent, nature and genetic implications*, in *Evolving genes and proteins*, V. Bryson and H.J. Vogel, Editors. 1965, Academic Press: New York. p. 377-397.
5. Jestin, J.L. and A. Kempf, *Chain-termination codons and polymerase-induced frameshift mutations.* FEBS Letters, 1997. **419**: p. 153-156.
6. Freeland, S.J., et al., *Early fixation of an optimal genetic code.* Mol. Biol. Evol., 2000. **17**(4): p. 511-518.
7. Woese, C.R., *On the evolution of the genetic code.* Proc. Natl. Acad. Sci. USA, 1965. **54**(6): p. 1546-1552.
8. Goldberg, A.L. and R.E. Wittes, *Genetic code: aspects of organization.* Science, 1966. **153**(734): p. 420-424.
9. Wong, J.T., *Coevolution theory of the genetic code at age thirty.* Bioessays, 2005. **27**(4): p. 416-25.
10. Wong, J.T., *A co-evolution theory of the genetic code.* Proc. Natl. Acad. Sci. USA, 1975. **72**(5): p. 1909-1912.
11. Conway, J. and N. Sloane, *Sphere packing, lattices and groups*. 1988, Berlin: Springer.
12. Serre, J.P., *A course in arithmetic*. 1973, New York: Springer.
13. Gutfraind, A. and A. Kempf, *Error-reducing structure of the genetic code indicates code origin in non-thermophile organisms.* Origins Life Evol. Biosph., 2008. **38**: p. 75-85.
14. Jestin, J.L., *Degeneracy in the genetic code and its symmetries by base substitutions.* C. R. Biol., 2006. **329**(3): p. 168-171.
15. Jestin, J.L. and C. Soulé, *Symmetries by base substitutions in the genetic code predict 2' and 3' aminoacylation of tRNAs.* J. Theor. Biol., 2007. **247**: p. 391-394.
16. Jestin, J.L. and A. Kempf, *Degeneracy in the genetic code: how and why?* Genes Genomes Genomics, 2007. **1**: p. 100-103.
17. Jestin, J.L., *A rationale for the symmetries by base substitutions of degeneracy in the genetic code.* Biosystems, 2010. **99**: p. 1-5.
18. Koonin, E.V. and A.S. Novozhilov, *Origin and evolution of the genetic code: the universal enigma.* IUBMB Life, 2009. **61**: p. 99-111.
19. Sella, G. and D.H. Ardell, *The impact of message mutation on the fitness of a genetic code.* J. Mol. Evol., 2002. **54**(5): p. 638-651.
20. Seligmann, H. and D.D. Pollock, *The ambush hypothesis: hidden stop codons prevent off-frame gene reading.* DNA Cell Biol., 2004. **23**(10): p. 701-705.
21. Trevors, J.T. and D.L. Abel, *Chance and necessity do not explain the origin of life.* Cell Biol. Int., 2004. **28**: p. 729-739.

TABLE

Probabilities that codons encoding an amino acid such as His, Gly or Arg differ by unique base substitutions from other codons in the set (His, Gly and Arg) by chance only, (A) in the case cysteine is included in the codon set, (B) in the case cysteine is not included in the codon set. C_n^k.= n!/k!/(n-k)!

	A		B	
His	$C_{14}^{2}\, C_{48}^{4} / C_{62}^{6}$	=.29	$C_{14}^{2}\, C_{48}^{8} / C_{62}^{10}$	=.32
Gly	$C_{24}^{8}\, C_{36}^{2} / C_{60}^{10}$	=.0061	$C_{24}^{6}\, C_{36}^{2} / C_{60}^{8}$	=.033
Arg	$C_{28}^{8}\, C_{30}^{0} / C_{58}^{8}$	=.0016	$C_{28}^{6}\, C_{30}^{0} / C_{58}^{6}$	=.0093

