ISSN 2075-1729
www.mdpi.com/journal/life

Peer-Review Record:

Three-Dimensional Algebraic Models of the tRNA Code and the 12 Graphs for Representing the Amino Acids

Marco V. José, Eberto R. Morgado, Romeu Cardoso Guimarães, Gabriel S. Zamudio, Sávio Torres de Farías, Juan R. Bobadilla and Daniela Sosa

***Life* 2014, *4*, 341-373, doi:10.3390/life4030341**

Reviewer 1: Edward Trifonov (University of Haifa, Israel)
Reviewer 2: Anonymous
Reviewer 3: Anonymous
Editor: Niles Lehman (Portland State University, USA, Guest Editor of Special Issue "The Origins and Early Evolution of RNA")

Received: 26 April 2014
Revision Received: 23 July 2014
Accepted: 24 July 2014
Published: 11 August 2014

First Round of Evaluation

Round 1: Reviewer 1 Report

This is yet another educated viewpoint on the structure and origin of genetic code, based on its symmetry properties.

Round 1: Author Response to Reviewer 1

Thanks for your positive report.

Round 1: Reviewer 2 Report

Major comments:

1. The paper introduces the mathematical theory very well and allows the reader to understand basic mathematical concepts (like groups) directly.
2. Regarding the introduction: There exists a body of literature on codes the authors do not cite, e.g.,

Michel, C.J. Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes. *Computational Biology and Chemistry* 2012, 37, pp. 24–37.
Michel, C.J.; Seligmann, H. Bijective transformation circular codes and nucleotide exchanging RNA transcription. *BioSystems* 2014, 118, pp. 39–50.
(and references therein).

3. Since the approaches are very similar, please elaborate on these other approaches in the introduction to enable the reader to get an overview of the field.
4. The development of the mathematical theory seems to be sound, but the relation to biology is vague. The authors list many properties of their mathematical structures, especially the "genetic hotels" they report as results, but the biological relevance is hardly discussed.
5. A good direction is taken in section 8, where phenotypic graphs of amino acids are generated based on the developed theory. Still, a stronger link towards biological relevant properties would increase the theory's impact much more. The approaches to explain/hypothesize some biological implications should be much more elaborate and best supported by evidence.

In general I support the publication of this manuscript, since it can be the basis of more "biologically related" results.

Minor comments:

- Table 2 is quite unreadable in b/w. Since it is "just" another representation of a graph it might also go into the supplement to save space.
- Table 3 is a little bit "compressed", as is Figure 11B.

Round 1: Author Response to Reviewer 2

Answers to points 2 and 3:

We now provide an explicit connection between evolutionary routes for deriving the whole SGC, starting from a primeval RNY code with the circular codes. First, we added in Discussion a biological explanation of the Genetic Hotel of the SGC that includes the new Figure 13:

"The RNY codon model that was utilized for the algebraic procedure of the Hotels has biological backing in the observations of Eigen's group on tRNA sequences [10], supposed to be ancient genes. The abiotic support comes from the observations on abundant amino acids in Miller's sets [28,29]; and in Trifonov's review [30] that concentrate on the GNC row of the matrix. Symmetry procedures are added by the Hotels procedures to expand the initial RNY set to reach the full set of codes. In fact the evolution of the SGC from a primeval RNY code can be easily be visualized from the Genetic Hotel as illustrated in Figure 13. The set of codons RNY comprises what is called the RNA World (magenta). At this stage there were 16 RNY triplets encoding 8 amino acids. The products were ribozymes and coenzymes that were used for obtaining energy (NADP, FAD, ATP synthase). Most of these molecules are RNY sequences. Simple translations of this RNY condominium leads to the sets YNY (blue) and RNR (yellow) that altogether formed the Ribo-Nucleoprotein World. The second tRNA code currently known as the Second Operational tRNA code appeared at this stage. This second code is an analog and nondegenerate code

(20 amino acids charged by 20 tRNAs). The duplication of the first half of proto-tRNA minihelixes gave rise to the third tRNA code of digital nature: codon–anti-codon interactions. Finally, a translation of any of the condominiums YNY and/or RNR lead to the set of YNR codons (green) which altogether originates the DNA-Protein World (the 4 condominiums). The Last Universal Common Ancestor is considered to be a population of organisms possessing this frozen code [26,31].”

