JOURNAL OF SYSTEMS CHEMISTRY

RESEARCH ARTICLE Open Access

Testing for adaptive signatures of amino acid alphabet evolution using chemistry space

Melissa A Ilardo[1*] and Stephen J Freeland[1,2]

Abstract

Background: Multidisciplinary consensus indicates that half of the genetically amino acids are likely to have been available on the prebiotic earth, which implies certain adaptive expectations for the relationship between those amino acids and later additions to the genetic code. Chemistry space a concept that translates molecules to corresponding points in multidimensional space provides a framework for investigating these relationships. We therefore developed three tests to explore these implications using chemistry space to quantify otherwise qualitative questions.

Results: All three of our tests individually, as well as combined, provide quantitative evidence to support an adaptive expansion of the genetically encoded amino acid alphabet from 10 prebiotically plausible ("early") amino acids to the full set of 20 amino acids found within the standard genetic code.

Conclusions: We present three logically independent, novel tests of the adaptive growth of the amino acid alphabet from a smaller, functionally cohesive alphabet of only 10 amino acids to the 20 amino acids of the standard genetic code. While similar tests in the past have compared the genetically encoded amino acids to an external context of amino acids that were not incorporated into the genetic code our tests focus on the internal context of the 20 genetically encoded amino acids and find strong support. Of particular note one of these tests for the first time moves beyond consideration of amino acids as monomers and begins to explore polypeptides by considering the chemistry space of amino acid dimers.

Keywords: Amino acid alphabet, Genetic code, Chemistry space, Natural selection

Background

In recent years, chemistry space has revolutionized the search for pharmaceutically relevant molecules (e.g. [1-4]). The utility of chemistry space as a concept, however extends far beyond the pharmaceutical industry and is here used to investigate adaptive properties of the genetically encoded amino acids. The fundamental point of chemistry space is to assign molecules with numeric values that define specific aspects of their physical and/or chemical attributes. This simple step transforms a collection of unique molecules into a set of points in multi-dimensional space, which are therefore amenable to powerful visualization and quantitative analysis. Key to this transformation is replacing conceptual properties (such as "hydrophobicity"), with precisely defined, measurable molecular descriptors that quantify them (e.g. LogP).

Previously, chemistry space has been used to measure adaptive properties of the genetically encoded amino acids compared to an "external" context of alternative amino acids that were not incorporated into the standard genetic code [5]. This investigation produced strong evidence that the standard amino acid alphabet is indeed unusual in its coverage of size charge and hydrophobicity. Here we seek to develop further evidence for or against this adaptive interpretation of the amino acid alphabet. This time we focus upon the widespread consensus that the genetically encoded amino acids may be meaningfully divided into those that were abiotically available to the earliest life (*early* amino acids) versus those that were inventions derived by life itself (*late* amino acids) [6-10]. We therefore (use the data in Tables 1 and 2 to) test three specific hypotheses about the 'internal' adaptive logic of the standard amino acid alphabet in the

* Correspondence: melissailardo@gmail.com
[1]NASA Astrobiology Institute, University of Hawaii, Honolulu, USA
Full list of author information is available at the end of the article

Table 1 Raw Data: molecular descriptor values for each of the 20 genetically encoded amino acids

	ACD LogP	ACD MolVol	pI
		Early amino acids	
A	-0.574	76.736	6.06
D	-1.075	87.868	2.81
E	-0.969	104.375	3.42
G	-0.928	59.852	6.06
I	0.799	126.632	6.13
L	0.8	126.632	6.195
P	-0.06	97	6.725
S	-1.49	74.241	4.98
T	-1.136	91.125	5.85
V	0.289	110.126	6.39
		Late amino acids	
R	-0.999	118.724	11.87
N	-1.88	94.039	4.89
C	0.085	90.783	4.98
Q	-1.576	110.546	5.73
H	-1.418	108.983	7.33
K	-0.734	129.933	9.82
M	0.217	123.718	5.77
F	0.24	137.437	5.77
W	0.704	149.875	6.26
Y	-0.418	135.867	6.07

context of the chemistry space of the "early" versus "late" amino acids:

(i) If the concept of *early* amino acids is correct, then a subset of 10 genetically encoded amino acids formed a functionally cohesive protein-building alphabet of some earlier genetic code. We might expect this particular subset to exhibit similar adaptive qualities to those reported previously for the entire set of 20 genetically encoded amino acids [5].

(ii) If the amino acid alphabet grew from this smaller subset to its current size by adaptively adding new members (the *lates*), then we might anticipate that the late amino acids contribute the sort of literal expansion of chemistry space implied by previous qualitative statements such as "*The driving force* [in the growth of the amino acid alphabet] *is the possibility to produce fitter proteins when the repertoire of amino acids is enlarged*" [11].