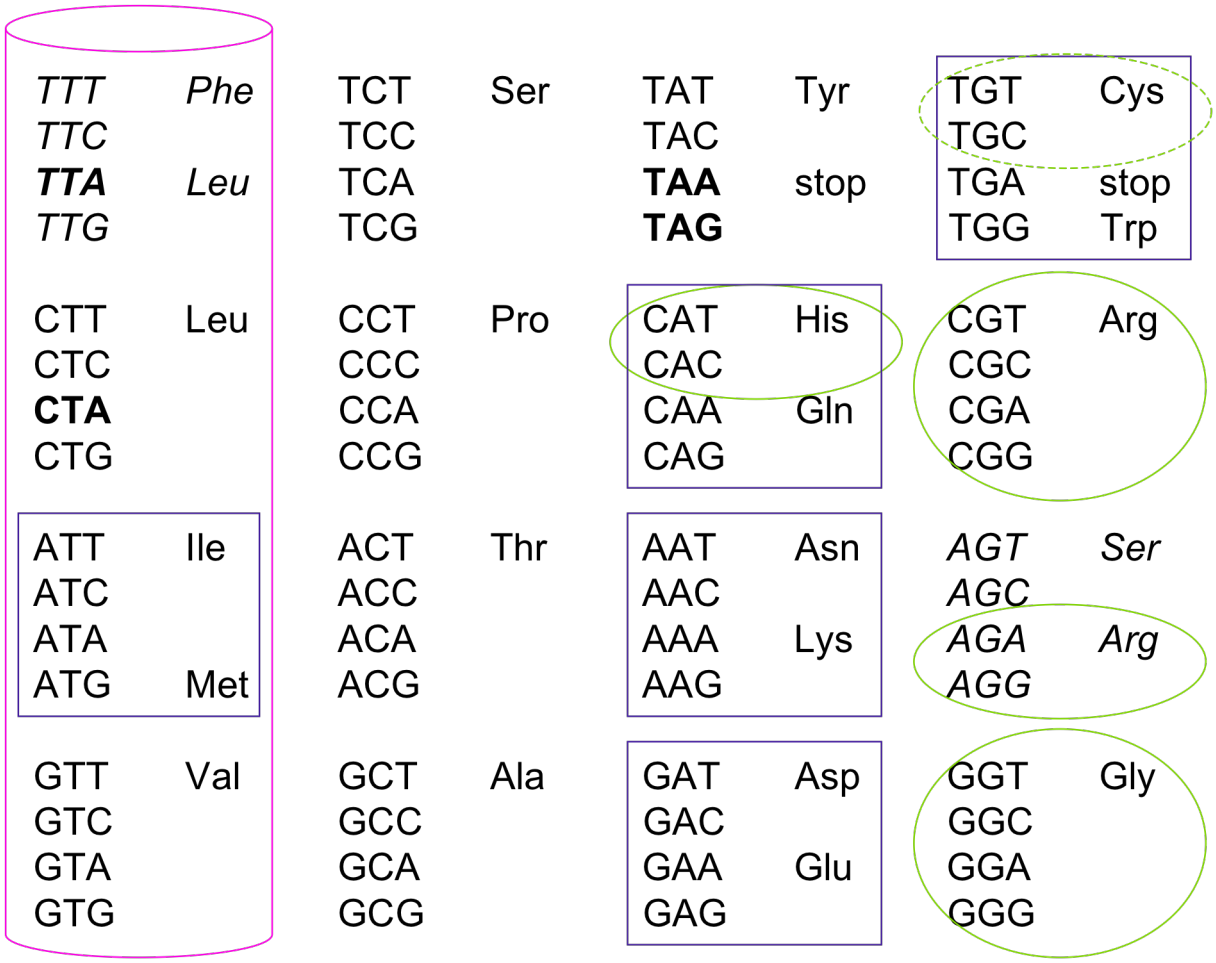
TTT Phe
TTC
TTA Leu
TTG
CTT Leu
CTC
CTA
CTG
ATT Ile
ATC
ATA
ATG Met
GTT Val
GTC
GTA
GTG
TCT Ser
TCC
TCA
TCG
CCT Pro
CCC
CCA
CCG
ACT Thr
ACC
ACA
ACG
GCT Ala
GCC
GCA
GCG
TAT Tyr
TAC
TAA stop
TAG
CAT His
CAC
CAA Gln
CAG
AAT Asn
AAC
AAA Lys
AAG
GAT Asp
GAC
GAA Glu
GAG
TGT Cys
TGC
TGA stop
TGG Trp
CGT Arg
CGC
CGA
CGG
AGT Ser
AGC
AGA Arg
AGG
GGT Gly
GGC
GGA
GGG

Fig. 1

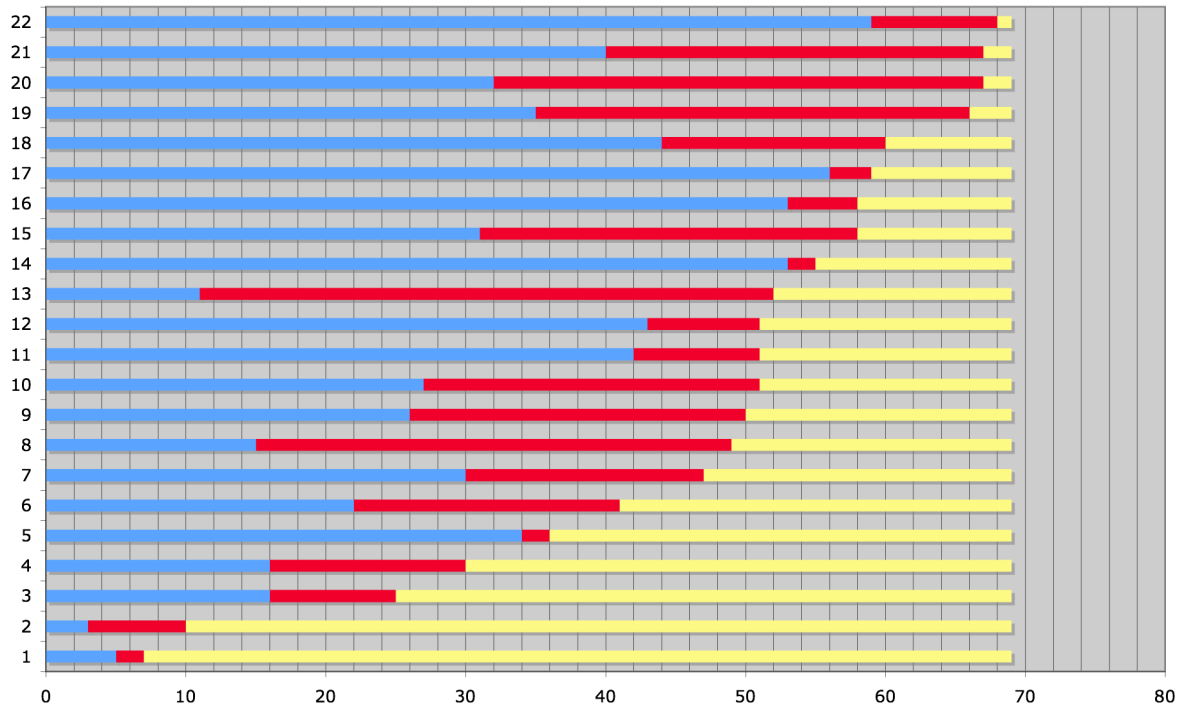
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80

Fig. 2