After this section the explicit relation is made:

“There are two ways of deriving the SGC from the primeval RNY code. First, we considered not a strict comma-less code as proposed by Crick et al. [32] but rather a degenerate RNA code which can be translated in the 1st (RNY), 2nd (NYR), and the 3rd (YRN) reading frames (frame-shift reading mistranslations that allow for the so-called statistical proteins [24]). The second pathway, shown in Figure 13, is derived by allowing transversions in the 1st (YNY) and 3rd (RNR) nucleotide bases of the 16 codons of the RNA code. It is worth to mention the discovery of the so-called circular codes [33] which have several remarkable properties [34 and references therein]. For example, it has recently been shown that circular codes show imprints in motifs of tRNAs and in 16S rRNA [35]. A maximal circular code X_0 of 20 trinucleotides was identified statistically on a large gene population of eukaryotes and prokaryotes [33]:

$$X_0 = \left\{ \begin{matrix} \text{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,} \\ \text{GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC} \end{matrix} \right\}$$

We remark that there is an interesting and intriguing connection between the circular codes with our approach of frame-shift reading mistranslations for obtaining the SGC from RNY codons. Note in fact that there are 12 out of the 16 RNY codons in X_0 and this circular code has been related to the origin of a primeval genetic code [36].”

Please note that we have added salient references regarding the circular codes, such as references 33–36.

Answers to points 4:

Besides the previous biological explanation of our Genetic Hotels, we added several biological issues such as the following two paragraphs placed in the Discussion:

“We remark that the polar requirement values of amino acids have been assigned to their respective codons in the classical table of the genetic code [24, 26]. Therefore, to our knowledge, this is the first time that polar requirement values are directly assigned to amino acids in a network.”

and:

“Polar requirement is an abiotic property of free amino acid molecules in solution; hydropathy is a substitute for that term but referring to properties of amino acid residues in protein tertiary structures, such as belonging to hydrophilic or hydrophobic stretches or segments or regions of molecules that, due to this, will be internal or external in the globular proteins or transmembranal or extramembranal in membrane proteins. Therefore, hydropathy is a biologic property [27]. There are three groups in the hydropathy correlation: the initial is of non-correlated hydroapathetic amino acids; the first correlated are in the extremes, highly hydrophilic together

with their partners in the dimers that are highly hydrophobic; the third belongs to the mixed sector of triplets and presents a nice line filled all through its length. The dimer rationale presupposes the 'phenotypic groups' to be hydropathy-discrepant, as we have found in this work."

Answers to points 5:

As now mentioned in the Abstract a whole new set of calculations was carried out in order to obtain a new biological result:

"The averages of statistical centrality measures of the 12 graphs for each of the 3 codes are carried out and they are statistically compared."

This new set of results, including the new Figure 7, is found at the end of Section *8.2.3. Centrality measures*:

"In order to compare the statistical properties of the 12 graphs for each code, we calculated the centrality measures for each of them and calculated their averages to perform a one-way anova test to determine if they are statistically similar (Figure 7 and SI-4). We consider only the 14 amino acids which are common to all 3 types of graphs."

and:

"We test the null hypothesis that the patterns of all curves for each centrality measure were statistically indistinguishable. By means of a simple one-way anova, we were unable to reject the hypothesis that the betweenness ($p < 0.89$) and eigenvector ($p < 0.703$) are statistically similar whereas we were able to reject the hypothesis that the degree $p < 0.012$) and closeness ($p < 0.022$) are statistically similar. This simple finding means that the hubs (eigenvector) and the shortest paths (betweenness) are preserved among the 3 graphs whereas the nodes are not all equally connected to the other nodes (degree) and the spread of information from a given node to all the other nodes (closeness) is not the same for all graphs."

The relevant new biological result is clearly stated in the Discussion:

"The averages of the centrality measures of the 12 graphs for each of the 3 codes were calculated. A simple one-way anova test showed that the eigenvector and betweenness are statistically indistinguishable in the 3 types of graphs while the degree and closeness differ from each other. Therefore, while the topology of the 3 graphs is statistically the same, the differences in degree and closeness essentially capture the idiosyncrasies of wobbling of the S-tRNA-C and the H-tRNA-C in regard to the SGC."

Answers to minor comments:

We eliminated Tables 2 and 3 since they can be found in Supplementary Information. Figure 11B is now Figure 12B. We were unable to decompress Figure 12B.

Round 1: Reviewer 3 Report

The authors use algebraic tools to study tRNA. I very much like an approach to combine modern algebraic tools to the study of biological systems and am very happy to read manuscript outlining research in that area. I think that the research is relevant and timely.
I do have, however, some concerns about this particular manuscript:

(1) It appears that the authors fail to cite and potentially are unaware of much relevant literature and research in that area, e.g.,:

   (i) "Genetic Coding: Approaches to Theory Construction", Findley, Findley and McGlynn. *J. Theor. Biol.* **1982**, *97*, 299–318
   (ii) "Review and application of group theory to molecular systems biology", Rietman, Karp, and Tuszynski. Theoretical Biology and Medical Modelling (2011) 8: 21
   (iii) "Symmetries of Genetic Code-Doublets", Danckwerts, Neubert, J. Mol. Evol. (1975) 5, 327
   (iv) "Global aspects in the algebraic approach to the genetic code", Forger, Hornos, Hornos, Physical Review E 1997, Vol 56 (6)

There are many more references which are relevant and have not been discussed. The authors do not cite/discuss any of these. It therefore does also not become clear how their work relates to existing work.