(iii) If the concepts of early and late amino acids are correct, then the growth of the amino acid alphabet tested in (i) and (ii) should make adaptive sense when considering the use of amino acids in polymers. In particular, a smaller, early alphabet of 10 amino acids implies a library of 55 different amino acid dimers that were available for use in early proteins. An adaptive interpretation would predict that the addition of late amino acids should not overlap with this already-populated chemistry space. That is, the late amino acids would be expected to populate empty regions of chemistry space, and therefore fill functional roles that neither the early amino acids nor their dimers could perform.

Experimental

Testing for 'internal' adaptive properties of the genetically encoded amino acids involves two steps. First, any such test must define an appropriate chemistry space for the amino acids. This requires careful selection of molecular descriptors that accurately depict amino acids in terms of their relationships with one another and their roles within proteins. Given an appropriately defined chemistry space, it becomes possible to perform quantitative tests for the adaptive logic of the genetically encoded amino acids. For this second step, we test the three hypotheses outlined in the introduction. The first test probes the concept of early amino acids, asking whether this specific subset of 10 amino acids distinguishes itself relative to other possible subsets of the standard amino acid alphabet. The second test complements the first by turning to assess the idea of late amino acids, asking whether this subset amounts to a literal expansion of chemistry space. The third and final test places these two investigations into a broader, unified context by asking whether the early and late amino acids make adaptive sense when considered alongside a third component: dimers constructed from the early amino acids.

Defining amino acid chemistry space

Defining an appropriate chemistry space of amino acids is essential for any quantitative analysis of the amino acid alphabet. This conceptually simple step is rendered challenging by the vast array of amino acid molecular properties that have been measured. For example, the Amino Acid Index (or AAindex) comprises an extensive collection of such measures for the genetically encoded amino acids drawn from the scientific literature [12]. Currently, the database lists over 600 molecular descriptors. Though few of these descriptors are entirely independent of one another, the question remains: which subset best reflects relevance to the role of building proteins? To address this challenge, we start by noting that three key properties are commonly acknowledged to dominate amino acids' biochemical roles within protein structure and function: size, hydrophobicity, and charge [13-15].

Table 2 Raw Data: molecular descriptor values for early amino acid dimers

	ACD LogP	ACD MolVol
	Early amino acid dimers	
GG	−2.291	98.801
GA	−1.937	115.685
GS	−2.483	113.191
GT	−2.129	130.074
GV	−1.074	149.075
GL	−0.565	165.582
GI	−0.565	165.582
GP	−0.602	126.817
GD	−2.509	126.817
GE	−2.43	143.324
AA	−1.584	132.568
AS	−2.129	130.074
AT	−1.776	146.957
AV	−0.72	165.958
AL	−0.211	182.465
AI	−0.211	182.465
AP	−0.249	143.839
AD	−2.155	143.701
AE	−2.076	160.207
SS	−2.879	127.58
ST	−2.289	144.463
SV	−0.927	163.464
SI	−0.396	179.971
SP	−1.32	141.344
SD	−1.796	141.206
SE	−2.979	157.713
TT	−1.942	161.347
TV	−1.117	180.348
TL	−0.049	196.854
TI	−0.049	196.854
TP	−0.973	158.228
TD	−2.551	158.09
TE	−2.632	174.597
VV	0.143	199.348
VL	0.652	215.8
VI	0.652	215.855
VP	0.614	177.229
VD	−0.267	177.091
VE	−1.45	193.597
LL	1.162	232.362
LI	1.162	232.362
LP	1.124	193.735
LD	0.264	193.597
LE	−0.704	210.1
II	1.162	232.3
IP	1.124	193.735
ID	0.264	193.597
IE	−0.704	210.1
PP	1.415	164.077
PD	−0.931	163.963
PE	−0.852	180.47
DD	−2.334	154.833
DE	−2.255	171.34
EE	−2.314	187.846

Each property contributes to the biochemical interactions of amino acids in unique and essential ways. The size or bulk of amino acids' sidechains has long been recognized as an important factor in defining amino acid similarity [13]; hydrophobicity is likewise widely acknowledged as a fundamental determinant of folding pathways of nascent peptides and has been previously linked to the genetic code through the adaptive hypothesis [16,17]; and electrostatic interactions between amino acids have been shown to play a crucial role in inter- and intra-protein molecular interactions [15,18]. For these reasons, and because extensive tests have probed the reliability of various measures of these properties [19], especially in relation to the genetically encoded amino acids [5], these are the three dimensions we chose to investigate.

Having selected the properties of size, hydrophobicity, and charge, it remains to choose specific molecular descriptors with which to measure each of these conceptual properties. Size is the most straightforward of the three, owing to strong agreement across a variety of different descriptors. We elected to use ACD Molar Volume because it is freely available for a wide range of molecules, including amino acid dimers (see Test 3) via ChemSpider (www.chemspider.com). To represent hydrophobicity, we selected ACD LogP, also freely available through ChemSpider. LogP represents a subtly different, related property of lipophilicity, which is essentially hydrophobicity with the added consideration of polarity [20]. Specifically, LogP measures the logarithm partition coefficient, which represents the ratio of a compound's concentration in organic versus aqueous-phase solvents of a two compartment system (i.e. a the measure of the molecule's relative solubility in each of the two solvents). We chose to represent the charge (or electrostatic interactions) of a compound using Kowin isolectric point (pI). Whereas other measures of charge are highly dependent on the pH at which the measurement is taken, pI records the pH at which the concentration of the anionic and cationic forms of an amino acid are equal. It can be derived theoretically by calculating the pKa (or dissociation constant) values for the ionized states of the amino acid that exist one positive and one negative charge away from the neutral state of the amino acid [19].