(2) The authors make several "modelling decisions", e.g., the assignment in line 128. The following assignments are made C<->00, U<->01, A<->10, G<->11. Table 1 shows that this then leads to a Four Klein Group. Not all elements in that group are "equal". C is the neutral element. The question is why is one of the four letters (C in this case) given a special character? It seems to change the relevance of that one element. What are the implications of this choice on their observations?
(3) I think that the paper is largely a summary of very elementary elements of algebra and some fundamental facts in biology. Then several (arbitrary) assignments are made and a number of observations are made. It does not become clear in how far these observations have any relevance or are just an artifact of the particular choices made at the beginning.
(4) The authors present what they call a "model" in the paper. In fact, it seems to be a set of assignments they make, which lead so some structure. Whether that structure is relevant in biology remains open. Also it does not seem that the authors make predictions based on their models that are then experimentally verified, which is a crucial step for the building of a good model.
(5) It does not become clear what this (experimentally not tested) "model" contributes to our understanding. The authors would need to make that much clearer.
(6) I do not think that the authors make the relevance TO BIOLOGY clear. Is this just a mathematical exercise?
(7) The manuscript is long, which is partially justifiable because of the mathematical and biological content, but reading the text it appears not well structured and often one does not know where one is going.
(8) There are many statements made which are strong generalisations or at least a bit misleading , e.g. (line 596) "the importance of a node or edge is commonly determined by its centrality and this depends on the characteristics or specific properties we are interested in." What depends on the

properties we are interested in? The fact that the importance of a node is commonly determined by its centrality?

(9) If often appears that the authors calculate things "because they can", e.g., Table 3. The biological relevance of this is hardly clear.

(10) The language needs to be polished.

(11) The references are not consistent, e.g., reference 1 is Crick FHC and 2 is Crick F.H.C.

Usually, the number of problems in this manuscript would make me reject the manuscript outright, but I am very keen on reading more papers in this area, so I hope that with substantial work the authors can improve the manuscript sufficiently to warrant publication.

Round 1: Author Response to Reviewer 3

Answers to comment 1:

We are grateful to the reviewer for the recommended references. We have included some of them in the following paragraphs in the Discussion:

"The SRM is only now starting to be examined by the mathematical procedures. Danckwerts and Neubert [37] used the Four-Klein group to partition the set of dinucleotides into the doublets that would match with a 3rd base for a triplet that has no influence on the coded amino acid (M1 set) and those doublets that do not code for amino acids without knowledge of the 3rd base in the triplet (M2 set). This approach is directly applicable to the SRM which is based on pDiN. In the SRM the set M1={GU, GG, AG, CG, GA, GC, AC, CC} read in 3' to 5' for anticodons clearly forms a Four Klein Group. In the SRF the set subset {GU, GC, AC, CC} belongs to the mixed YR pDiN whereas the set subset {GG, AG, CG, GA} corresponds to the homogeneous sector of RR pDiN.

The SRM starts exactly on the basis of symmetries derived from triplet complementarity and is very simple, with encodings occurring in four modules of four boxes each, dictated neatly by

ΔG-values. A chronology of amino acid encodings was generated on the bases of various protein functional properties. The chronology indicated a starting couple of amino acids which was not suspected by any of the previous models but found support in a simple amino acid biosynthesis pathway. The suggestion received further consistency from independent phylometabolic examinations [38] but still waits for experimental tests on the dimer mechanism."

And comparisons of the Genetic Hotels with other algebraic models can now be found in the Discussion:

"Algebraic models for the formation of the code are many. A pioneering work of the use of group theory for extracting the symmetries and evolution of the SGC was developed by Hornos and Hornos in 1993 [39]. Their approach, which uses Lie algebras, is based on analogies with particle physics and symmetry breaking from higher-dimensional space [40]. They showed that the current SGC must be slightly broken and they successfully predicted experimental values of amino acids polarities. Another most recent model which also uses Lie algebras of the genetic code over the Galois field of 4 DNA bases has been developed [41]. With our model, based on

elementary algebra, we have shown that with simple translations of the RNY condominium we can derive the whole SGC."

Answers to comment 2:

In *Section 2.2.*, we have now added the following:

"There is nothing special about this matching of the nucleotides. In fact the set N can be partitioned into two disjoint binary classes in three different ways, according to chemical criteria: strong-weak, amino-keto and pyrimidine-purine and similar results can be obtained [13]."

Answers to comment 3:

See reply to the previous point 2.