Test 1: are the early amino acids an adaptive subset?

Our hypothesis predicts that if the 10 amino acids designated as "early" were indeed used by some earlier genetic code as a protein-building alphabet, they should exhibit unusual qualities as a subset that are analogous to the unusual properties detected for the entire set of 20 genetically encoded amino acids. In particular, the concept of amino acid coverage was used previously to measure the adaptive value of an amino acid alphabet in terms of its protein building potential [5]. Coverage considers both the range of values covered within a particular descriptor and how evenly these values are distributed within that range (Figure 1). An adaptive set of amino acids comprises members evenly distributed across a broad range for key physicochemical properties of size (molar volume), charge (Isoelectric Point) and hydrophobicity (LogP). Such a set of amino acids can combine within an evolving protein to approximate any suite of properties required by shifting environmental conditions.

A simple way to test the adaptive value of the early amino acids is therefore to take the 20 amino acids and ask: if we were to select 10 amino acids at random and measure their coverage in size (Molar Volume), hydrophobicity (ACD LogP) and charge (Isoelectric point), then how often would these random subsets exhibit a better coverage of amino acid chemistry space than that of the "earlies"? We therefore wrote a script to measure the coverage of the 10 early amino acids and of 1 million random samples of 10 amino acids chosen randomly from within the twenty encoded amino acids. This script was run separately for each dimension of amino acid chemistry space and then for each possible combination of 2 and 3 dimensions simultaneously.

Test 2: do the late amino acids expand the chemistry space of the early amino acids?

If the "late" amino acids were added to expand the universe of genetically encoded proteins, then we might expect them to be associated with some measurable expansion of amino acid chemistry space. In other words, we may predict that the late amino acids lie further from the earlies than would be expected for an arbitrary division of the amino acids.

To test this idea, we wrote a further script that first calculated the mean of the cluster of early amino acids for a given molecular descriptor and then measured the distance between this mean and each of the lates as a summed total distance (Figure 2). It then replicated this measurement 1 million times, each time randomly designating 10 of the 20 amino acids as early and 10 as late in order to record how often the true late amino acids show greater expansion from the earlies than occurs by chance. As with Test 1 above, this second test was performed for each individual dimension of amino acid chemistry space, and for combinations of 2 and 3 dimensions simultaneously. As an additional test of the robustness of our results, we repeated all calculations having first removed Glycine and Alanine, which, owing

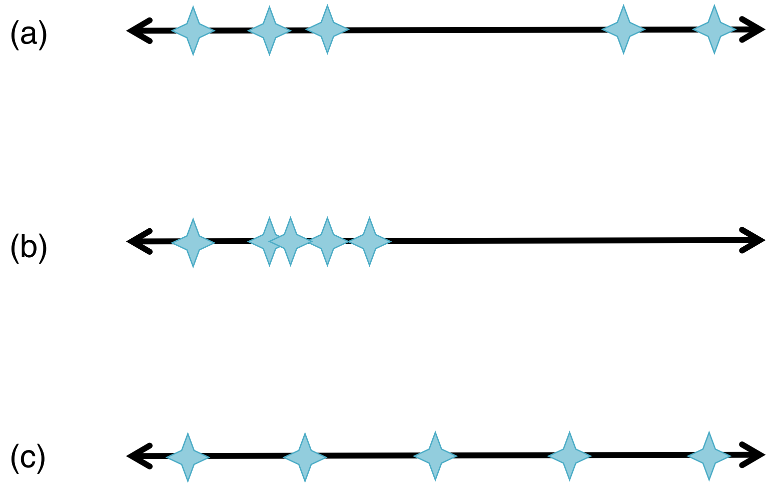

Figure 1 Defining amino acid coverage: examples of range and evenness for hypothetical amino acid sets. Amino acids are shown as dots on a linear descriptor (e.g. "LogP"). The set defined in **(a)** displays high range, but poor evenness. In **(b)**, the set is highly even, but over a small range. The set in **(c)** is optimized for both range and evenness. Adapted from figure in Philip and Freeland [5].

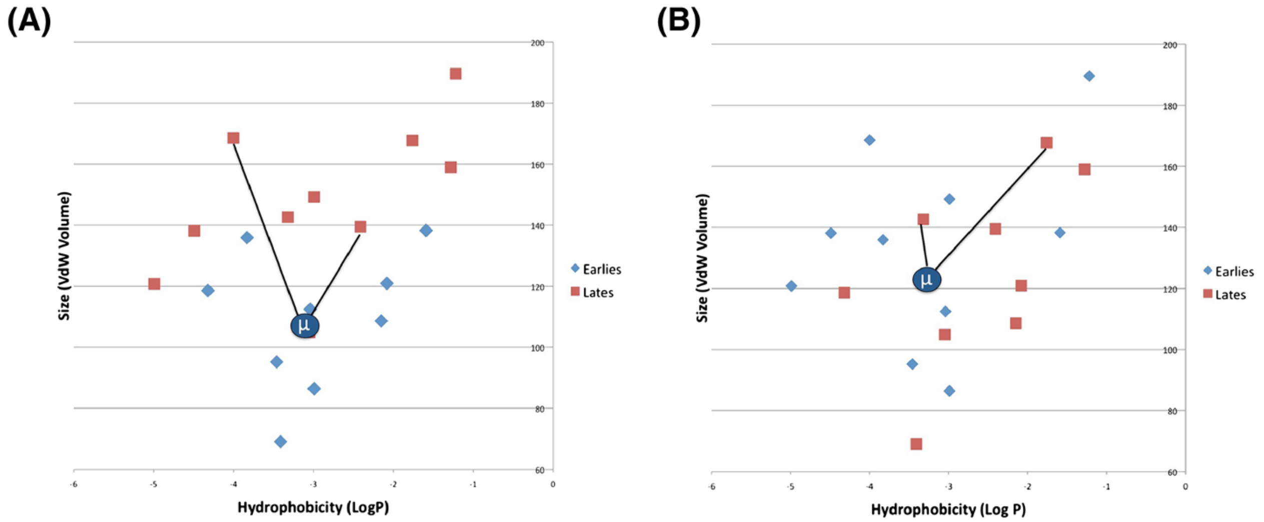

Figure 2 Calculating the expansion of the late amino acids from the cluster of earlies. This test first calculates the one, two, or three-dimensional mean of the true subset of 10 early amino acids (μ) and then calculates the average distance between each late amino acid and this mean **(panel A)**. The same calculation is then repeated when amino acids are re-designated as 10 "earlies" and 10 "lates" at random **(panel B)**.

to their unique structural simplicity, can arguably be considered outliers rather than meaningful degrees of freedom of amino acid possibility space.

Test 3: do the late amino acids fill an adaptive gap?

Our general adaptive hypothesis is that the late amino acids were added to the genetic code by natural selection because they expanded the protein-building chemistry space available to some simpler precursor of the standard genetic code. Test 2 therefore considers whether the late amino acids populate new points in chemistry space that were unavailable to the early amino acids. However, amino acids do not act in isolation: they are polymerized to form proteins. This implies that the biologically relevant chemistry space of the early amino acids also includes the chemistry space of their dimers, trimers, etc. For the late amino acids to be adaptively advantageous, they must not only expand the chemistry space of the early amino acids, but also do so in such a way that they are performing a novel role; that is, one not already fulfilled by existing amino acids or their oligomers. Within one dimension, early amino acid dimers and late monomers overlap considerably in range for size and hydrophobocity (Figure 3). Any adaptive separation must therefore be found in combinations of these properties.

In order to test this, we first defined an area around all points in chemistry space (late amino acid monomers and early amino acid dimers) to represent the region of chemistry space populated by each molecule. This area was calculated as a circle centered on each point with a radius equal to the average distance between all points (see Figure 4). Using the designation of late amino acids and early-dimers, we measured the frequency with which dimers were found to overlap with the chemistry space of late amino acids. We then randomized the designation of late amino acids and early dimers and

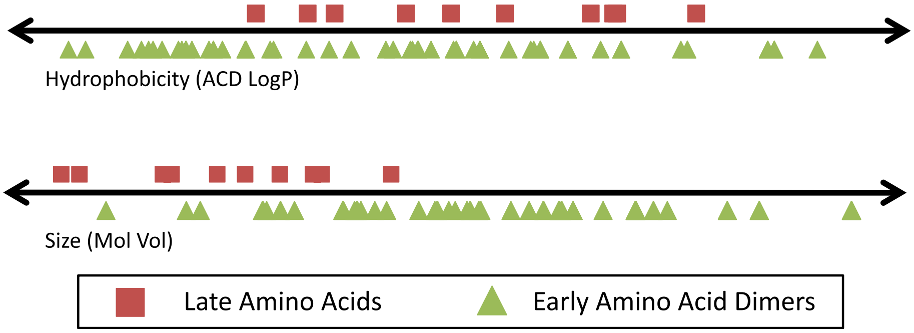

Figure 3 One-dimensional chemistry space of late amino acids and early amino acid dimers. This simple plot clearly shows that for the dimensions of size and hydrophobicity, the late amino acids overlap in chemistry space with early amino acid dimers.