Answers to comment 4:

We hope that with the new result in which we show that differences in degree and closeness, but not betweenness and eigenvector, essentially capture the idiosyncrasies of wobbling of the S-tRNA-C and the H-tRNA-C in regard to the SGC, the added information to the Abstract, Introduction, Results, and Discussion, in particular Figure 13, the manuscript can be more readable and streamlined. Certainly our biological results are more clearly stated.

We also added in the Discussion just before Figure 13 the following biological result:

"The predictions of critical scale invariance (using renormalization group techniques) associated to symmetry breaking of the different stages of the evolution of the SGC have been verified with actual data of current genomes of Eubacteria and Archaea [15]."

In addition, our previous novel result of connecting polar requirement values to graphs of amino acids is

"We remark that the polar requirement values of amino acids have been assigned to their respective codons in the classical table of the genetic code [24, 26]. Therefore, to our knowledge, this is the first time that polar requirement values are directly assigned to amino acids in a network."

The SRM is a biological model based upon the phylogenies of the amino acids. The SRM was greatly helpful, throughout the manuscript, and in particular in the explanation of the results of polarity versus hydropathy. See 2nd paragraph of the Answer to point 4 of reviewer 2.

Answers to comment 5:

In the Introduction we have added how the Genetic Hotels have been tested against actual genomic data:

"The SGC has been theoretically derived from a primeval RNY genetic code under a model of sequential symmetry breakings [12-14], and vestiges of this primeval RNY genetic code were found in current genomes of both Eubacteria and Archaea [15]. All distance series of codons showed critical scale invariance not only in RNY sequences (all ORFs concatenated discarding the non-RNY triplets), but also in all codons of two intermediate steps of the genetic code and in all kind of codons in the current genomes [15]. Such scale invariance has been preserved for at least 3.5 billion years, beginning with an RNY genetic code to the SGC throughout two

evolutionary pathways. These two likely evolutionary paths of the genetic code were also analyzed algebraically and can be clearly visualized in 3, 4 and 6 dimensions [13, 14].”

Answers to comment 6:

Please see reply to reviewer 2, in particular Answers to points 2 and 3. See also reply to points 4 and 5.

Answers to comment 7:

See reply to the previous point 4.

Answers to comment 8:

We have omitted the sentence criticized by the reviewer and we have replaced it by:

“In order to characterize the different graphs of the amino acids for the different codes we use the following statistical properties.”

Please see also reply to reviewer 2, point 4.

Answers to comment 9:

We have omitted Table 3. Please see reply to reviewer 2, point 4.

Answers to comment 10:

There were indeed some typos and sentences that required polishing. We thank the referees for their suggestions and criticisms of the manuscript. There were indeed some typos and sentences that required to be amended and we worked in improving the exposition and in citing most of the references indicated by the reviewers. Thirteen references were added.

Second Round of Evaluation

Round 2: Reviewer 2 Report

The authors made significant changes to the manuscript to answer to the reviewers’ concerns. I appreciate their effort so far and would suggest to publish the paper even if clarification of the relevance to biology could still be improved, but this could also be done in a follow up paper. I also appreciate that the authors now included a statistical analysis to compare the 12 graphs.

Lines 675 reads as follows:

“In order to compare the statistical properties of the 12 graphs for each code, we calculated the centrality measures for each of them and calculated their averages to perform a one-way anova test to determine if they are statistically similar (Figure 7 and SI-4).”

Lines 687 reads:

“We test the null hypothesis that the patterns of all curves for each centrality measure were statistically indistinguishable. By means of a simple one-way anova, we were unable to reject the hypothesis that the betweenness ($p < 0.89$) and eigenvector ($p < 0.703$) are statistically similar whereas we were able to reject the hypothesis that the degree ($p < 0.012$) and closeness ($p < 0.022$) are statistically similar.”

Could you please clarify how you constructed the test for equality. A simple one-way ANOVA (please capitalize, it's an abbreviation) is a test for difference not for equality. If you constructed a test for equality from the ANOVA, that's fine, but then please provide a little more detail which construction method you choose (e.g., by using confidence intervals?), since equality tests are rather non-standard in scientific research papers.

Please provide the information which statistical software was used for the analysis.

Round 2: Author Response to Reviewer 3

We have added the following paragraph stating that we indeed used an ANOVA test for equality that considers confidence intervals as displayed by whiskers plots. We also mention the statistical software that we used and we capitalize the word ANOVA whenever used.

> "We consider only the 14 amino acids which are common to all 3 types of graphs in order to have a balanced ANOVA. This test was carried out using Matlab version R2012a which uses the command "anova1". If X is a matrix, anova1 treats each column as a separate group, and determines whether the population means of the columns are equal. Note that the whisker plots for this test provide a test group of medians, and this is not to be confused with the F test for different means in the classical ANOVA table."