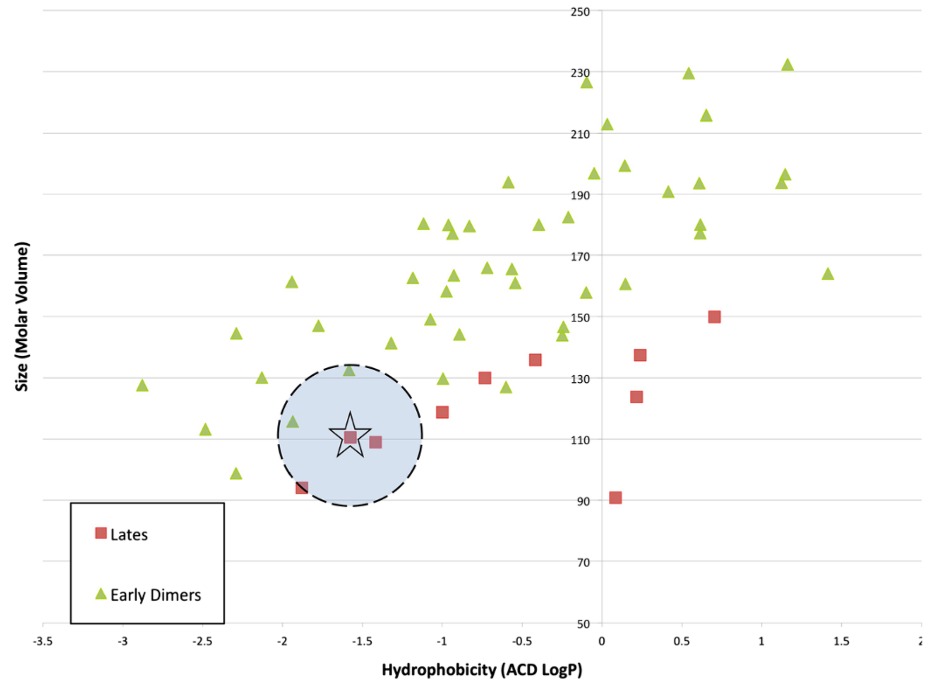

Figure 4 Measuring the overlap in chemistry space between late amino acids and early amino acid dimers. The orb centered on a late amino acid represents the chemistry space that molecule is able to approximate. Adaptively advantageous late amino acids would be expected to show lower overlap with existing molecules (i.e. the early amino acid dimers) than would be expected by chance.

measured equivalently how often dimers were found to share chemistry space with late amino acids. This allowed us to measure the overlap in chemistry space between these two sets of molecules compared to what could be expected by chance and to therefore quantify the novelty contributed by the late amino acids.

Results

Test 1

Our first test examined the adaptive properties of a functional set of amino acids used by a putative precursor to the standard genetic code relative to an internal context: alphabets of equivalent size randomly selected from the genetically encoded amino acids. Figure 5 reports how often randomly selected sets of 10 amino acids 'cover' chemistry space better than the 10 early amino acids.

Test 2

Our second test assessed the adaptive value of the late amino acids as additions to the early amino acids in terms of literal addition (expansion) of chemistry space. Figure 6 summarizes how often the true late amino acids lie further from the true cluster of early amino acids than when we randomly assign an equivalent number of amino acids as early and the remainder as late. The figure also includes the same results with the outliers Glycine and Alanine removed.

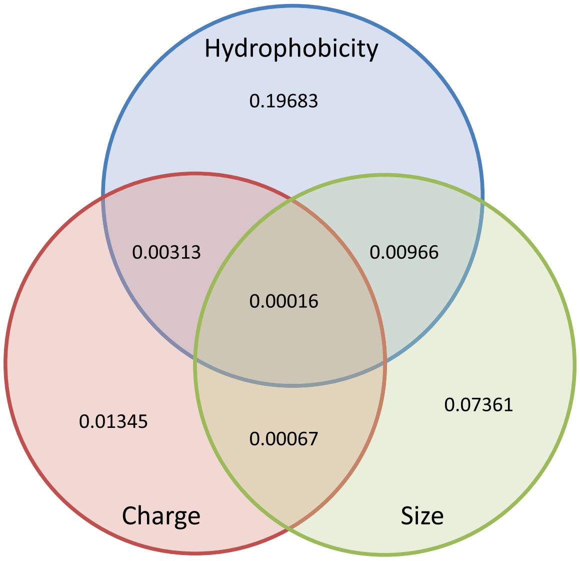

Figure 5 The frequency with which random sets of 10 amino acids exhibit better coverage (range and evenness) than the early amino acids in each of three molecular descriptors and all combinations of these descriptors. Of particular note, when two or more dimensions of amino acid chemistry space are considered simultaneously, far less than 1% of random amino acid sets of size 10 match the coverage of the early amino acids.

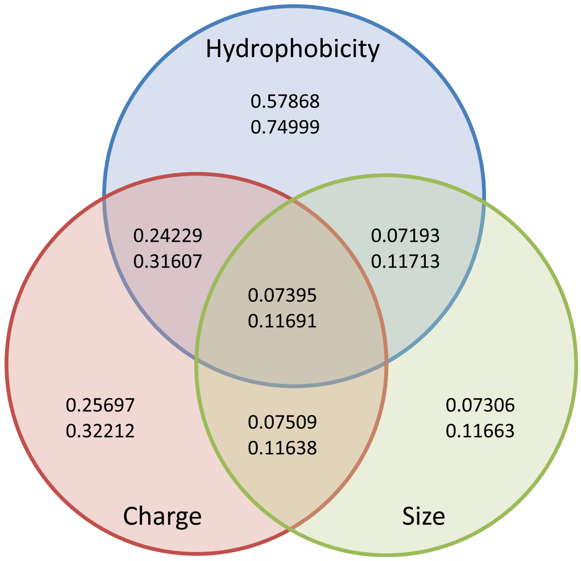

Figure 6 The proportion of trials in which randomly designated "late" amino acids represent a greater expansion from randomly designated "earlies" than the true late amino acids from the true earlies. The second row number in each section of the Venn diagram represents the results of the same test performed with Glycine and Alanine removed as outliers. The observed effect of late amino acids expanding the chemistry space of the early amino acids appears strongest in the dimension of size, both as an individual measure of chemistry space and in combination with other descriptors.

Test 3

Our third test evaluates whether "late" amino acids fill a gap by expanding the chemistry space of monomeric "early" amino acids without occupying an area of chemistry space not already filled by dimers made from early amino acids. In the 2-dimensional chemistry space of size and hydrophobicity, we found the "late" amino acids and "early" dimers are more separated (exhibit less overlap) than random designations of these molecules 99.77% of the time.

Discussion

Our aim here was to apply the concept of chemistry space in order to corroborate or refute previous claims for adaptive properties of the set of 20 genetically encoded amino acids. Whereas previous claims tested the genetically encoded amino acids against a background of plausible alternatives that were never, as far as we can tell, incorporated into genetic coding, our focus here was on the internal logic of the 20 genetically encoded amino acids, building from the premise of *early* versus *late* amino acids.

Our first test asked whether the 10 amino acids that are thought to have formed a simpler, earlier genetic code exhibit, as a set, similar adaptive properties as have been recorded for the full amino acid alphabet. Our results show a weak adaptive signal for any individual dimension of chemistry space. However, if we accept that amino chemistry space becomes more meaningful for protein structure and function as it is measured in two or three dimensions, then our results provide strong support for this notion of an internal adaptive logic to the early amino acids. In terms of their coverage of chemistry space, the early amino acids prove to be a highly unusual subset of the twenty genetically encoded amino acids. This is consistent with the inference that, at some point before the emergence of the standard genetic code, they could have constituted a cohesive and functional alphabet. In the same way as the standard 20 were shown to be adaptively advantageous compared to a broader pool of alternatives [5], the "early" amino acids also appear to cover chemistry space exceptionally well.

Our findings complement a body of recent research spanning multiple approaches that suggests the early amino acids contain enough chemical information to form a coherent functional set. This includes a study that suggests the set of early amino acids are sufficient to enable protein folding by reducing the amino acid composition of a protein while maintaining its foldability [10]. We draw attention to the unusual terminology of this paper, which is consistent with similar efforts by others but misleading to those outside the field. The bold claim that a functional protein has been made entirely from early amino acids actually refers to a protein sequence that comprises 80% early amino acids, with the remaining 20% of the sequence drawing from the full alphabet of 20 amino acids. More straightforwardly, our results also agree with a study that used phylogenetic analysis of amino acid compositional bias in ribosomal proteins to conclude that "at a more primitive state, the code would still contain a similar diversity of physiochemical properties" [21]. A third study that examined modern protein sequences deficient in late amino acids of functional significance (the basic amino acids Lysine and Arginine) also supported the idea of "small proteins without basic amino acids performed important functions in the prebiotic chemistry of early Earth" [22].

Test 2 asked whether the 10 amino acids that are thought to have been later additions to the genetic code represent a quantifiable expansion of amino acid chemistry space. Here we see a weak signal that the late amino acids lie further in chemistry space from the earlies than expected by chance. It appears, however, that most of this effect is coming from the dimension of size. In other words, the later additions to the standard amino acid alphabet differ from the early amino acids primarily in that they were larger. Considering that the early amino acids include the smallest L-alpha amino acids that are chemically possible, this is perhaps not surprising. Nevertheless, so long as we accept that size is an important component of an adaptive chemistry space, then these results support an adaptive explanation: the late amino acids show a literal expansion of two and three-dimensional chemistry space. This assumption is bolstered by the results of the same test performed with Glycine and Alanine removed, where we still find that, in almost 90% of trials, the true late amino acids show a greater expansion in chemistry space than would be expected by chance. Our findings again are in general agreement with previous inference drawn from a different approach, where the late amino acids are believed to have been advantageous precisely because of their "unique specialized" structure [21].

Our third test offers clarification for the otherwise somewhat ambiguous results of Test 2. If the concept of early versus late amino acids is correct, then it implies that all dimers comprising two early amino acids were present before any late amino acid was incorporated into the genetic code. We therefore tested whether the expansion of amino acid chemistry space brought about by the addition of late amino acids was steered by an additional consideration: avoiding overlap with regions already occupied by dimers made from the earlies. In accordance with this adaptive hypothesis, we find a strong signal that the lates indeed filled a gap in chemistry space that was not already occupied by the early amino acids or their dimers.

This third test represents a qualitative expansion of thinking about adaptive amino acid chemistry space in that it is the first time molecules larger than monomers have been considered. Indeed, it is largely thanks to the availability of a free database, Chemspider, which includes relevant molecules and their key descriptors that these measurements were made possible. In this context, it is unfortunate that the molecular descriptor isoelectric point is not readily available for dimers, but this implies a simple, logical future step to verify or undermine our current interpretation.

Another clear question raised by our analysis is whether other molecular structures might have complemented early amino acids to form the functional chemistry space in which the genetic code emerged. Many variations of the RNA-world model predict that metabolism during the time of genetic code evolution would have involved a heavy presence by enzymatic cofactors featuring nucleobases and their derivatives [23]. Indeed, it has been specifically proposed that "aromatic amino acids could have been selected for as RNA "replacements" in a dying RNA world, mimicking nucleotides in important structural roles (e.g., stacking interactions)"

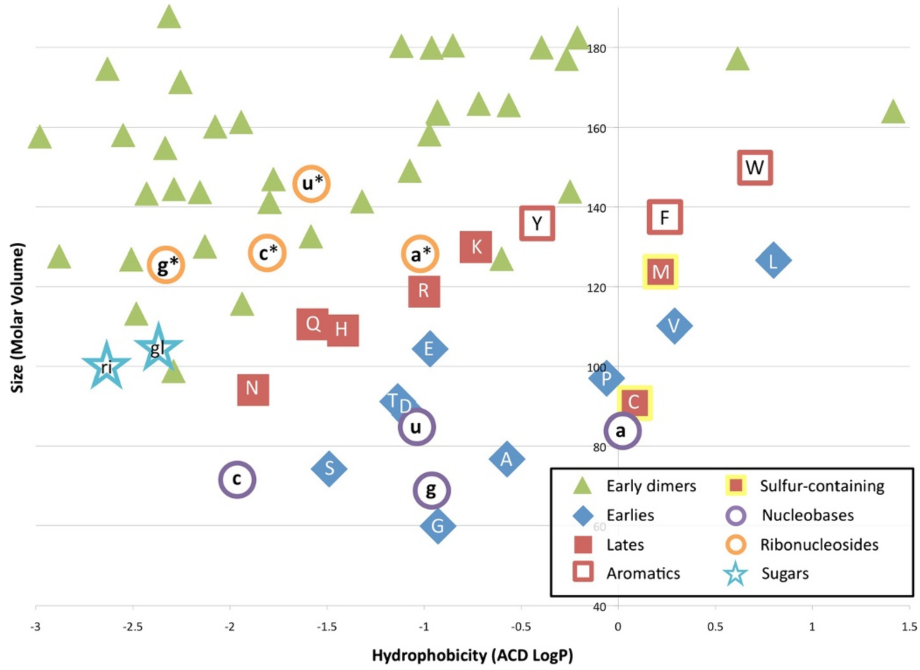

Figure 7 A visual exploration of fundamental metabolites and the chemistry space of the genetically encoded amino acids. A 2-dimensional plot of chemistry space using size (ACD Molar Volume) against hydrophobicity (ACD LogP) for the 20 amino acids of the standard genetic code, indicated by their standard one-letter abbreviations. Early amino acid dimers are represented as green triangles. Other core metabolites are plotted including nucleobases (a = adenine, c = cytosine, g = guanine, u = uracil), ribonucleosides (a* = adenosine, c* = cytidine, g* = guanosine, u* = uridine), and sugars (gl = glucose; ri = ribose).

[21]. We therefore considered visually a simple 2-dimensional plot of the chemistry space analyzed in test 3 (that is the chemistry space of the earlies, lates, and early dimers) in order to add some obvious examples of molecules that suggest themselves to this way of thinking (Figure 7). To our best inspection, nothing here suggests an obvious complementary role for these cofactors, which may well indicate the need of more sophisticated, higher-dimensional analysis.

Conclusion

Here we present three, logically independent tests that each generates quantitative evidence to corroborate or challenge previous claims for adaptive properties of the standard amino acid alphabet. Each test operates in terms of a simple, 3-dimensional chemistry space built from amino acid charge (pI), size (ACD Molar volume), and hydrophobicity (ACD LogP). Whereas previous claims focus on the external chemistry space of amino acids that are *not* part of the genetic code, we turn to consider the internal logic of the 20 genetically encoded amino acids. In particular, we consider the surprisingly strong, multidisciplinary consensus that has emerged in recent years to suggest that the genetic code may have begun with only half the 20 amino acids currently found in the standard genetic code. This division of the standard amino acid alphabet into “early” and “late” amino acids allows us to make three predictions based upon an adaptive hypothesis: (i) the “early” amino acids should form a cohesive sub-set with similar properties to the final, full-sized amino acid alphabet; (ii) the “late” amino acids should demonstrate quantifiable expansion of amino acid chemistry space in terms of dimensions that define protein-building potential, and (iii) the expansion of “late” amino acids should populate regions of chemistry space that were not already available to a genetic code that build dimers from the early amino acids. Taken together, our results provide strong support for these predictions.

Methods

In order to implement our methods, we wrote our source code in Java version 6. We then ran it on Mac OS X Version 10.6.8. Source code is available upon request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

MAI developed and implemented the algorithm. Both authors were involved in conception and writing of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA09DA77A issued through the Office of Space Science. We thank James Stephenson for helpful and stimulating discussion.

Author details

[1]NASA Astrobiology Institute, University of Hawaii, Honolulu, USA.
[2]Interdisciplinary Studies Program, University of Maryland, Baltimore, USA.

Received: 15 June 2013 Accepted: 2 August 2013
Published: 21 January 2014

References

1. Dobson CM: **Chemical space and biology.** *Nature* 2004, **432:**824–828. 7019.
2. Barker A, Kettle JG, Nowak T, Pease JE: **Expanding medicinal chemistry space.** *Drug Discov Today* 2012, **18**(5-6):298–304.
3. Lloyd DG, Golfis G, Knox AJ, Fayne D, Meegan MJ, Oprea TI: **Oncology exploration: charting cancer medicinal chemistry space.** *Drug Discov Today* 2006, **11.3:**149–159.
4. Reymond JL, Mahendra A: **Exploring chemical space for drug discovery using the chemical universe database.** *ACS Chem Neurosci* 2012, **3**(9):649–57.
5. Philip GK, Freeland SJ: **Did evolution select a nonrandom “alphabet” of amino acids?** *Astrobiology* 2011, **11.3:**235–240.
6. Wong JT, Bronskill PM: **Inadequacy of prebiotic synthesis as origin of proteinous amino acids.** *J Mol Evol* 1979, **13.2:**115–125.
7. Trifonov EN: **Consensus temporal order of amino acids and evolution of the triplet code.** *Gene* 2000, **261.1:**139–151.
8. Higgs PG, Pudritz RE: **A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code.** *Astrobiology* 2009, **9.5:**483–490.
9. Cleaves HJ: **The origin of the biologically coded amino acids.** *J Theor Biol* 2010, **263.4:**490–498.
10. Longo LM, Lee J, Blaber M: **Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein.** *Proc Natl Acad Sci* 2013, **110.6:**2135–2139.
11. Weberndorfer G, Hofacker IL, Stadler PF: **On the evolution of primitive genetic codes.** *Orig Life Evol Biosph* 2003, **33.4-5:**491–514.
12. Kawashima S, Ogata H, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 1999, **27**(1):368–369.
13. Grantham R: **Amino acid difference formula to help explain protein evolution.** *Science* 1974, **185**(4154):862.
14. Ladunga I, Smith RF: **Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties.** *Protein Eng* 1997, **10**(3):187–196.
15. Müller-Späth S: **Charge interactions can dominate the dimensions of intrinsically disordered proteins.** *Proc Natl Acad Sci* 2010, **107**(33):14609–14614.
16. Kauzmann W: **Of protein denaturation.** *Adv Protein Chem* 1959, **14:**1.
17. Baussand J, Deremble C, Carbone A: **Periodic distributions of hydrophobic amino acids allows the definition of fundamental building blocks to align distantly related proteins.** *Proteins* 2007, **67**(3):695-–708.
18. Gilson MK, Honig BH: **Calculation of electrostatic potentials in an enzyme active site.** *Nature* 1987, **330**(6143):84–86.
19. Lu Y, Freeland SJ: **On the evolution of the standard amino-acid alphabet.** *Genome Biol* 2006, **7**(1):102.
20. Van der Waterbeemd H, Karajiannis H, Tayar NE: **Lipophilicity of amino acids.** *Amino Acids* 1994, **7**(2):129–145.
21. Fournier GP, Gogarten JP: **Rooting the ribosomal tree of life.** *Mol Biol Evol* 2010, **27**(8):1792–1801.
22. McDonald GD, Storrie-Lombardi MC: **Biochemical constraints in a protobiotic earth devoid of basic amino acids: the “BAA (−) World”.** *Astrobiology* 2010, **10**(10):989–1000.
23. White HB III: **Coenzymes as fossils of an earlier metabolic state.** *J Mol Evol* 1976, **7**(2):101–104.

doi:10.1186/1759-2208-5-1
Cite this article as: Ilardo and Freeland: **Testing for adaptive signatures of amino acid alphabet evolution using chemistry space.** *Journal of Systems Chemistry* 2014 **5**:1.